Allan S. Domingues	9293290
Douglas Seiti Kodama	9277131
Ryan Marçal S. M. Martinez	7612072
Thauan Leandro Gonçalves	9293543

Simplificação de Obras de Machado de Assis com PLN

ICMC - USP
São Carlos

1 Introdução

Pretende-se, com este projeto, o desenvolvimento de um processo automatizado para simplificação de obras de Machado de Assis com técnicas de Processamento de Linguagem Natural e a criação de uma aplicação para acesso aos resultados de tal processo. Encontra-se aqui a continuação de um trabalho sobre o qual este será construído: um córpus contendo a identificação de palavras a serem simplificadas em romances do autor supracitado, tal como sinônimos e estruturas equivalentes adequados ao contexto.

Este trabalho está dividido em múltiplas seções. A seção 2 é destinada a contextualizar o projeto, explicando os motivos para a escolha do autor, como é o córpus que será usado e o que já foi feito com ele, incluindo resultados positivos e problemas encontrados: a seção 3 sumariza nossos objetivos; a quarta seção trata do desenvolvimento em si; na quinta e última seção avalia-se os resultados obtidos com diferentes técnicas incluídas no processo e exemplos.

2 Contextualização

2.1 Da escolha do autor

Machado de Assis foi um autor brasileiro que viveu no século XIX (1839-1908), cuja produção contempla prosa, poesia, teatro e crítica. Embora tenha se iniciado no romantismo, suas obras mais importantes pertencem à fase realista, trabalhos que constituem o córpus aqui utilizado. Trata-se de uma escolha feita no trabalho anterior, de criação do córpus, e neste mantida para que não fossem criadas dificuldades adicionais ao trabalho, evitando fugir de seu escopo e facilitando a reutilização.

Os motivos para a escolha do autor anteriormente, no entanto, são mais fortes: sendo reconhecidamente um dos maiores autores brasileiros e leitura obrigatória na formação básica de estudantes (até recentemente, ao menos), é importante que a leitura de suas obras mantenha-se acessível para qualquer público. Considerando a dificuldade que o tempo pode colocar àqueles que leem suas obras, originalmente voltadas ao grande público (Machado não escrevia obras consideradas difíceis pela forma à sua época), a simplificação de seus textos com base em critérios linguísticos pode atrair novos leitores, sem necessariamente piorar ou tornar simplória alguma obra.

Além disso, visando dar prioridade ao conteúdo do texto original e ao aprendizado de seu léxico e estrutura, optamos por construir uma aplicação que não substitua o original, mas baseie-se em sugestões de palavras semelhantes, geradas automaticamente, evitando assim críticas quanto à degeneração das obras.

Há também motivos técnicos para a limitação a um único autor no momento de criação e anotação do córpus. Lidar com um único autor em uma estética e gênero específicos diminui a esparsidade do córpus em se tratando de fenômenos linguísticos, facilitando disponibilidade de e concordância entre anotadores.

2.2 Do córpus

O córpus utilizado é sincrônico (de uma única época), histórico (com o português usado no Brasil durante o século XIX) e obtido de livros públicos (não infringe direitos autorais); tem como gênero romances e possui representatividade e balanceamento suficientes para o propósito de usá-lo como medida de avaliação para o presente trabalho.

Ele é composto por uma amostra de frases retirada de cinco romances: Memórias Póstumas de Brás Cubas, Quincas Borba, Dom Casmurro, Esaú e Jacó e Memorial de Aires. Essas frases foram escolhidas após uma primeira filtragem pela frequência dos lemas de classes flexionáveis (exceto verbos) no Córpus Brasileiro (SARDINHA et al, 2008).

Isto é, para criação do córpus, houve primeiramente análise morfossintática das sentenças de todos os livros, utilizando-se o *tagger* da biblioteca *NLTK*[1] (BIRD et al, 2009), treinado utilizando o Córpus Mac-Morpho; posteriormente, foram lematizadas todas as palavras que pertencessem às classes gramaticais de substantivo, adjetivo e advérbio, utilizando-se o Unitex. As palavras no Córpus Brasileiro também foram lematizadas e aquelas que possuíam lema igual tiveram suas frequências somadas.

