Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 12, Número 2 (2020)

ISSN: 1647-0818

linguaMATICA

Volume 12, Número 2 – 2020

LinguaMÁTICA

ISSN: 1647–0818

Editores

Alberto Simões
José João Almeida
Xavier Gómez Guinovart

Conteúdo

Editorial

Ainda há pouco publicávamos a primeira edição da Linguamática e, subitamente, eis-nos a comemorar uma dúzia de anos. A todos os que colaboraram durante todos estes anos, tenham sido leitores, autores ou revisores, a todos o nosso muito obrigado.

Não é fácil manter um projeto destes durante tantos anos, sem qualquer financiamento. Todo este sucesso é graças a trabalho voluntário de todos, o que nos permite ter artigos de qualidade publicados e acessíveis gratuitamente a toda a comunidade científica.

Nestes doze anos muitos foram os que nos acompanharam, nomeadamente na comissão científica. Alguns dos primeiros membros, convidados em 2008, continuam ativamente a participar neste projeto, realizando revisões apuradas.

Outros, por via do seu percurso académico e pessoal, já não colaboram connosco. Como sinal de agradecimento temos mantido os seus nomes na publicação. No entanto, não podemos estar presos ao passado, pelo que deixamos aqui um agradecimento especial a alguns dos membros que irão deixar de constar na comissão científica, de forma ativa, e que passarão a ser mencionados, no sítio da revista, como anteriores colaboradores da Linguamática: Ana Frankenberg-Garcia, Antón Santamarina, Arantza Díaz de Ilarraza, Belinda Maia, Iñaki Alegria, Joseba Abaitua, Maria das Graças Volpe Nunes *e* Tony Berber Sardinha. *O nosso muito obrigado!*

Xavier Gómez Guinovart
José João Almeida
Alberto Simões

Comissão Científica

Alberto Álvarez Lugrís,
Universidade de Vigo

Alberto Simões,
Instituto Politécnico do Cávado e Ave

Aline Villavicencio,
Universidade Federal do
Rio Grande do Sul

Álvaro Iriarte Sanroman,
Universidade do Minho

Ana Frankenberg-Garcia,
University of Surrey

Anselmo Peñas,
Universidad Nacional de
Educación a Distancia

Antón Santamarina,
Universidade de Santiago de
Compostela

Antoni Oliver González,
Universitat Oberta de Catalunya

Antonio Moreno Sandoval,
Universidad Autónoma de Madrid

António Teixeira,
Universidade de Aveiro

Arantza Díaz de Ilarraza,
Euskal Herriko Unibertsitatea

Arkaitz Zubiaga,
Dublin Institute of Technology

Belinda Maia,
Universidade do Porto

Bruno Martins,
Instituto Superior Técnico

Carmen García Mateo,
Universidade de Vigo

Diana Santos,
Linguateca/Universidade de Oslo

Ferran Pla,
Universitat Politècnica de València

Gael Harry Dias,
Université de Caen Basse-Normandie

Gerardo Sierra,
Universidad Nacional
Autónoma de México

German Rigau,
Euskal Herriko Unibertsitatea

Helena de Medeiros Caseli,
Universidade Federal de São Carlos

Horacio Saggion,
University of Sheffield

Hugo Gonçalo Oliveira,
Universidade de Coimbra

Iñaki Alegria,
Euskal Herriko Unibertsitatea

Irene Castellón Masalles,
Universitat de Barcelona

Iria da Cunha,
Universidad Nacional de
Educación a Distancia

Joaquim Llisterri,
Universitat Autònoma de Barcelona

José João Almeida,
Universidade do Minho

José Paulo Leal,
Universidade do Porto

Joseba Abaitua,
Universidad de Deusto

Juan-Manuel Torres-Moreno,
Université d´Avignon et
des Pays du Vaucluse

Kepa Sarasola,
Euskal Herriko Unibertsitatea

Karin Becker,
UFRGS, Brasil

Laura Plaza,
Complutense University of Madrid

Lluís Padró,
Universitat Politècnica de Catalunya

Luís Morgado da Costa,
Nanyang Technological University

Manex Agirrezabal,
University of Copenhagen

Marcos Garcia,
Universidade da Corunha

María Inés Torres,
Euskal Herriko Unibertsitatea

Maria Graças Volpe Nunes,
Universidade de São Paulo

Mário Rodrigues,
Universidade de Aveiro

Mercè Lorente Casafont,
Universitat Pompeu Fabra

Miguel Solla Portela,
Universidade de Vigo

Mikel Forcada,
Universitat d'Alacant

Pablo Gamallo Otero,
Universidade de Santiago de
Compostela

Patrícia Cunha França,
Universidade do Minho

Patricia Martin Rodilla
Universidade de Santiago de
Compostela

Ricardo Rodrigues
Instituto Politécnico de Coimbra

Rui Pedro Marques,
Universidade de Lisboa

Susana Afonso Cavadas,
University of Exeter

Tony Berber Sardinha,
Pontifícia Universidade
Católica de São Paulo

Xavier Gómez Guinovart,
Universidade de Vigo

Revisor Convidado:

Nuno Escudeiro, Instituto Superior de Engenharia do Porto

Artigos de Investigação

Proposta recebida em Fevereiro 2020 e aceite para publicação em Agosto 2020.

Adaptação Lexical Automática em Textos Informativos do Português Brasileiro para o Ensino Fundamental

Automatic Lexical Adaptation in Brazilian Portuguese Informative Texts for Elementary Education

Nathan Siegle Hartmann
Instituto de Ciências Matemáticas e Computação
Universidade de São Paulo, Brasil
nathansh@icmc.usp.br

Sandra Maria Aluísio
Instituto de Ciências Matemáticas e Computação
Universidade de São Paulo, Brasil
sandra@icmc.usp.br

Resumo

A Adaptação Textual é uma grande área de pesquisa do Processamento de Línguas Naturais (PLN), bastante conhecida como prática educacional, e possui duas grandes abordagens: a Simplificação e a Elaboração Textual. Não há muitos trabalhos na literatura de PLN que tratam todas as fases da Adaptação Lexical para implementação de sistemas. Vários trabalhos tratam independentemente as tarefas de Simplificação e Elaboração Lexicais, trazendo contribuições parciais, já que cada uma das tarefas possuem seus próprios desafios. Este trabalho propôs um *pipeline* para a Adaptação Lexical e apresenta contribuições para três das quatro etapas do *pipeline*, sendo elas: (i) proposta e avaliação de métodos para a tarefa de Identificação de Palavras Complexas; (ii) análise de córpus para levantamento de padrões de Elaboração Lexical do tipo definição; (iii) disponibilização do córpus SIMPLEX-PB 3.0, contendo em sua nova versão definições curtas extraídas de dicionário que foram revisadas manualmente, anotações de termos técnicos extraídas de dicionário, e métricas linguísticas de complexidade lexical; e (iv) proposta e avaliação de métodos para Simplificação Lexical, estabelecendo um novo *SOTA* para a tarefa aplicada no Português Brasileiro.

Palavras chave

adaptação textual, simplificação lexical, elaboração lexical, auxílio à leitura de crianças

Abstract

Text Adaptation is a large Natural Language Processing (NLP) research area, well known as educational practice and has two main approaches: Simplification and Text Elaboration. There is not much work in the NLP literature that addresses all phases of Lexical Adaptation for systems implementation. Several works independently deal with the Lexical Simplification and Elaboration tasks, bringing partial contributions, since each task has its own challenges. This work proposed a pipeline for Lexical Adaptation and presents contributions in three of the four stages of the Lexical Adaptation pipeline: (i) proposal and evaluation of methods for the Complex Word Identification task; (ii) corpus analysis to survey Lexical Elaboration word definition standards; (iii) the SIMPLEX-PB 3.0 corpus, containing in its new version short definitions extracted from dictionaries that were manually revised, annotations of technical terms extracted from a dictionary, and linguistic metrics of lexical complexity; and (iv) proposal and evaluation of methods for Lexical Simplification, establishing a new SOTA for the task applied in Brazilian Portuguese.

Keywords

text adaptation, lexical simplification, lexical elaboration, reading aid for children

1. Introdução

Dada a importância do ensino da leitura e compreensão de textos em âmbito mundial e aos constantes progressos na área de Processamento de Línguas Naturais (PLN) nos últimos anos, tem havido um grande interesse de pesquisa na adaptação automática de textos escritos a fim de torná-los acessíveis para um número maior de pessoas (Siddharthan, 2006; Bulté et al., 2018; Pasqualini, 2018; Štajner et al., 2019), como adultos com baixa escolaridade (Max, 2006; Watanabe et al., 2010; Amancio, 2011; Aluísio & Gasperin, 2010; Barlacchi & Tonelli, 2013; Pasqualini, 2018), crianças (Mihalcea & Csomai, 2007; De Belder & Moens, 2010; Trieschnigg & Hauff, 2011; Kajiwara et al., 2013), aprendizes de uma segunda língua (Gardner & Hansen, 2007; Petersen & Ostendorf, 2007; Paetzold & Spe-

DOI: 10.21814/lm.12.2.323
This work is Licensed under a Creative Commons Attribution 4.0 License

cia, 2017), indivíduos com deficiências cognitivas (Bott et al., 2012), indivíduos surdos (Inui et al., 2003; Chung et al., 2013), afásicos (Devlin & Tait, 1998; Devlin & Unthank, 2006; Rello et al., 2013b) e disléxicos (Rello et al., 2013b,a).

A Adaptação Textual é uma área de pesquisa do PLN bastante conhecida como prática educacional e possui duas grandes abordagens - a Simplificação e a Elaboração Textual (Mayer, 1980; Young, 1999; Saggion, 2017; Štajner & Saggion, 2018; Arfé et al., 2018). A primeira pode ser definida como qualquer tarefa que reduza a complexidade lexical ou sintática de um texto enquanto tenta preservar seu significado (Siddharthan, 2006, 2014), tendo um grande impacto na leiturabilidade (ou inteligibilidade) de um texto; pode ser dividida nas técnicas: (i) Simplificação Lexical, (ii) Simplificação Sintática e (iii) Sumarização Automática. A segunda tem impacto na compreensibilidade de um texto, isto é, na facilidade com que um texto pode ser compreendido e também no aumento do vocabulário do leitor, pois se utiliza de um conjunto de técnicas para inserir material redundante, por exemplo: (i) adição de sinônimos/antônimos ao lado de palavras ou expressões complexas, (ii) definição de conceitos, ou ainda (iii) tornar explícitas as conexões entre as ideias do texto (Mayer, 1980; Aluísio & Gasperin, 2010).

A Adaptação Lexical, foco deste trabalho, é uma subárea da Adaptação Textual, trazendo as técnicas de Elaboração e Simplificação Lexicais, apresentadas a seguir.

A Elaboração Lexical tem a função de auxiliar a compreensibilidade de um texto, familiarizando termos ou palavras desconhecidas para um dado leitor. Ela é realizada com a adição de informações redundantes como uso de definições via *links* nas próprias palavras[1] ou ao lado das palavras complexas via uso de informações parentéticas, paráfrases, e apostos (Urano, 2000; Bulté et al., 2018). Essa redundância de informação aumenta a coesão do texto e, consequentemente, torna-o mais compreensível (Crossley et al., 2011). Já a Simplificação Lexical se realiza com a troca de palavras ou expressões por variações (geralmente sinônimos) que podem ser entendidas por um maior número de pessoas (Štajner & Saggion, 2018). Ao utilizarmos palavras menos frequentes/raras, não estamos apenas auxiliando o leitor a compreendê-la, mas também a compreender todo o texto, que ficará mais simples (Crossley et al., 2007, 2011).

[1] Um exemplo deste tipo pode ser visto na Wikipédia e em Amancio (2011).

Um sistema automático para simplificação lexical realiza os seguintes passos em *pipeline*, conforme apresentado na Figura 1:

1. dada uma sentença, selecionam-se as palavras ou expressões que são consideradas complexas para um leitor e/ou tarefa computacional;
2. buscam-se substitutos, geralmente usando repositórios como as *wordnets*;
3. filtram-se os substitutos para se recuperar apenas os sinônimos com o mesmo sentido usado no contexto da sentença original; e
4. ranqueiam-se os substitutos segundo o critério de simplicidade para o leitor e/ou tarefa. Normalmente, a frequência em um grande córpus da língua alvo e o tamanho das palavras são utilizados como critério de simplicidade.
5. Após a escolha do sinônimo adequado, há a troca da palavra em foco pelo sinônimo selecionado, que pode pedir ajustes na escrita das palavras da oração, como a adequação de gênero, número e grau.

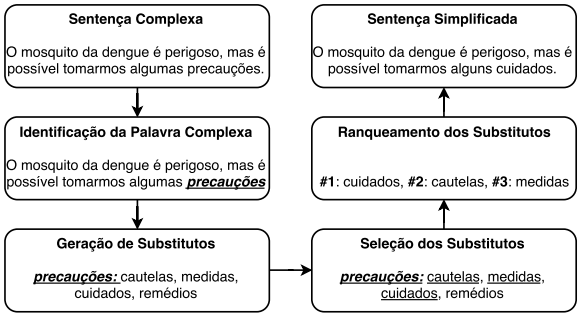


Figura 1: *Pipeline* tradicional para a tarefa de Simplificação Lexical (Specia et al., 2012), ilustrado com exemplos do Português.

Assim como o *pipeline* apresentado para a tarefa de Simplificação Lexical, podemos idealizar um fluxo de processamento para a grande tarefa de Adaptação Lexical (Figura 2), que tem como propósito decidir quando elaborar ou simplificar as palavras de um texto, segundo as necessidades do leitor e cenário de uso do texto. Para um dado texto, ou sentença: (i) identificam-se as palavras complexas; (ii) decide-se qual a melhor abordagem de adaptação para cada palavra (simplificação ou elaboração); e (iii) adapta-se cada palavra complexa identificada, segundo a abordagem de adaptação lexical selecionada.

A Adaptação Lexical é importante porque, para os leitores compreenderem o contexto do trecho que estão lendo, eles precisam relacionar o seu conhecimento léxico-semântico para inferirem o significado das palavras (de Sousa et al.,

2020). Quando o foco é um público alvo específico como, por exemplo, crianças, devemos considerar que as limitações da compreensão leitora acarretam em uma dificuldade no estabelecimento de relações semânticas das palavras no texto. Essa dificuldade impossibilita os leitores desconsiderarem as informações irrelevantes e manterem as informações relevantes na memória de trabalho (Henderson et al., 2013). Em resumo, as dificuldades no nível lexical acarretam complicações na compreensão global de um texto, evidenciando a importância da adaptação lexical de textos para certos públicos (de Sousa et al., 2020).

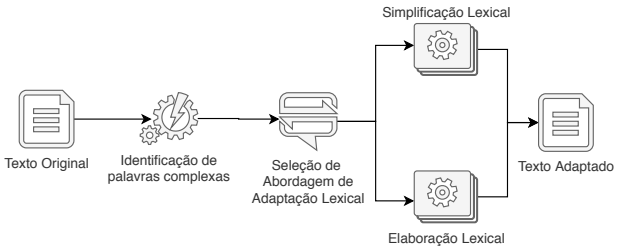


Figura 2: *Pipeline* para a tarefa de Adaptação Lexical.

Este trabalho apresenta contribuições em três das quatro etapas do *pipeline* de Adaptação Lexical, sendo elas:

- **Identificação de Palavras Complexas** - Proposta de 12 métodos, contabilizando 153 diferentes variações para a tarefa de Identificação de Palavras Complexas, avaliação e apresentação dos melhores resultados obtidos;
- Análise de córpus para levantamento de padrões de Elaboração Lexical do tipo definição e apresentação dos padrões e ocorrências identificadas para apoiar a decisão de quando usar Elaboração Lexical;
- **Elaboração Lexical** - Disponibilização do córpus SIMPLEX-PB 3.0[2] contendo, além dos recursos do SIMPLEX-PB 2.0 (Hartmann et al., 2020), definições curtas extraídas de dicionário e revisadas manualmente, anotações de termos técnicos extraídas de dicionário, e métricas linguísticas de complexidade lexical, proporcionando também o uso do recurso em estudos de Elaboração Lexical; e
- **Simplificação Lexical** - Proposta de métodos para Simplificação Lexical e estabelecimento de um novo estado da arte (*SOTA*, em inglês) para a tarefa aplicada no Português Brasileiro.

[2] https://github.com/nathanshartmann/SIMPLEX-PB-3.0

A Seção 2 apresenta trabalhos relacionados da área de Adaptação Lexical e o estado da arte para as tarefas de Elaboração e Simplificação Lexicais. A Seção 3 apresenta a última versão do SIMPLEX-PB, um córpus de suporte à Adaptação Lexical que ao longo dos últimos anos foi evoluindo e hoje contém, além de outras informações: sentenças, suas palavras complexas, sinônimos ranqueados de acordo com sua complexidade pelo público alvo (crianças), e definições curtas para as palavras complexas. A Seção 4 apresenta o levantamento de padrões de Elaboração Lexical do tipo definição, via análise de córpus, para apoiar a decisão de quando usar Elaboração Lexical. A Seção 5 apresenta os *datasets* compilados (Seção 5.1), a descrição dos métodos desenvolvidos para Identificação de Palavras Complexas (Seção 5.2), a avaliação desses métodos, usando as métricas F1 e AUC (do inglês *Area Under Curve ROC* - também do inglês *Receiver Operating Characteristic*) (Seção 5.3), e como aplicamos esses métodos na tarefa de Simplificação Lexical no cenário do público infantil, avaliando eles no dataset SIMPLEX-PB 3.0 (Seção 5.4). A Seção 5.5 apresenta o Adapt2Kids, um site para demonstrar a aplicação dos métodos e recursos de simplificação lexical avaliados neste artigo. Por fim, a Seção 6 apresenta as conclusões deste estudo e os trabalhos futuros.

2. Trabalhos Relacionados

Não há muitos trabalhos na literatura de PLN sobre Adaptação Lexical, na atualidade. Ao invés disso, tratam independentemente as tarefas de Simplificação e Elaboração Lexicais, já que cada uma possui seus próprios desafios. O tutorial sobre Simplificação Textual apresentado por Štajner e Saggion no Coling 2018 (Štajner & Saggion, 2018), inclusive, traz a tarefa de elaboração como um dos módulos da Simplificação Textual, juntando a ela também a tarefa de redução/eliminação de conteúdo. Neste artigo, preferimos, entretanto, tratá-las de forma independente.

O trabalho de Bulté et al. (2018), descrito abaixo, foi, no melhor do nosso conhecimento, um desses raros casos que abordou a grande tarefa de Adaptação Lexical, desenvolvendo um *pipeline* de Adaptação Lexical para o Holandês.

Primeiro, a complexidade de cada palavra é estimada pela consulta da sua Idade de Aquisição em recursos psicolinguísticos e a sua frequência no córpus Wablief, de holandês simplificado. Uma palavra é complexa se ela ultrapassar um limiar pré-estabelecido dessas duas *features*. Os

sinônimos para a palavra complexa são consultados no Cornetto (Vossen et al., 2013) e esses são considerados mais simples quando a Idade de Aquisição e Frequência forem inferiores à palavra complexa. Por fim, um modelo de língua é acionado para verificar se a palavra de substituição selecionada se encaixa no contexto em questão. Os autores elaboram: (i) as palavras identificadas como difíceis que não foram simplificadas pelo sistema, (ii) as palavras compostas, e (iii) os nomes próprios, enriquecendo o texto com informações da Wikipedia (definições e links nas palavras).

2.1. Trabalhos de Elaboração Lexical

Não há muitos trabalhos na área de Elaboração Textual que propuseram métodos computacionais. Trabalhos da área da educação com foco no ensino do inglês como segunda língua (Tsang, 1987; Yano et al., 1994; Urano, 2000) avaliaram o seu uso no auxílio da leitura em vez de automatizar a tarefa. Três trabalhos que propuseram métodos para a tarefa de Elaboração Lexical são os de Mihalcea & Csomai (2007), Amancio (2011) e Trieschnigg & Hauff (2011). Esses três trabalhos fizeram uso de recursos obtidos da Web para elaborar palavras complexas.

Em um cenário educacional, é importante que os alunos tenham um fácil acesso à informação adicional relacionada ao material de estudo. O trabalho de Mihalcea & Csomai (2007) propõe um método de Elaboração Lexical que pode ser utilizado para relacionar parte do conteúdo do material escolar a enciclopédias ou notas de aula, por exemplo. O método proposto pelos autores, chamado de *Wikify!*, identifica conceitos importantes em um texto e conecta esses conceitos a uma página relacionada da Wikipedia, permitindo ao leitor entender um termo ou conceito desconhecido ou simplesmente se informar melhor sobre o assunto.

Amancio (2011) desenvolvou um trabalho de Elaboração Lexical para adultos com baixa escolaridade, no escopo do projeto PorSimples (Aluísio & Gasperin, 2010). Os autores propuseram dois métodos de Elaboração Lexical. O primeiro é baseado em um modelo que gera perguntas que explicitam a ligação existente entre o verbo da sentença (evocador) e suas constituintes (ou argumentos). O segundo modelo apresenta definições para Entidades Nomeadas a partir de consultas na Wikipédia.

Trieschnigg & Hauff (2011) analisaram livros de literatura infantil produzidos entre os anos 1.800 e 1.925 e perceberam que a escrita desses livros infantis variou, pois a língua varia com o passar do tempo, e esse processo faz com que palavras caiam em desuso. A divergência de palavras utilizadas chega a 42% comparando textos produzidos em 1.825 e textos produzidos em 1.925. Essa diferença de vocabulário indica que crianças no século 21 podem encontrar dificuldades ao ler esse tipo de material literário. Nesse cenário, Trieschnigg & Hauff (2011) propuseram um método de Elaboração Lexical que busca a definição de uma palavra difícil no Wiktionary[3] e apresentam ela ao leitor, auxiliando as crianças nativas da língua inglesa na leitura dessas histórias dos séculos passados.

2.2. Trabalhos de Simplificação Lexical

Devlin & Tait (1998) coordenaram o projeto PSET (*Practical Simplification of English Text*), o primeiro trabalho da literatura com foco na simplificação automática de textos, para torná-los acessíveis para pessoas com afasia. Os autores focaram na simplificação lexical de substantivos e adjetivos complexos, apenas. Para a identificação de palavras complexas, os autores consultaram a frequência das palavras no *Oxford Psycholinguistic Database* (Quinlan, 1992). Para a seleção de palavras candidatas a substituir a palavra complexa, buscaram todos os sinônimos da palavra complexa na WordNet (Fellbaum, 1998).

O projeto PorSimples[4] (Aluísio & Gasperin, 2010) foi o pioneiro em simplificação lexical e sintática para o Português. O público alvo do projeto foram adultos com baixa escolaridade. Os autores criaram uma lista de palavras simples composta de palavras do Dicionário Ilustrado de Português (Biderman, 2003) para crianças de 6 a 11 anos, acrescida de uma lista de palavras dos textos do jornal Zero Hora, Seção *Para seu Filho Ler* e de palavras concretas do trabalho de Janczura et al. (2007). Os autores definiram palavras complexas como aquelas não contidas na lista de palavras simples.

O trabalho de Bott et al. (2012) desenvolveu o sistema LexSiS, um sistema de Simplificação Lexical baseada em substituição por sinônimos, para textos em espanhol. Os autores também modelaram as palavras do léxico por meio de modelagem vetorial treinada sobre um córpus de 8M de palavras extraído da Web. Primeiro, os autores buscaram sinônimos para a palavra complexa no OpenThesaurus[5]. Então, eles geraram

[3] https://www.wiktionary.org

[4] Simplificação Textual do Português para Inclusão e Acessibilidade Digital.

[5] http://openoffice-es.sourceforge.net/

uma representação vetorial para o 10-gram centrado na palavra complexa, de forma a modelar o contexto em que ela ocorre. O substituto escolhido é o sinônimo cuja representação vetorial tenha o maior valor de similaridade pelo cosseno com a representação do contexto da palavra original. Essa abordagem também obteve resultados superiores a *baselines* que utilizavam apenas conhecimento lexical, selecionando o sinônimo mais frequente do OpenThesaurus como substituto. LexSiS foi incorporado no sistema Simplext (Saggion et al., 2015), que trata também a Simplificação Sintática para o espanhol, e é composto por três módulos: o módulo de Simplificação Sintática, o módulo LexSiS, que faz a Simplificação Lexical baseada em sinônimos, e um módulo de Simplificação Lexical baseado em regras que cobre simplificações não resolvidas pelos módulos anteriores como normalização de verbos de elocução, redução do conteúdo de sentenças, e redução, explicação ou normalização de informação numérica.

Kajiwara et al. (2013) desenvolveram um trabalho em Simplificação Lexical para o Japonês. Sabe-se que crianças estão em fase de aprendizado e exposição à língua falada e escrita e, portanto, elas possuem um vocabulário menor do que o dos adultos, em geral. Nesse sentido, há a necessidade de adequar textos de alguns gêneros para que as crianças estejam aptas a lê-los. Os autores simplificaram textos jornalísticos utilizando apenas palavras contidas no dicionário BVL (*Basic Vocabulary to Learn*)(Mutsuro & Toshihiro, 2002), em que as palavras complexas são as chaves (palavras de consulta) do dicionário e todas as palavras contidas na descrição da palavra complexa são candidatas a substituí-la.

O trabalho de Glavaš & Štajner (2015) apresenta o sistema Light-LS, que trata a tarefa de Simplificação Lexical por meio de uma abordagem não supervisionada de *Machine Learning*. Os autores advogam que o uso de córpus simplificados e recursos como *wordnets* para busca de sinônimos dificulta a implementação de sistemas de Simplificação Lexical naquelas línguas que não possuem córpus e recursos robustos como os encontrados para o inglês. Os autores ainda argumentam que não deveria ser necessário o uso de córpus específicos para a tarefa de Simplificação Lexical, já que palavras simples também são encontradas em córpus de propósito geral. Portanto, os autores propõem o uso de *word embeddings* (Mikolov et al., 2013) para a tarefa de Simplificação Lexical. A abordagem dos autores é independente de língua e necessita unicamente de um grande córpus para treinamento do modelo de *embeddings*. Foi usado o modelo de *embeddings* GloVe (Pennington et al., 2014) e os candidatos a substituirem a palavra complexa eram aqueles com menor distância do cosseno entre a *embedding* da palavra candidata e a da palavra complexa.

Apesar do aumento das iniciativas de pesquisas em Simplificação Lexical após a SemEval 2012 *Text Simplification shared task*, ainda não havia ferramentas que dessem apoio ao desenvolvimento de sistemas de Simplificação Lexical ponta a ponta (como o *pipeline* apresentado na Figura 1). O trabalho de Paetzold & Specia (2015) veio suprir essa demanda com o sistema LEXenstein, um *framework* para desenvolvimento de sistemas para Simplificação Lexical. Em um trabalho subsequente (Paetzold & Specia, 2017), os autores propuseram um método estado da arte para a tarefa de Simplificação Lexical, aperfeiçoando o método de Glavaš & Štajner (2015) e propondo um *ranker*, que recebe duas palavras e retorna qual delas é mais simples.

3. O córpus SIMPLEX-PB 3.0

O SIMPLEX-PB (Hartmann et al., 2018) (ou somente SIMPLEX) é um córpus originalmente concebido para avaliação de métodos de Simplificação Lexical em Português Brasileiro, criado e disponibilizado ao público como um esforço para fomentar a pesquisa na área. Ele contém 1.719 instâncias segundo a proporção de palavras de conteúdo encontradas no corpus (Hartmann et al., 2016): 56 % substantivos, 18% adjetivos, 18% verbos e 6% advérbios. A partir dessa distribuição, há ainda uma subdivisão igualmente distribuída para favorecer: palavras mais frequentes, palavras com maior número de sinônimos e palavras com mais sentidos. Ao todo, 757 palavras distintas foram identificadas como complexas para crianças em idade para cursar o Ensino Fundamental.

O córpus contém uma lista de sinônimos para cada palavra complexa. A geração dessa lista de sinônimos foi feita a partir de um processo de anotação realizado por três especialistas em linguística que trabalham com crianças. Dois deles possuem mestrado e o terceiro possui doutorado. Cada anotador filtrou, de uma lista de sinônimos previamente capturada do TeP (Thesaurus eletrônico para o Português do Brasil) (Maziero et al., 2008), quais palavras eram apropriadas para substituir a palavra complexa original. Eles também sugeriram substituições que não foram listadas. O especialista doutor ano-

thesaurus/

tou todas as frases e cada um dos especialistas com mestrado anotou metade delas em um procedimento duplo-cego. O Cohen Kappa (Cohen, 1960) foi de 0,74 para o primeiro par de anotadores e 0,72 para o segundo par.

No entanto, a primeira versão do SIMPLEX possuia várias limitações que impediam sua aplicabilidade, como palavras incorretamente marcadas como sinônimos para uma palavra complexa em seu contexto, baixa quantidade de sinônimos e a ausência da ordenação dos sinônimos pela sua simplicidade. Assim, surgiu o SIMPLEX-PB 2.0 (Hartmann et al., 2020), uma versão ampliada e aprimorada do SIMPLEX que foi submetida a várias rodadas de anotação manual para capturar com precisão as necessidades de simplificação de crianças de escolas de periferia. O córpus foi aprimorado com um incremento no número de sinônimos para as palavras-alvos complexas (7,31 sinônimos em média) e a introdução da ordenação dos sinônimos pela sua simplicidade, produzidas pelo próprio público-alvo — crianças entre 10 e 14 anos estudando em instituições públicas de periferia no Brasil.

Neste trabalho, disponibilizamos uma nova versão do córpus, denominada SIMPLEX-PB 3.0. Nessa nova versão, o córpus foi enriquecido com *features* linguísticas, *proxies* de complexidade lexical. A nova versão do SIMPLEX ainda conta com definições de suas palavras complexas e anotações de termos técnicos, informações que fazem com que o córpus também possa ser utilizado para estudos em Elaboração Lexical. Atualmente, o córpus conta com 52 colunas de informação. Um exemplo da estruturação do SIMPLEX-PB 3.0 pode ser visto na Figura 3 e detalhes dos três tipos de enriquecimento de dados são mostrados nas próximas seções.

3.1. Definições de palavras complexas

Tomando como base os trabalhos de Mihalcea & Csomai (2007), Amancio (2011) e Bulté et al. (2018), que trabalharam com a Elaboração Lexical por meio da inserção da definição das palavras complexas, decidimos estender o SIMPLEX com definições curtas para cada palavra complexa, possibilitando assim o uso do recurso para estudos de Elaboração Lexical.

Watanabe et al. (2010) mostrou que aproximadamente 73,5% dos artigos da Wikipedia em Português (Wikipédia) trazem a definição do conceito chave do artigo logo na primeira sentença. No entanto, verificamos que nem todas as palavras complexas do SIMPLEX são contempladas pela Wikipédia, o que não nos garantiria uma cobertura completa. Portanto, optamos por utilizar o Dicio[6], um dicionário de português contemporâneo, composto por mais de 400 mil palavras e que, para cada palavra, apresenta a sua definição, classificação gramatical, etimologia, divisão silábica, plural, sinônimos, antônimos, transitividade verbal, conjugação de verbos e rimas. O Dicio comtempla 100% das palavras complexas do SIMPLEX. Fazendo uso do Dicio, recuperamos a definição de cada palavra complexa do SIMPLEX. Uma etapa de pós-processamento foi necessária a fim de garantir que a definição inserida no SIMPLEX seja curta, direta e simples, removendo apostos e orações que não sejam as principais.

3.2. Anotações de termos técnicos

Enriquecemos, ainda, o SIMPLEX com anotações de quais palavras complexas são termos técnicos ou possuem um contexto específico de aplicação. Para isso, consultamos o Priberam[7], que retorna uma anotação e descrição do uso específico de certas palavras consultadas, por exemplo:

- Sanguíneo – [**Liturgia católica**]
  Pano que serve para limpar o cálice, na missa (purificador, sanguinho);
- Câmara – [**Anatomia**]
  Cavidade ou espaço anatómico (ex.: câmara do olho);
- *Hardware* – [**Informática**]
  Material físico de um computador;
- Figurado – [**Figurado**]
  Dócil, brando.

Entendemos que essas anotações podem ser bons indicativos de quais palavras devem ser elaboradas, mas um estudo com base em córpus é necessário para comprovar a hipótese.