Então, para cada palavra considerada rara (quando a sua frequência após lematizada não fosse superior a 5000, 3000 e 6000 para cada uma das classes gramaticais mencionadas, respectivamente), os anotadores definiam se a palavra deveria ser considerada difícil ou não para um público infanto-juvenil. A interface para anotação das palavras que os anotadores consideravam difíceis está representada na Figura 1.

[1] *Natural Language ToolKit*

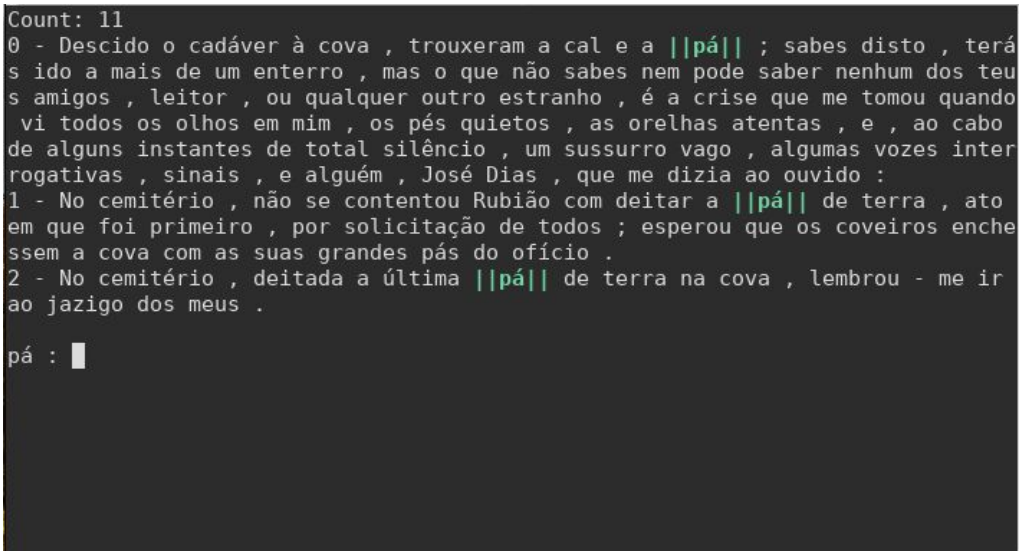


Figura 1: Interface para anotação da dificuldade das palavras. Como mostrado acima, são fornecidas algumas frases que contenham a palavra alvo.

Finalmente, com o uso de um programa, foram selecionadas no máximo 3 frases (menos do que isso quando não havia ocorrências suficientes) para as quais os anotadores deveriam selecionar um conjunto de sinônimos (um por frase, apenas) mais simples fornecidos de um *thesaurus*, o TeP 2.0 (MAZIERO et al, 2009), para criar uma nova versão da frase, conforme exemplo:

> “Era isso motivo de renhidas contendas em nossa casa , porque meu tio João , não sei se por espírito de classe e simpatia de ofício , perdoava no déspota o que admirava no general , meu tio padre era inflexível contra o corso ; os outros parentes dividiam - se : daí as controvérsias e as [rusgas] .”

Importante notar que a frase acima está com apenas uma palavra simplificada, mas no córpus cada frase pode possuir múltiplas simplificações. A interface para se realizar as a anotação de sinônimos está representada na Figura 2.