3.3. *Features* linguísticas de complexidade lexical

Com o intuito de disponibilizar insumos úteis para pesquisas nas áreas de Identificação de Palavras Complexas e Simplificação Lexical, enriquecemos o SIMPLEX com 38 *features* linguísticas implementadas neste trabalho e que já foram utilizadas com sucesso por trabalhos da literatura. As *features* são:

[6] https://www.dicio.com.br
[7] https://dicionario.priberam.org

	palavra_dificil	sentença	sinônimos_ranqueados	termo_técnico	AoA	synsets
0	raias	O que você sabe sobre as *raias* ?	[arraias, raias]	False	6.390000	0
1	derme	`` A *derme* suína foi escolhida porque sua composição é 78 % compatível com a nossa" , conta a cientista .	[derme, pele, tecido]	True	6.850653	1
2	prestes	Você está *prestes* a conhecer uma ciência chamada astroquímica .	[próximo, em vias de, perto de, prestes]	False	6.900000	2
3	fabulosos	Eram seres *fabulosos* da mitologia grega , metade homem e metade cavalo , que habitavam as regiões da Arcádia (Peloponeso Central) e Tessália (sul da Macedônia) .	[incríveis, fantásticos, fabulosos]	False	6.770000	4
4	vilãs	`` As plantas não são *vilãs* .	[perigosos, vilãs, desprezíveis, vis, más]	False	6.540000	2
5	brusco	`` O movimento não precisa ser *brusco* para machucar .	[súbito, repentino, indelicado, violento, brusco]	False	5.650000	5
6	vizinhança	Com a ajuda dessa *vizinhança* seca , um deserto se formaria mais facilmente .	[concunvizinhança, imediação, proximidade, arredor, vizinhança]	False	7.000000	0
7	retém	Ele não *retém* a água e , por isso , é seco .	[segura, conserva, retém]	False	5.700000	2
8	ferocidade	Apesar do tamanho e da *ferocidade* do bicho , são raros os casos de ataques contra humanos .	[ferocidade, brabeza, braveza]	False	5.930000	0
9	fórmula	A *fórmula* química da água é H20 porque na sua composição há duas par tes de hidrogênio e uma parte de oxigênio .	[dosagem, fórmula, receita, récipe, prescrição]	True	6.850653	0

Figura 3: Recorte de dez linhas e seis colunas do SIMPLEX 3.0 para fins ilustrativos.

- **Contagem** – Frequência e diversidade contextual[8] das palavras e de seus lemas no córpus Leg2Kids (Hartmann & Aluísio, 2019) e no córpus de textos informativos infantis (Hartmann et al., 2016);
- **Lexicais** – Quantidade de caracteres e sílabas das palavras e de seus lemas (Devlin & Tait, 1998). Utilizamos o pacote *Pyphen* para cálculo do número de sílabas;
- **Wordnet** – Quantidade de sentidos, hiperônimos e hipônimos das palavras e de seus lemas na OpenWordNet-PT (Paiva et al., 2012; Crossley et al., 2011);
- **Psicolinguísticas** – Frequência subjetiva[9], idade de aquisição (Age of Acquisition — AoA, em Inglês), concretude e imageabilidade das palavras, do repositório *Psycholinguistic Properties of Brazilian Portuguese*[10] (dos Santos et al., 2017), ou de seus lemas quando a palavra não estiver no repositório (Hartmann et al., 2018);
- **Modelo de língua** – A $\log_{10}$ probabilidade das palavras e seus lemas em córpus, considerando como contexto as 3 palavras precedentes, no córpus de Hartmann et al. (2016).

Mais detalhes sobre as *features* linguísticas são apresentados na Seção 5.2.1.

[8]Diversidade contextual é número de documentos em que a palavra ou expressão ocorre.

[9]Frequência subjetiva é a estimativa do número de vezes que uma palavra é encontrada por indivíduos em sua forma escrita ou falada.

[10]http://nilc.icmc.usp.br/portlex/

4. Elaboração Lexical

Intuitivamente, podemos supor que palavras da língua geral devam ser simplificadas e as palavras técnicas ou de conceitos enciclopédicos devem ser elaboradas. O trabalho de Bulté et al. (2018), que propôs um método para a decisão sobre qual abordagem de Adaptação Lexical utilizar em cada caso de um texto, fez uso de Elaboração Lexical somente para palavras compostas, nomes próprios e para aquelas palavras identificadas como complexas mas que não tiveram sinônimos mais simples encontrados na base lexical Cornetto[11].

Nesta seção, apresentamos uma primeira análise em córpus de ocorrências de Elaboração Lexical por definição e estudo de quais palavras do SIMPLEX-PB 3.0 estão presentes entre essas palavras elaboradas.

4.1. Análise de Córpus para casos de Elaboração Lexical

Com o intuito de entender quais palavras devem ser elaboradas para crianças do Ensino Fundamental, realizamos um estudo em um recorte do córpus formado por matérias e artigos de novembro de 1990 a novembro de 2015 da revista Ciência Hoje das Crianças (CHC)[12], em busca de padrões de Elaboração Lexical, via definições. A CHC é uma revista de divulgação científica para crianças (entre 8 a 13 anos), criada em 1986 e editada pelo Instituto Ciência Hoje sob a responsabilidade da Sociedade Brasileira para o Progresso da Ciência (SBPC).

[11]http://wordpress.let.vupr.nl/cornetto/.

[12]http://chc.org.br/

O trabalho de Aluísio (1995) mostra que trazer uma definição curta ao lado de uma palavra/termo ajuda na familiarização do conceito da palavra do texto. A autora lista 5 tipos de definições:

- **Definição Formal** – apresenta os elementos semânticos termo, classe e caracteristicas;
- **Definição Semi-formal** – similar à formal, mas não apresenta a classe;
- **Definição por Substituição** introduz uma nova informação com um significado similar ao termo que foi introduzido, isto é, apresenta uma reformulação do termo;
- **Definicão por Ilustração** – pode ser subclassificada em Definição por Exemplificação e Definição por Particularização. Ambas as subclassificações são orientadas ao uso de aposto mas a primeira traz um exemplo ao contexto e a segunda especifica o elemento para auxiliar no seu entendimento;
- **Definição por Estipulação** – é encontrada unida aos outros tipos de definições e seu propósito é colocar limites de tempo, lugar, área de pesquisa ou de significado para a definição que a acompanha.

Analisamos 187 dos 2.503 artigos do córpus CHC e 26 das 72 reedições disponibilizadas em busca de ocorrências de elaboração por definição. Identificamos 126 ocorrências de definições, distribuidas ao longo de 59 artigos. Na Tabela 1, são apresentadas as estatísticas de cada tipo de Elaboração Lexical por definição identificado.

Com base nas 126 ocorrências de Elaboração Lexical identificadas manualmente na amostra do córpus CHC, levantamos padrões de elaboração para facilitar a busca de definições em novos córpus. A Tabela 2 lista os padrões identificados para cada tipo de Elaboração Lexical por definição.

Fazendo uso desses padrões, buscamos no córpus de textos informativos voltados para crianças do Ensino Fundamental compilado em Hartmann et al. (2016) por ocorrências de palavras complexas do SIMPLEX que casem com os padrões de elaboração elencados. Esse córpus contém 124.993 sentenças, das quais 23.790 apresentam ocorrências de alguma das 715 palavras difíceis (ou de suas flexões) do SIMPLEX.

Para geração das flexões, fizemos uso do UNITEX-DELAF (Dicionário de Palavras Simples Flexionadas para o Português Brasileiro) (Muniz, 2004).

Com base nos padrões de Elaboração Lexical por definição listados, identificamos apenas 294 sentenças contendo a palavra complexa na janela de até 5 *tokens* anteriores ao padrão identificado. Fizemos esse relaxamento, pois desejavamos ter uma maior cobertura inicial para depois filtrarmos manualmente os casos válidos. Esse relaxamento não foi aplicado para os padrões "palavra seguida por parêntese" e "palavra seguida por travessão".

Após uma análise das ocorrências identificadas, verificamos que somente 41 das 294 ocorrências eram de fato casos de Elaboração Lexical de uma palavra complexa do SIMPLEX. O padrão mais comumente encontrado foi o uso de parênteses e travessões para introduzir a definição de uma palavra (ver Tabela 3), o que implica no uso mais comum de definições por substituição (ver Tabela 4).

O córpus compilado em Hartmann et al. (2016) é composto, por exemplo, por livros didáticos e revistas como a Superinteressante e o Mundo Estranho. Estas revistas, por definição, apresentam conceitos e conhecimento de mundo para as crianças, sendo necessário e, inclusive, é parte do propósito do material apresentar e explicar o significado de palavras/conceitos.

Entendemos que as palavras complexas do SIMPLEX, extraídas de dicionários que são um recorte do léxico trabalhado nos ciclos escolares do Ensino Fundamental, limitaram nossa análise. Além disso, realizamos a busca com uma lista de apenas 31 padrões (cf. na Tabela 2), o que limitou a identificação de certos tipos de definição, como a formal, por exemplo, cujo padrão mais comum também traz casos não definitórios, como mostrado nos dois exemplos ilustrativos a seguir:

- O menino é legal. (não é um caso de Elaboração Lexical);
- Os mamutes eram quadrúpedes enormes, muito pesados e pouco agéis. (exemplo de Elaboração Lexical).

Das 41 ocorrências de elaboração identificadas, filtramos aquelas que possuem anotação de termo técnico no córpus SIMPLEX. Ao todo, 395 das 1.582 entradas do SIMPLEX (aproximadamente 25%) foram marcadas como termos complexos pela consulta ao Priberam. Verificamos que 22 ocorrências de elaboração possuem essa anotação, ou seja, aproximadamente metade dos

Tipo de definição	Ocorrências	Exemplo em córpus
Definição Formal	60	A **alavanca** é simplesmente uma barra rígida apoiada sobre um ponto fixo.
Definição por Substituição	25	(...) **perímetro**, isto é, qual é o resultado da soma dos lados do triângulo.
Definição Semi-formal	22	(...) **compostos voláteis** conhecidos como ácidos graxos de cadeia curta, ...
Definição por Substituição + Definição por Estipulação	8	(...) **Phalloceros**, que pode ser traduzido como "pênis com chifres" em grego...
Definição Formal + Definição por Estipulação	4	Na mitologia grega, **Medusa** era um monstro com o rosto de mulher...
Definição por Ilustração	3	(...) **aves de rapina**, como os gaviões, as corujas e os falcões...
Definição Semi-formal + Definição por Estipulação	3	A **sucuri-de-Marajó**, como o nome já diz, habita a ilha de Marajó...
Definição por Ilustração + Definição por Estipulação	1	(...) apresenta características tanto de dinossauros quanto de aves (...) Trata-se de uma espécie de **dino-ave**.

Tabela 1: Ocorrências de Elaboração Lexical identificadas em estudo em amotra do córpus CHC.

Tipo de definição	Padrão
Definição Formal	é a ideia de é considerado/(da) são considerados/(das)
Definição Semi-formal	(é/são) caracterizada(s) por (é/são) caracterizado(s) por pode ser definido(a) como podem ser definidos(das) como recebe esse nome
Definição por Substituição	isto é que quer dizer em outras palavras conhecido(a) como conhecidos(das) como chama(m) de "palavra seguida por parênteses" "palavra seguida por travessão"
Definição por Exemplificação	por exemplo tal como tais como
Definição por Particularização	em particular
Definição por Semi-formal com Particularização	como o nome já diz
Definição por Substituição com Estipulação	pode ser traduzido como

Tabela 2: Padrões de Elaboração Lexical utilizados na consulta de ocorrências de elaboração.

Padrão	Ocorrências
"palavra seguida por parêntese"	20
"palavra seguida por travessão"	14
conhecido como	3
conhecidas como	1
por exemplo	1
é considerado	1
é considerada	1

Tabela 3: Ocorrências de padrões de Elaboração Lexical por definição de palavras complexas do SIMPLEX em córpus.

Tipo de definição	Ocorrências
Definição por Substituição	38
Definição Formal	2
Definição por Exemplificação	1

Tabela 4: Ocorrências de padrões de Elaboração Lexical por tipo de definição de palavras complexas do SIMPLEX em córpus.

casos elaborados possuem a marcação de termo técnico no SIMPLEX. Temos um bom indicativo de que palavras técnicas costumam ser elaboradas. Entretanto, esse estudo precisa ser aprofundado via análise em grandes córpus que tenham definições já anotadas, como o Newsela[13], por exemplo, mesmo sendo este na língua inglesa.

5. Simplificação Lexical

Após o sucesso da primeira *shared task* da tarefa de Identificação de Palavras Complexas (CWI – *Complex Word Identification*, em Inglês) no *SemEval* de 2016, em 2018 aconteceu a segunda edição da tarefa na NAACL-HLT, no *BEA Workshop* (Tetreault et al., 2018).

A segunda edição da CWI foi uma competição na qual os participantes poderiam participar de duas tarefas: (i) classificar automaticamente palavras como sendo complexas, ou não, isto é, uma tarefa de classificação binária, ou (ii) predizer o grau de complexidade de uma palavra, ou seja, uma tarefa de classificação probabilística. A competição foi disponibilizada com *datasets* para 4 línguas (inglês, espanhol, alemão e francês). Para a anotação desses *datasets*, 10 falantes nativos de cada língua e 10 não-nativos deveriam ler um parágrafo e anotar as palavras que consideravam difíceis de serem compreendidas por crianças, falantes não nativos e pessoas com problemas de linguagem. Ao final, os *datasets* disponibilizavam sentenças (contextos), as palavras anotadas de cada sentença e dois rótulos: (i) o primeiro para a tarefa de classificação, com valor 1 caso a maioria dos anotadores tivesse identificado a palavra como difícil, e 0 caso contrário; e (ii) a média das anotações para a tarefa de classificação probabilística.

Com base na nossa experiência no CWI 2018[14] (Hartmann & dos Santos, 2018), em que obtivemos a segunda melhor colocação na tarefa de classificação e terceira melhor colocação na tarefa de classificação probabilística para a língua

[13] https://newsela.com/.

[14] Optamos por competir nas duas tarefas para a língua inglesa.

inglesa (Yimam et al., 2018), trouxemos tanto o conhecimento adquirido para o Português Brasileiro (PB), como as *features* utilizadas e os métodos que melhor desempenharam, realizando as adaptações necessárias em relação aos recursos disponíveis. A discussão segue na Seção 5.2.

No Brasil, atualmente, o Ensino Fundamental é dividido em duas etapas – do 1º ao 5º ano, e do 6º ao 9º ano. Os Parâmetros Curriculares Nacionais (1998) subdividem essas duas fases em quatro ciclos: 1º ao 3º ano, 4º e 5º anos, 6º e 7º anos e 8º e 9º anos. O PNLD (Programa Nacional do Livro Didático)[15], criado em 1985 pelo Ministério da Educação do Brasil, é uma iniciativa de amplo impacto na educação, pois objetiva a escolha, aquisição, e distribuição gratuita de livros didáticos para os alunos das escolas públicas do Ensino Fundamental. Desde 2001, o Programa passou a contemplar a lexicografia (da Graça Krieger, 2012), selecionando e adquirindo dicionários para os alunos dessa etapa de ensino. O PNLD, por sua vez, subdivide o Ensino Fundamental em 3 níveis de complexidade lexical, sendo eles: 1º ano (nível 1), 2º ao 5º anos (nível 2) e 6º ao 9º anos (nível 3). Além disso, o PNLD disponibilizou uma série de dicionários que contemplam o léxico a ser aprendido em cada etapa escolar. O trabalho de Hartmann et al. (2016) selecionou uma amostra dos dicionários recomendados pelo PNLD para compilar três recursos lexicais representativos dos léxicos desses níveis escolares:

- **Dicionário Tipo 1** – Composto pelas entradas do Dicionário Caldas Aulete Turma do Cocoricó, Lexikon, contabilizando 1.371 palavras;
- **Dicionário Tipo 2** – Composto pelas entradas do Dicionário Ilustrado de Português, compilado por Maria Tereza Camargo Biderman, da Editora Ática e Dicionário Escolar da Língua Portuguesa Ilustrado com a Turma do Sítio do Picapau Amarelo, Editora Globo, contabilizando 8.171 palavras diferentes;
- **Dicionário Tipo 3** – Composto pelas entradas do Minidicionário Contemporâneo da Língua Portuguesa de Caldas Aulete, Lexikon Editorial, contabilizando 29.970 palavras.

Eventuais interseções entre os léxicos dos dicionários foram tratadas. Se uma palavra é complexa para um ano escolar T+2, ela naturalmente é complexa para os anos escolares T e $T+1$. Assim, nos casos de interseções, mantivemos a palavra apenas no léxico referente ao dicionário de mais alto nível. Com isso, a volumetria final dos três léxicos indicativos dos anos escolares é:

- **Dicionário Tipo 1** – 1.363 palavras;
- **Dicionário Tipo 2** – 6.836 palavras;
- **Dicionário Tipo 3** – 22.085 palavras;

O mapeamento do conhecimento adquirido no CWI 2018 para o cenário do PB pôde ser feito graças aos dicionários do PNLD, alinhados com os níveis escolares do Ensino Fundamental, tomando como premissa que uma criança consulta um dicionário quando ela desconhece uma palavra. Assim, podemos assumir que os dicionários direcionados a um dado ano escolar contêm as palavras difíceis/complexas para as crianças neste nível.

Sabendo, ainda, que há uma progressão natural na aquisição lexical conforme os anos escolares avançam (Hartmann et al., 2016), é natural afirmarmos que as palavras do dicionário do tipo 3 são mais complexas que as palavras dos dicionários do tipo 2, e que essas são mais complexas do que as palavras do dicionários do tipo 1. Para a tarefa de Simplificação Lexical, podemos utilizar essas diferenças para aprendermos, com uso de métodos de *Machine Learning*, quais são as características que determinam a gradação da complexidade de uma palavra e, consequentemente, ranqueá-la de acordo com a sua complexidade frente a outras palavras.

5.1. *Datasets* compilados para a tarefa de Identificação de Palavras Complexas

Para capturarmos a complexidade lexical a partir de diferentes visões dos nossos três dicionários que representam os níveis de complexidade lexical do Ensino Fundamental apontados pelo PNLD, pareamos os dicionários para rotular suas palavras como "fáceis" ou "difíceis": Dicionário Tipo 1 com Dicionário Tipo 2 (Tipo1-Tipo2); Dicinário Tipo 1 com Dicionário Tipo 3 (Tipo1-Tipo3); e Dicionário Tipo 2 com Dicionário Tipo 3 (Tipo2-Tipo3). Os pareamentos dos dicionários nos dão diferentes perspectivas para mensurar a gradação da complexidade lexical conforme os anos escolares avançam e, com isso, há um maior espaço a ser explorado por métodos de *Machine Learning*.

Para cada um desses pares, criamos *datasets* com as palavras dos dicionários e suas ocorrências em sentenças do córpus de Hartmann et al. (2016), um córpus de textos escritos para serem material de leitura de crianças no Ensino Fundamental. Para cada *dataset*, anotamos aquelas pa-

[15] http://portal.mec.gov.br/seb/arquivos/pdf/relatorio_internet.pdf

lavras do dicionário de menor nível lexical como fáceis (valor 0) e as palavras do dicionário de maior nível lexical foram anotadas como difíceis (valor 1). As volumetrias dos *datasets* compilados e a quantidade de instâncias com palavras fáceis e difíceis são apresentados na Tabela 5. Dois exemplos de instâncias do *dataset* Tipo2-Tipo3 podem ser vistos na Tabela 6.

Dataset	Instâncias	Fáceis	Difíceis
Tipo1-Tipo2	201.902	100.525	101.377
Tipo1-Tipo3	142.157	100.525	41.632
Tipo2-Tipo3	148.174	103.820	44.354

Tabela 5: Estatísticas dos *datasets* compilados para a tarefa de Identificação de Palavras Complexas.

Sentença	Palavra do Dicionário	É complexa?
O livro conta com a história de um salvamento muito importante (...)	salvamento	0
A data coincide com o dia de Nossa Senhora Aparecida (...)	coincide	1

Tabela 6: Exemplos de anotação da complexidade lexical para criação de *dataset* para a tarefa de Identificação de Palavras Complexas.

Os *datasets* criados não são balanceados, assim, fizemos o balancemento dos dados por meio do *subsampling* da classe majoritária. Esse balanceamento consiste na seleção de instâncias da classe com maior ocorrência até equilibrarmos a volumetria com a da classe com menos ocorrências (He & Garcia, 2009).

Dividimos nosso córpus em três partes (Tabela 7): córpus de treinamento (≈60% das instâncias), córpus de desenvolvimento (≈20% das instâncias) e córpus de teste (≈20% das instâncias).

Dataset	**Treino**	**Desenvolvimento**	**Teste**
Tipo1-Tipo2	130.818	37.050	34.884
Tipo1-Tipo3	52.044	15.500	15.716
Tipo2-Tipo3	53.350	17.084	18.274

Tabela 7: Estatísticas dos *datasets* compilados para as etapas de treino, desenvolvimento e teste de métodos de classificação para a tarefa de Identificação de Palavras Complexas.

5.2. Identificação de Palavras Complexas

5.2.1. Método com features linguísticas

Assim como no CWI 2018, desenvolvemos uma solução de *Machine Learning* utilizando *features* linguísticas a partir da palavra alvo e do seu contexto. Fazendo o devido mapeamento de recursos lexicais do PB dos quais calculamos as *features*, foi possível replicar para o PB o conjunto de *features* que obtiveram boa performance no inglês:

- **Contagem** – Frequência e diversidade contextual das palavras e de seus lemas no córpus Leg2Kids (Hartmann & Aluísio, 2019) e no córpus de textos informativos infantis (Hartmann et al., 2016);
- **Lexicais** – Quantidade de caracteres e sílabas das palavras e de seus lemas;
- **Wordnet** – Quantidade de sentidos, hiperônimos e hipônimos das palavras e de seus lemas na OpenWordNet-PT (Paiva et al., 2012);
- **Psicolinguísticas** – Frequência subjetiva, idade de aquisição, concretude e imageabilidade das palavras (dos Santos et al., 2017), ou de seus lemas quando a palavra não estiver contabilizada no recurso;
- **Modelo de língua** – A $\log_{10}$ probabilidade das palavras e seus lemas em córpus, considerando como contexto as três palavras precedentes, no córpus de Hartmann et al. (2016).

A literatura também aponta as *features* selecionadas como bons indicativos de complexidade lexical. Devlin & Tait (1998) utilizou métricas de contagem básicas, como a frequência das palavras e a quantidade de caracteres. Essas *features* se mostraram bons *proxies* para a tarefa de Simplificação Lexical. De Belder & Moens (2010) avaliou o uso da frequência das palavras na simplificação de textos para crianças nativas da língua inglesa. Hartmann & Aluísio (2019) fez uso da frequência e diversidade contextual para a tarefa de Identificação de Palavras Complexas no PB e também obteve bons resultados. Crossley et al. (2011) comenta que palavra pouco polissêmicas (que possuem poucos sentidos) e palavras muito específicas (que possuem poucos hipônimos associados) são indicativos de complexidade lexical e, assim, motiva o uso de *features* de *wordnets*. Hartmann et al. (2018) utilizaram *features* psicolinguísticas na tarefa de Simplificação Lexical. Horn et al. (2014) e Paetzold & Specia (2016) utilizaram modelos de língua na tarefa de Simplificação Lexical.

Um total de 19 *features* foram desenvolvidas. Aplicamos o *zipf score* ($log_{10}(x)$+3) para todas as *features* desenvolvidas, exceto aquelas de modelo de língua. Com isso, iniciamos o treinamento de métodos de *Machine Learning* com um total de 38 *features*.

Sabemos que nem todas essas informações são necessariamente úteis. Algumas podem não explicar o evento de uma palavra ser simples ou complexa. Outras podem ser correlacionadas entre si, ou seja, redundantes. Portanto, rodamos o método Boruta (Kursa et al., 2010; Kursa & Rudnicki, 2010) de seleção de variáveis. O Boruta verifica quais *features* são mais informativas para explicarem o evento de interesse do que uma variável aleatória produzida a partir do embaralhamento da própria *feature*. Se uma *feature* explica um evento, ela está correlacionada com o fato de uma palavra ser simples ou complexa, mas se embaralharmos essa *feature*, ela perde a correlação com o evento e passa a não mais explicá-lo. O Boruta eliminou 8 *features* para os três datasets.

A justificativa de escolher o Boruta dentre outros métodos de seleção foi devido ao fato do algoritmo ser projetado para classificar o que o artigo original chama de "problema todas relevantes": encontrar um subconjunto de *features* que são relevantes para uma determinada tarefa de classificação. Isso é diferente do "problema mínimo-otimo", que é o problema de encontrar o subconjunto mínimo de *features* que têm desempenho em um modelo. Embora os modelos de aprendizado de máquina em produção devam, em última análise, visar a seleção de *features* mínimas otimas, a tese de Boruta é que, para fins de exploração, a otimização mínima vai longe demais. Além disso, o método é robusto à correlação de *features*. Em cenários com uma quantidade grande de *features*, tratar a correlação delas pode ser uma tarefa demasiadamente custosa. Assim, utilizar o Boruta pode também acelerar a etapa de preparação de *features*, justificando sua escolha nesta pesquisa.

Em seguida, calculamos a Correlação de Pearson entre cada par de *feature* para identificarmos *features* correlacionadas. Nos casos em que a correlação foi maior ou igual a 0,9, mantivemos apenas uma *feature* do par analisado. Com isso, removemos mais 14 *features* dos *datasets* Tipo1-Tipo2 e Tipo2-Tipo3; e 12 *features* do *dataset* Tipo1-Tipo3.

Assim, 16 *features* linguísticas foram selecionadas para o treinamento de métodos de *Machine Learning* nos *datasets* Tipo1-Tipo2 e Tipo-Tipo3; e 18 *features* foram selecionadas para o *dataset* Tipo2-Tipo3. Como os *datasets* são distintos, não necessariamente as mesmas *features* foram selecionadas em todos eles. Na Tabela 8, são listadas as *features* selecionadas para representar cada *dataset*.

Selecionamos quatro métodos de classificação baseados em *Machine Learning*: a Regressão Logística (método tradicional e comumente utilizado como *baseline* de soluções), o SVM, a Random Forest (*ensemble* do tipo *bagging*) e XGBoost (*ensemble* do tipo *boosting*). Foi realizada otimização bayesiana de hiper-parâmetros para todos os métodos com uso do pacote Hyperopt (Bergstra et al., 2013).

5.2.2. Método com word embeddings

Utilizamos a mesma arquitetura de rede neural (ver Figura 4) implementada para o método que fez uso de *word embeddings* no CWI 2018, tendo sido apenas necessária a substituição do modelo de *word embeddings* por um treinado no PB.

O fluxo de processamento de uma dada palavra pela rede neural é o seguinte: a *embedding* de uma palavra alimenta a entrada da rede, que segue com 2 camadas densas de 100 neurônios cada e função de ativação ReLu (Nair & Hinton, 2010). Por fim, uma última camada com um único neurônio e função de ativação sigmóide que retorna um *score* em termos de probabilidade entre 0 e 1. Quanto mais próximo de 1, mais complexa é a palavra. A rede é treinada por 10 épocas.

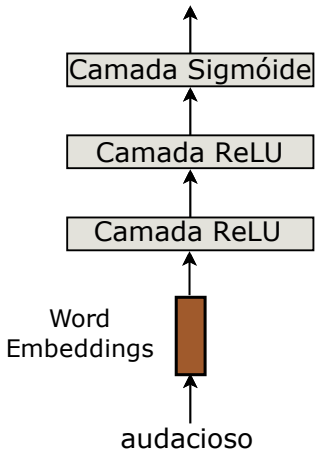


Figura 4: Arquitetura de rede neural com *word embeddings* de Hartmann & dos Santos (2018).

Para o CWI 2018, fizemos uso do modelo de *word embeddings* GloVe (Pennington et al., 2014) com 100 dimensões, sem avaliar variações do modelo com mais ou menos dimensões (cada dimensão pode ser interpretada como uma *feature*), nem outras abordagens de *word embeddings*.

Tipo1-Tipo2	Tipo1-Tipo3	Tipo2-Tipo3
Idade de aquisição	Idade de aquisição	Idade de aquisição
Concretude	Concretude	Concretude
Imageabilidade	Imageabilidade	Imageabilidade
Modelo de língua	Modelo de língua	Modelo de língua
Modelo de língua lema	Modelo de língua lema	Modelo de língua lema
Frequência subjetiva	Frequência subjetiva	Frequência subjetiva
Diversidade contextual no Leg2Kids	Diversidade contextual no Leg2Kids	Diversidade contextual no Leg2Kids
Diversidade contextual no Leg2Kids lema	Diversidade contextual no Leg2Kids lema	Diversidade contextual no Leg2Kids lema
Caracteres	Caracteres	Caracteres
Caracteres do lema	Caracteres do lema	Caracteres do lema
Frequência no córpus informativo infantil	Frequência no córpus informativo infantil	Frequência no córpus informativo infantil
Frequência no córpus informativo infantil zipf	Frequência no córpus informativo infantil zipf	Frequência no córpus informativo infantil zipf
Sentidos	Frequência no córpus Leg2Kids	Frequência no córpus Leg2Kids
Sentidos lema	Frequência no córpus Leg2Kids zipf	Frequência no córpus Leg2Kids zipf
Sentidos lema zipf	Sentidos	Sentidos
Sentidos zipf	Sentidos lema	Sentidos lema
	Sentidos lema zipf	
	Sentidos zipf	

Tabela 8: *Features* linguísticas selecionadas de cada *dataset* para treinamento dos modelos de Identificação de Palavras Complexas.

O trabalho de Hartmann et al. (2017) mostrou que não é trivial inferir a performance global de um modelo de *word embeddings*, ou seja, sua performance deve ser analisada caso a caso, tarefa a tarefa. Portanto, avaliamos aqui todos os modelos pré-treinados[16] de *word embeddings* pelo grupo de pesquisa NILC (Hartmann et al., 2017), sendo eles: Word2Vec (Mikolov et al., 2013), Wang2Vec (Ling et al., 2015), GloVe (Pennington et al., 2014) e FastText (Joulin et al., 2016), com variações de 50, 100, 300, 600 e 1.000 dimensões. No geral, os modelos com maior dimensionalidade possuem maior custo computacional, o que limita o seu uso, mas empiricamente foi observado que esse custo se paga por apresentarem melhores resultados quando aplicados (Hartmann et al., 2017).

5.2.3. Método com embedding contextual - Elmo

No CWI 2018, avaliamos a LSTM (Gers et al., 1999; Le et al., 2017), rede neural estado da arte na época para representação contextual. A rede neural foi pré-treinada como modelo de língua no *One Billion Word dataset* (Chelba et al., 2014), o que lhe deu a capacidade de aprender a representar sentenças, ou seja, contextos.

Essa rede foi então utilizada para a tarefa de Identificação de Palavras Complexas (ver Figura 5). Usando palavra por palavra de uma sentença, a rede é alimentada até atingir a palavra alvo de interesse (aquela que desejamos classificar como fácil ou difícil). Nesse momento, fizemos uso da representação produzida pela LSTM (*embedding* que codifica o contexto analisado) e passamos essa *embedding* por uma camada composta por um único neurônio e função de ativação sigmóide,

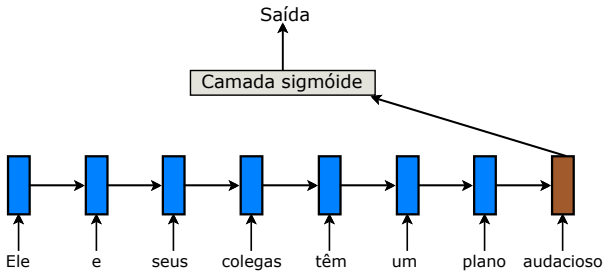


Figura 5: Arquitetura de rede neural com *embedding* contextual de Hartmann & dos Santos (2018).

retornando um *score* em termos de probabilidade entre 0 e 1. A rede é treinada por 10 épocas.

Atualmente, novos modelos para representação contextual foram propostos e um novo estado da arte para a representação contextual foi estabelecido, sendo a rede LSTM precurssora desses avanços. O trabalho de Peters et al. (2018) propôs o Elmo, o primeiro modelo de *embeddings* contextuais da literatura. A arquitetura do Elmo faz uso de uma rede convolucional para representação dos caracteres de cada palavra e, então, duas redes LSTM que representam o contexto analisado.