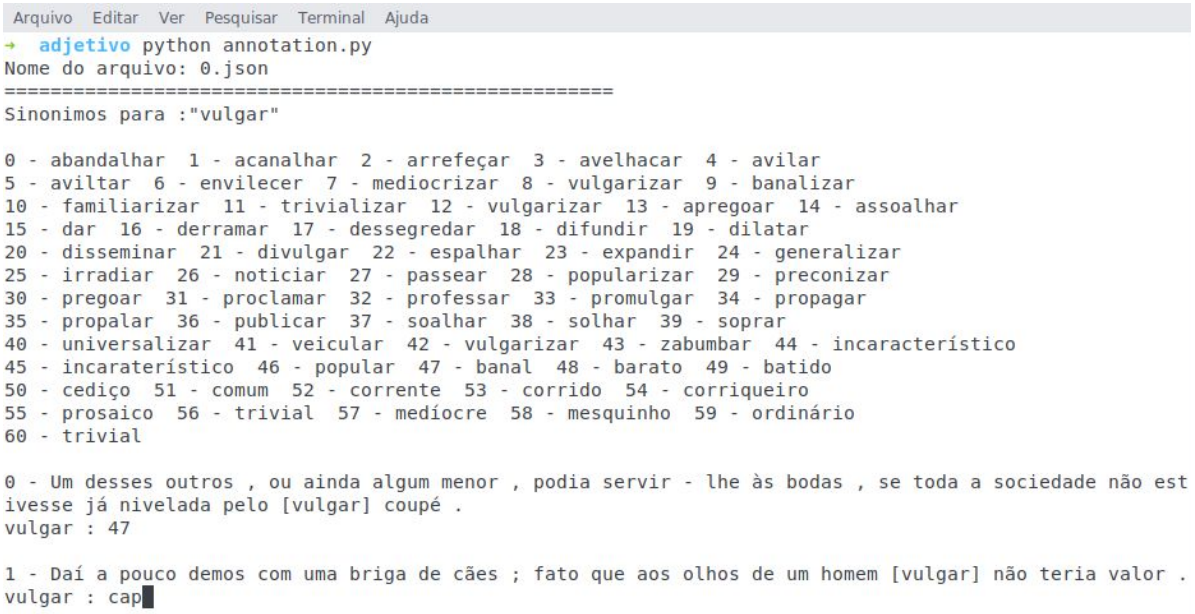


Figura 2: os sinônimos de cada palavra difícil eram associados a números. O usuário deveria digitar um deles para substituir as ocorrências na frase ou um número que não estivesse associado à palavra alguma, caso a frase não pudesse ser mais simples com uma das palavras fornecidas.

2.3 Concordância entre os anotadores

Tendo-se realizado a anotação manual das palavras que foram consideradas difíceis ou não, fez-se uma avaliação da concordância entre os anotadores. A etapa de anotação resultou em vários arquivos no formato *json* que continham uma sequência de 0 ou 1, sendo que o primeiro foi usado quando uma dada palavra foi considerada fácil pelo anotador e 1 caso contrário. Com esses arquivos, utilizou-se a medida de Jaccard de similaridade, que consiste no simples cálculo do número de entradas iguais de dois arquivos dividido pelo número total de entradas.

Seguindo o processo de corte pela frequência descrito anteriormente para os adjetivos que apareciam nas obras de Machado, foram gerados 3 arquivos *json* em comum para anotação. A concordância medida em cada um desses arquivos foi:

Arquivo de adjetivos 1

	Allan	Douglas	Ryan	Thauan
Allan	1	0.52	0.66	0.42
Douglas	0.52	1	0.5	0.54
Ryan	0.66	0.5	1	0.52
Thauan	0.42	0.54	0.52	1

Arquivo de adjetivos 2

	Allan	Douglas	Ryan	Thauan
Allan	1	0.58	0.86	0.7
Douglas	0.58	1	0.56	0.72
Ryan	0.86	0.56	1	0.72
Thauan	0.7	0.72	0.72	1

Arquivo de adjetivos 3

	Allan	Douglas	Ryan	Thauan
Allan	1	0.7	0.68	0.74
Douglas	0.7	1	0.62	0.72
Ryan	0.68	0.62	1	0.82
Thauan	0.74	0.72	0.82	1

Utilizando-se as medidas de concordância em cada arquivo, gerou-se uma medida média de concordância, dada pela média aritmética dos 3 arquivos:

	Allan	Douglas	Ryan	Thauan
Allan	1	0.6	0.73	0.62
Douglas	0.6	1	0.56	0.66
Ryan	0.73	0.56	1	0.686
Thauan	0.62	0.66	0.686	1

Repetindo-se o processo acima, obteve-se também as concordâncias para a classe gramatical dos advérbios:

	Allan	Douglas	Ryan	Thauan
Allan	1	0.89	0.82	0.65
Douglas	0.89	1	0.76	0.54
Ryan	0.82	0.76	1	0.71
Thauan	0.65	0.54	0.71	1

Finalmente, o mesmo processo foi feito para os substantivos, obtendo-se uma média alta de concordância:

	Allan	Douglas	Ryan	Thauan
Allan	1	0.989	0.988	0.985
Douglas	0.989	1	0.989	0.986
Ryan	0.988	0.989	1	0.989
Thauan	0.985	0.986	0.989	1

2.4 Especificação do córpus

O córpus foi armazenado em um arquivo *json*. Cada entrada segue um formato padrão, no qual observa-se a palavra considerada difícil, seguida de uma lista das frases nas quais essa palavra ocorre, sendo que a palavra difícil foi substituída por um sinônimo mais fácil indicado entre colchetes na frase.