Recentemente, Castro (2019) treinou dois modelos Elmo para o PB. Um desses modelos foi treinado numa coleta da Wikipédia, contendo aproximadamente 267 milhões de *tokens*. O outro modelo foi treinado no córpus BrWac (Wagner Filho et al., 2018), que é um córpus compilado a partir de textos da web e contém 2,7 bilhões de *tokens*. Avaliamos ambos os modelos neste trabalho.

A *embedding* produzida pelo Elmo contém três dimensões com 1.024 valores cada. Segundo os autores do artigo original, especula-se que

[16] `nilc.icmc.usp.br/embeddings`

a primeira dimensão captura informações morfológicas dado o seu processamento no nível dos caracteres das palavras; a segunda camada captura informações sintáticas; e a terceira camada captura informações semânticas. Apesar da tarefa de Identificação de Palavras Complexas ter um alto caráter morfológico, o contexto no qual a palavra está inserido é altamente relevante. Logo, avaliamos aqui o uso da representação de cada camada independentemente, bem como o uso agregado das três camadas (concatenação delas) e ainda a média das camadas. Fazemos uso da mesma arquitetura de rede neural utilizada com as *embeddings* da LSTM no CWI 2018, substituindo essas *embeddings* pelas produzidas pelo Elmo.

5.2.4. Método com embedding contextual - BERT

As redes neurais recorrentes, como a LSTM e a GRU (Cho et al., 2014), que conseguiram grande destaque na literatura por possuirem a capacidade de representar um contexto e terem inclusive inspirado a criação do Elmo, possuem uma limitação para representar sequências longas (Devlin et al., 2018). Isso se dá porque essas redes possuem um conceito de memória e, para continuarem mantendo o contexto atual processado (memória curta), elas acabam deixando de representar as palavras mais antigas do texto (memória longa). A memória dessas redes limita seu uso na representação de sentenças longas e, principalmente, parágrafos ou texto.

Para suprir essa deficiência, o trabalho de Vaswani et al. (2017) introduziu um novo conceito de rede neural: a Rede com Atenção. Essa rede possui um mecanismo que verifica, para cada palavra, qual a sua relevância na representação de todo o conteúdo processado (sentença, parágrafo ou texto). Esse conceito motivou o trabalho de Devlin et al. (2018), que introduziu o BERT, um novo modelo de *embedding* contextual. Juntamente com o artigo, os autores disponibilizaram dois modelos pré-treinados em um córpus composto pelas Wikipedias de 104 línguas: o BERT Base, modelo treinado a partir de uma rede neural com 12 camadas e que produz uma *embedding* de 768 valores; e o BERT Large, uma rede neural mais profunda, com 24 camadas, e que produz uma *embedding* de 1.024 valores.

Recentemente, Souza et al. (2019) treinou essas duas versões do BERT no córpus BrWAC (Wagner Filho et al., 2018), o mesmo utilizado para treinamento de uma das *embeddings* do Elmo para o PB, e disponibilizou os modelos.

Neste trabalho, avaliamos o uso de ambas as *embeddings* do BERT (Base e Large) na tarefa de Identificação de Palavras Complexas. Fazemos uso da mesma arquitetura de rede neural utilizada com as *embeddings* do Elmo, substituindo essas *embeddings* pelas produzidas pelo BERT.

5.3. Avaliação dos Métodos propostos para Identificação de Palavras Complexas

Todo método de classificação binária (aquele que classifica uma instância como 0 ou 1) entrega um número real (um *score*) entre 0 e 1, assim como a tarefa de classificação probabilística do CWI 2018. A conversão desse *score* para os valores 0 ou 1 (uma tarefa de classificação tradicional da literatura) é feita a partir de um ponto de decisão (comumente, esse valor é 0,5), em que valores abaixo do ponto de decisão são arredondados para 0 e os valores iguais ou acima são arredondados para 1.

Para ordenarmos uma lista de palavras pela sua simplicidade, precisamos fazer uso dos *scores* gerados pelo modelo e não das predições arredondadas. Para tanto, analisamos a métrica AUC, pois ela nos dá uma interpretabilidade quanto a qualidade da ordenação dos *scores* de predição da complexidade de cada palavra: é esperado que palavras simples possuam um *score* menor que os das palavras complexas. Além de averiguar a qualidade da ordenação, a AUC é uma métrica de avaliação da qualidade de classificadores pois, se os *scores* preditos estão ordenados pela complexidade das palavras, podemos ter um ponto de decisão para arredondar esses *scores*, mapeando as predições para as classes "fácil" (valor 0) ou "difícil" (valor 1).

Calculamos também a métrica F1, por ela ser a métrica mais utilizada na avaliação de classificadores. Para o cálculo da F1, fazemos uso do ponto de decisão que maximiza a métrica no *dataset* de desenvolvimento. As métricas AUC e F1 são reportadas nos *datasets* de teste, que são aqueles não utilizados no treinamento dos modelos nem na identificação do melhor ponto de decisão.

As performances dos métodos com *features* linguísticas são apresentadas na Tabela 9. Nos três *datasets*, os métodos foram capazes de classificar a complexidade das palavras com uma alta acertividade. Os métodos de *ensemble* foram os que desempenharam melhor, apesar da boa performance dos métodos tradicionais. Os melhores resultados nos três *datasets* foram obtidos pelo método XGBoost, o que é de se esperar, dado que

mais da metade das melhores soluções em competições do Kaggle (até meados de 2016, pelo menos) foram obtidas com uso desse método (Chen & Guestrin, 2016).

Dataset	Reg. Logística		SVM		R. Forest		XGBoost	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Tipo1-Tipo2	0,82	0,75	0,83	0,77	0,88	0,79	**0,93**	0,84
Tipo1-Tipo3	0,97	0,91	0,97	0,92	**0,99**	0,95	**0,99**	0,96
Tipo2-Tipo3	0,88	0,78	0,87	0,78	0,89	0,79	**0,91**	0,81

Tabela 9: Performance dos métodos treinados com *features* linguísticas na tarefa de Identificação de Palavras Complexas.

Os resultados dos modelos treinados com uso de *word embeddings* como *features* são apresentados na Tabela 10. Como avaliamos 5 variações de dimensionalidade (50, 100, 300, 600 e 1.000 dimensões) de 4 diferentes técnicas para geração de *word embeddings* (Word2Vec, Wang2Vec, GloVe e FastText) em três *datasets* diferentes, totalizamos o treinamento e teste de 60 modelos. Considerando a volumetria de resultados produzidos, optamos por apresentar somente o melhor resultado de cada técnica de *word embedding* nos três *datasets* de teste.

Percebemos que não houve a predominância de técnica de *word embedding* entre os modelos que desempenharam melhor. Para os *datasets* Tipo1-Tipo2 e Tipo1-Tipo3, os modelos que utilizaram FastText (CBOW com 600 dimensões e SkipGram com 300 dimensões, respectivamente) obtiveram os melhores resultados. Para o *dataset* Tipo2-Tipo3, o melhor resultado foi obtido pelo modelo que utilizou GloVe com 600 dimensões. Essa não predominância está alinhada com os resultados de Hartmann et al. (2017), que mostrou a não trivialidade em inferir a performance global de uma *word embedding*, fazendo-se necessária a avaliação do seu uso em cada tarefa de interesse.

Em relação à dimensionalidade das *embeddings* utilizadas, os melhores modelos fizeram uso de *embeddings* com 300 dimensões ou mais. Em sua maioria, os modelos que desempenharam melhor utilizaram *embeddings* de 600 ou 1.000 dimensões, o que reforça a maior informatividade das *embeddings* com mais dimensões.

Os resultados dos métodos treinados com *embeddings* do Elmo são apresentados na Tabela 11. Treinamos modelos com uso da primeira, segunda e terceira camadas de *embedding* produzidas pelo Elmo, bem como com a concatenação e média dessas camadas. Avaliamos dois modelos pré-treinados de *embeddings* do Elmo, um treinado na Wikipédia e outro treinado no BrWac. Isso totaliza 30 modelos treinados e avaliados em nossos três *datasets*. Assim como feito na apresentação dos resultos dos modelos de *word embeddings*, apresentamos aqui apenas os resultados dos melhores experimentos para cada um dos dois modelos pré-treinados de *embeddings* do Elmo.

Dataset	Word2Vec		Wang2Vec		FastText		GloVe	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Tipo1-Tipo2	SkipGram 600		SkipGram 300		CBOW 600		GloVe 600	
	0,88	0,86	0,92	0,86	**0,94**	0,87	0,89	0,81
Tipo1-Tipo3	SkipGram 300		SkipGram 600		SkipGram 1000		GloVe 600	
	0,95	0,88	0,97	0,90	**0,98**	0,94	**0,98**	0,92
Tipo2-Tipo3	SkipGram 600		SkipGram 1000		SkipGram 1000		GloVe 600	
	0,84	0,79	0,84	0,77	0,88	0,81	**0,91**	0,83

Tabela 10: Performance dos métodos treinados com *word embeddings* na tarefa de Identificação de Palavras Complexas.

Por unanimidade, os melhores modelos foram aqueles que fizeram uso da concatenação das três camadas de *embeddings* do Elmo. Isso mostra que a mensuração da complexidade lexical de uma palavra extrapola o nível morfológico, dependendo também da validade da palavra no contexto inserido. Esse resultado está alinhado com as questões levantadas nos trabalhos Henderson et al. (2013) e de Sousa et al. (2020), em que argumentam sobre a importância da manutenção de um léxico adequado ao público alvo e que a não compreensão desse léxico pode levar o leitor a não compreender o contexto lido como um todo.

Dataset	Elmo Wikipédia		Elmo BrWac	
	AUC	F1	AUC	F1
Tipo1-Tipo2	3 camadas concatenadas		3 camadas concatenadas	
	0,98	0,94	**0,98**	0,92
Tipo1-Tipo3	3 camadas concatenadas		3 camadas concatenadas	
	0,99	0,94	0,97	0,90
Tipo2-Tipo3	3 camadas concatenadas		3 camadas concatenadas	
	0,96	0,89	0,95	0,84

Tabela 11: Performance dos métodos treinados com *embeddings* do Elmo na tarefa de Identificação de Palavras Complexas.

Os resultados dos métodos treinados com *embeddings* geradas pelo BERT são apresentados na Tabela 12. Como aqui não houve muitas combinações de experimentos, apresentamos as performances de todos os 6 modelos treinados. Os melhores resultados nos três *datasets* foram obtidas pelas *embedding* do BERT Large, o que é de se esperar, já que a rede neural usada no treinamento dessas *embeddings* possui o dobro de camadas em relação ao BERT Base, o que lhe dá maior poder de aprendizado.

Na Tabela 13, apresentamos o consolidado dos métodos que melhor desempenharam em cada uma das 4 categorias de *features* avaliadas: featu-res linguísticas, *word embeddings*, as *embeddings* contextuais do Elmo e as do BERT.

Dataset	BERT Base		BERT Large	
	AUC	F1	AUC	F1
Tipo1-Tipo2	0,91	0,83	**0,93**	0,86
Tipo1-Tipo3	0,92	0,86	**0,97**	0,92
Tipo2-Tipo3	0,91	0,84	**0,93**	0,86

Tabela 12: Performance dos métodos treinados com *embeddings* do BERT na tarefa de Identificação de Palavras Complexas.

Consideramos a métrica AUC na seleção do método com melhor performance. Em caso de empate, selecionamos o método que obteve maior valor de F1.

Os métodos treinados com uso das *embeddings* obtidas pelo Elmo obtiveram, consistentemente, os melhores resultados nos três *datasets*. Nossa leitura desses resultados remete ao trabalho de Hartmann & dos Santos (2018), que contrasta as abordagens de *Feature Engineering*: engenharia de *features*, ou seja, a construção manual de variáveis que representem o evento desejado; e *Feature Learning*: o aprendizado automático das informações representativas do evento de interesse.

Em relação às etapas de Identificação de Palavras Complexas e Simplificação Lexical, trabalhos recentes têm mostrado que métodos que fazem uso de *Feature Learning* estão desempenhando melhor do que os métodos que utilizam *Feature Engineering* (Glavaš & Štajner, 2015; Paetzold & Specia, 2017; Hartmann & dos Santos, 2018; Štajner et al., 2019). Esse cenário está alinhado com os resultados obtidos nesta avaliação. Ainda assim, é importante destacar a boa performance dos métodos que utilizam *features* linguísticas. Esses métodos obtiveram resultados próximos (em alguns cenários melhores, em outros piores) aos dos métodos de *word embeddings* e também do BERT.

Em um primeiro momento, poderíamos esperar que os modelos que fizeram uso das *embeddings* do BERT obteriam os melhores resultados em nossos experimentos, já que esse modelo de *embeddings* contextuais veio suprir limitações que persistem no modelo do Elmo. No entanto, vale novamente a ressalva de que é difícil inferir a performance global de uma *embedding* (agora extrapolando para as *embeddings* contextuais). Enquanto o BERT é altamente contextual, já que foi desenvolvido para melhor representar textos longos em relação aos modelos anteriores, o Elmo nos mune de informações morfológicas (primeira camada), além de informações contextuais (segunda e terceira camadas). Assim, por mais que o BERT tenha a capacidade de capturar as relações contextuais em que a palavra está inserida, o Elmo ainda possui um alto teor morfológico que vai além das capacidades do BERT.

Dataset	*Features* Linguísticas		*Word Embed.*		Elmo		BERT	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Tipo1-Tipo2	0,93	0,84	0,94	0,87	**0,98**	0,94	0,93	0,86
Tipo1-Tipo3	0,99	0,96	0,98	0,94	**0,99**	0,94	0,97	0,92
Tipo2-Tipo3	0,91	0,81	0,91	0,83	**0,96**	0,89	0,93	0,86

Tabela 13: Performance dos melhores métodos de cada categoria explorada na tarefa de Identificação de Palavras Complexas.

5.4. Aplicação dos métodos de Identificação de Palavras Complexas na tarefa de Simplificação Lexical

Segundo os métodos mais bem sucedidos da literatura de Simplificação Lexical, para simplificarmos uma palavra devemos identificar sinônimos e ranqueá-los pela sua adequação ao contexto e, claro, pela sua simplicidade (Paetzold & Specia, 2017; Štajner et al., 2019). Assim, propomos a aplicação dos métodos de Identificação de Palavras Complexas, desenvolvidos neste trabalho, na tarefa de Simplificação Lexical, já que eles levam tanto a complexidade de uma palavra quanto o seu contexto em consideração.

Para avaliação da tarefa de Simplificação Lexical, fazemos uso do SIMPLEX-PB 3.0. Como descrito na Seção 3, a nova versão do SIMPLEX passou por uma etapa de ordenação dos sinônimos das palavras complexas, realizada pelo próprio público alvo - crianças cursando o Ensino Fundamental. Sabendo quais palavras são complexas, qual o contexto em que as palavras estão inseridas e a ordenação dos sinônimos, podemos avaliar quais métodos produzem resultados mais alinhados com a expectativa das crianças. Para isso, embaralhamos as palavras complexas entre os seus sinônimos e predizemos a complexidade das listas de palavras com os melhores métodos obtidos na Seção 5.3.

Nosso método com *features* linguísticas representa a abordagem clássica de Simplificação Lexical (Devlin & Tait, 1998; Aluísio & Gasperin, 2010), que até o inicío da segunda década do século, representou o que havia de melhor na literatura. Nosso método que faz uso de *word embeddings* representa uma evolução do trabalho de Glavaš & Štajner (2015) que, pela primeira vez, obteve resultados melhores do que aqueles que faziam uso de *features* linguísticas e mostrou o potencial da abordagem de *Feature Learning*. Os métodos que fazem uso das *embeddings* contex-

tuais do Elmo e do BERT representam o atual estágio do PLN e, como apresentado por Štajner et al. (2019), são os com maior potencial para a tarefa de Simplificação Lexical.

Nossos métodos são comparados com 5 *baselines*, sendo eles:

- Frequência da palavra no córpus Leg2Kids (Hartmann & Aluísio, 2019);
- Frequência da palavra no córpus de textos informativos para crianças do Ensino Fundamental (Hartmann et al., 2016);
- Diversidade Contextual da palavra no córpus Leg2Kids;
- Idade de Aquisição da palavra (dos Santos et al., 2017);
- Método de Glavaš & Štajner (2015).

A frequência e diversidade contextual são métricas conhecidas por serem bons *proxies* de simplicidade, como atestado por Hartmann & Aluísio (2019). Para ambas as métricas, fazemos sua consulta no córpus Leg2Kids, um córpus de legendas de desenhos animados e filmes do gênero familiar e, para a frequência, também fazemos sua consulta no córpus de textos informativos voltados para crianças do Ensino Fundamental. A idade de aquisição é outro *proxy* de simplicidade (Hartmann et al., 2018) e também o utilizamos como *baseline*. Por fim, para termos um método robusto para comparação, implementamos a solução de Glavaš & Štajner (2015), que ranqueia os sinônimos pela similaridade pelo cosseno entre a *embedding* Glove de 300 dimensões dos sinônimos com a da palavra complexa.

Aplicamos todos os nossos métodos individualmente, ranqueando as palavras e comparando-os com a ordenação fornecida pelas crianças. Também avaliamos algumas combinações de métodos, conceito aplicado por Hartmann et al. (2018), ao calcularmos o *ranking* médio de diferentes soluções. A combinação de modelos visa avaliar se um cenário desempenha melhor do que outro e também se o uso combinado desses modelos aumenta a performance geral da predição. Os cenários de combinação de métodos são:

- Média dos modelos com *features* linguísticas (modelos com *Feature Engineering*);
- Média dos modelos com uso de *embeddings* (*word embeddings*, Elmo e BERT – *Feature Learning*);
- Média de todos os modelos treinados no *dataset* Tipo1-Tipo2;
- Média de todos os modelos treinados no *dataset* Tipo1-Tipo3;
- Média de todos os modelos treinados no *dataset* Tipo2-Tipo3.

A Tabela 14 apresenta os resultados do nosso experimento, avaliando os métodos de Identificação de Palavras Complexas na tarefa de Simplificação Lexical. Todos os métodos foram avaliados em 4 critérios:

- **Top 1** – A palavra ranqueada como mais simples coincide com a palavra identificada como mais simples pelas crianças;
- **Top 1 e 2** – A palavra ranqueada como mais simples está entre as duas palavras mais simples identificadas pelas crianças;
- **Top 1, 2 e 3** – A palavra ranqueada como mais simples está entre as três palavras mais simples identificadas pelas crianças;
- ρ – Correlação de Spearman entre o *ranking* de palavras produzido e o *ranking* esperado pelas crianças. Valores próximos a -1 indicam correlação inversa, valores próximos a zero indicam a falta de correlação e valores próximos a 1 indicam correlação perfeita.

Ao analisarmos a correlação de Spearman (ρ) dos nossos resultados, percebemos que, em linhas gerais, os valores estão muito próximos a zero, o que indica falta de correlação com o ranqueamento produzido pelas crianças. No entanto, é importante destacarmos que a correlação obtida entre as próprias crianças no SIMPLEX foi muito baixa (Hartmann et al., 2020). Nas instâncias em que 2 crianças ranquearam a palavra complexa e seus sinônimos, o ρ foi de 0,05. Já nos casos em que 3 crianças ranquearam uma mesma instância, o ρ foi de 0,16. Assumindo o ρ das crianças como um *upperbound*, não é esperado que tenhamos um ρ com o ranqueamento das crianças maior do que as crianças obtiveram entre elas.

Sabemos que, durante a anotação do SIMPLEX, as crianças concordaram em 81,5% das palavras ranqueadas como as mais simples e em 58,5% das 2 palavras mais simples. Nesse cenário de maior concordância entre as crianças, temos mais garantias para comparação da performance das nossas predições em relação a expectativa das crianças.

Analisando o Top 1, o *baseline* que melhor desempenhou foi a Idade de Aquisição, ainda que por uma margem muito pequena. Esse resultado está correlacionado com os próprios ciclos do Ensino Fundamental, pois conforme a idade

Modelo	Categoria	Top 1	Top 1 ou 2	Top 1, 2 ou 3	ρ
Média Tipo2-Tipo3	Média dos rankings	**0.392**	0.640	0.817	0.155
Média Tipo1-Tipo2	Média dos rankings	0.387	**0.646**	0.811	**0.164**
Média Tipo1-Tipo3	Média dos rankings	0.375	0.630	0.812	0.129
Média	Média dos rankings	0.359	0.606	0.804	0.094
Elmo Tipo1-Tipo2	Elmo Wikipedia	0.354	0.623	0.809	0.115
Média Embeddings	Média dos rankings	0.352	0.638	0.811	0.046
Linguísticas Tipo2-Tipo3	XGBoost	0.350	0.638	0.809	0.051
Média Elmo	Média dos rankings	0.350	0.616	0.809	0.115
BERT Tipo2-Tipo3	BERT Large	0.349	0.640	**0.821**	0.082
Linguísticas Tipo1-Tipo2	XGBoost	0.346	0.601	0.786	0.055
Média Linguísticas	Média dos rankings	0.344	0.615	0.799	0.084
Linguísticas Tipo1-Tipo3	XGBoost	0.341	0.613	0.796	0.068
Elmo Tipo2-Tipo3	Elmo Wikipedia	0.341	0.591	0.807	0.031
Word Embedding Tipo2-Tipo3	GloVe 600	0.341	0.611	0.801	0.036
Idade de Aquisição	Baseline	0.336	0.588	0.786	0.044
Word Embedding Tipo1-Tipo2	FastText CBOW 600	0.336	0.618	0.796	0.036
Frequência textos informativos	Baseline	0.329	0.633	0.796	0.055
Média BERT	Média dos rankings	0.326	0.613	0.802	0.036
Diversidade Contextual Leg2Kids	Baseline	0.324	0.596	0.796	0.026
BERT Tipo1-Tipo3	BERT Large	0.321	0.608	0.801	0.005
Elmo Tipo1-Tipo3	Elmo Wikipedia	0.321	0.590	0.801	0.055
Glavaš & Štajner (2015)	Baseline	0.319	0.638	0.806	0.025
Frequência Leg2Kids	Baseline	0.316	0.595	0.797	0.024
Word Embedding Tipo1-Tipo3	FastText SkipGram 1000	0.314	0.626	0.806	-0.032
BERT Tipo1-Tipo2	BERT Large	0.306	0.590	0.789	-0.028

Tabela 14: Performance dos métodos propostos e *baselines* na tarefa de Simplificação Lexical. Modelos ordenação pela coluna Top 1.

da criança avança e ela progride pelos anos escolares, o léxico a ser desenvolvido aumenta e novas palavras/desafios surgem.

O *baseline* que melhor desempenhou no Top 1 ou 2 e Top 1, 2 ou 3 foi o Glavas. O uso de *word embeddings* apresentou uma melhor cobertura de atuação do que os demais *baselines* que utilizam contagens ou informações psicolinguísticas. O método de Glavaš & Štajner (2015) não foi o que melhor identificou a palavra mais simples de acordo com a expectativa das crianças, mas a palavra mais bem ranqueada pelo método estava entre as 2 ou 3 palavras mais simples na visão das crianças.

Dentre os métodos desenvolvidos neste trabalho, aquele que obteve a melhor performance no Top 1 foi o Elmo Tipo1-Tipo2. Os métodos desenvolvidos com *embeddings* do Elmo já haviam sido os melhores na tarefa de Identificação de Palavras complexas e entendemos que, assim como a Idade de Aquisição (*baseline* que obteve melhor desempenho no Top1), o Elmo Tipo1-Tipo2 consegue capturar as informações intrínsicas daquelas palavras mais simples, ou seja, o que as tornam as mais simples, já que este método foi treinado no *dataset* composto pelos dicionários de menor nível lexical.

O método que obteve o melhor desempenho no Top 1 ou 2 e Top 1, 2 ou 3 foi o BERT Tipo2-Tipo3. Entendemos que a inferência da gradação da simplicidade/complexidade de palavras extrapola o nível morfológico, dependendo também do contexto no qual a palavra está inserida. O BERT é um modelo contextual, desenvolvido para representar textos. Assim, por mais que nossa correlação com as crianças seja tão baixa quanto a correlação delas próprias, o Tipo2-Tipo3 é o modelo que selecionou, como mais simples, palavras que estão comumente entre as três mais simples na seleção das crianças.

Entendemos que as crianças não divergem completamente quanto ao entendimento da complexidade das palavras, do contrário não haveria léxicos no PB, informações psicolinguísticas nem estudos sobre Adaptação Textual. Embora a correlação de Spearman das próprias crianças seja baixa, temos que entender que estamos lidando com crianças em fase de formação e que certamente possuem muito mais incertezas do que certezas. Essa incerteza reflete nos baixos valores de correlação, mas ao focar as análises somente nas palavras mais bem ranqueadas pelas crianças, percebemos que há concordância. Essa concordância é refletida aqui, quando verificamos a alta performance de nossos métodos ao restringirmos a três palavras mais bem ranqueadas.

Finalmente, avaliamos as combinações de modelos. Analisando o Top 1, as três melhores soluções foram as médias dos modelos que melhor performaram em cada um dos nossos três *datasets*, ou seja, a combinação de todas as abordagens avaliadas. Essa combinação de métodos se beneficia dos diferentes viéses capturados por cada uma das abordagens avaliadas: uso de *features* lexicais, *word embeddings* e *embeddings* contextuais do Elmo e também do BERT. Assim como em Hartmann et al. (2018), a combinação de diferentes *features*/soluções proporcionou uma performance superior na Simplificação Lexical do que o uso individual de cada uma das soluções.

Para realçar a performance da combinação das melhores soluções de cada uma das quatro abordagens de Identificação de Palavras Complexas exploradas, a média dos 3 modelos com *embeddings* do Elmo, abordagem que obteve os melhores resultados na Identificação de Palavras Complexas, obteve apenas a 8ª melhor colocação entre os modelos, analisando o Top 1. Percebemos, portanto, que de fato o uso de diversos viéses enriqueceu a solução de ranqueamento de palavras pela sua simplicidade ao contexto e que, por mais que as redes neurais atuais (*deep learning*) sejam detentoras dos holofotes do *Machine Learning*, o uso de soluções clássicas, quando combinadas a essas redes, ainda trazem ganhos às aplicações.

5.5. Adapt2Kids: Demonstrando os métodos de Simplificação Lexical para Português do Brasil

A fim de disponibilizar uma ferramenta para exemplificação do *pipeline* de Simplificação Lexical e experimentação por parte da comunidade, desenvolvemos o Adapt2Kids[17]. Este sistema exemplifica o processo de Simplificação Lexical para sentenças, não atendendo a demandas reais de simplificação de um texto inteiro, que exigiria uma escolha de quais palavras deveriam ser simplificadas e quais poderiam ser elaboradas. A execução consiste em três passos:

1. A entrada de uma sentença pelo usuário (limitamos na unidade "sentença", pois foi a mesma utilizada ao longo deste artigo;
2. A delimitação de um limiar de classificação para distinção entre palavras simples e complexas; e
3. A submissão do processamento ao clicar no botão "Processar".

[17] http://www.nilc.icmc.usp.br/adapt2kids/

O processo interno realizado pelo sistema consiste em três grandes etapas:

1. A identificação de palavras complexas por meio da aplicação de um classificador neural que faz uso das *embeddings* contextuais do Elmo, abordagem que apresentou melhor performance na Identificação de Palavras Complexas (ver Seção 5.3). Trabalhamos apenas com as palavras do UNITEX-DELAF (Muniz, 2004) para focar nas palavras da língua geral, excluindo nomes próprios, por exemplo;
2. A seleção de palavras candidatas a substituição é dada por meio de uma consulta aos sinônimos disponibilizados em http://www.sinonimos.com.br (recurso também utilizado na expansão de sinônimos do SIMPLEX 2.0 (Hartmann et al., 2020)); ou por meio de uma consulta aos sinônimos do TeP (Maziero et al., 2008) (recurso utilizado na busca por sinônimos na primeira versão do SIMPLEX (Hartmann et al., 2018)); ou por meio do cálculo da similaridade pelo cosseno entre a embedding GloVe (Hartmann et al., 2017) da palavra complexa e as demais palavras do vocabulário, limitando as palavras com mesma PoS *tag* por meio de verificação ao *tagger* nlpnet (Fonseca & Rosa, 2013); ou a combinação das abordagens; e
3. O ranqueamento das palavras candidatas por meio da aplicação do classificador de complexidade (o mesmo utilizado na identificação de palavras complexas) ao substituir a palavra complexa pela candidata.

Utilizamos a média das previsões dos três modelos neurais desenvolvidos baseados nas *embeddings* contextuais do ELMo. Esses modelos foram treinados com diferentes visões do léxico que contempla o Ensino Fundamental (*datesets* Tipo1-Tipo2, Tipo1-Tipo3 e Tipo2-Tipo3).

Apresentamos até 5 palavras mais bem ranqueadas. Demais candidatos são filtrados para facilitar a visualização do resultado.

6. Conclusão e Trabalhos Futuros

Este trabalho investigou várias etapas do *pipeline* de Adaptacão Textual, trazendo várias contribuições para a área de PLN e Educação.

Em relação à Identificação de Palavras Complexas, avaliamos desde abordagens clássicas até as mais modernas, passando por *word embeddings* e também *embeddings* contextuais. Trouxemos a *expertise* obtida ao trabalharmos com

a tarefa para o inglês, desenvolvemos e avaliamos vários métodos para a tarefa. Apesar da boa performance dos métodos clássicos, que fazem uso de *features* linguísticas, o uso de *embeddings* contextuais do Elmo foi a solução que desempenhou melhor na tarefa de Identificação de Palavras Complexas. Como trabalhos futuros, pretendemos avaliar o silibificador utilizado no projeto ReGra (Nunes et al., 1999) e incluído na ferramenta Coh-Metrix-Port (Scarton & Aluísio, 2010). Neste trabalho, fizemos uso do pyphen[18], que utiliza os dicionários do Hunspell (dicionários utilizados em soluções de mercado, como o LibreOffice, OpenOffice.org, Mozilla Firefox e Google Chrome), para o qual identificamos erros produzidos na silabificação de palavras (uma das *features* linguísticas utilizadas neste trabalho).

Para a etapa de Seleção de Abordagem de Adaptação Lexical, fizemos um estudo baseado em córpus para identificar padrões de elaboração e também quais são as palavras do SIMPLEX mais comumente elaboradas. Aproximadamente, 50% das palavras elaboradas tinham anotação de termos técnicos no dataset SIMPLEX, o que é um bom indicativo de que palavras técnicas devem ser elaboradas, corroborando nossa hipótese inicial. No entanto, devido ao conjunto limitado de regras estabelecidas, não conseguimos tirar conclusões definitivas a respeito de quais *features* das palavras indicam um processo de elaboração ou simplificação. Como trabalhos futuros nessa frente, o estudo com foco no desenvolvimento de um método que selecione a melhor abordagem de Adaptação Lexical passa possivelmente pelo uso de grandes córpus de textos simplificados, como o Newsela ou a Wikipédia em Inglês, pois nesses recursos encontramos definições que ajudam a capturar *features* destas palavras que foram elaboradas. A Newsela, especialmente, possui anotações de elaborações e simplificações, o que possibilitaria o treinamento de métodos de *Machine Learning* para aprender quando simplificar ou elaborar. Uma vez que tenhamos o método de seleção das abordagens de Adaptação Lexical funcional seria interassante realizar novamente a avaliação com crianças, para assim avaliar um sistema completo de Adaptação Lexical.

Em relação à etapa de Simplificação Lexical, aplicamos os métodos desenvolvidos para a tarefa de Identificação de Palavras Complexas e observamos que eles desempenham melhor do que *baselines* da área e, inclusive, são melhores do que uma das abordagens mais bem sucedidas da literatura, por exemplo, o método de Glavaš & Štajner (2015). Verificamos ainda que os melhores resultados foram obtidos ao combinarmos nossas abordagens pelo *ranking* médio delas. Nossa solução é o novo *SOTA* para a tarefa de Simplificação Lexical aplicada ao PB.

Para a tarefa de Elaboração Lexical, enriquecemos o córpus SIMPLEX com definições curtas, revisadas manualmente, das palavras complexas, o que possibilita a sua aplicação como método de Elaboração Lexical por definição. Como trabalhos futuros, pretendemos criar um recurso mais sofisticado que contemple uma quantidade maior de palavras e definições curtas, podendo ser utilizado até mesmo para o aprendizado da geração de definições.

Por fim, neste trabalho, disponibilizamos publicamente o SIMPLEX-PB 3.0. Nessa nova versão, o córpus foi enriquecido com *features* linguísticas, que são *proxies* de complexidade lexical, definições de suas palavras complexas e anotações de termos técnicos, informações que fazem com que o córpus também possa ser utilizado para estudos em Elaboração Lexical. O site Adapt2Kids[19] apresenta uma demonstração dos recursos e métodos desenvolvidos e relatados no artigo.

Agradecimentos

O presente trabalho foi realizado com o apoio da FAPESP, proc. nº 2016/00500-1. Agradecemos também às crianças do Projeto Pequeno Cidadão do campus USP em São Carlos, por terem feito a avaliação de sentenças do corpus SIMPLEX-PB 3.0 e as equipes gestora e profissional do Projeto Pequeno Cidadão pelo apoio.

Referências

Aluísio, Sandra Maria & Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: the PorSimples project for simplification of Portuguese texts. Em *NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, 46–53.

Aluísio, Sandra Maria. 1995. *Ferramentas de auxílio a escrita de artigos científicos em Inglês como língua estrangeira*: Universidade Estadual de São Paulo, Brasil. Tese de Doutoramento.

Amancio, Marcelo Adriano. 2011. *Elaboração textual via definição de entidades mencionadas*

[18] https://pyphen.org/

[19] http://nilc.icmc.usp.br/adapt2kids/

e de perguntas relacionadas aos verbos em textos simplificados do português: Universidade de São Paulo, Brasil. Tese de Mestrado.

Arfé, Barbara, Lucia Mason & Inmaculada Fajardo. 2018. Simplifying informational text structure for struggling readers. *Reading and Writing* 31(9). 2191–2210. 10.1007/s11145-017-9785-6.

Barlacchi, Gianni & Sara Tonelli. 2013. ERNESTA: A sentence simplification tool for children's stories in Italian. Em *Computational Linguistics and Intelligent Text Processing (CICLing)*, 476–487. 10.1007/978-3-642-37256-8_39.

Bergstra, James, Daniel Yamins & David Daniel Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. Em *International Conference on International Conference on Machine Learning*, I–115–I–123.

Biderman, Maria Tereza Camargo. 2003. Dicionários do português: da tradição à contemporaneidade. *ALFA: Revista de Linguística* 47(1). 53–69.

Bott, Stefan, Luz Rello, Biljana Drndarevic & Horacio Saggion. 2012. Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. Em *International Conference on Computational Linguistics (COLING)*, 357–374.

Bulté, Bram, Leen Sevens & Vincent Vandeghinste. 2018. Automating lexical simplification in dutch. *Computational Linguistics in the Netherlands Journal* 8. 24–48.

Castro, Pedro Vitor Quinta. 2019. *Aprendizagem profunda para reconhecimento de entidades nomeadas em domínio jurídico*: Universidade Federal de Goiás, Brasil. Tese de Mestrado.

Chelba, Ciprian, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn & Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. ArXiv:1312.3005 [cs.CL].

Chen, Tianqi & Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. Em *International Conference on Knowledge Discovery and Data Mining (KDD)*, 785–794. 10.1145/2939672.2939785.

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk & Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. ArXiv:1406.1078 [cs.CL].

Chung, Jin-Woo, Hye-Jin Min, Joonyeob Kim & Jong C Park. 2013. Enhancing readability of web documents by text augmentation for deaf people. Em *3rd International Conference on Web Intelligence, Mining and Semantics*, 1–10. 10.1145/2479787.2479808.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 37–46. 10.1177/001316446002000104.

Crossley, Scott A., David B. Allen & Danielle S. McNamara. 2011. Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a foreign language* 23(1). 84–101.

Crossley, Scott A., David F. Dufty, Philip M. McCarthy & Danielle S. McNamara. 2007. Toward a new readability: A mixed model approach. Em *Annual Meeting of the Cognitive Science Society*, 197–202.

De Belder, Jan & Marie-Francine Moens. 2010. Text simplification for children. Em *SIGIR workshop on accessible search systems*, 19–26.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv:1810.04805 [cs.CL].

Devlin, Siobhan & John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases* 161–173.

Devlin, Siobhan & Gary Unthank. 2006. Helping aphasic people process online information. Em *International Conference on Computers and Accessibility (SIGACCESS)*, 225–226. 10.1145/1168987.1169027.

Fellbaum, Christiane. 1998. *WordNet: An electronic lexical database*. MIT Press.

Fonseca, Erick R. & João Luís G. Rosa. 2013. A two-step convolutional neural network approach for semantic role labeling. Em *International Joint Conference on Neural Networks (IJCNN)*, 1–7. 10.1109/IJCNN.2013.6707118.

Gardner, Dee & Elizabeth C. Hansen. 2007. Effects of lexical simplification during unaided reading of english informational texts. *TESL Reporter* 40(2). 27–59.

Gers, Felix A, Jürgen Schmidhuber & Fred Cummins. 1999. Learning to forget: Continual prediction with LSTM. *Neural Computation* 12(10). 10.1162/089976600300015015.

Glavaš, Goran & Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? Em *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, vol. 2, 63–68. 10.3115/v1/P15-2011.

da Graça Krieger, Maria. 2012. Dicionários escolares e ensino de língua materna. *Estudos Linguísticos* 41(1). 169–180.

Hartmann, Nathan, Livia Cucatto, Danielle Brants & Sandra Aluísio. 2016. Automatic classification of the complexity of nonfiction texts in Portuguese for early school years. Em João Silva, Ricardo Ribeiro, Paulo Quaresma, André Adami & António Branco (eds.), *Computational Processing of the Portuguese Language (PROPOR)*, 12–24. 10.1007/978-3-319-41552-9_2.

Hartmann, Nathan, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jessica Rodrigues & Sandra Aluisio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. ArXiv:1708.06025 [cs.CL].

Hartmann, Nathan & Leandro Borges dos Santos. 2018. NILC at CWI 2018: Exploring feature engineering and feature learning. Em *Workshop on Innovative Use of NLP for Building Educational Applications*, 335–340. 10.18653/v1/W18-0540.

Hartmann, Nathan S., Gustavo H. Paetzold & Sandra M. Aluísio. 2018. SIMPLEX-PB: A lexical simplification database and benchmark for Portuguese. Em *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 272–283. 10.1007/978-3-319-99722-3_28.

Hartmann, Nathan S, Gustavo H Paetzold & Sandra M Aluísio. 2020. A dataset for the evaluation of lexical simplification in portuguese for children. Em *Conference on Computational Processing of the Portuguese Language (PROPOR)*, 55–64. 10.1007/978-3-030-41505-1_6.

Hartmann, Nathan Siegle & Sandra Maria Aluísio. 2019. Avaliação do uso da diversidade contextual e da frequência para a tarefa de identificação de palavras complexas em simplificação lexical. Em *Symposium in Information and Human Language Technology (STIL)*, 294–302.

He, Haibo & Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21(9). 1263–1284. 10.1109/TKDE.2008.239.

Henderson, Lisa, Margaret Snowling & Paula Clarke. 2013. Accessing, integrating, and inhibiting word meaning in poor comprehenders. *Scientific Studies of Reading* 17(3). 177–198. 10.1080/10888438.2011.652721.

Horn, Colby, Cathryn Manduca & David Kauchak. 2014. Learning a lexical simplifier using Wikipedia. Em *Annual Meeting of the Association for Computational Linguistics*, 458–463. 10.3115/v1/P14-2075.

Inui, Kentaro, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida & Tomoya Iwakura. 2003. Text simplification for reading assistance: a project note. Em *International Workshop on Paraphrasing (IWP)*, 9–16. 10.3115/1118984.1118986.

Janczura, Gerson A., Goiara M. Castilho, Nelson O. Rocha, Terezinha de Jesus C. Van Erven & Tin Po Huang. 2007. Normas de concretude para 909 palavras da língua portuguesa. *Psicologia: Teoria e Pesquisa* 23(2). 195–204. 10.1590/S0102-37722007000200010.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou & Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv:1612.03651 [cs.CL]* .

Kajiwara, Tomoyuki, Hiroshi Matsumoto & Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. Em *Conference on Computational Linguistics and Speech Processing (ROCLING)*, 59–73.

Kursa, Miron B., Aleksander Jankowski & Witold R. Rudnicki. 2010. Boruta–a system for feature selection. *Fundamenta Informaticae* 101(4). 271–285. 10.3233/FI-2010-288.

Kursa, Miron B. & Witold R. Rudnicki. 2010. Feature selection with the Boruta package. *Journal of Statistical Software* 36(11). 1–13. 10.18637/jss.v036.i11.

Le, Minh, Marten Postma & Jacopo Urbani. 2017. Word sense disambiguation with lstm: Do we really need 100 billion words? ArXiv:1712.03376 [cs.CL].

Ling, Wang, Chris Dyer, Alan Black & Isabel Trancoso. 2015. Two/too simple adaptations of word2vec for syntax problems. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1299–1304. 10.3115/v1/N15-1142.

Max, Aurélien. 2006. Writing for language-impaired readers. Em *7th Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, 567–570. 10.1007/11671299_59.

Mayer, Richard E. 1980. Elaboration techniques that increase the meaningfulness of technical text: An experimental test of the learning strategy hypothesis. *Journal of Educational Psychology* 72(6). 770–784. 10.1037/0022-0663.72.6.770.

Maziero, Erick G., Thiago A.S. Pardo, Ariani Di Felippo & Bento C. Dias-da Silva. 2008. A base de dados lexical e a interface web do TeP 2.0: thesaurus eletrônico para o português do Brasil. Em *Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, 390–392. 10.1145/1809980.1810076.

Mihalcea, Rada & Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. Em *Conference on Information and Knowledge Management (CIKM)*, 233–242. 10.1145/1321440.1321475.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. Em *Advances in neural information processing systems*, 3111–3119.

Muniz, Marcelo Caetano Martins. 2004. *A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB*: Universidade de São Paulo. Tese de Mestrado.

Mutsuro, Kai & Matsukawa Toshihiro. 2002. *Method of vocabulary teaching: Vocabulary table version*. Mitsumura Tosho Publishing.

Nair, Vinod & Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. Em *International Conference on Machine Learning (ICML)*, 807–814.

Nunes, Maria das G. V., Denise C. Kuhn, Ana Raquel Marchi, Ana Cláudia Nascimento, Sandra M. Aluísio & Osvaldo N. de Oliveira Junior. 1999. Novos rumos para o ReGra: extensão do revisor gramatical do português do Brasil para uma ferramenta de auxílio à escrita. Em *Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR)*, s/p.

Paetzold, Gustavo & Lucia Specia. 2017. Lexical simplification with neural ranking. Em *Conference of the European Chapter of the Association for Computational Linguistics*, 34–40.

Paetzold, Gustavo H. & Lucia Specia. 2016. Simplenets: Evaluating simplifiers with resource-light neural networks. Em *LREC Workshop & Shared Task on Quality Assessment for Text Simplification*, 42–46.

Paetzold, Gustavo Henrique & Lucia Specia. 2015. LEXenstein: A framework for lexical simplification. *International Joint Conference on Natural Language Processing 2015: System Demonstrations (ACL-IJCNLP)* 85–90. 10.3115/v1/P15-4015.

Paiva, Valeria de, Alexandre Rademaker & Gerard de Melo. 2012. OpenWordNet-PT: An open brazilian wordnet for reasoning. Em *COLING: Demonstration Papers*, 353–360.

Pasqualini, Bianca Franco. 2018. *CorPop: um corpus de referência do português popular escrito do Brasil*: Universidade Federal do Rio Grande do Sul. Tese de Doutoramento.

Pennington, Jeffrey, Richard Socher & Christopher D. Manning. 2014. Glove: Global vectors for word representation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. 10.3115/v1/D14-1162.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee & Luke Zettlemoyer. 2018. Deep contextualized word representations. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2227–2237. 10.18653/v1/N18-1202.

Petersen, Sarah E. & Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. Em *Workshop on Speech and Language Technology for Education (SLaTE)*, 69–72.

Quinlan, Philip T. 1992. *The Oxford psycholinguistic database*. University Press.

Rello, Luz, Ricardo Baeza-Yates, Stefan Bott & Horacio Saggion. 2013a. Simplify or help?: text simplification strategies for people with dyslexia. Em *International Cross-Disciplinary Conference on Web Accessibility (W4A)*, 1–10. 10.1145/2461121.2461126.

Rello, Luz, Ricardo Baeza-Yates, Laura Dempere-Marco & Horacio Saggion. 2013b. Frequent words improve readability and short words improve understandability for people with dyslexia. Em *International Conference on Human-Computer Interaction (INTERACT)*, 203–219. 10.1007/978-3-642-40498-6_15.

Saggion, Horacio. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies* 10(1). 1–137. 10.2200/S00700ED1V01Y201602HLT032.

Saggion, Horacio, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello & Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Compututing* 6(4). s/p. 10.1145/2738046.

dos Santos, Leandro Borges, Magali Sanches Duran, Nathan Siegle Hartmann, Arnaldo Candido, Gustavo Henrique Paetzold & Sandra Maria Aluisio. 2017. A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. Em *International Conference on Text, Speech, and Dialogue*, 281–289. 10.1007/978-3-319-64206-2_32.

Scarton, Carolina Evaristo & Sandra Maria Aluísio. 2010. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do Coh-Metrix para o Português. *Linguamática* 2(1). 45–61.

Siddharthan, Advaith. 2006. *Syntactic simplification and text cohesion*: University of Cambridge, Inglaterra. Tese de Doutoramento.

Siddharthan, Advaith. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics* 165(2). 259–298. 10.1075/itl.165.2.06sid.

de Sousa, Lucilene Bender, Lilian Cristine Hübner & Roselaine Berenice Ferreira da Silva. 2020. Lexical-semantic integration by good and poor reading comprehenders. *Ilha do Desterro* 73(1). 63–78. 10.5007/2175-8026.2020v73n1p63.

Souza, Fabio, Rodrigo Nogueira & Roberto Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. ArXiv:1909.10649.

Specia, Lucia, Sujay Kumar Jauhar & Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. Em *First Joint Conference on Lexical and Computational Semantics (SEM)*, 347–355.

Štajner, Sanja & Horacio Saggion. 2018. Data-driven text simplification. Em *International Conference on Computational Linguistics: Tutorial Abstracts (COLING)*, 19–23.

Štajner, Sanja, Horacio Saggion & Simone Paolo Ponzetto. 2019. Improving lexical coverage of text simplification systems for spanish. *Expert Systems with Applications* 118. 80–91. 10.1016/j.eswa.2018.08.034.

Tetreault, Joel, Jill Burstein, Ekaterina Kochmar, Claudia Leacock & Helen Yannakoudakis (eds.). 2018. *Proceedings of the 13th workshop on innovative use of nlp for building educational applications.*

Trieschnigg, Dolf & Claudia Hauff. 2011. Classic children's literature-difficult to read? Em *European Conference on Information Retrieval (ECIR)*, 691–694. 10.1007/978-3-642-20161-5_72.

Tsang, Wai King. 1987. Text modifications in ESL reading comprehension. *RELC journal* 18(2). 31–44. 10.1177/003368828701800203.

Urano, Ken. 2000. *Lexical simplification and elaboration: Sentence comprehension and incidental vocabulary acquisition*: University of Hawai'i, EUA. Tese de Doutoramento.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. ArXiv:1706.03762 [cs.CL].

Vossen, Piek, Isa Maks, Roxane Segers, Hennie Van Der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang & Maarten De Rijke. 2013. Cornetto: a combinatorial lexical semantic database for Dutch. Em *Essential Speech and Language Technology for Dutch*, 165–184. Springer. 10.1007/978-3-642-30910-6_10.

Wagner Filho, Jorge, Rodrigo Wilkens, Marco Idiart & Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. Em *International Conference on Language Resources and Evaluation (LREC)*, 4339–4344.

Watanabe, Willian Massami, Arnaldo Candido Jr, Marcelo Adriano Amâncio, Matheus De Oliveira, Thiago Alexandre Salgueiro Pardo, Renata PM Fortes & Sandra M Aluísio. 2010. Adapting web content for low-literacy readers by using lexical elaboration and named entities labeling. *New Review of Hypermedia and Multimedia* 16(3). 303–327. 10.1080/13614568.2010.542620.

Yano, Yasukata, Michael H Long & Steven Ross. 1994. The effects of simplified and elaborated texts on foreign language reading comprehension. *Language learning* 44(2). 189–219. 10.1111/j.1467-1770.1994.tb01100.x.

Yimam, Seid Muhie, Chris Biemann, Shervin Malmasi, Gustavo H. Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack & Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. ArXiv:1804.09132 [cs.CL].

Young, Dolly J. 1999. Linguistic simplification of SL reading material: Effective instructional practice? *The Modern Language Journal* 83(3). 350–366. 10.1111/0026-7902.00027.

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

Avaliando entidades mencionadas na coleção ELTeC-por

Assessing named entities in the ELTeC-por collection

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

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

Marcin Wlodek
Linguateca
martimwlodek@hotmail.com

Resumo

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

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

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

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

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

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

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

Palavras chave

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

Abstract

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

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

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

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

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

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

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

Keywords

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

1. Introdução

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

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

DOI: 10.21814/lm.12.2.336
This work is Licensed under a Creative Commons Attribution 4.0 License

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

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

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

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

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

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

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

2. Caracterização da coleção

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

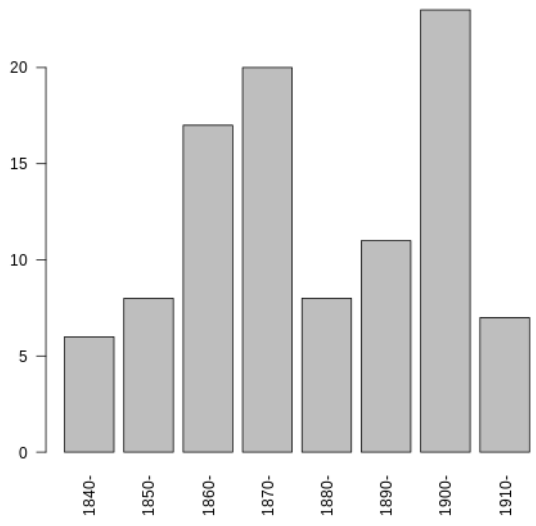


Figura 1: As obras do ELTeC-por por década

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

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

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

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

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

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

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

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

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

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

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

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

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

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

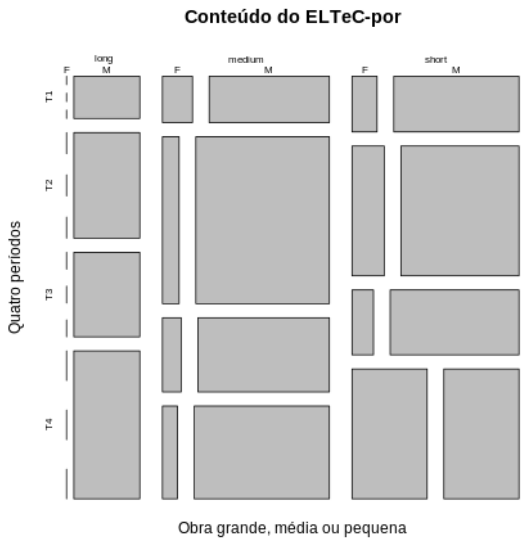


Figura 2: Visualizando o ELTeC-por de acordo com o período, o tamanho e o género do autor

3. Entidades mencionadas em textos literários

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

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

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

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

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

3.1. O PALAVRAS-NER

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

3.1.1. Identificação de entidades mencionadas e sua segmentação

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

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

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

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

3.1.2. Classificação de entidades mencionadas

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

3.1.3. O conjunto de categorias

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

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

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

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

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

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

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

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

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

3.1.4. A sequência de comandos do PALAVRAS-NER

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

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

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

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

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

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

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

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

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

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

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

3.1.5. Tradução do PALAVRAS-NER para este projeto

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

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

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

3.2. A anotação e sua revisão

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

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

Figura 4: As entidades mencionadas codificadas pelo BRAT

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

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

3.3. Textos processados

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

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

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

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

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

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

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

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

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

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

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

3.4. Primeiras tarefas

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

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

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

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

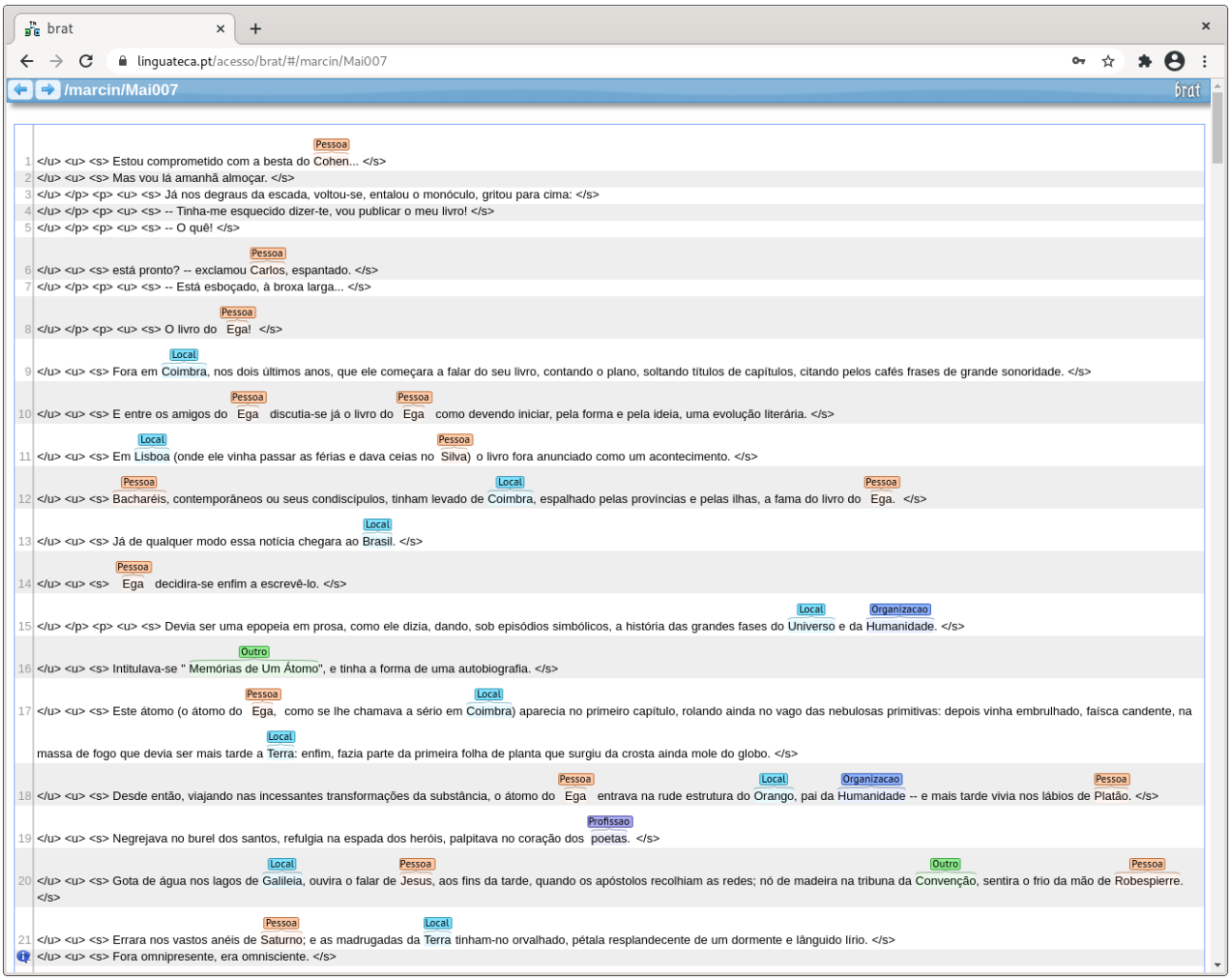


Figura 5: Sistema BRAT de revisão de anotação, com uma parte de *Os Maias*

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

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

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

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

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

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

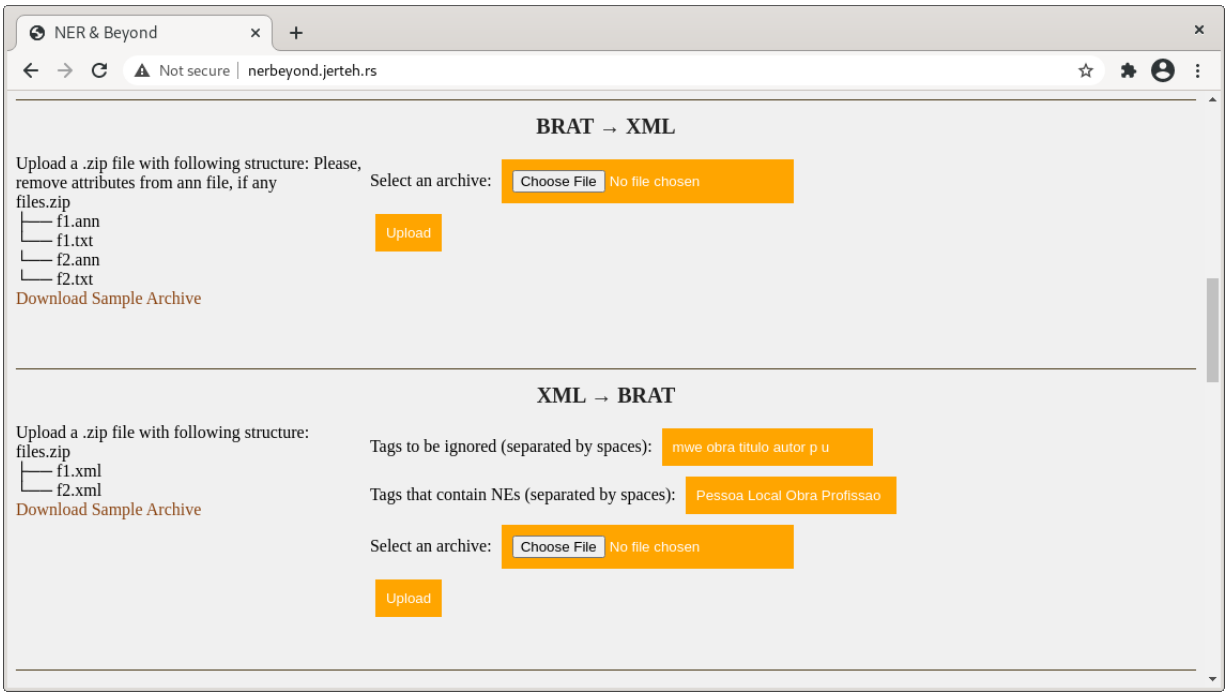


Figura 6: Parte do sistema para conversão acessível do servidor sérvio

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

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

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

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

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

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

3.5. Relato de um trabalho intelectual

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.5.1. Erro ou divergência?

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

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

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

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

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

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

4. Resultados detalhados

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

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

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

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

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

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

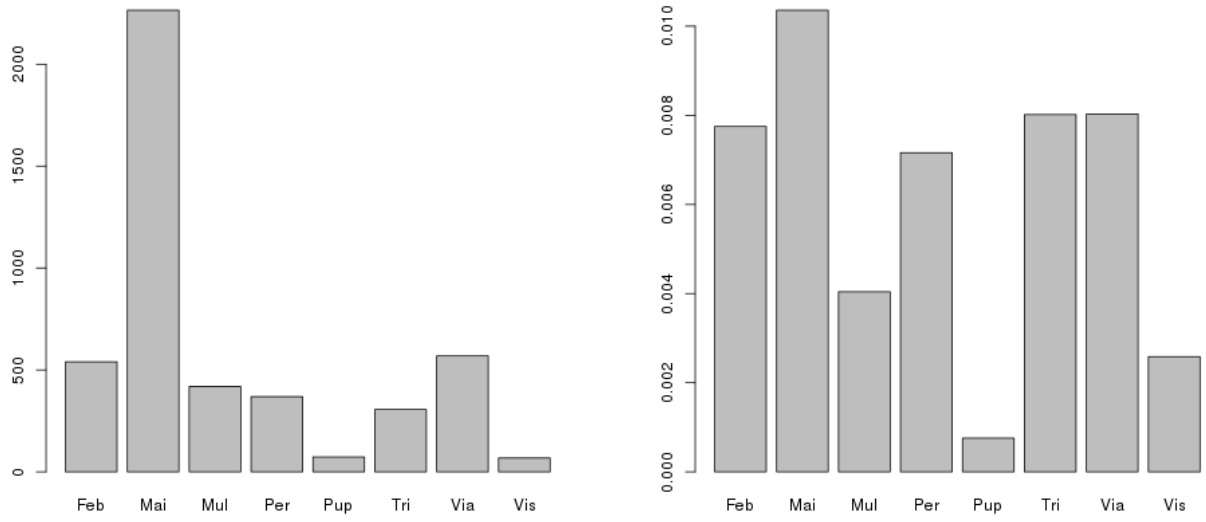
Vemos assim que aqueles textos que mais mencionam lugares, também são, talvez não surpreendentemente, aqueles que os repetem mais: Mais uma vez os Maias estão no topo. Exceção para o caso da Mulata que, talvez por ser passada no Brasil, menciona bastantes lugares mas pouco os repete.

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

Tabela 7: As 15 entidades mais frequentes em cada obra

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

Tabela 8: Quais os lugares mais mencionados por obra



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

Figura 7: Distribuição dos lugares

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

Tabela 9: Lugares diferentes e repetição

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

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

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

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

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

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

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

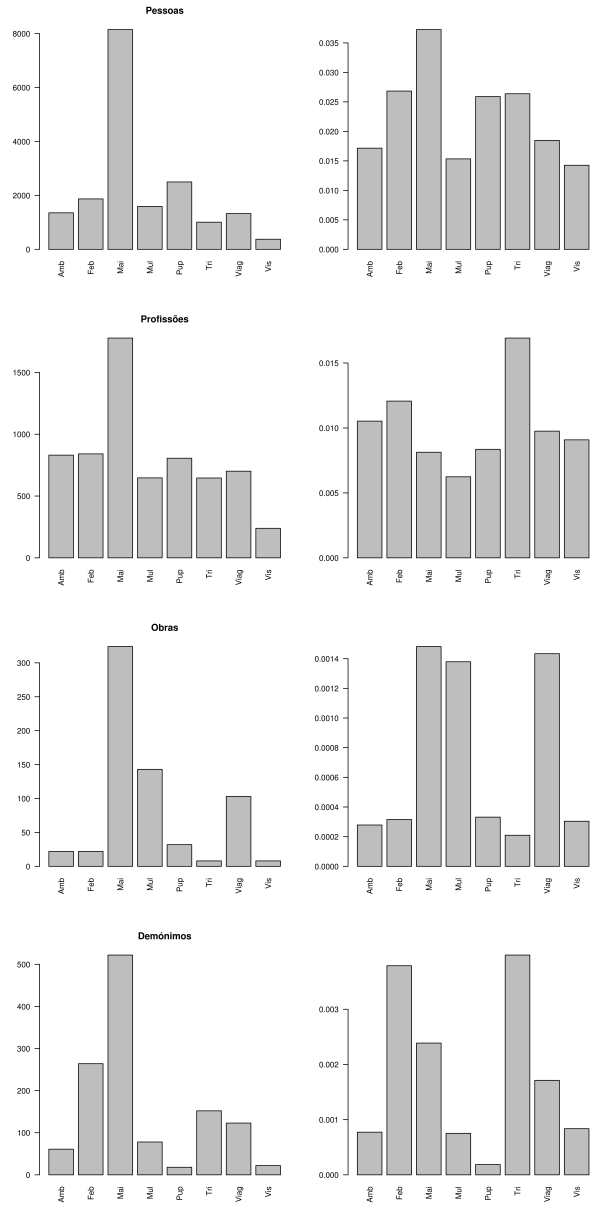


Figura 8: Distribuição em números absolutos, e relativos, das pessoas, das profissões, das obras e dos gentílicos

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

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

5. Avaliação do desempenho do PALAVRAS-NER

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Tabela 12: Demónimos e profissões mais frequentes em cada obra

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

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

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

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

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

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

6. Hipóteses sobre os textos literários

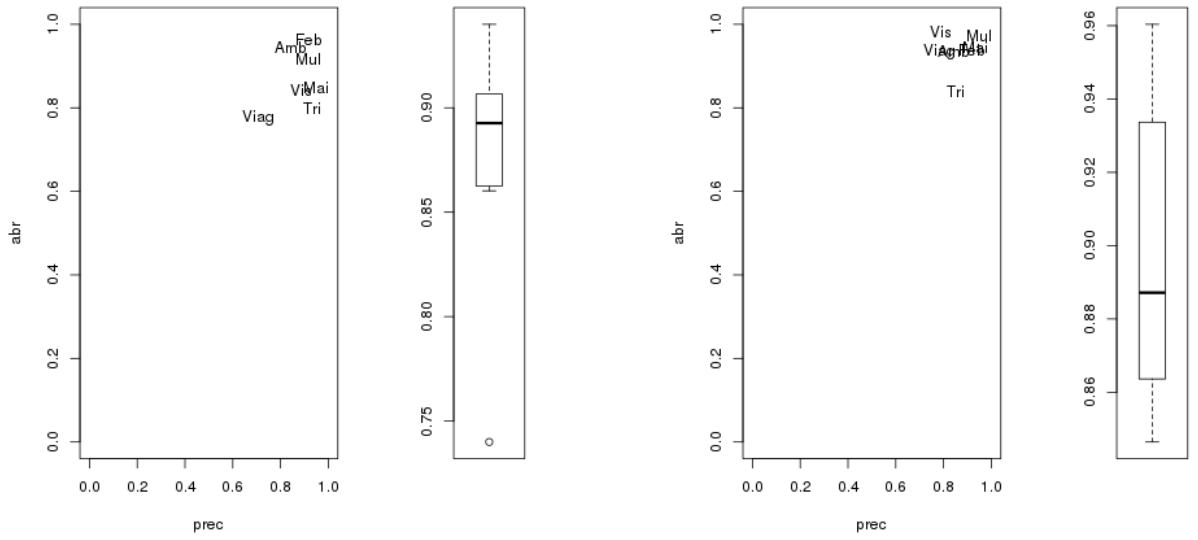
As seguintes perguntas parecem ser pertinentes:

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

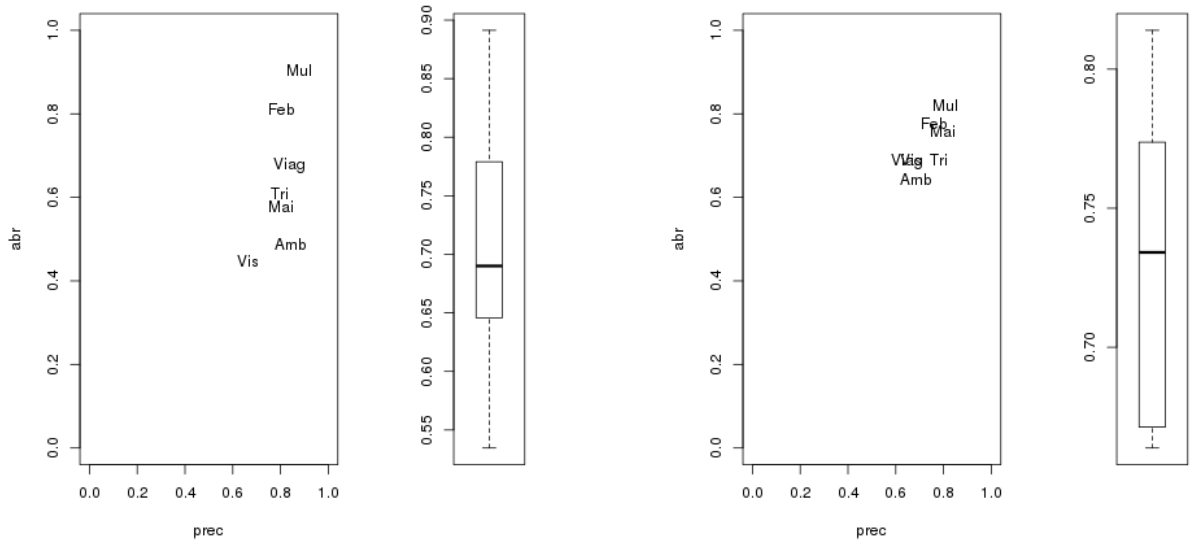
Não conseguimos responder positivamente a nenhuma destas perguntas. Em caso nenhum conseguimos encontrar uma propriedade que separasse as obras em duas classes obviamente distintas, quer entre canónicas e não canónicas, quer entre romances históricos ou contemporâneos.

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

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



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



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

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

7. Comparação com outros trabalhos

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

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

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

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

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

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

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

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

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

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

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

8. Conclusões e trabalho futuro

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

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

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

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

[18] https://spacy.io/

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

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

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

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

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

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

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

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

Agradecimentos

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

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

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

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

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

Referências

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proposta recebida em Maio 2020 e aceite para publicação em Novembro 2020.

Avaliação de recursos computacionais para o português

Evaluating computational resources for Portuguese

Matilde Gonçalves
Instituto Superior Técnico
Universidade de Lisboa
INESC-ID Lisboa
matilde.do.carmo.lages.goncalves@tecnico.ulisboa.pt

Luísa Coheur
Instituto Superior Técnico
Universidade de Lisboa
INESC-ID Lisboa
luisa.coheur@tecnico.ulisboa.pt

Jorge Baptista
Universidade do Algarve
INESC-ID Lisboa
jbaptis@ualg.pt

Ana Mineiro
Instituto de Ciências da Saúde
Universidade Católica Portuguesa
Centro de Investigação Interdisciplinar em Saúde
amineiro@ics.lisboa.ucp.pt

Resumo

Têm sido desenvolvidas várias ferramentas para o processamento da língua portuguesa. No entanto, devido a escolhas variadas na base dos comportamentos destas ferramentas (diferentes opções de pré-processamento, diferentes conjuntos de etiquetas morfossintáticas e de dependências, etc.), torna-se difícil ter uma ideia do desempenho comparativo de cada uma. Neste trabalho, avaliamos um conjunto de ferramentas gratuitas e publicamente disponíveis, que realizam as tarefas de Etiquetação Morfossintática e de Reconhecimento de Entidades Mencionadas, para a língua portuguesa. São tidos em conta doze modelos diferentes para a primeira tarefa e oito para a segunda. Todos os recursos usados nesta avaliação (tabelas de mapeamento de etiquetas, *corpora* de referência, etc.) são disponibilizados, permitindo replicar/afinar os resultados. Apresentamos ainda um estudo qualitativo de dois analisadores de dependências. Não temos conhecimento de nenhum trabalho similar recente, isto é, que tenha em conta as ferramentas atuais disponíveis, realizado para a língua portuguesa.

Palavras chave

processamento da linguagem natural, avaliação de recursos, língua portuguesa, análise morfosintática, reconhecimento de entidades mencionadas, análise de dependências

Abstract

There are several tools for the Portuguese language. However, and due to different choices at the basis of these tools' behaviour (different preprocessing, different labels, etc.), it becomes difficult to have an idea of each one's comparative performance. In this work, we propose an evaluation of tools, publicly available and free, that perform the tasks of Part-of-Speech Tagging and Named Entity Recognition, for the Portuguese language. We evaluate twelve different models for the first task and eight for the second. All the resources used in this evaluation (mapping tables between labels, testing *corpora*, etc.) will be made available, allowing to replicate/fine-tune the results here presented. We also present a qualitative analysis of two dependency parsers. To the best of our knowledge, no recent work that considers the recent available tools, was carried out for the Portuguese language.

Keywords

natural language processing, evaluation of resources, portuguese language, part-of-speech tagging, named entity recognition, dependency parsing

1. Introdução

A área do Processamento da Linguagem Natural (PLN) encontra-se em profunda expansão e em Portugal não é excepção. Se há alguns anos os trabalhos computacionais ligados à língua portuguesa eram de pura investigação, hoje em dia várias empresas têm projetos neste campo, desenvolvendo sistemas de pesquisa em dados médicos, agentes virtuais, sistemas de tradução, etc. Do mesmo modo, vários agentes interessados utilizam ferramentas que operam sobre o português e, apesar de a língua inglesa continuar a ser imbatível em termos de recursos disponíveis, existem actualmente várias ferramentas gratuitas que oferecem modelos pré-treinados (ou facilmente treináveis) para a língua portuguesa, em

DOI: 10.21814/lm.12.2.331
This work is Licensed under a Creative Commons Attribution 4.0 License

especial para as variantes do português europeu e do Brasil. Coloca-se, então, a questão de escolher a ferramenta mais adequada para a tarefa em mãos. Vários trabalhos têm-se focado na avaliação de ferramentas (Gamallo & Garcia, 2013), além das que têm sido levadas a cabo no quadro de diferentes *fora* de avaliação conjunta para a língua portuguesa, tendo como objectivo determinar o estado da arte em várias tarefas de PLN. Algumas destas avaliações conjuntas focaram-se na avaliação de ferramentas que realizam tarefas de base, nomeadamente Etiquetação Morfossintática (EMS) — Morfo-olimpíadas (Santos et al., 2003)[1] — e Reconhecimento de Entidades Mencionadas (REM) — MiniHAREM, primeiro e segundo HAREM (Santos & Cardoso, 2007; Santos et al., 2008), bem como IberLEF-2019 (Collovini et al., 2019). Outras avaliações visaram tarefas de mais alto nível, tais como determinar a similaridade semântica entre duas frases (Fonseca et al., 2016). Todas estas competições são tipicamente montadas por equipas de peritos em PLN, responsáveis pelos dados de treino, coleções douradas, etc. (Santos et al., 2003); do mesmo modo, em todas estas

competições participam usualmente equipas que fazem a sua investigação em PLN.

Ora, nos dias de hoje, mais do que saber exatamente qual o estado da arte na realização de uma tarefa, os não-especialistas precisam de poder decidir (rapidamente) que ferramentas usar. Dada toda a oferta actual, perdemo-nos facilmente na avaliação de ferramentas. Assim, neste artigo, focamo-nos na avaliação de ferramentas que realizam as tarefas de EMS e REM para a língua portuguesa (português europeu). No entanto, não é nosso objectivo comparar detalhadamente as várias ferramentas e escolher a vencedora com base em sofisticados *corpora* de referência anotados, tal como realizado em avaliações anteriores, nomeadamente nas levadas a cabo pela Linguateca[2], mas mostrar como, com base numa metodologia simples mas correta, se pode ter uma ideia da utilidade de uma ferramenta e/ou modelos associados, consoante as necessidades da aplicação final em vista. Também não é nosso objetivo trazer os utilizadores para uma avaliação. No entanto, além da facilidade de instalação e de utilização de cada ferramenta, mostramos como estas podem ser testadas, dando pistas sobre a sua aplicabilidade. Todo o código usado, bem como recursos linguísticos criados, estão disponíveis[3], com especial destaque para as tabelas de mapeamento de etiquetas. Estas permitem avaliar os sistemas sobre a mesma referência, sendo o ponto que maior dificuldades acarreta ao levar a cabo uma tarefa de avaliação com sistemas tão variados, os quais retornam etiquetas diferentes, em especial na tarefa de EMS. De notar, igualmente, que neste artigo vão ser avaliadas apenas ferramentas disponíveis publica e gratuitamente para a língua portuguesa. Além disso, apresentamos um estudo qualitativo de duas ferramentas que realizam a tarefa de Análise de Dependências (AD), tendo em conta vários factores, tais como a pontuação, a segmentação e a etiquetagem morfossintática, dos quais dependem as dependências obtidas. De notar que se trata de um estudo qualitativo das saídas de dois sistemas, sem qualquer pretensão de generalidade ou de quantificar essas observações.

As contribuições deste trabalho são, pois:

- a construção de dois *corpora* de referência, um para cada uma das tarefas (EMS e REM);
- a adaptação dos *corpora* de referência tendo em conta os diferentes pré-processamentos dos dados, realizados pelas diversas ferramentas;
- os *scripts* de conversão entre as etiquetas de cada ferramenta e as etiquetas dos *corpora* de referência;
- a avaliação de nove ferramentas (doze modelos diferentes) na tarefa de EMS;
- a avaliação de oito modelos distintos na tarefa de REM;
- a avaliação (qualitativa) de dois analisadores na tarefa de AD.

Apesar de, ao longo dos anos, vários trabalhos se focaram na avaliação de ferramentas para a língua portuguesa, não temos conhecimento de um trabalho similar a esta escala.

Este artigo está organizado como se segue: na Secção 2, apresentamos as ferramentas em análise e, na Secção 3, todo o *setup* experimental. A Secção 4 trata os resultados relativos à EMS e a Secção 5, os resultados relativos à tarefa de REM. Finalmente, na Secção 6, apresenta-se o estudo relativo à AD e, na Secção 7, resume-se as principais conclusões e discute-se o trabalho futuro.

[1] https://www.linguateca.pt/Morfolimpiadas/
[2] http://www.linguateca.pt
[3] https://gitlab.hlt.inesc-id.pt/lcoheur/ptools

2. Ferramentas em análise

Nesta secção, são descritas as ferramentas e modelos pré-treinados, desenvolvidos para o processamento de português, para as tarefas de EMS e REM (e, em dois casos, de AD), e disponibilizados gratuitamente. A maioria destas ferramentas fornece modelos pré-treinados para as tarefas em estudo. As linguagens de programação destas ferramentas alternam entre o Java, o C++ e o Python, e, de um modo geral, apresentam documentação, o que torna relativamente fácil a sua instalação e utilização. Algumas destas ferramentas podem ser usadas em linha de comando, não exigindo muitos conhecimentos de programação.

2.1. FreeLing

O FreeLing (Padró, 2012)[4] é uma biblioteca *open source* em C++, que disponibiliza modelos pré-treinados para o português, para (entre outras) as tarefas de EMS e REM. Encontra-se disponível um manual de utilizador[5] completo e bem estruturado, no qual são descritos os procedimentos de instalação, importação para outras linguagens de programação, o sistema de etiquetas, etc. Apesar de a maioria das tarefas de FreeLing estar disponível através de linha de comandos, algumas funcionalidades apenas são acessíveis usando a ferramenta como biblioteca.

2.2. NLTK

O NLTK (Bird et al., 2009)[6] é uma das plataformas mais utilizadas em PLN. Não dispõe de modelos pré-treinados para o português, mas oferece várias funcionalidades (implementadas em Python), bem como vários *corpora* (incluindo em português europeu), que permitem criar modelos para várias tarefas de PLN. Os *corpora* disponíveis para o processamento de texto em português fazem parte do projeto Floresta Sintática[7]. De notar que os *corpora* do Floresta Sintática permitem treinar modelos capazes de realizar a tarefa de EMS, mas não contêm informação sobre entidades nomeadas, pelo que a realização da tarefa de REM depende da existência de outros *corpora*.

O uso desta plataforma requer algum conhecimento em programação. Contudo, existe um conjunto de programas para a linha de comandos chamado NLTK-Trainer[8] que permite ao utilizador abstrair-se da programação, facilitando o treino de modelos presentes na ferramenta, a avaliação desses modelos e a análise de *corpora*. A instalação da plataforma NLTK encontra-se documentada para cada sistema operativo[9]. Por outro lado, a utilização das componentes que a ferramenta oferece é facilitada com exemplos de utilização, incluindo a realização de tarefas para a língua portuguesa[10].

2.3. OpenNLP

O OpenNLP (Apache Software Foundation, 2014)[11] é uma biblioteca para Java, que fornece modelos pré-treinados, inclusive para português, para a tarefa de EMS. A documentação para a instalação da ferramenta não se encontra referenciada na página da ferramenta[12], o que dificultou a instalação. No entanto, no site da ferramenta, existe um guia de referência bem estruturado, que descreve os modos de utilização da ferramenta, o procedimento para treino de modelos e a execução de cada componente com base em modelos pré-treinados. Estas informações, são acompanhadas por exemplos. O OpenNLP oferece ainda um modo de utilização baseado na execução de programas na linha de comandos. Assim, para treinar, testar e aplicar esta ferramenta com modelos pré-treinados nas diferentes tarefas de PLN, não são exigidos conhecimentos de programação. Por ser uma biblioteca direcionada para a linguagem de programação Java, a sua importação para Python requer uma componente que faça a ligação. Para tal, neste trabalho usou-se a interface NLTK-OPENNLP[13].

2.4. NLPyPort

O NLPyPort (Ferreira et al., 2019b,a)[14] (Python) realiza as tarefas de EMS e REM, disponibilizando os modelos criados. Os recursos usados pelo NLPyPort, à excepção dos da tarefa de identificação do lema, são baseados em modelos e funções da ferramenta NLTK, previamente apresentada. À função genérica de divisão em tokens

[4] http://nlp.lsi.upc.edu/freeling/index.php/

[5] https://freeling-user-manual.readthedocs.io/en/latest/

[6] http://www.nltk.org

[7] https://www.linguateca.pt/Floresta/.

[8] Acessível em https://github.com/japerk/nltk-trainer e documentado em https://nltk-trainer.readthedocs.io/en/latest/

[9] https://www.nltk.org/install.html

[10] http://www.nltk.org/howto/portuguese_en.html

[11] http://opennlp.apache.org/

[12] O procedimento da instalação encontra-se na página https://opennlp.apache.org/building.html

[13] https://github.com/paudan/opennlp_python

[14] https://github.com/jdportugal/NLPyPort

da ferramenta NLTK (word_tokenize), o NLPyPort adiciona uma função de identificação de pronomes clíticos e contrações. De notar que existe uma versão anterior desta ferramenta compatível com projetos desenvolvidos na linguagem de programação Java (Rodrigues et al., 2018)[15].

2.5. Polyglot

O Polyglot (Al-Rfou, 2015)[16] é uma biblioteca para Python que suporta funcionalidades que permitem realizar várias tarefas de PLN para diversas línguas, incluindo as de EMS e REM. Dispõe ainda de modelos pré-treinados capazes de levar a cabo ambas as tarefas em análise. Por consistir numa biblioteca (Python), a sua utilização pressupõe alguma experiência com programação. A documentação[17] desta ferramenta encontra-se bem estruturada e os procedimentos para a sua instalação e importação estão descritos de forma clara. A documentação inclui também tutoriais e informações sobre cada tarefa, o que simplifica o seu uso.

2.6. SpaCy

O SpaCy (Honnibal et al., 2020)[18] é uma biblioteca que incorpora modelos pré-treinados de várias línguas, inclusive de português, para as tarefas de EMS, REM e AD. Esta ferramenta foi pensada para ser importada como biblioteca em programas Python e não disponibiliza outras opções de uso. A sua documentação contém, entre outras informações, procedimentos para a sua instalação, importação para outras linguagens e realização das diferentes tarefas de PLN, acompanhados de exemplos, o que contribui para a sua usabilidade. No entanto, não existe documentação sobre as etiquetas usadas, o que complicou a sua avaliação no que respeita a tarefa de EMS.

2.7. StanfordNLP

O StanfordNLP (Qi et al., 2018)[19] é uma biblioteca para Python, que permite realizar várias tarefas de PLN, incluindo EMS e REM. Tal como o SpaCy, também realiza AD. O StanfordNLP oferece modelos pré-treinados para 53 línguas, inclusive para português. Na realidade, o StanfordNLP é uma interface para Python da ferramenta Stanford CoreNLP, em Java, que o grupo *Stanford NLP* disponibiliza. A interface para Python requer conhecimentos em programação. No entanto, as mesmas funcionalidades que o StanfordNLP apresenta podem ser executadas por via da linha de comandos, usando a ferramenta principal (Stanford CoreNLP). Existem tutoriais[20] que descrevem os passos da instalação e utilização desta biblioteca, assim como exemplos para as diferentes tarefas de PLN, nomeadamente para as de EMS e REM[21].

2.8. TreeTagger

Das tarefas em estudo, o TreeTagger (Màrquez & Rodríguez, 1998)[22] realiza apenas a tarefa de EMS para o português europeu, oferecendo modelos pré-treinados. Os procedimentos relativos à sua instalação e uso são descritos de forma clara no site da ferramenta. O modo de utilização de TreeTagger não implica conhecimentos de programação, pois pode ser realizado através da linha de comandos. A ferramenta pode ser igualmente importada para Python por via de uma componente intermediária, como por exemplo, treetagger-python[23]. São também disponibilizados tutoriais[24].

2.9. LinguaKit

O LinguaKit (Gamallo & Garcia, 2017a) une várias ferramentas de processamento da língua natural permitindo a realização de tarefas como a lematização, análise de sentimentos, análise morfossintática, análise sintática, reconhecimento e classificação de entidades nomeadas, entre outras. As diferentes tarefas podem ser executadas através da linha de comandos ou pela versão web[25]. No repositório[26] do LinguaKit encontram-se as instruções de instalação, procedimentos e exemplos de utilização da ferramenta através da linha de comandos.

[15] https://github.com/rikarudo/NLPPORT.
[16] https://draquet.github.io/PolyGlot/
[17] https://polyglot.readthedocs.io/en/latest/Installation.html
[18] http://www.spacy.io
[19] http://stanfordnlp.github.io/stanfordnlp/
[20] https://stanfordnlp.github.io/stanfordnlp/installation_usage.html#getting-started
[21] Por falta de memória RAM foi necessário usar o servidor da Google (Google Colabs) para instalar a biblioteca (seguir estes passos: https://stanfordnlp.github.io/stanfordnlp/).
[22] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[23] https://github.com/miotto/treetagger-python
[24] https://freeling-tutorial.readthedocs.io/en/latest/
[25] https://www.linguakit.com/es/analisis-completo
[26] https://github.com/citiususc/Linguakit

2.10. Modelos pré-treinados para a tarefa de Reconhecimento de Entidades Mencionadas

O trabalho descrito por Pires (2017) fornece modelos pré-treinados para a tarefa de REM, para o português europeu. Estes modelos – doravante modelos-SIGARRA – foram treinados com as ferramentas OpenNLP, StanfordNLP, SpaCy[27] e NLTK, com base no *corpus* SIGARRA NEWS. Neste trabalho, vamos também avaliar estes modelos tendo em conta a tarefa de REM.

3. *Setup* Experimental

O primeiro passo foi criar um *corpus* de referência (ou *colecção dourada* ou *corpus* de teste) para as duas tarefas em análise. Esse *corpus* é descrito na Secção 3.1. O segundo passo foi atender a especificidades de cada ferramenta, o que levou a modificações do *corpus* de referência (Secção 3.2) no que respeita à segmentação feita pelas diferentes ferramentas. De seguida, foram ainda definidos vários *scripts* que transformam etiquetas do *corpus* criado nas etiquetas usadas pelas diferentes ferramentas (Secções 3.3 e 3.4 para a conversão de etiquetas na tarefa de EMS e REM, respetivamente). Finalmente, apresentam-se as medidas de avaliação, na Secção 3.6.

3.1. Construção do *Corpus* de Referência

O *corpus* de referência é composto por 101 frases retiradas de revistas, jornais e livros portugueses, disponíveis *on-line*. Para efeitos deste estudo considerou-se um *corpus* de referência reduzido, ainda que construído tendo em conta alguma diversidade e proporção das fontes utilizadas, nomeadamente, por ser constituído por diferentes tipos de texto e, dentro de cada tipo, com fontes diversas, para abranger diferentes usos da língua. Dessas frases, 59% pertencem a revistas como a Visão[28] e a Exame Informática[29], 29% são de jornais (jornal Observador[30] e as restantes do Público[31]) e 13% pertencem ao livro "O Príncipe com Orelhas de Burro", de José Régio (Tabela 1).

Antes da anotação das frases, foram corrigidos pequenos erros ortográficos, pois este não era um ponto de avaliação. Assume-se assim que os textos que chegam às diferentes ferramentas estão limpos de erros ortográficos.

Tipo	Fonte	%
Revistas	Visão	30
	Exame Informática	29
Jornais	Observador	27
	Público	2
Livro		13

Tabela 1: Composição do *corpus* de referência.

Seguidamente, foram anotadas as classes gramaticais e informações de flexão associadas a cada palavra e respetiva classe. Por exemplo, para um substantivo, além do seu tipo (comum ou próprio), foram igualmente anotados o seu género e número. Da mesma forma, para os verbos foram anotados o modo, o tempo, a pessoa e o número. Repetiu-se este processo para as outras classes gramaticais. A Tabela 2 apresenta a frequência de cada classe e a Tabela 3 lista a frequência das entidades nomeadas na coleção dourada.

De notar que esta tarefa de anotação não foi realizada por um perito. A ideia não é ter um corpus perfeito sob o ponto de vista da anotação, mas um corpus que representasse o que um não-perito consideraria interessante/correto anotar.

3.2. Adaptação às diferentes ferramentas

Como foi dito, cada ferramenta apresenta as suas particularidades no que respeita ao modo como processa texto, o que implica que a coleção dourada anterior não possa ser usada sem antes ser pré-processada de acordo com as características de cada ferramenta. Por exemplo, a maior parte destas ferramentas não divide contrações. Assim, a coleção dourada teve de ser adaptada a algumas ferramentas: Polyglot, NLTK, SpaCy e OpenNLP. Por outro lado, as ferramentas FreeLing e LinguaKit permitem unir locuções e nomes compostos, sendo uma opção que decidimos explorar. Por exemplo, a locução "a partir de", apesar de ter 3 palavras, é vista como uma unidade lexical, o que é representado ligando os seus elementos por *underscore*: "a_partir_de". Esta particularidade implica que, para estas ferramentas, sejam anotadas como uma unidade as locuções e nomes compostos da colecção dourada, ao invés de serem anotados separadamente os elementos gramaticais que compõem essas expressões.

[27] As instruções para o carregamento do modelo da ferramenta SpaCy não se ajustaram ao ficheiro, pelo que não foi possível usar esse modelo. A razão pode ter ficado a dever-se à incompatibilidade da versão mais recente da ferramenta com o modelo.

[28] https://visao.sapo.pt

[29] https://visao.sapo.pt/exameinformatica/

[30] https://observador.pt.

[31] https://www.publico.pt

Classes	Sub-classes	Frequência
Adjetivo (Adj)		179
Advérbio (Adv)	Negação (N)	15
	Normal (G)	73
Conjunção (Conj)	Coordenativa (C)	93
	Subordinativa (S)	58
Determinante (Det)	Artigo (Art)	468
	Indefinido (Ind)	24
	Possessivo (Poss)	14
	Demonstrativo (Dem)	35
	Interrogativo (Int)	2
Nome (Nom)	Comum (C)	699
	Próprio (P)	109
Numeral (Num)		77
Preposição (Prep)		525
Pronome (Pron)	Pessoal (Pes)	23
	Indefinido (Ind)	4
	Relativo (Rel)	43
	Demonstrativo (Dem)	16
Verbo (Verb)	Auxiliar (Aux)	82
	Indicativo (Ind)	142
	Condicional (Cond)	3
	Conjuntivo (Conj)	5
	Gerúndio (Ger)	13
	Particípio Passado (PP)	98
	Infinitivo (Inf)	74

Tabela 2: Frequência de cada classe morfossintática na coleção dourada.

Classes	Frequência
Organização	30
Localização	22
Pessoa	8
Data	21

Tabela 3: Frequência de cada classe de entidades nomeadas na coleção dourada.

A forma como as unidades textuais são segmentadas (*segmentação*) varia igualmente.. Por exemplo, algumas ferramentas consideram "quarta-feira" como um único token e outras separam cada elemento da palavra ("quarta", "-" e "feira"). Já o OpenNLP considera os verbos ligados a pronomes clíticos como um único token, como sucede em "reuniram-se" e em "escondê-lo". Esta variedade de características, tornou árdua a compatibilidade da coleção dourada principal com todas as ferramentas, pelo que foram criadas diferentes coleções douradas, específicas para cada ferramenta.

No que respeita à tarefa de Reconhecimento de Entidades Mencionadas, foram consideradas as entidades específicas de cada uma das ferramentas, de modo a que as saídas das diversas ferramentas fossem compatibilizadas com as entidades consideradas na coleção dourada.

3.3. Sobre as Classes Morfossintáticas

Cada ferramenta tem o seu sistema de classes ou etiquetas morfossintáticas para as tarefas de PLN. O acesso a esta informação facilita a conversão automática das anotações da coleção dourada nas etiquetas das diferentes ferramentas. Esta tarefa não é, de todo, trivial devido às diferenças nas etiquetas das ferramentas e, em alguns casos, devido à indisponibilidade de um manual de utilizador que descrevesse o conjunto de símbolos possíveis. É o caso da ferramenta SpaCy, que não possui uma descrição das etiquetas morfossintáticas. No entanto, foi possível, na maioria dos casos, mapear sem problemas de maior, as etiquetas usadas pelo *corpus* de re-

ferência nas etiquetas das diferentes ferramentas. De notar ainda que algumas ferramentas contribuem com etiquetas com uma maior granularidade (por exemplo, o FreeLing, tal como se verá adiante).

3.3.1. NLTK, OpenNLP e NLPyPort

O sistema de etiquetas do NLTK corresponde ao conjunto de etiquetas com a categoria gramatical utilizado na anotação dos *corpora* da Floresta Sintática[32]. Quanto ao OpenNLP, não se encontrou, na documentação da ferramenta, uma descrição do conjunto de etiquetas para o modelo em português, apenas para a língua inglesa. No entanto, verificou-se que o sistema de etiquetas do OpenNLP é semelhante ao da ferramenta NLTK, existindo uma ligeira diferença entre ambos no tratamento de sinais de pontuação: o OpenNLP agrupa os sinais de pontuação sob uma única etiqueta, ao contrário do NLTK, que considera cada sinal de pontuação como uma classe distinta. Desta forma, para a conversão das anotações, procedeu-se de maneira semelhante à da ferramenta NLTK, à exceção dos sinais de pontuação que precisaram de um outro processamento. Por outro lado, as etiquetas consideradas nestas duas ferramentas não contêm informação de flexão. Além disso, o NLTK e o OpenNLP distinguem apenas verbos no infinitivo, gerúndio e particípio passado; todos as restantes formas verbais são classificadas de forma indiferenciada como "verbos finitos". Relativamente aos tipos de Pronomes só são identificados 3 tipos (v.g. determinativos, pessoais e independentes). A existência de uma descrição do conjunto de etiquetas facilitou a conversão das anotações da coleção dourada, apesar de não ser totalmente clara a categorização de alguns determinantes e pronomes, como por exemplo a classe *pron-det* (pronome determinativo). Por esta razão, as etiquetas relativas aos pronomes e determinantes serão tratadas à parte.

Quanto ao NLPyPort, avaliou-se apenas o modelo pré-treinado para a tarefa de EMS. Dado que os recursos da ferramenta NLPyPort se baseiam na ferramenta NLTK e o modelo para esta tarefa foi treinado com os *corpora* Bosque da Floresta Sintática e Mac-Morpho (Fonseca & Rosa, 2013), o conjunto de etiquetas morfossintáticas assemelha-se ao conjunto de etiquetas da ferramenta NLTK. Contudo, tal como no OpenNLP, os sinais de pontuação são agrupados numa mesma etiqueta, *punc*. Existe ainda uma etiqueta adicional (etiqueta por omissão) atribuída a um token quando o sistema não consegue identificar a sua classe gramatical. Essa etiqueta ("N") corresponde à classe gramatical nome. Assim, existem duas etiquetas diferentes ("N" e "n") para a classe nome comum e decidimos tratar estas duas etiquetas como sendo a mesma.

3.3.2. Polyglot e StanfordNLP

As ferramentas Polyglot e StanfordNLP baseiam as suas etiquetas nas da CoNLL-U (Buchholz & Marsi, 2006), compostas por vários elementos que descrevem morfossintaticamente uma palavra. Um dos elementos corresponde à classe gramatical universal, UPOS[33]; outro dos elementos vem do FEATS[34], que descreve informações morfológicas associadas à palavra. Com este conjunto universal de etiquetas, a correspondência entre as anotações da coleção de referência e as classes gramaticais realizou-se de uma forma mais simples. No entanto, o Polyglot só apresenta informações sobre a classe gramatical universal. Em particular, o conjunto de etiquetas morfossintáticas usado nos modelos do Polyglot corresponde às classes gramaticais principais da Tabela 2, como adjetivo, determinante, advérbio, etc.

3.3.3. SpaCy

No que respeita ao SpaCy, e tal como dito anteriormente, apesar de ter uma função que permite obter a descrição de uma determinada etiqueta, o sistema não disponibiliza um glossário completo com a descrição das etiquetas relativas à análise de texto em português, o que dificultou a compreensão dos resultados obtidos. Assim, o seu sistema de etiquetas não é claro, apesar de, aparentemente, se basear igualmente nos mesmos formalismos standard do Polyglot e StanfordNLP. No entanto, partilha as etiquetas relativas aos pronomes e determinantes do NLTK e OpenNLP. A conversão automática tornou-se mais complexa para esta ferramenta.

3.3.4. FreeLing, LinguaKit e TreeTagger

No manual de utilizador da ferramenta FreeLing encontram-se descritos os conjuntos de etiquetas morfossintáticas e de entidades nomeadas. Além das classes gramaticais principais, as etiquetas contêm informações sobre outros aspetos morfológicos como o género, número, modo e tempo

[32] https://www.linguateca.pt/

[33] https://universaldependencies.org/u/pos/.

[34] https://universaldependencies.org/u/feat/index.html

verbal. Por exemplo, para os adjetivos, as classes existentes têm em conta o tipo, o género, grau e número. O sistema de etiquetas desta ferramenta é extenso, tornando a conversão automática complexa. Porém, a descrição clara de cada classe no manual de utilizador evitou dificuldades na correspondência entre as anotações da coleção dourada e as etiquetas. De acordo com os autores[35], o LinguaKit tem as mesmas etiquetas que o FreeLing[36]. No entanto, foram encontradas pequenas diferenças.

O TreeTagger utiliza um sistema de etiquetas semelhante, no qual, além das classes gramaticais principais, são representadas outras informações morfológicas das palavras[37]. No entanto, as classes não possuem o mesmo nível de especificidade que as do FreeLing. Por exemplo, apenas a classe dos nomes incorpora informações sobre flexão em género e número. A descrição do conjunto das classes também é clara. Quanto às etiquetas dos verbos, só apresentam informação sobre o tipo do verbo (auxiliar ou principal) e os modos verbais. Tal como na ferramenta FreeLing, as contrações são resolvidas automaticamente e identificadas com o sinal "+". Por exemplo, a contração "das" é identificada com "SPS + DA" por ser a contração de uma Preposição (SPS) com o Artigo Definido (DA) ("de + as").

3.4. Sobre as Entidades Mencionadas

Como dito anteriormente, consideraremos os modelos pré-treinados para a tarefa de REM, para o português europeu, tais como descritos em Pires (2017) (os modelos-SIGARRA) e treinados com as ferramentas OpenNLP, StanfordNLP e NLTK, com base no *corpus* SIGARRA NEWS[38]. Este *corpus* é composto por 1.000 artigos retirados da secção de notícias do sistema de informação da Universidade do Porto (SIGARRA) e, de acordo com Pires (2017), o seu conjunto de etiquetas corresponde a oito classes relacionadas com o domínio dos *corpora*: Hora, Evento, Organização, Curso, Pessoa, Localização, Data e UnidadeOrganica.

No que respeita às restantes ferramentas, o FreeLing e o LinguaKit apresentam as seguintes classes de entidades e as respetivas etiquetas: Pessoa (NP00SP0), Localização (NP00G00), Organização (NP00O00) e Outros (NP00V00). Esta última etiqueta corresponde a entidades nomeadas que não se integram em nenhuma das categorias anteriores. No entanto, por ser uma classe ambígua, tokens que não são reconhecidos como entidades na coleção dourada como "Verão" são classificados como Outros. Para facilitar a avaliação automática, foi adicionada mais uma etiqueta ("O") que identifica os tokens que não são entidades. Finalmente, quanto ao Polyglot, este deteta apenas 3 classes de entidades: Pessoa (I-PER), Localização (I-LOC) e Organização (I-ORG). Tal como para a ferramenta anterior foi adicionada a etiqueta "O", com o mesmo significado.

3.5. Sobre a Análise de Dependências

No que respeita à tarefa de AD, cinco frases foram aleatoriamente retiradas do corpus e processadas pelas ferramentas SpaCy e StanfordNLP. As frases são as seguintes:

1. Num comunicado enviado às redacções, o gabinete do primeiro-ministro fazia saber que considerava que a interpretação da lei que defendia a demissão imediata de um governante por negócios de empresas de familiares com entidades públicas, mesmo que estas nada tivessem a ver com o titular de cargo político, "ultrapassa largamente, no seu âmbito e consequências, o que tem sido a prática corrente ao longo dos anos".
2. Em 2016, vários membros do governo americano que estavam a prestar serviços em Havana, Cuba , assim como os seus familiares , começaram a queixar-se de uma série de sintomas neurológicos, incluindo dificuldades de concentração e memória, tonturas e problemas visuais e de equilíbrio.
3. Os sintomas foram associados à exposição a sons repentinos e de grande intensidade e volume de uma fonte desconhecida que os pacientes reportam ter ouvido nas suas casas e quartos de hotel.
4. Nem o seu marido, porém, voltou a reparar nesse indício que uma tarde lhe gelara nos lábios palavras de exprobração e cólera.
5. Um homem de 44 anos foi identificado.

Seguidamente, os resultados do processamento foram dados a um especialista, para análise, sem que fossem identificadas as ferramentas. Este analisou os resultados das frases dadas tendo em conta a pontuação, segmentação, etiquetagem morfossintática e, finalmente, as dependências.

[35] https://gramatica.usc.es/pln/tools/CitiusTools.html

[36] https://github.com/citiususc/Linguakit/blob/master/tagger/tagset_pt-es-gl.html

[37] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Portuguese-Tagset.html.

[38] rdm.inesctec.pt/ro/dataset/cs-2017-006

3.6. Medidas de Avaliação

Estando as etiquetas da coleção dourada em conformidade com as etiquetas de cada ferramenta, o próximo passo consistiu em decidir que medidas usar na avaliação. Optámos por usar a Micro- e a Macro-Média da medida F1, tal como descritas por Jurafsky & Martin (2019) e implementadas no SCIKIT-LEARN[39], pois também são usadas por outros trabalhos nesta área, tais como o proposto por Garcia & Gamallo (2015).

4. Etiquetação Morfossintática

Começamos por discutir alguns aspectos relativos aos modelos criados pelo NLTK. De seguida, apresentamos e discutimos os resultados obtidos pelos diferentes modelos e ferramentas.

4.1. Sobre o Modelos do NLTK

Todas as ferramentas com a excepção do NLTK oferecem modelos pré-treinados. Assim, para esta ferramenta, foram treinados vários modelos. Um desses modelos é o modelo de Bigramas, tal como apresentado nos manuais do NLTK. Foram ainda implementados modelos baseados na Máxima Entropia e ainda um modelo denominado Perceptrão, pois o OpenNLP fornece modelos baseados nestes dois algoritmos e considerámos interessante a comparação. De notar que poderiam ter sido implementados modelos mais adequados à predição de sequências, como os HMM ou os CRF (Lafferty et al. (2001)) ou ainda modelos baseados em redes neuronais. No entanto, a ideia foi testar o que a ferramenta oferece directamente. Todos estes modelos foram treinados nos *corpora* "Floresta Sintática", disponíveis com o NLTK[40]. De notar que a versão disponível no NLTK da Floresta Sintática não contém alguns sinais de pontuação como dois-pontos <:>, reticências $< \cdots >$, parênteses curvos e o símbolo de percentagem $< \% >$, pelo que estes items não são previstos pelos diferentes modelos. A divisão em tokens das frases foi realizada pela função *word_tokenize*[41]. A criação de qualquer um dos modelos não requer um grande esforço na ótica de um informático, pois consiste em importar uma biblioteca própria para o efeito. No entanto, alguns modelos têm as suas especificidades. Assim, para o modelo de bigramas, cada bigrama é um *token*, tal como identificado pelo *tokenizador*; no caso de aparecer um *token* nunca visto no treino, decidiu-se atribuir, por omissão, a etiqueta "n" (classe gramatical Nome), por ser a classe mais comum. Nos casos em que os modelos requeriam a recolha de características (*features*) dos dados, foram usadas características muito simples, como a própria palavra, a palavra anterior e a seguinte, se a palavra corrente começava com maiúscula, etc. A seleção das *features* dita o desempenho da ferramenta. No entanto, está fora do âmbito deste trabalho desenvolver um estudo exaustivo sobre as *features* a utilizar.

4.2. Resultados Globais

A Tabela 4 apresenta os resultados globais dos diferentes modelos, tendo em conta a totalidade das etiquetas (de cada ferramenta) e considerando a Micro- e a Macro-média relativas à medida F1, tal como definidas anteriormente. Devemos realçar que estes valores não permitem uma comparação totalmente justa dos diferentes modelos, pois, como dito anteriormente, os valores de Micro- e Macro-Média da F1 são calculados com base no conjunto de etiquetas de cada ferramenta, que varia entre estas. Assim, estes valores devem dar apenas uma ideia geral. De modo a comparar de forma justa estes modelos, a secção seguinte detalha os valores para estas medidas sem ter em conta as informações de flexão.

4.3. Resultados por Classe Gramatical

A Tabela 5 permite comparar os diferentes modelos tendo em conta as categorias que têm em comum e sem ter em conta as informações de flexão. De notar que são apenas mostradas as classes gramaticais mais relevantes ('Bg' representa o modelo baseado em Bigramas, 'P' o Perceptrão, 'ME' a Máxima Entropia, 'NLPy' o NLPyPort, 'PG' o Polyglot, 'TT' o TreeTagger, 'FL' o FreeLing, 'StfNLP' o StanfordNLP e 'ONLP' o OpenNLP). As Tabelas 6 e 7 apresentam os resultados para os dois grupos de modelos que partilham algumas etiquetas específicas.

4.4. Discussão

Os melhores resultados globais são obtidos pelo OpenNLP, Máxima Entropia, 91% de Macro-Média e aos 94% de Micro-Média, porém, é importante realçar que o nível de detalhe do conjunto de etiquetas da ferramenta não é tão fino

[39] https://scikit-learn.org

[40] Os modelos disponibilizados pelo OpenNLP, SpaCy e StanfordNLP também utilizam este *corpus*.

[41] Esta função tem a particularidade de mudar as aspas iniciais <"> de uma frase para dois acentos graves <``> e as finais <"> para duas plicas <''>. Por isso, é necessária uma reconversão para aspas, antes do treino e teste dos dados.

Ferramenta	Micro-Média	Macro-Média
NLTK (Bigramas)	0.87	0.69
NLTK (Perceptrão)	0.89	0.71
NLTK (Máxima Entropia)	0.93	0.74
NLPyPort	0.86	0.83
Polyglot	0.79	0.68
Spacy	0.90	0.47
FreeLing	0.90	0.58
LinguaKit	0.83	0.48
TreeTagger	0.91	0.73
StanfordNLP	0.91	0.61
OpenNLP (Perceptrão)	0.93	0.77
OpenNLP (Máxima Entropia)	**0.94**	**0.91**

Tabela 4: Micro- e Macro-Média relativos a F1 para os diferentes modelos.

Classes		NLTK Bg	NLTK P	NLTK ME	NLPy	PG	SpaCy	TT	FL	LinguaKit	StfNLP	ONLP ME	ONLP P
Adj		0.76	0.82	0.83	0.75	0.63	0.89	0.92	**0.96**	0.81	0.94	0.88	0.85
Adv	G	0.81	0.77	0.81	0.80	0.46	0.85	**0.91**	0.84	0.85	0.89	0.82	0.81
	N							**0.97**	**0.97**	0.93	**0.97**		
Conj	C	0.97	0.95	0.96	0.97	0.94	0.95	0.97	0.94	0.97	0.97	0.97	**0.98**
	S	0.56	0.72	0.60	0.60	0.56	0.68	0.87	**0.88**	0.75	0.74	0.70	0.70
Det	Art	0.95	0.93	0.96	0.97	-	0.97	0.94	**0.99**	0.95	0.98	0.96	0.97
Nom	C	0.82	0.91	0.93	0.81	0.93	0.95	0.94	**0.98**	0.92	**0.98**	0.94	0.94
	P	0.33	0.33	0.81	0.33	0.68	0.85	0.51	0.96	**0.97**	0.85	0.77	0.75
Num		0.84	0.95	**0.98**	0.83	0.72	0.96	**0.98**	0.95	0.80	0.97	0.94	0.97
Prep		0.95	0.95	0.96	0.97	0.85	0.96	0.97	**0.99**	0.95	0.96	0.96	0.96
Pron	Pes	0.97	0.90	0.90	**1.00**	-	0.83	0.98	0.84	0.70	**1.00**	0.91	0.94
Verb	Aux					0.00	**0.84**	0.00	0.00	0.00	0.73		
	Ind	0.88	0.94	0.97	0.88		**0.88**	0.82	0.83	0.81	0.85	0.96	0.97
	Cond						0.50	*		*	**0.80**		
	Conj					0.62	0.36	0.74	0.71	0.78	**0.92**		
	Ger	0.91	1.00	0.91	0.91		0.95	0.91	0.91	0.76	0.95	**1.00**	0.96
	PP	0.82	0.95	0.92	0.82		0.94	**0.96**	0.95	0.79	0.93	**0.96**	**0.96**
	Inf	0.89	0.93	0.95	0.89		0.91	0.92	0.94	0.91	0.95	**0.96**	0.95

Tabela 5: Valores de F1 para as categorias comuns. O * indica que o Condicional é visto como Indicativo por estes sistemas

Classes		Polyglot	TreeTagger	FreeLing	LinguaKit	StanfordNLP
Det	Ind	0.79	0.96	**1.00**	0.96	0.91
	Poss		1.00	1.00	0.70	1.00
	Dem		0.89	1.00	0.89	0.97
	Int		0.00	0.00	0.00	1.00
Pron	Ind	0.70	0.46	**0.89**	0.43	0.75
	Rel		0.93	**0.94**	0.88	0.94
	Dem		0.79	0.86	0.76	**0.90**

Tabela 6: Valores de F1 para as ferramentas que têm as categorias **Det** e **Pron** mais finas

Classes	NLTK			NLPy	SpaCy	OpenNLP	
	Bigramas	Perc.	ME			ME	Perc.
Pron-det	0.83	0.79	0.86	0.88	0.81	**0.89**	0.88
Pron-indp	0.80	0.88	0.82	0.79	**0.91**	0.86	0.88

Tabela 7: Valores de F1 para as etiquetas **Pron-det** e **Pron-indp**. A primeira contém os determinantes, pronomes demonstrativos, pronomes interrogativos, pronomes possessivos e pronomes relativos; a segunda os pronomes indefinidos e outros pronomes de outras categorias que expressam imprecisão.

quanto nos modelos anteriores, pois, tal como referido anteriormente, o conjunto de etiquetas do OpenNLP não contém informações de flexão de número, género e tempos verbais (ao contrário de outras ferramentas como o FreeLing, o StanfordNLP e o TreeTagger, que apresentam etiquetas mais finas).

Quanto à avaliação por classe, a ferramenta StanfordNLP é aquela que apresenta maiores valores na previsão de verbos no conjuntivo e no condicional. O FreeLing apresenta melhor desempenho na classificação de adjetivos e conjunções subordinativas. O LinguaKit ultrapassa Freeling na previsão de nomes próprios, alcançando 97% de F1-measure. Por fim, o TreeTagger destaca-se na previsão de advérbios (de notar que nesta ferramenta o condicional não é considerado modo mas sim um tempo verbal do modo indicativo).

Existem ainda algumas particularidades das ferramentas, que merecem ser discutidas. Por exemplo, como dito anteriormente, o FreeLing e o LinguaKit reconhecem nomes compostos como um único *token* (por exemplo, "Eduardo Cabrita" é tratado como um *token* único "Eduardo_Cabrita"). Esta particularidade auxilia a classificação de nomes próprios compostos, evitando assim que preposições e nomes próprios sejam incorretamente classificados, o que é constante nas outras ferramentas. Desta forma, estas duas ferramentas obtiveram o melhor desempenho na previsão de nomes próprios. Infelizmente estas ferramentas não classificam corretamente verbos auxiliares (VA). Neste aspecto, a melhor ferramenta é o SpaCy (84%), apesar de o StanfordNLP também conseguir identificar verbos auxiliares (73%). O FreeLing e a LinguaKit têm também outras características que as tornam interessantes: palavras relacionadas com datas são unidas e consideradas como um só token, à semelhança dos nomes próprios, como "Novembro de 2018", que é tratado como "Novembro_de_2018". Para a ferramenta FreeLing, o critério de atribuição desta classe é, contudo, pouco claro, por não haver uma descrição desta etiqueta. Estas ferramentas também unem numerais, por exemplo, consideram "10 mil milhões" como um único *token*. No que diz respeito apenas ao FreeLing, estas uniões não são totalmente precisas, levando a erros de classificação de alguns *tokens*, como acontece na expressão "15 e 35", que é considerado como um só *token* "15_e_35". Esta situação leva, posteriormente, a uma incorreta classificação, neste caso atribuindo a classe Interjeição.

5. Reconhecimento de Entidades

Nesta subsecção serão apresentados os resultados e conclusões sobre o desempenho das ferramentas na tarefa de REM.

5.1. Sobre os modelos-SIGARRA

No que diz respeito ao NLTK, tal como previamente indicado, estudaram-se três modelos-SIGARRA, pré-treinados no corpus SIGARRA NEWS: modelo baseado em Árvores de Decisão, no Naïve Bayes e na Máxima Entropia. Os modelos avaliados associados ao OpenNLP e StanfordNLP são também os referidos modelos-SIGARRA, treinados no mesmo corpus, mas com estas ferramentas.

5.2. Resultados Globais

A Tabela 8 mostra os resultados globais dos diferentes modelos, tendo em conta a Micro- e a Macro-Média relativas à medida F1. O Freeling e o LinguaKit são os melhores sistemas quanto à Micro-Média e o StanfordNLP quanto à Macro-Média. De notar que, sendo os modelos do NLTK, StanfordNLP e OpenNLP, modelos-SIGARRA, há uma diferença substancial de valores quanto à Macro-Média (o modelo Naïve Bayes com 0.18 e o StanfordNLP 0.78), assunto que se discutirá mais à frente.

Ferramenta	Micro-Média F1	Macro-Média F1
NLTK (Árvore de Decisão)	0.97	0.47
NLTK (Naïve Bayes)	0.92	0.18
NLTK (Máxima Entropia)	0.97	0.35
Polyglot	0.98	0.76
FreeLing	**0.99**	0.77
LinguaKit	**0.99**	0.69
StanfordNLP	0.98	**0.78**
OpenNLP	0.97	0.46

Tabela 8: Micro- e Macro-Média relativos a F1 para os diferentes modelos.

5.3. Resultados por Tipo de Entidade

A Tabela 9 apresenta a comparação entre os valores de F1-measure de cada classe e para cada ferramenta na tarefa de REM. O StanfordNLP é a ferramenta com bons (ou os melhores) em praticamente todas as classes, à exceção das classes Localização e Organização, nas quais é a ferramenta FreeLing que se destaca.

5.4. Discussão

Como dito anteriormente, há uma grande diferença de valores quanto à Micro- e Macro-Média em alguns modelos. Na verdade, os resultados apresentados mostram quão enganadora pode ser a Micro-Média (F1). Esta medida tem em conta a soma de todos os Verdadeiros/Falsos Positivos/Negativos de todas as classes, sendo calculada posteriormente a Precisão e a Cobertura, e, finalmente, a F1. Ora, todos os casos que não são considerados como entidades mencionadas e não são de facto entidades mencionadas (isto é, o grosso das palavras, pois a maioria das palavras de um texto não são entidades mencionadas) contam como Verdadeiros Positivos, indicando que esta métrica não é nada informativa neste cenário em que as classes não são balanceadas.

Por outro lado, e como referido, os algoritmos usados têm um papel extremamente relevante na tarefa de NER. Os resultados dos diferentes algoritmos na base dos modelos-SIGARRA treinados com o NLTK ilustram bem essa situação.

É também interessante verificar como a estratégia de segmentação usada pelas ferramentas pode fazer a diferença. No caso do FreeLing e do LinguaKit, a divisão em tokens destas ferramentas foi um alicerce na classificação correta de entidades nomeadas, pelo simples facto de preservar nomes compostos, tornando possível a identificação de entidades compostas como , "Estados Unidos da América" e "Diário de Notícias". Outra característica que se realça corresponde à sua capacidade para identificar entidades nomeadas estrangeiras como "Sujoy Ghosh", "Einstein" e "Xuekun Fang". Em contrapartida, verificou-se que a ferramenta FreeLing considerou incorretamente alguns *tokens* presentes no início das frases como entidades, talvez por começarem com letra maiúscula. Segue-se um exemplo desse caso: "*Contactada* pela Lusa...". Em comparação ao FreeLing, o LinguaKit apresenta piores resultados, principalmente na classificação de localizações e organizações. Estes são alguns exemplos de entidades classificadas erroneamente: "ANA"(Organização) e "América do Sul"(Localização) foram considerados como "Pessoa", "Xuekun Fang" (Pessoa) e "Ásia do Sul" (Localização) como "Outros".

A análise da classificação do Polyglot revelou que esta considera nacionalidades tais como "mexicanos", "dinamarquesa" e "americanos" sempre como Localização. No entanto, esta ferramenta consegue identificar entidades nomeadas estrangeiras como "Einstein", "Kaiserslautern" e "Sujoy Ghosh" e algumas entidades nomeadas compostas como "Estados Unidos", "Universidade de Aveiro" e "Physical Review". Já o OpenNLP tem dificuldades a identificar entidades nomeadas estrangeiras, como por exemplo, "Netflix", "Maximilian Gunther", "Einstein"; por outro lado, não é totalmente capaz de classificar nomes compostos. Por exemplo, "Associação de Proteção e Socorro" e "Universidade de Londres" são corretamente identificadas. Contudo, outras entidades como "Universidade da Califórnia", "Diário de Notícias", "Estados Unidos" e "Eduardo Cabrita" já não o são. Outra conclusão retirada é que não identifica siglas como "EUA", "MIT", "PSML", "EHT", etc.

O StanfordNLP, por seu turno, consegue identificar entidades nomeadas noutras línguas, incluindo siglas. No entanto, verificou-se que, quando uma sigla está entre parênteses, a fer-

				NLTK				
Classes	FreeLing	LinguaKit	Polyglot	AD	NB	EM	OpenNLP	StanfordNLP
Data	–	–	–	0.78	0.11	0.74	0.76	**0.92**
Localizacao	**0.92**	0.82	0.73	0.51	0.06	0.17	0.00	0.34
Organizacao	**0.84**	0.63	0.61	0.62	0.34	0.58	0.70	0.75
Pessoa	0.67	0.63	0.70	0.42	0.00	0.35	0.32	**0.88**

Tabela 9: Valores de F1 de cada ferramenta, tendo em conta as entidades nomeadas da coleção dourada.

ramenta identifica os parênteses como fazendo parte da entidade nomeada. Por exemplo, dada a sigla MIT neste formato "(MIT)", além de MIT ser considerada como Organização, o parêntese "( )" também o é. De entre as ferramentas analisadas para esta tarefa, o StanfordNLP apresenta um melhor desempenho em praticamente todas as classes, à excepção da classe Localização e Organização, onde o vencedor é o FreeLing, como anteriormente referido.

6. Análise de Dependências

Nesta secção, apresentamos o estudo qualitativo relativo à tarefa de AD. Vários parâmetros são analisados: pontuação, segmentação, etiquetagem morfossintática e dependências. Como dito anteriormente, este trabalho foi executado sem conhecimento dos sistemas que produziram as anotações. Assim, doravante, o Sistema A corresponde ao SpaCy e Sistema B ao StanfordNLP.

6.1. Pontuação

Os dois sistemas adotam diferentes estratégias de segmentação relativamente aos sinais de pontuação e à sua análise.

No Sistema A, os sinais de pontuação (vírgulas, aspas e pontos finais) aparecem ligados ao token anterior, e.g. (Frase 1) *redacções* **,** ; *públicas* **,** ; *político* **,"** ; *anos* **".** (negritos nossos), e não parecem receber qualquer análise, não havendo nenhuma dependência associada. Note-se, além disso, que as aspas junto a *político* emparelham com as aspas ligadas a *anos*, mas este emparelhamento não parece ser feito pelo sistema.

Por outro lado, no Sistema B, os sinais de pontuação são tratados como tokens independentes e sobre eles recai (sobretudo) uma dependência específica: `punct`. No entanto, não é evidente o significado que a dependência exprime, quando se considera o elemento que opera sobre o sinal de pontuação, a sua função na frase ou outros sinais com que se encontra articulado. Assim, por exemplo, na Frase 1 (Figura 1), o adjunto deslocado para o início de frase (v.g. *Em um comunicado enviado a as redações* **,**) apresenta a dependência `punct` ligando o particípio *enviado* à vírgula; este particípio depende de/modifica (dependência `acl`, *adjectival clause*, *clausal modifier of noun*) o nome *comunicado*, que é a cabeça ou núcleo deste constituinte, pelo que todos os outros elementos flecham direta ou indiretamente sobre este nome. Por sua vez, este nome depende (`obl`; complemento oblíquo) do verbo principal da oração seguinte (*fazia*). Ora, se a função da vírgula neste exemplo é assinalar a deslocação (anteposição) do adjunto complemento de *fazer*, a partir da sua posição básica, para o início da frase; esta forma de representação, não dá conta de forma clara da função sintática que a vírgula exerce, já que não faz depender a vírgula do núcleo sintático do complemento (*comunicado*), mas sim de um dos seus modificadores (*enviado*).

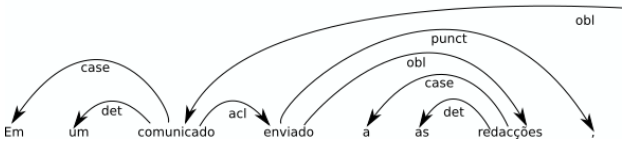


Figura 1: "Num comunicado enviado às redacções, . . . ", Sistema B (Frase 1).

Mais adiante, e ainda no Sistema B e na mesma frase, duas vírgulas delimitam a oração subordinada concessiva (**,** *mesmo que estas nada tivessem a ver com. . .* **,**) e encontram-se (e bem!) "emparelhadas", estando ambas dependentes do verbo finito desta oração (*tivessem*). O mesmo sucede (Figura 2), no complemento adjunto (. . . *político* **,** **"** *ultrapassa largamente* **,** *no seu âmbito e consequências* **,**), também delimitado por vírgulas, e que são emparelhadas e postas na dependência do verbo anterior (*ultrapassa*). Contudo, no caso das aspas, as primeiras (de abertura), estas são feitas depender do verbo *ultrapassa* que está à sua direita e (estranhamente) com a dependência `nsubj`, que corresponde à função de sujeito; enquanto as segundas (de fecho), são colocadas na dependência de *anos*

(no fim da frase, fora da Figura 2), mas agora com a etiqueta `amod`, que corresponde à função de modificador (adjetival) de nome.

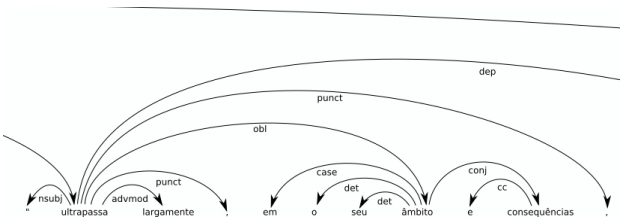


Figura 2: “ **“** ultrapassa largamente **,** em o seu âmbito e consequências **,** ”, Sistema B (Frase 1).

É, assim, óbvia a inconsistência da anotação das dependências sobre os sinais de pontuação, quer pelo uso de etiquetas que não podem ser aplicadas a este tipo de tokens (`nsubj` e `amod`), quer pelos elementos textuais de que se faz depender os sinais de pontuação, quer ainda pelo incorreto emparelhamento de sinais que funcionam em conjunto (aspas e vírgulas).

6.2. Segmentação e etiquetagem morfossintática

Ambos os sistemas apresentam diversos problemas de segmentação de texto (delimitação de tokens) e de marcação de categorias morfossintáticas. Várias unidades lexicais multipalavras (ou palavras compostas/locuções) são analisadas como tokens distintos. No Sistema A, por exemplo, encontramos *primeiro-*/`adj` e *ministro*/`noun` e não o nome composto *primeiro-ministro*/`noun`; *mesmo*/`adv` e *que*/`conj` e não a locução conjuntiva (ou conjunção composta *mesmo que*/`conj`; *cargo*/`noun` e *político*/`adj` e não o composto *cargo político*/`noun`; e a sequência *a*/`prep` *o*/`art` *longo*/`adj` *de*/`prep` não é tratado como uma locução preposicional (ou preposição composta), *ao longo de*/`prep`. Refira-se, ainda, a misteriosa etiqueta `x` (Figura 3) atribuída ao pronome da (assim chamada) oração relativa sem antecedente (Veloso (2013)), v.g. *o*/`X` *que tem sido*.

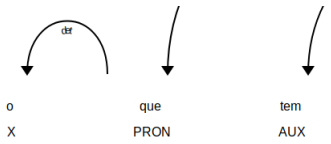


Figura 3: Etiqueta `x`, Sistema A (Frase 1).

Já o Sistema B trata como um só token o composto *primeiro-ministro* e liga por uma dependência `fixed` o elemento *mesmo* a *que* na conjunção composta *mesmo que*. As restantes locuções desta frase também não são identificadas por este sistema. Na frase 2, a locução conjuncional *assim como*/`conj` também não foi reconhecida pelos sistemas. As restantes frases não apresentam expressões multipalavras, exceto, talvez, o composto *quarto de hotel*, na frase 3, que nenhum dos sistemas analisa como um único *token*.

Alguns problemas de etiquetagem morfossintática podem ainda ser observados na saída do Sistema A e parecem difíceis de explicar. Na frase 2, o nome *membros* é marcado como `propn` (nome próprio); também o elemento *como* (da locução *assim como*) é curiosamente etiquetado como `noun`. O nome *anos*, no fim da frase 1 e o o adjetivo-nome *familiares*, na frase 2, a que se encontram ligados um ponto final e uma vírgula, respetivamente, aparecem etiquetados como `sym` (symbol), talvez pelo facto de os sinais de pontuação não terem sido tratados como *tokens*. Contudo, tal só ocorre nestes dois nomes, havendo várias outras situações idênticas, com estes sinais de pontuação, em que tal não sucede.

6.3. Dependências

Os sistemas apresentam análises muito semelhantes, pelo que analisaremos os problemas com base na saída do Sistema A, indicando as diferenças mais relevantes. Nesta forma representação, as dependências são marcadas sobre os arcos do grafo de dependências e estabelecem-se sempre entre palavras/*tokens*, isto é, entre os elementos que são núcleos (cabeça) dos constituintes sintáticos e os elementos que deles dependem, e.g. *a prática*: *a* ←(`def`) *prática*. Os constituintes da frase (e.g. grupos nominais, preposicionais, etc.) não são diretamente delimitados, mas podem ser deduzidos a partir do grafo de dependências, já que todos os elementos se ligam direta ou indiretamente ao seu núcleo. Repare-se que se adota uma representação *top-down* da dependência, em que um elemento hierarquicamente superior (*governor*) flecha sobre o elemento diretamente abaixo (*dependent*), independentemente da sua disposição linear na frase, e.g. *enviado às redacções*: *enviado* (`obl`)→ *redaccções*, *redacções* (`det`)→ *as*.

Para maior facilidade de comparação, em ambas as saídas, foram utilizadas as já referidas dependências universais (de Marneffe et al., 2014), em cuja descrição nos baseámos para a análise destas saídas. Assim na Frase 1 (Figura 1), o complemento indireto (dativo) *às redações* encontra-se (corretamente!) ligado ao

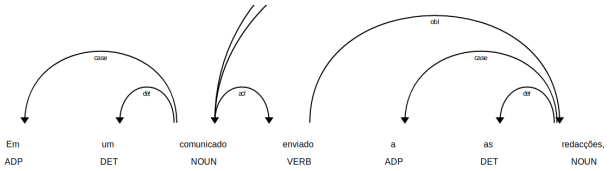


Figura 4: "Num comunicado. . . ", Sistema A

particípio *enviado*, mas pela dependência `obl` (*oblique*, oblíquo) e não, como se esperaria, por `iobj` (*indirect object*).

Ambos os sistemas fazem depender o complemento preposicional *Em comunicado enviado às redacções*, deslocado para o início da frase, do verbo *fazia* por meio da dependência `obl`. Contudo, o sistema A extrai também uma dependência `appos`, que consideramos espúria, entre *comunicado* e *gabinete*. Tal pode dever-se ao facto de *primeiro-ministro* não ter sido tratado como um único *token*. Efetivamente, o Sistema B reconhece o composto e analisa-o corretamente como complemento de nome (`nmod`) de *gabinete*, enquanto o Sistema A apenas estabelece essa dependência com o elemento *primeiro*. Por essa razão também, o sujeito (`nsubj`) de *fazia*, no sistema B, é (corretamente) *gabinete*, ao passo que no Sistema A, que não considera o composto, o sujeito deste verbo é *ministro*.

Outro aspeto relacionado com a incorreta segmentação é o facto de a preposição composta *ao longo de*, que forma a expressão temporal *ao longo dos anos*, não ter sido identificada, pelo que o seu elemento *longo*/`noun` é analisado, por ambos os sistemas, como um mero complemento de nome (`ncomp`) de *prática*. O facto de, aparentemente, não se ter tratado a expressão como uma entidade mencionada de tempo (Hagège et al., 2010) leva a supor que, se este nome não estivesse construído com um verbo copulativo (*sido*), a expressão temporal iria ficar a depender de qualquer constituinte anterior e não do verbo (pleno) de tal frase.

Ainda neste sentido, repare-se que a locução *tivessem a ver* (praticamente sinónima de *estar relacionado com*) não deverá ter sido reconhecida por nenhum dos sistemas, que a analisam como uma mera construção de auxiliar (Baptista et al., 2010), a qual, aliás, aparentemente só existe em português com um valor modal: "Só tenho a/posso/devo dizer que. . . ", e que, claramente não é a construção aqui empregue. Vemos, assim, a importância do reconhecimento das expressões multipalavra e do impacto que têm na análise sintática (dependências) das frases.

Outro aspeto interessante é a análise da oração relativa sem antecedente, introduzida por *o* (marcado em A com a categoria `x`). Este elemento é tratado como determinante do pronome *que* (não se indica a subclasse de pronome, nomeadamente, se é relativo ou de outro subtipo). Esta análise poderia ser considerada correta no caso das *verdadeiras* relativas sem antecedente, resultantes da redução de um nome elidido que fosse cabeça desse constituinte e antecendente do pronome relativo (e.g. *Havia vários livros. Comprei* ***o*** [***livro***] *que tinha capa azul*). O determinante poderia, nesse caso variar em género e número consoante o nome que determina (e.g.*Havia várias camisas. Comprei* ***as*** [***camisas***] *que não tinham colarinho*). Ora, neste caso, trata-se de uma construção diferente, em que *o* é invariável (****as*** *que têm sido prática corrente*), e cuja análise pode ser aproximada do pronome relativo *o qual*.

Um outro problema detetado, envolvendo também determinantes, é a dependência `det` entre *nada* e *estas* (v.g. *mesmo que estas nada tivessem a ver com*), o que faz considerar que as duas palavras foram analisadas como um só constituinte. Trata-se aqui do chamado "emprego pronominal" de *estas* (= *estas entidades públicas*) e não do "emprego determinativo", que forma um constituinte autónomo, sujeito de *tivessem*; e do pronome indefinido *nada* que é aqui complemento deste verbo. Provavelmente, por essa razão, a dependência de sujeito (`nsubj`) estabelece-se entre *nada* e o verbo, apesar de a flexão deste último não autorizar essa relação, já que viola a concordância gramatical sujeito-verbo.

Um dos problemas mais difíceis de tratar é a imbricação de elementos coordenados. Na formalização em grafo aqui adotada, a dependência `conj` liga os elementos coordenados, sendo sempre orientada da esquerda para a direita, enquanto uma dependência `cc` liga o segundo membro da coordenação à conjunção coordenativa, sendo igualmente sempre orientada, mas da direita para a esquerda. Podemos ver um caso complexo deste problema na Frase 2, cuja hierarquia representamos por parênteses numerados: (. . .) *incluindo* $[_1[_2$*dificuldades de* $[_3$*concentração*$]_3$ *e* $[_4$*memória*$]_4]_2$, $[_5$*tonturas*$]_5]$ *e* $[_6$*problemas* $[_7$*visuais*$]_7$ *e* $[_8$*de equilíbrio*$]_8]_6]_1$. Esta estrutura leva a considerar os seguintes pares de elementos coordenados (cuja numeração mantemos, para maior clareza): *concentração*$_3$ (`conj`)→ *memória*$_4$, *visuais*$_3$ (`conj`)→ *de equilíbrio*$_4$; e a tripla coordenação, assinalada com uma vírgula e a conjunção *e*, que se repre-

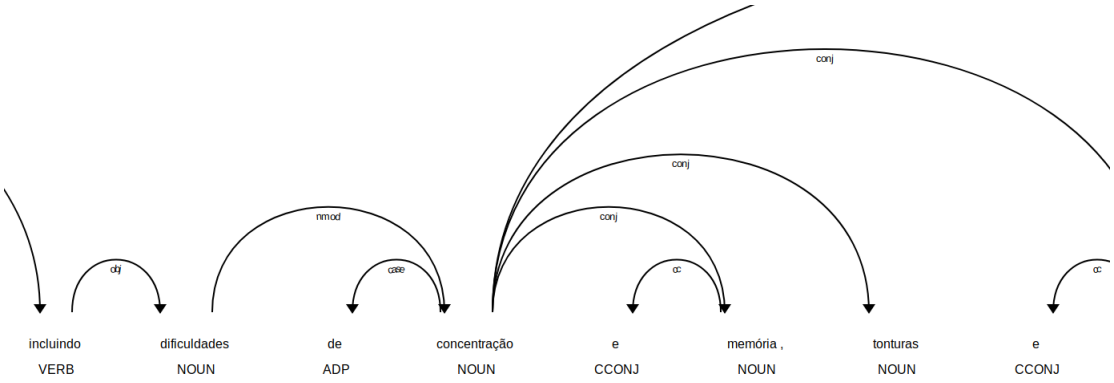


Figura 5: "... incluindo dificuldades..."

senta por duas dependências com foco no primeiro elemento: *dificuldades*$_2$ (`conj`)→ *tonturas*$_5$ e *dificuldades*$_2$ (`conj`)→ *problemas*$_6$.

Ora, os sistemas apresentam uma saída igual (Figura 5), onde apenas o primeiro par é corretamente capturado. A Frase 3 apresenta um problema similar. Aparentemente, o facto de apenas um dos sistemas tratar explicitamente a pontuação (neste caso, a vírgula) não parece fazer diferença nos resultados.

Noutras situações, os erros parecem resultar de dificuldades de análise de dependência a longa distância entre os elementos relacionados. Assim, na oração relativa explicativa da Frase 3 (Figura 6), v.g. *sons* (...), *que os pacientes reportam ter ouvido*, o pronome relativo *que*, referente a *sons*, deveria ter sido analisado como complemento direto (obj) de *ouvido* e não de *reportam*.

Ao seguir a definição e exemplos das dependências universais (de Marneffe et al., 2014), a descrição dos advérbios parece particularmente problemática, já que ignora distinções substanciais, algumas bem conhecidas pelas gramáticas mais tradicionais (para uma síntese moderna, v. (Molinier & Levrier, 2000)). Assim, não se distinguem: (i) o valor determinativo de *nem* no grupo nominal *nem o seu marido* (frase 4); (ii) o advérbio (de quantificação) *largamente*, que funciona como mero modificador verbal de *ultrapassa*, em *ultrapassa largamente* (frase 4); e o advérbio conjuntivo (com valor adversativo) *porém* (frase 4), que deve ser considerado como modificador de toda a frase, ligando-a à frase anterior. Ainda que os argumentos da dependência possam ser considerados corretos, todos estes advérbios são representados pela mesma dependência (`advmod`), sem os diferenciar. Por outro lado, nenhum dos sistemas reconhece o valor adverbial da expressão temporal *uma tarde* (frase 5), o que dá origem a um conjunto de dependências incorretas.

6.4. Discussão

Em jeito de síntese, é possível dizer que: (i) os problemas de segmentação e de etiquetagem morfossintática estão na origem não só das diferenças entre os sistemas comparados, mas também dos principais erros de análise sintática (dependências) encontrados. (ii) Neste sentido, reveste-se de especial importância a identificação das unidades lexicais compostas e expressões multipalavras, praticamente ignoradas por um dos sistemas e muito incompleta no outro. (iii) Nos restantes casos, os sistemas têm um comportamento muito semelhante ou mesmo idêntico. (iv) A pontuação é um elemento fundamental na análise sintática, mas um dos sistemas aparentemente ignora-a, enquanto o outro parece não a usar, do que resultam diversos erros.

Entre os diversos problemas de análise sintática detetados, salientamos (v) o tratamento da coordenação e da imbricação de elementos coordenados, bem como (vi) a ausência de distinção entre funções sintáticas muito diferentes, desempenhadas por vários tipos sintático-semânticos de advérbios, e todas colapsadas sob a mesma dependência "universal". Trata-se de fenómenos linguísticos bem conhecidos, por vezes complexos e dificilmente tratáveis, mas que o desenvolvimento de sistemas de processamento computacional de português terá enfrentar. O levantamento destas e de outras situações problemáticas é fundamental para construir um *roadmap* dos desafios a vencer.

7. Conclusões e Trabalho Futuro

Neste trabalho fizemos a avaliação de várias ferramentas (a maioria através de modelos pré-treinados) para as tarefas de EMS e REM, para a língua portuguesa. Apresentámos ainda um estudo qualitativo sobre duas ferramentas que re-

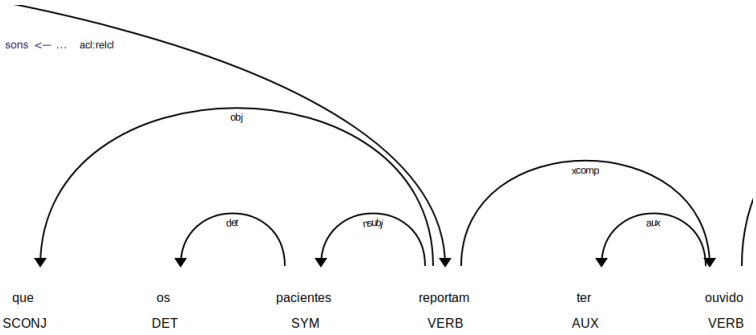


Figura 6: "sons … que os pacientes reportam ter ouvido…"

alizam a tarefa de AD. Não há uma ferramenta claramente vencedora, pois, dependendo das necessidades de cada um, uma ferramenta pode ser mais ou menos apropriada, de acordo com vários factores. Fica, no entanto, claro, que, para além da utilização de colecções de etiquetas variadas, há uma grande diferença entre as ferramentas se considerarmos o modo como a segmentação é feita. Por outro lado, há ferramentas que são mais finas em algumas classes de etiquetas, o que torna mais difícil a obtenção de bons resultados. No entanto, mais uma vez, dependendo das necessidades dos seus utilizadores, pode fazer sentido ou não usar essas etiquetas mais finas em detrimento de melhores resultados. De notar também que as decisões tomadas a nível da segmentação, bem como os resultados atingidos em termos de EMS (e REM), vão afectar a AD. Todos os *corpora* usados, bem como o mapeamento de etiquetas, serão tornados públicos[42], permitindo a replicação das experiências e facilitando futuras avaliações na mesma linha. Como trabalho futuro, o LinguaKit (Gamallo & Garcia, 2017b; Gamallo et al., 2018) deverá ser avaliado na tarefa de AD. Por outro lado, podem também ser explorados mais modelos, por exemplo, com o NLTK. Uma outra experiência a realizar, seria usar o mesmo corpus de treino para treinar vários modelos de maneira a poder comparar diretamente e apenas os diferentes algoritmos.

Agradecimentos

Este trabalho foi parcialmente suportado pela Fundação para a Ciência e a Tecnologia através dos projectos UIDB/50021/2020 e PTDC/LLT-LIN/29887/2017, financiando este último a bolsa de Matilde Gonçalves.

[42] https://gitlab.hlt.inesc-id.pt/lcoheur/ptools

Referências

Al-Rfou, Rami. 2015. *Polyglot: A massive multilingual natural language processing pipeline*: State University of New York at Stony Brook. Tese de Doutoramento.

Apache Software Foundation. 2014. OpenNLP natural language processing library. http://opennlp.apache.org/.

Baptista, Jorge, Nuno Mamede & Fernando Gomes. 2010. Auxiliary verbs and verbal chains in European Portuguese. Em *Computational Processing of the Portuguese Language (PROPOR)*, 110–119. doi 10.1007/978-3-642-12320-7_14.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media, Inc.

Buchholz, Sabine & Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. Em *Conference on Computational Natural Language Learning*, 149–164.

Collovini, Sandra, Joaquim Francisco Santos Neto, Bernardo Scapini Consoli, Juliano Terra, Renata Vieira, Paulo Quaresma, Marlo Souza, Daniela Barreiro Claro & Rafael Glauber. 2019. IberLEF 2019 Portuguese named entity recognition and relation extraction tasks. Em *Proceedings of the Iberian Languages Evaluation Forum (IberLEF)*, 390–410.

Ferreira, João, Hugo Gonçalo Oliveira & Ricardo Rodrigues. 2019a. Improving NLTK for processing Portuguese. Em *Symposium on Languages, Applications and Technologies (SLATE)*, 18:1–18:9. doi 10.4230/OASIcs.SLATE.2019.18.

Ferreira, João, Hugo Gonçalo Oliveira & Ricardo Rodrigues. 2019b. NLPyPort: Named entity recognition with CRF and rule-based relation

extraction. Em *Iberian Languages Evaluation Forum (IberLEF)*, 468–477.

Fonseca, Erick, Leandro Borges dos Santos, Marcelo Criscuolo & Sandra Aluísio. 2016. Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática* 8(2). 3–13.

Fonseca, Erick Rocha & João Luís G. Rosa. 2013. Mac-Morpho revisited: Towards robust part-of-speech tagging. Em *Brazilian Symposium in Information and Human Language Technology (STIL)*, s/pp.

Gamallo, P., M. Garcia, C. Piñeiro, R. Martinez-Castaño & J. C. Pichel. 2018. LinguaKit: A big data-based multilingual tool for linguistic analysis and information extraction. Em *Conference on Social Networks Analysis, Management and Security (SNAMS)*, 239–244. 10.1109/SNAMS.2018.8554689.

Gamallo, Pablo & Marcos Garcia. 2013. Freeling e treetagger: um estudo comparativo no âmbito do Português. Relatório técnico. Universidade de Santiago de Compostela.

Gamallo, Pablo & Marcos Garcia. 2017a. Linguakit: A multilingual tool for linguistic analysis and information extraction. *Linguamática* 9. 19–28. 10.21814/lm.9.1.243.

Gamallo, Pablo & Marcos Garcia. 2017b. LinguaKit: uma ferramenta multilingue para a análise linguística e a extração de informação. *Linguamática* 9(1). 19–28. 10.21814/lm.9.1.243.

Garcia, Marcos & Pablo Gamallo. 2015. Yet another suite of multilingual NLP tools. Em *Languages, Applications and Technologies*, 65–75. 10.1007/978-3-319-27653-3_7.

Hagège, Caroline, Jorge Baptista & Nuno Mamede. 2010. Caracterização e processamento de expressões temporais em português. *Linguamática* 2(1). 63–76.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem & Adriane Boyd. 2020. spaCy: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. 10.5281/zenodo.1212303.

Jurafsky, Dan & James H. Martin. 2019. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson.

Lafferty, John D., Andrew McCallum & Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Em *International Conference on Machine Learning*, 282–289.

de Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre & Christopher D. Manning. 2014. Universal stanford dependencies: A cross-linguistic typology. Em *Conference on Language Resources and Evaluation (LREC)*, 4585–4592.

Màrquez, Lluís & Horacio Rodríguez. 1998. Part-of-speech tagging using decision trees. Em *Machine Learning*, 25–36. 10.1007/BFb0026668.

Molinier, Christian & Françoise Levrier. 2000. *Grammaire des adverbes: description des formes en* -ment. Genève: Droz.

Padró, Lluís. 2012. Analizadores multilingües en FreeLing. *Linguamática* 3(2). 13–20.

Pires, André Ricardo Oliveira. 2017. *Named Entity Extraction from Portuguese web text*: Faculty of Engineering of University of Porto. Tese de Mestrado.

Qi, Peng, Timothy Dozat, Yuhao Zhang & Christopher D. Manning. 2018. Universal dependency parsing from scratch. Em *CoNLL Shared Task: Multilingual Parsing from Raw Text to Universal DependenciesCoNLL Shared Task*, 160–170. 10.18653/v1/K18-2016.

Rodrigues, Ricardo, Hugo Gonçalo Oliveira & Paulo Gomes. 2018. NLPPort: A pipeline for Portuguese NLP. Em *Symposium on Languages, Applications and Technologies (SLATE)*, 18:1–18:9. 10.4230/OASIcs.SLATE.2018.18.

Santos, Diana & Nuno Cardoso. 2007. Balanço do primeiro HAREM e perspectivas de trabalho futuro. Em *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 87–94. Linguateca.

Santos, Diana, Luís Costa & Paulo Rocha. 2003. Cooperatively evaluating portuguese morphology. Em *Computational Processing of the Portuguese Language (PROPOR)*, 259–266. 10.1007/3-540-45011-4_40.

Santos, Diana, Hugo Gonçalo Oliveira, Cláudia Freitas, Cristina Mota & Paula Carvalho. 2008. Segundo HAREM: Balanço e perspectivas de futuro. Apresentação no Encontro do Segundo HAREM.

Veloso, Rita. 2013. *Gramática do Português*, vol. 2 chap. Subordinação relativa, 2061–2134. Fundação Calouste Gulbenkian.

Projetos, Apresentam-se!

Proposta recibida en xullo 2020 e aceptada para publicación en decembro 2020.

Aplicación de WordNet e de word embeddings no desenvolvemento de prototipos para a xeración automática da lingua

Application of WordNet and word embeddings in the development of prototypes for automatic language generation

María José Domínguez Vázquez
Universidade de Santiago de Compostela-ILG
majo.dominguez@usc.es

Resumo

Esta presentación de dous prototipos de xeración automática de lingua natural achega unha visión de conxunto da metodoloxía aplicada na descrición e procesamento dos datos lingüísticos, así como das técnicas e ferramentas xa existentes ou desenvolvidas co fin de garantir o funcionamento dos simuladores en alemán, español e francés.

Palabras chave

Gramática de valencias, Patróns argumentais, Procesamento da linguaxe natural (PLN) e Xeración da linguaxe natural (XLN), WordNet, word embeddings

Abstract

This presentation of two prototypes of automatic natural language generation provides an overview of the methodology applied in the description and processing of linguistic data, as well as of the techniques and tools already existing or developed in order to guarantee the functioning of the simulators in German, Spanish and French.

Keywords

Valency Grammar, Argument patterns, Natural Language Processing (NLP) and Natural Language Generation (NLG), WordNet, word embeddings

1. Introdución: estado do problema

Un importante hándicap na intelixencia artificial é converter a información lingüístico-semántica en información codificable e reproducible por non humanos, isto é, dotar as máquinas de coñecemento léxico (Navigli & Ponzeto, 2012) coa pretensión de que as computadoras interactúen coa súa contorna de modo humanizado utilizando as linguas naturais como código de intercambio.[1] Nesta liña, resulta especialmente relevante o deseño de léxicos computacionais — lexibles por máquinas e dotados de información semántica —, que faciliten a desambiguación dos diferentes significados (Agirre & Edmonds, 2006). As investigacións lexicográficas poden impulsar significativamente esta tarefa, se entendemos que "[b]y far the best existing semantic descriptions of language are dictionaries [...]" (Trap-Jensen, 2018, p.34). Nesta dobre aproximación — a lexicográfica e computacional — enmárcanse os prototipos de xeración automática de frases nominais en contexto para o español, francés e alemán, *Xera* e *Combinatoria* (*vid.* 2.2).[2]

Estes xeradores nacen ligados ao dicionario online multilingüe PORTLEX[3] , un recurso valencial, semicolaborativo, modular, multilingüe e *cross-lingual* sobre o potencial combinatorio da frase nominal en 5 linguas (Domínguez & Valcárcel, 2020). As principais dificultades no seu desenvolvemento non so atinxen á notable inversión de tempo na análise e compilación de todos e cada uns dos esquemas argumentais así como do seu potencial combinatorio (Domínguez & Valcárcel, 2020), senón tamén ás limitacións observadas no uso de córpora:[4]

- Na extracción dos datos tirados dos córpora non é posible aplicar ningún filtro de pre-

[1] A día de hoxe xa existen asistentes virtuais como Siri, Alexa ou Google Home, isto é, máquinas que "falan", pero tamén temos máquinas que aprenden e son adestradas.

[2] http://portlex.usc.gal/combinatoria. Os recursos son gratuitos e de libre acceso. Para máis información, *vid.* a información sobre as condicións de uso na anterior ligazón.

[3] http://portlex.usc.gal/portlex/

[4] Cómpre dicir que os córpora ofrecen paulatinamente vías máis precisas en prol da análise das propiedades distributivas e sintagmáticas do léxico. Serven de exemplo Sketch Engine (https://www.sketchengine.eu/), *COSMAS II* (https://cosmas2.ids-mannheim.de/cosmas2-web/), CORGA (http://corpus.cirp.es/corga/) ou TLFi (http://atilf.atilf.fr/).

DOI: 10.21814/lm.12.2.337
This work is Licensed under a Creative Commons Attribution 4.0 License

selección semántico-argumental nin un envorcado de datos atendendo as acepcións de significado.[5] Así, o lexema *x* no composto *x* + UMZUG alemán na súa acepción 'cambio de domicilio ou lugar de traballo' pode expresar un rol semántico AXENTE — "Aquel ou aquilo que realiza unha acción". Non obstante, a listaxe de frecuencia obtida mediante unha busca *Corpus Query Language* (CQL) en *Sketch Engine* recolle maioritariamente exemplos de compostos doutra acepción de significado, en concreto de "cabalgata" ou "desfile".[6] Xunto coa necesidade, pois, de cribar os compostos atendendo ao seu significado relacional, engádese a de atopar de xeito automático as súas posibles combinatorias cos outros actantes argumentais nas diferentes realizacións — combinatorias, ademais, comúns na lingua.

- Os resultados envorcados non cumpren os requisitos para a súa inclusión nun dicionario de valencias ao non exemplificar complementos argumentais. Un claro exemplo desta casuística son as realizacións adxectivais nas diferentes linguas contempladas no recurso. Unha análise semántica dos 70 primeiros adxectivos obtidos de *Sketch Engine* para o esquema argumental AUSENCIA + *Adxectivo*, permite concluír que soamente unha porcentaxe moi reducida serve para o propósito do noso recurso. De entre esta listaxe só un 4 % dos adxectivos representa un rol semántico e soamente un único exemplo pode ser adxudicado ao rol AXENTE (*a ausencia escolar*), sendo o mesmo, porén, ambiguo.

Xa que a frecuencia léxica no eixo sintagmático non está en necesaria correlación coa súa función específica ou argumental, atopar candidatos léxicos que cubran o actante valencial que se exemplifica supón a análise manual dun número significativo de coocorrencias, e isto, para as 5 linguas contempladas no dicionario. Constatadas estas dificultades, que ademais atrasaban enormemente o traballo, decidimos proceder á inversa: xerar directamente as estructuras argumentais aplicando previamente filtros semántico-combinatorios (*vid.* 2.2), no canto de buscalas nos córpora. Este é o punto de partida dos simuladores.

2. Os xeradores lingüísticos: metodoloxía e fases

2.1. Visión de conxunto

É a partir dos anos 90 cando se intensifica a investigación no campo da xeración automática da lingua natural. Nun principio os xeradores non priorizaban aspectos importantes como a integrabilidade, a portabilidade ou a eficiencia, nin lle dedicaban especial atención a aspectos de natureza máis lingüística, como a coherencia semántica ou textual, factores que, xunto coa fluidez e a variación nas posibles realizacións, son elementos centrais na avaliación dos xeradores de lingua natural (Hashimoto et al., 2019; Horacek & Zock, 2015; Jiménez et al., 2020; Vicente et al., 2015). Xunto con protocolos e estudos de avaliación da calidade dos resultados, xa se poden xerar de xeito automático textos e frases con diferentes características (Vicente et al., 2015; Nallapati et al., 2016; Sordoni et al., 2015) — incluso case sen input de partida (Roemmele, 2016) —, así como imaxes a partir de textos e viceversa (Otter et al., 2020). As aproximacións á xeración automática dende ou para a aplicación lexicográfica non é ampla.[7] Existen diferentes propostas para a xeración automática de dicionarios (Bardanca Outeiriño, 2020; Kabashi, 2018; Delli Bovi & Navigli, 2017), de artigos lexicográficos (Geyken et al., 2017) ou ben dalgunha das súas partes (os exemplos, en Kosem et al. (2019)). Estes recursos, porén, non perseguen os mesmos obxectivos que os xeradores que aquí presentamos.

2.2. Os xeradores *Xera* e *Combinatoria*

Os dous prototipos para a xeración de patróns argumentais de frases simples (*Xera*) e complexas (*Combinatoria*) en español, francés e alemán

[5] Recursos de diferente tipoloxía achegan descricións semántico-argumentais, en especial, para o inglés — *Verbnet* (http://verbs.colorado.edu/~mpalmer/projects/verbnet.html), Propbank (http://verbs.colorado.edu/~mpalmer/projects/ace.html), *VerbAtlas* (http://verbatlas.org/), *CPA* (http://www.pdev.org.uk/). Para outras linguas, como español e catalán, *vid.* AnCora (http://clic.ub.edu/corpus/es/ancora).

[6] Por cuestión de espazo amósanse soamente os 10 primeiros lemas e a súa frecuencia: *Festumzug* (18742), *Laternenumzug* (5568), *Faschingsumzug* (5191), *Karnevalsumzug* (4669), *Martinsumzug* (2838), *Rosenmontagsumzug* (2441), *Serverumzug* (1952), *Lampionumzug* (1638), *Fackelumzug* (1626) e *Fastnachtsumzug* (1467).

[7] Si que se aplican diferentes ferramentas de análise e técnicas, por exemplo, FreeLing (Padró, 2012), estudos para a extracción automática de diferentes tipos de datos (Gamallo & Pichel, 2007; Kilgarriff et al., 2008; Renau & Nazar, 2016), así como con softwares que permiten a compilación de dicionarios, como, por exemplo, TshwaneLex (https://tshwanedje.com/). Así mesmo, créanse aplicacións de dicionarios para dispositivos móbiles a partir de redes semánticas como WordNet, por exemplo o Dicionario GalNet (http://sli.uvigo.gal/digalnet/) de Gómez Guinovart en Google Play.

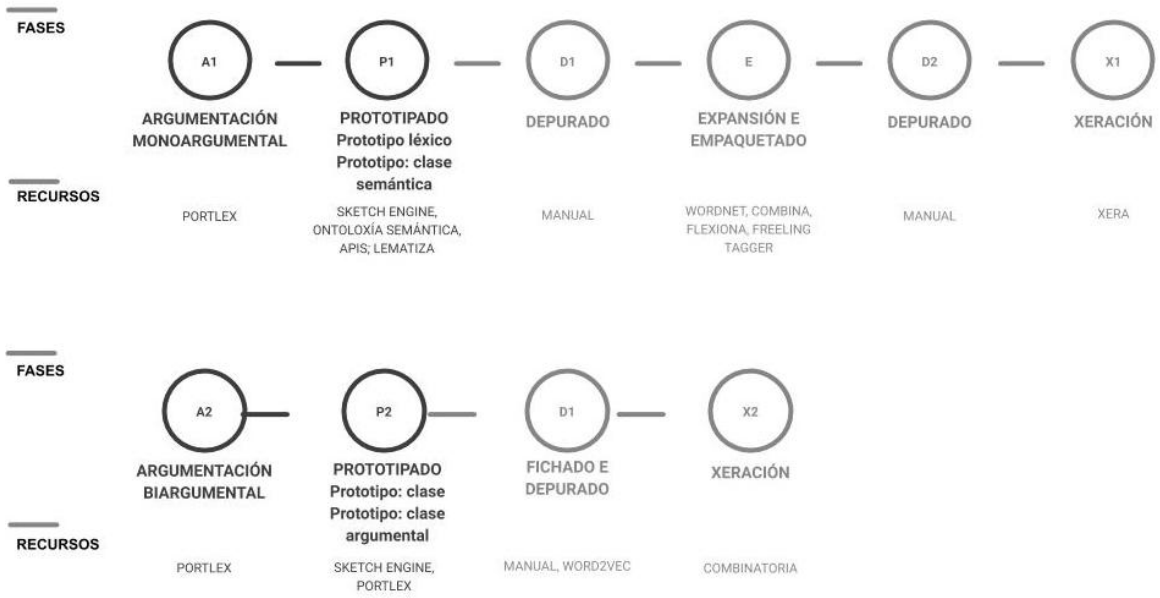


Figura 1: Metodoloxía e recursos de xeración automática argumental

están en funcionamento. Para o seu deseño precisamos unha metodoloxía que nos permitise describir e procesar os datos lingüísticos con información sintáctico-semántica combinatoria argumental, de tal xeito que estes se puideran programar e xerar automaticamente. Para tal fin, combinamos propostas teóricas e metodolóxicas recorrendo á gramática e semántica valencial, á teoría dos prototipos léxicos, ao PLN (recuperación e extracción) e XLN, así como a WordNet. Na segunda das ferramentas, aplicamos o método predictivo Word2vec (*vid.* 3). A Figura 1 recolle a metodoloxía xeral e as ferramentas que dan sustento aos xeradores no proceso de xeración da estrutura argumental; a Figura 2 resume o procedemento da contextualización, que xa se desenvolveu para a frase adxectival, pero está en curso no caso da oración.

O primeiro paso para a xeración automática da argumentación e o potencial combinatorio nominal consiste en establecer que actantes argumentais son específicos de cada substantivo[8] e que caudal léxico pode cubrir o eixo paradigmático dos devanditos actantes. Tomando como referencia a estrutura argumental do dicionario PORTLEX, analizamos os trazos ontolóxicos-categoriais (Engel, 2004) das (co)aparicións argumentais extraídas seguindo criterios de frecuencia de *Sketch Engine*. Atopámonos na fase de prototipado léxico: buscamos, por tanto, candidatos prototípicos para cubrir actantes funcionais concretos, isto é, exemplares léxicos, como, por exemplo, os lexemas *suor*, *tabaco* ou *pólvora* no rol CLASIFICATIVO de OLOR + *A*. Séguelle a segunda fase de prototipado, que consiste na determinación das clases semánticas prototípicas dun argumento concreto. Para o caso citado, por exemplo, [+Material] [+Substancia]: *suor*, *tabaco*, *pólvora* — [+Animado] [+Planta]: *flor*, *rosa* — [+Animado] [+Animal]: *porco*, *can* etc. A Figura 3 amosa un exemplo.

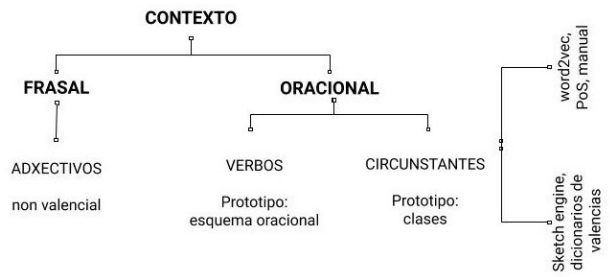


Figura 2: Metodoloxía xeral de contextualización

A descrición dos trazos categoriais e o prototipado léxico son conceptos esenciais non só dende un punto de vista puramente descritivo, senón tamén porque nos permiten acometer o seguinte estadio de traballo: a expansión léxica recorrendo a WordNet. Esta ampliación do número de candidatos léxicos atribuíbles a cada actante funcional conséguese mediante os trazos categoriais,

[8] A escolla dos 10 substantivos dos prototipos segue dous criterios centrais: (a) o seu estatus de portadores valenciais, isto é, a súa capacidade de abrir casillas funcionais e (b) a súa pertenza a diferentes campos semánticos, como Locación (*presenza*), Expresión (*pregunta*, *discusión* e *texto*), Afección (*morte*, *aumento* e *dor*) e Clasificación (*olor* e *sabor*). Deste xeito obtemos un amplo abanico de esquemas sintáctico-argumentais e combinatorias.

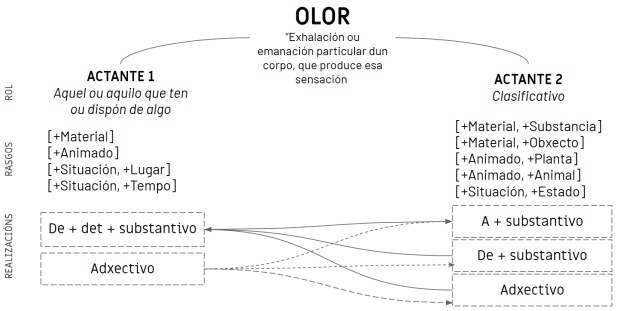


Figura 3: Argumentación e clases semánticas prototípicas

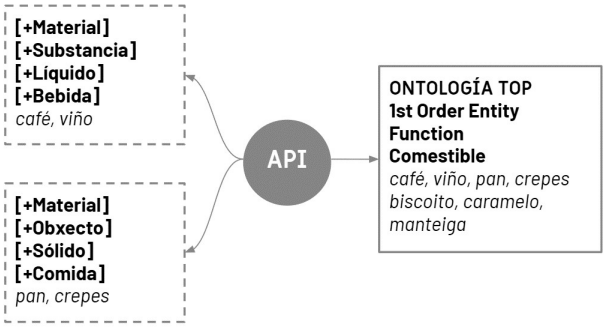


Figura 4: Expansión léxica usando WordNet

que fan de nexo de unión coas clases e atributos das categorías de WordNet (*vid.* Figura 4). Domínguez et al. (2019) indican que "the synsets of the wordnets following the EuroWordNet model of the Multilingual Central Repository (MCR) (...) are associated with semantic or cognitive features categorized in different ontologies".[9]

Aínda que esta organización cognitiva das ontoloxías de WordNet non coincide exactamente co inventario categorial da lexicografía valencial, tamén se constata que os trazos categoriais de tipo xeral si que conforman clases xerais nas ontoloxías do MCR. Por tanto, mediante este procedemento de expansión léxica[10], e posterior depuración e etiquetaxe seguindo unha ontoloxía sumativa de elaboración propia, acádase unha selección de caudal léxico no eixo paradigmático que comparte as características semánticas do prototipo léxico-semántico que tomamos como punto de partida, isto é, que conforma unha clase semántica concreta. Por tanto, a partir dos prototipos léxicos determinamos clases semánticas prototípicas. Unha vez obtido este caudal léxico, realízase a flexión e o empaquetado, no que temos en conta cuestións de etiquetaxe morfosintáctica e semántica. Estes repertorios léxicos atribuídos a cada argumento nominal son fundamentais para a xeración.

3. Ferramentas e recursos

No desenvolvemento dos xeradores operamos nas diferentes fases con (a) recursos existentes ou con ferramentas creadas *ad hoc* (*vid.* Figuras 1 e 5), entre as que diferenciamos (b) as que dan sustento aos xeradores, así como (c) os xeradores en si mesmos, que envorcan os datos en formato JSON e CSV.[11]

(a) **Recursos existentes:** Cabe subliñar aquí Sketch Engine, FreeLing e WordNet: i) **Sketch Engine** permítenos extraer, mediante consultas CQL, a frecuencia de cooaparicións en diferentes estruturas argumentais que tomamos do dicionario PORTLEX; ii) para o desenvolvemento do código de flexión partimos dos dicionarios do etiquetador **FreeLing**. Xa que na extracción léxica a partir de WordNet tamén se obteñen formas compostas — sendo estas unha posible realización argumental — recorremos tamén a este recurso para obter a división en lemas. iii) A rede semántica **WordNet** (Gómez Guinovart & Solla, 2018) posibilita a obtención do caudal léxico atendendo a clases ontolóxico-categoriais (*vid.* b).

(b) **Ferramentas para a análise lingüística desenvolvidas *ad hoc*:**[12]

- Coa finalidade de extraer datos léxicos das consultas que recorren ás relacións semánticas de WordNet e ás ontoloxías vinculadas aos synsets no modelo EuroWordNet desenvolvéronse tres **APIs**, unha para cada lingua obxecto de estudo.[13].

[9] Dado que ao inicio do traballo unicamente o español contaba cun wordnet vinculado ás ontoloxías mencionadas como parte do MCR (http://adimen.si.ehu.es/web/MCR), foi preciso crear bases de datos para o francés e o alemán. O procedemento explícase en Domínguez et al. (2019, p. 61–62): "This was done by extracting the alignment between lexical variants and identifying offsets of the meaning from the WordNet Libre du Français (WOLF) (Sagot & Fišer, 2008) and with data from the Extended Open Multilingual WordNet (Bond & Foster, 2013). Both have been made available on the GalNet interface after being converted to the EuroWordNet format of the MCR.". Dende o 2017 tamén está en desenvolvemento o Open German WordNet (https://github.com/hdaSprachtechnologie/odenet).

[10] Para máis información *vid.* a descripción da ferramenta *Combina* no apartado 3.

[11] Poden ser, por tanto, de aplicación como léxicos computacionais lexibles por máquinas con información semántica.

[12] Unha descrición detallada das ferramentas atópase en (Domínguez et al., 2019).

[13] http://portlex.usc.gal/develop/de/api/; http://portlex.usc.gal/develop/es/api/; http://portlex.usc.gal/develop/fr/api/

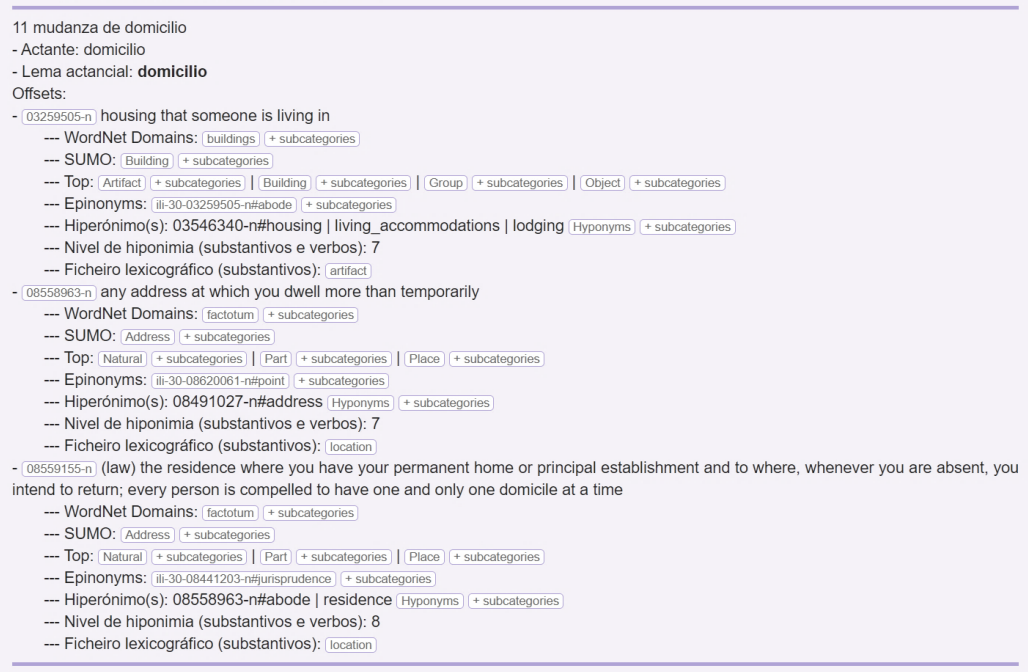


Figura 5: Imaxe de Lematiza

- A partir dos datos envorcados do corpus Sketch Engine, **Lematiza**[14] amosa os lemas coas súas variantes de significado (**synsets**) ligadas ás diferentes ontoloxías de WordNet (*vid.* Figura 5). Tamén é posible acceder a cada unha das categorías das ontoloxías de WordNet directamente premendo nos diferentes apartados.
- A ferramenta **Combina**[15] permite realizar consultas ás ontoloxías de WordNet (*vid.* 2.2) e, deste xeito, o envorcado semiautomático de datos compartidos ou combinados da *Top Concept Ontology*[16] (Álvez et al., 2008), dos *WordNet Domains*[17] (Bentivogli et al., 2004), da *Suggested Upper Merged Ontology*[18] (Niles & Pease, 2001), dos *Basic Level Concepts* (Izquierdo et al., 2007) e dos *Epinónimos* (Gómez Guinovart & Solla, 2018), ademais dos primitivos semánticos de Miller et al. (1990). Tras un procedemento de depuración e etiquetaxe, este caudal léxico expandido conforma os paquetes léxicos. Por tanto, se Lematiza amosa as variantes de significado (por exemplo para "domicilio" na Figura 5)[19], Combina presenta o repertorio léxico expandido no eixo paradigmático. Así, na busca de exemplares léxicos que, por exemplo, ocupen o actante locativo para un esquema argumental do tipo AUSENCIA + DE + *Locación* unha busca compartida nesta ferramenta[20] envorca un repertorio léxico en consonancia cos prototipos léxicos establecidos, por exemplo:
  - `[03259505-n` casa`]`,
  - `[04172107-n` chalet adosado`]`,
  - `[03259505-n` domicilio`]`,
  - `[03259505-n` piso`]`,
  - `[03259505-n` residencia`]` ou
  - `[04517408-n` segunda residencia`]`.
- A ferramenta **Flexiona** realiza a flexión dos lemas seleccionados.

(c) **Ferramentas de xeración: *Xera* e *Combinatoria*.**

Na actualidade ambas ferramentas contemplan a análise de 10 substantivos en 3 linguas e aportan datos para 429 patróns sintácticos, 2536 estruturas sintáctico-semánticas (que resultan da interface entre a realización formal e a clasificación semántico-relacional), así como uns 2479 exemplos estándar. Un total de 60001 formas e 15718 lemas están adxudicados a diferentes clases semánticas.

[14] http://portlex.usc.gal/develop/lematiza
[15] http://portlex.usc.gal/develop/combina.php
[16] https://adimen.si.ehu.es/web/WordNet2TO
[17] http://wndomains.fbk.eu/
[18] http://www.adampease.org/OP/
[19] Os resultados da Figura 5 fan referencia á busca CQL en Sketch Engine `[lemma=''mudanza''] [lemma=''de''] [tag=''D.*'']? [tag=''A.*'']? [tag=''N.*''] .`
[20] API http://portlex.usc.gal/develop/es/api?ontology=top&category=Building e http://portlex.usc.gal/develop/es/api?ontology=epinonyms&category=ili-30-03259505-n&subcategories=on

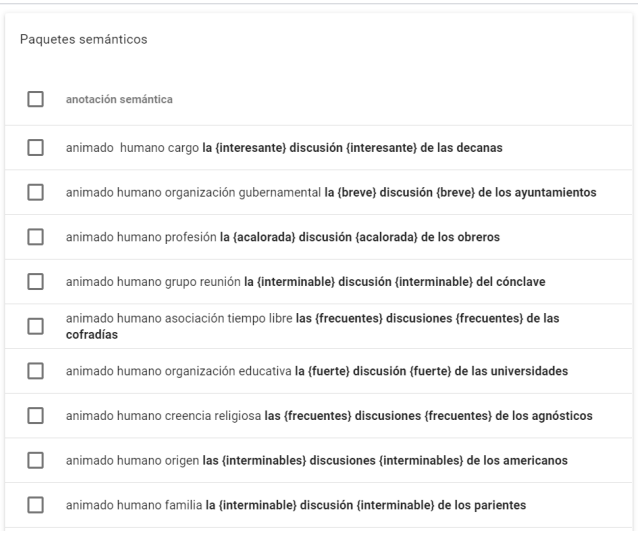


Figura 6: Clases semánticas para N1 de DISCUSIÓN

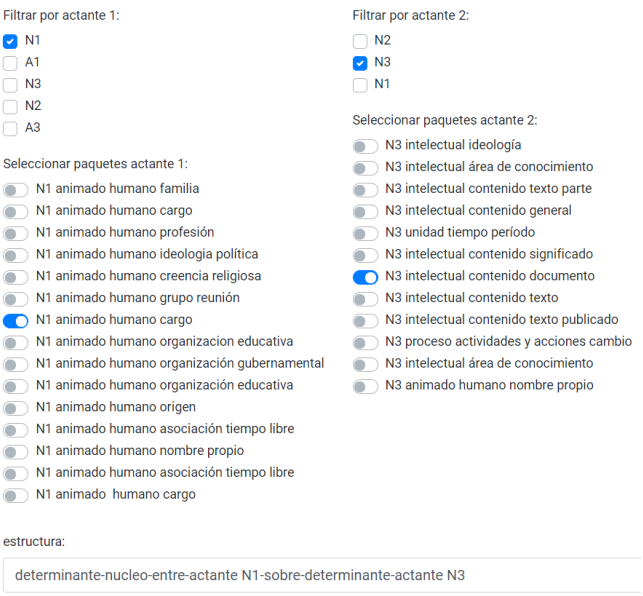


Figura 7: Interface de usuario en Combinatoria

Na primeira interface de acceso a **Xera**, o usuario selecciona nun despregable o idioma, o núcleo — neste caso DISCUSIÓN — e a estrutura argumental — N1: AXENTE ("Aquel ou aquilo que realiza unha acción") —, así como a clase ou clases semánticas prototípicas para o argumento concreto dun substantivo en cuestión (Figura 6)[21]. Cómpre engadir que os datos xerados seguen un principio de aleatoria predeterminada. Isto significa que as clases semánticas da cada argumento seguen un filtro semántico e a aleatoriedade atinxe aos representantes léxicos de cada clase, non ao rol semántico.

Un proceso semellante séguese na consulta do xerador de combinatoria biargumental, Combinatoria, que permite ao usuario seleccionar os actantes, paquetes e a combinatoria argumental (Figura 7), obtendo datos como os que recolle a Figura 8).

Porén, o tipo de datos envorcados non é a única diferenza entre ámbalas dúas ferramentas. **Combinatoria** tamén ofrece a posibilidade de cribar as combinatorias mediante o uso de *word embeddings*, representacións vectoriais dunha palabra en contexto, por tanto, un filtrado seguindo criterios de frecuencia de coaparición contextual (*vid.* parte superior da Figura 8). Para o desenvolvemento destes vectores recorremos ao método predictivo Word2vec (Mikolov et al., 2013), que fai uso dunha RNN (*Recurrent Neural Network*) de 2 niveis con dúas implementacións diferentes: o algoritmo Skip-gram tenta predicir o contexto máis adecuado dunha palabra analizando o seu vector e os vectores posteriores das palabras vinculadas ao vector da palabra orixe. O modelo CBOW tenta adiviñar a palabra máis adecuada para un contexto específico, é dicir, a palabra que máis frecuentemente ocupa un espazo: *a dor de* [espazo] *do neno*. Para a aplicación de Word2vec partimos dun modelo preadestrado de *Sketch Engine* e gardamos os resultados no formato propio de Word2vec (en matrices de N tamaño, onde N é o tamaño do vector preconfigurado para cada palabra). Para calcular a similitude entre os *tokens* calculamos a similitude entre os cosenos de dous vectores, tal e como representa a seguinte función matemática:

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|} \qquad (1)$$

O exemplo, tomado de (Bardanca Outeiriño, 2020), amosa así a distancia dos cosenos ou a similitude entre os lemas *rose* e *tulip* (Figura 9). Coa aplicación da fórmula a estes vectores obtemos un grao de similitude de 0.73:

$$\begin{aligned} a^T\cdot b &= 1\times1+1\times1+1\times1+1\times1+\\ &\quad 0\times1+1\times0+1\times0=4,\\ \|\vec{a}\| &= \sqrt{1^2+1^2+1^2+1^2+0^2+1^2+1^2}=2{,}44,\\ \|\vec{b}\| &= \sqrt{1^2+1^2+1^2+1^2+1^2+0^2+0^2}=2{,}23, \end{aligned}$$

$$similaridade = \frac{4}{2{,}44\times2{,}23} = 0{,}73$$

[21] Isto obsérvase comparando as clases semánticas da Figura 6 cos resultados dunha busca do argumento N3 ("AFECTADO: THEMA") coa mesma estrutura sintáctica [*determinante* + *adxectivo_o* + *discusión* + *adxectivo_o* + *de* + *determinante* + *actante:* N3]

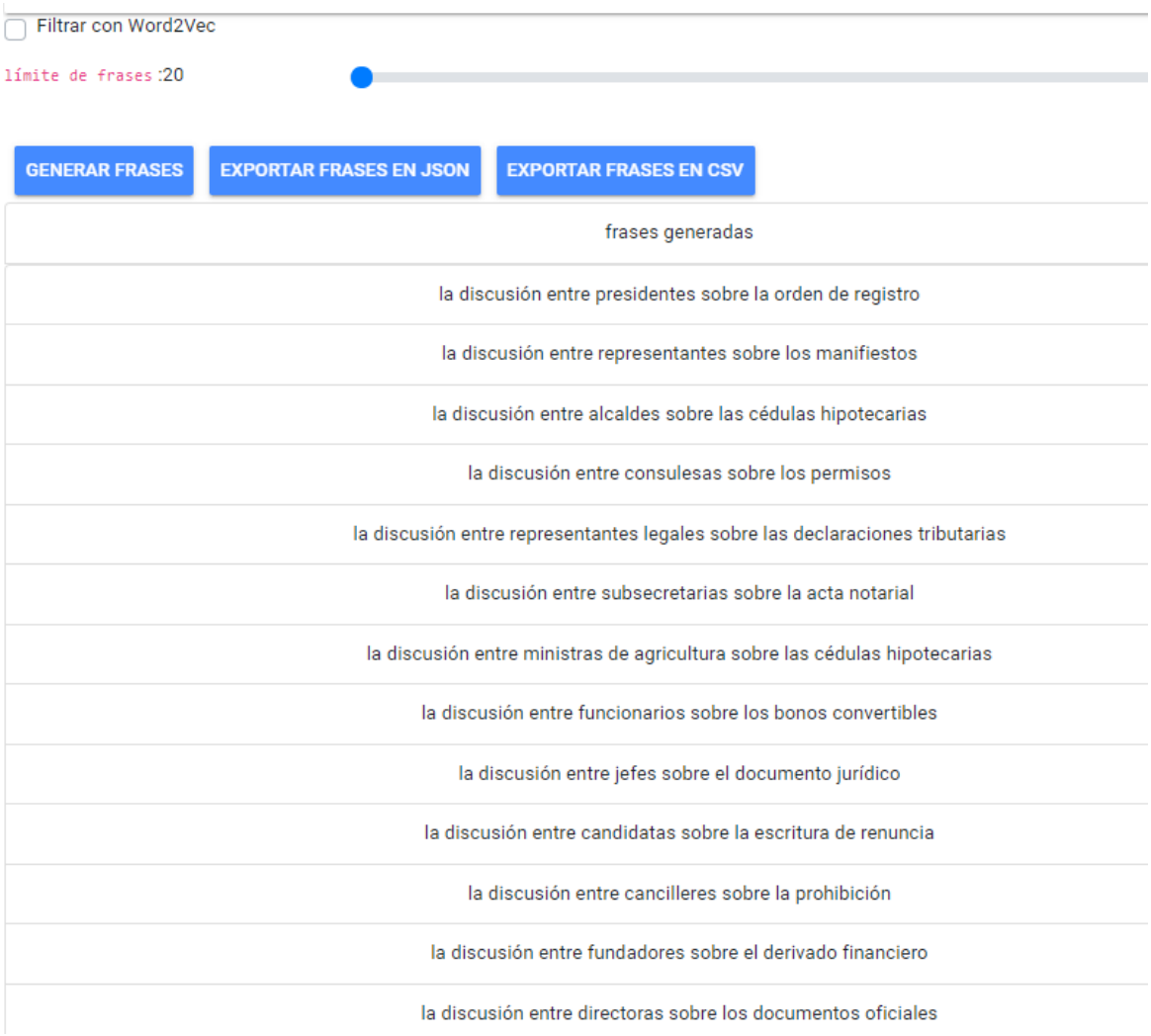


Figura 8: Envorcado de combinatoria

rose	0
carnation	1
peony	2
hydrangea	3
gerbera	4
flower	5
lily	6

$\vec{a}$

tulip	0	1	2	3	4

$\vec{b}$

rose	1	5	3	2	6

Figura 9: Similitude - Word2vec

Este mesmo procedemento é o que aplicamos para calcular a frecuencia de coaparición en contexto das palabras relacionadas cun *token*. Isto é, non usamos a análise vectorial para desambiguar significados, senón para filtrar os datos atendendo á compatibilidade semántico-contextual dos argumentos — ou sexa, a distribución no sentido de Firth (1957, p.11f) ou o significado combinatorio de Engel (2004, p.188). En definitiva, filtramos incompatibilidades do significado combinatorio, o que resulta especialmente relevante, xa que a corrección gramatical non implica directamente corrección ou adecuación semántico- comunicativa. Word2vec tamén se aplica na fase de contextualización (Figura 2) para cribar a coaparición en contexto da modificación adxectival (*vid.* 2.2).

4. Conclusións e traballos futuros

As principais dificultades no desenvolvemento dos prototipos foron as propias dos estudios multilingües e, en especial, a escolla da metodoloxía máis apropiada para a obtención e envorcado dos datos seguindo criterios non so formais, senón tamén semánticos. Establecer unha ontoloxía descriptiva que permitise combinar os trazos categoriais que servían de punto de partida — os da lexicografía valencial - coas ontoloxías de WordNet foi unha das tarefas máis laboriosas. Finalmente desenvolveuse unha ontoloxía propia seguindo unha aproximación *bottom up*.

Os prototipos xa están en funcionamento: da xeración de frases nominais simples — por tanto, frases con estrutura monoargumental correctas dende un punto de vista gramatical e aceptables semánticamente — encárgase o simulador *Xera*. *Combinatoria*, pola súa banda, aporta a xeración automática de frases nominais con estruturas biargumentais, tendo tamén entre as súas competencias a xeración do contexto frasal e do marco oracional. Perséguese, pois, o obxectivo de humanizar os resultados xerados. A curto prazo cómpre acometer:

1. tarefas de optimización dos propios recursos e a súa didactización, destacando aquí o aumento do número das unidades de análise así como de novos campos informativos. Queda por desenvolver o contexto oracional.

2. unha automatización dos procedementos analíticos. Hai que continuar traballando na aplicación combinada de ferramentas co fin de avanzar na automatización e optimización na extracción de información sintáctico-argumental e combinatoria. Nesta liña, uns dos aspectos que estamos a mellorar atinxe á extracción e tratamento do léxico expandido, que se viña facendo de xeito individualizado para cada unha das linguas e que supón, por exemplo, a reiteración de tarefas de depurado. Coa finalidade de evitar esta repetición de procedementos, desenvolveuse *TraduWord*, unha ferramenta de tradución do caudal léxico paradigmático a partir dos datos extraídos de WordNet. *TraduWord* xa foi aplicada no deseño dun prototipo de xeración automática para o galego e o portugués, *XeraWord*[22] (Bardanca Outeiriño et al., 2021), o cal se fundamenta na metodoloxía de *Xera* e *Combinatoria*. O deseño de ferramentas deste tipo abre, por tanto, unha nova canle de traballo.

Cara ao futuro poderíase explorar se a descripción da interface sintáctico-semántica mediante roles e clases semánticas, que da sustento aos xeradores, permitiría avanzar na obtención de resultados sistemáticos e regulares en prol da desambiguación de acepcións de significado. Algúns datos apuntan nesta dirección: así, no substantivo español DOLOR unha metaestructura do tipo [*de* + *determinante* + *nome*] é cómun á expresión do EXPERIMENTANTE (*el dolor del animal*), da ORIXE (*el dolor de la operación*) e da LOCACIÓN (*el dolor de espalda*). Porén, a aparición dun rol locativo e, por tanto, de clases semánticas como [animado humano órgano] ou [animado animal parte do corpo] é propia dunha acepción do tipo 'sensación molesta de tipo físico' e non dunha acepción relacionada co campo do sentimento. Por tanto, a análise dos exemplares léxicos xunto cos roles e clases semánticas podería ser un punto de partida. Ademais contamos con estudos preliminares sobre a modificación adxectival (López Iglesias, 2020) que permiten observar non só a especialización distribucional de determinados adxectivos, senón tamén a súa aparición preferente ou exclusiva segundo a acepción de significado.[23] Nesta liña, un estudo dos datos que envorca Word2vec (*vid.* 3) podería ser de interese para a análise da interface sintáctico-semántica.

[22] http://ilg.usc.gal/xeraword

Agradecementos

Esta investigación está en relación cos proxectos "Generación multilingüe de estructuras argumentales del sustantivo y automatización de extracción de datos sintáctico- semánticos" - MultiGenera (financiado pola Fundación BBVA) — Convocatoria de ayudas a equipos de investigación científica en Humanidades Digitales 2017), "Generador multilingüe de estructuras argumentales del sustantivo con aplicación en la producción en lenguas extranjeras" — MultiComb (financiado por FEDER/Ministerio de Economía, Industria y Competitividad — Agencia Estatal de investigación; 2018, FFI2017-82454-P) así como con "Ferramentas TraduWord e XeraWord: tradución de caudal léxico e xeración automática da linguaxe natural en galego e portugués" (2020-PU004, convocatoria de proxectos colaborativos, USC).

Referencias

Agirre, Eneko & Philip Edmonds. 2006. *Word sense disambiguation. algorithms and applications*. Dordrecht: Springer.

Álvez, Javier, Jordi Atserias, Jordi Carrera, Salvador Climent, Egoitz Laparra, Antoni Oliver & German Rigau. 2008. Complete and consistent annotation of wordnet using the top concept ontology. En *International Conference on Language Resources and Evaluation (LREC)*, 1529–1534.

Bardanca Outeiriño, Daniel. 2020. *Automatic generation of dictionaries*: Universidade de San-

[23] Así, por exemplo, para a expresión dunha dor de tipo físico son frecuentes en español adxectivos como *leve*, *persistente* ou *frecuente*, frente a *sincero*, *hondo* e *irreparable* que aparecen en realizacións do campo do sentimento.

tiago de Compostela. Trabajo de Fin de Máster.

Bardanca Outeiriño, Daniel, María Caíña Hurtado, María José Domínguez Vázquez, José Luis Iglesias Allones & Alberto Simões. 2021. Automatic generation of nominal phrases: Extending the tool *Xera* for portuguese and galician languages. No prelo.

Bentivogli, Luisa, Pamela Forner, Bernardo Magnini & Emanuele Pianta. 2004. Revising WordNet domains hierarchy: Semantics, coverage, and balancing. En *COLING Workshop on Multilingual Linguistic Resources*, 101–108.

Bond, Francis & Ryan Foster. 2013. Linking and extending an open multilingual WordNet. En *Meeting of the Association for Computational Linguistics (ACL)*, 1352–1362.

Delli Bovi, Claudio & Roberto Navigli. 2017. Multilingual semantic dictionaries for natural language processing: The case of BabelNet. En *Encyclopedia with Semantic Computing and Robotic Intelligence*, 149–163. World Scientific. 10.1142/9789813227927_0017.

Domínguez, María José, Miguel Anxo Solla & Carlos Valcárcel. 2019. Resources interoperability: exploiting lexicographic data to automatically generate dictionary examples. En *eLex Conference: Electronic lexicography in the 21st century*, 51–57.

Domínguez, María José & Carlos Valcárcel. 2020. PORTLEX as a multilingual and cross-lingual online dictionary. En *Studies on multilingual lexicography*, 135–158. 10.1515/9783110607659-008.

Engel, Ulrich. 2004. *Deutsche Grammatik – Neubearbeitung*. München: Iudicium.

Firth, John Rupert. 1957. *A synopsis of linguistic theory: 1930–1955*. Philological Society.

Gamallo, Pablo & Jose Ramom Pichel. 2007. Un método de extracción de equivalentes de traducción a partir de un corpus comparable castellano-gallego. *Procesamiento del Lenguaje Natural* 39. 241–248.

Geyken, Alexander, Frank Wiegand & Kay-Michael Würzner. 2017. On-the-fly generation of dictionary articles for the DWDS website. En *eLex Conference: Electronic lexicography in the 21st century*, 560–570.

Gómez Guinovart, Xavier & Miguel Anxo Solla. 2018. Building the Galician wordnet: methods and applications. *Language Resources and Evaluation* 52(1). 317–339. 10.1007/s10579-017-9408-5.

Hashimoto, Tatsunori, Hugh Zhang & Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. En *Conference of the North American Association for Computational Linguistics (NAACL)*, 1689–1701. 10.18653/v1/N19-1169.

Horacek, Helmut & Michael Zock. 2015. *New concepts in natural language generation: Planning, realization and systems*. London: Bloomsbury Academic.

Izquierdo, Rubén, Armando Suárez & German Rigau. 2007. Exploring the automatic selection of basic level concepts. En *International Conference on Recent Advances on Natural Language Processing (RANLP)*, 298–302.

Jiménez, Moreno, Luis Gil, Juan Manuel Torres-Moreno, Roseli S. Wedemann & Erich SanJuan. 2020. Generación automática de frases literarias. *Linguamática* 12(1). 15–30. 10.21814/lm.12.1.308.

Kabashi, Besim. 2018. A lexicon of Albanian for natural language processing. En *EURALEX International Congress: Lexicography in Global Contexts*, 855–862.

Kilgarriff, Adam, Milos Husák, Katy McAdam, Michael Rundell & Pavel Rychlý. 2008. GDEX: automatically finding good dictionary examples in a corpus. En *EURALEX International Congress*, 425–432.

Kosem, Iztok, Kristina Koppel, Tanara Zingano Kuhn, Jan Michelfeit & Carole Tiberius. 2019. Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography* 32. 119–137. 10.1093/ijl/ecy014.

López Iglesias, Nerea. 2020. *Analysing nominal phrase contexts for the automatic extraction of linguistic and lexicographic data*: Universidade do Minho. Trabajo de Fin de Máster.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. En *International Conference on Neural Information Processing Systems*, 3111–3119.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine J. Miller. 1990. WordNet: an on-line lexical database. *International Journal of Lexicography* 3(4). 235–244. 10.1093/ijl/3.4.235.

Nallapati, Ramesh, Bowen Zhou, Cicero dos Santos, Çağlar Gülçehre & Bing Xiang.

2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. En *Conference on Computational Natural Language Learning (CoNLL)*, 280–290. 10.18653/v1/K16-1028.

Navigli, Roberto & Simone Paolo Ponzeto. 2012. BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193. 217–250. 10.1016/j.artint.2012.07.001.

Niles, Ian & Adam Pease. 2001. Towards a standard upper ontology. En *International Conference on Formal Ontology in Information Systems*, 2–9. doi/10.1145/505168.505170.

Otter, Daniel. W., Julian R. Medina & Jugal K. Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems* 1–21. 10.1109/TNNLS.2020.2979670.

Padró, Lluís. 2012. Analizadores multilingües en FreeLing. *Linguamática* 3(2). 13–20.

Renau, Irene & Rogelio Nazar. 2016. Automatic extraction of lexical patterns from corpora. En *EURALEX International Congress: Lexicography and Linguistic Diversity*, 823–830.

Roemmele, Melissa. 2016. Writing stories with help from recurrent neural networks. En *Conference on Artificial Intelligence (AAAI)*, 4311–4312.

Sagot, Benoît & Darja Fišer. 2008. Building a free French wordnet from multilingual resources. En *OntoLex 2008 Workshop*, s/pp.

Sordoni, Alessandro, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao & Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. En *Annual Conference of the North American Chapter of the ACL*, 196–205. 10.3115/v1/N15-1020.

Trap-Jensen, Lars. 2018. Lexicography between nlp and linguistics: Aspects of theory and practice. En *EURALEX International Congress: Lexicography in Global Contexts*, 25–37.

Vicente, Marta, Cristina Barros, Francisco Agulló, Fernand S. Peregrino & Elena Lloret. 2015. La generación el lenguaje natural: análisis del estado actual. *Computación y Sistemas* 19(2). 721–756. 10.13053/CyS-19-4-2196.

http://www.linguamatica.com/

linguaMATICA

Artigos de Investigação

Adaptação Lexical Automática em Textos Informativos do Português Brasileiro
Nathan Siegle Hartmann & Sandra Maria Aluísio

Avaliando entidades mencionadas na coleção ELTeC-por
Diana Santos, Eckhard Bick & Marcin Wlodek

Avaliação de recursos computacionais para o português
Matilde Gonçalves, Luísa Coheur, Jorge Baptista & Ana Mineiro

Projetos, Apresentam-se!

Aplicación de WordNet e de word embeddings no desenvolvemento de prototipos para a xeración automática da lingua
María José Domínguez Vázquez