Pode-se observar esse padrão na Figura 3.

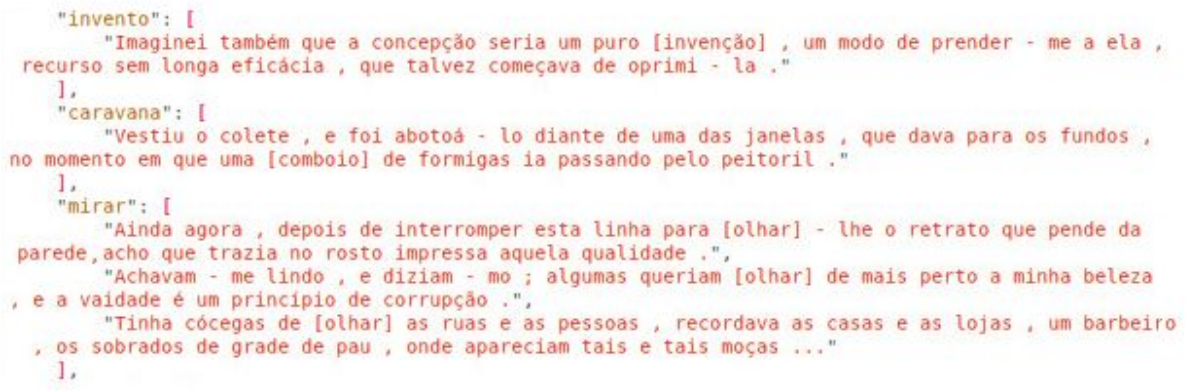


Figura 3: exemplo de palavras difíceis no corpus, destacadas entre colchetes nas frases em que aparecem

A quantidade de palavras difíceis, por classe gramatical, que foram substituídas nos textos, assim como a quantidade de frases em que elas aparecem, podem ser encontradas na tabela abaixo:

	Substantivos	Adjetivos	Advérbios
Ocorrências	26	38	15
Ocorrências em frases	45	75	33

2.5 Problemas encontrados ao construir o córpus

Houve, durante a construção do córpus, a expectativa de que ele fosse representativo, balanceado e refletisse as capacidades do público infantil. No entanto, houve obstáculos logo na definição de quais seriam as palavras difíceis.

A princípio, usar-se-ia a frequência das palavras no Córpus Brasileiro para determinar se elas seriam consideradas difíceis ou não, inclusive para que o viés humano fosse minimizado, mas não havia um limiar razoável que separasse claramente palavras fáceis de difíceis. Decidiu-se, portanto, que os cortes serviriam para fazer uma filtragem inicial, com posterior avaliação humana de dificuldade das palavras.

Além disso, houve problemas com as sentenças originais[2] em relação a caracteres estranhos ou palavras mal divididas. Para tratá-los foram utilizados

[2] Retiradas da biblioteca NLTK (*Natural Language ToolKit*) para Python, que possuiam o córpus do Machado em estruturas internas à linguagem (de programação)

códigos com expressões regulares para retirar caracteres estranhos. Porém, essas adições foram descartadas posteriormente ao verificar-se que o número de ocorrências de palavras sem erros atenuava o problema a ponto de torná-lo irrelevante.

3 Objetivos

Busca-se com este projeto desenvolver um método automatizado e simples para simplificação dos mesmos romances utilizados para a construção do córpus com técnicas de Processamento de Linguagem Natural e uma aplicação que permita ao usuário observar tais versões simplificadas em uma página web similar ao Simplifica (JUNIOR et al, 2009).

A escolha do tema coloca os já mencionados obstáculos para leitura de obras originalmente acessíveis a todos como matéria-prima para o estudo das técnicas de análise morfossintática, lematização, análise de frequências (dentre outras características presentes em corpora), *word embedding* e comparação espacial dos vetores originais para análise semântica.

Além disso, serão testados os resultados usando diferentes técnicas para avaliar o quão satisfatórias elas são atualmente para o Português e se o processo criado também mantém-se satisfatório quando usado em outros textos similares em gênero ou época.

4. Implementação

4.1. Simplificação lexical

Desenvolveu-se para a simplificação lexical um *pipeline* para tratamento dos textos de acordo com o fluxograma abaixo (Figura 4):

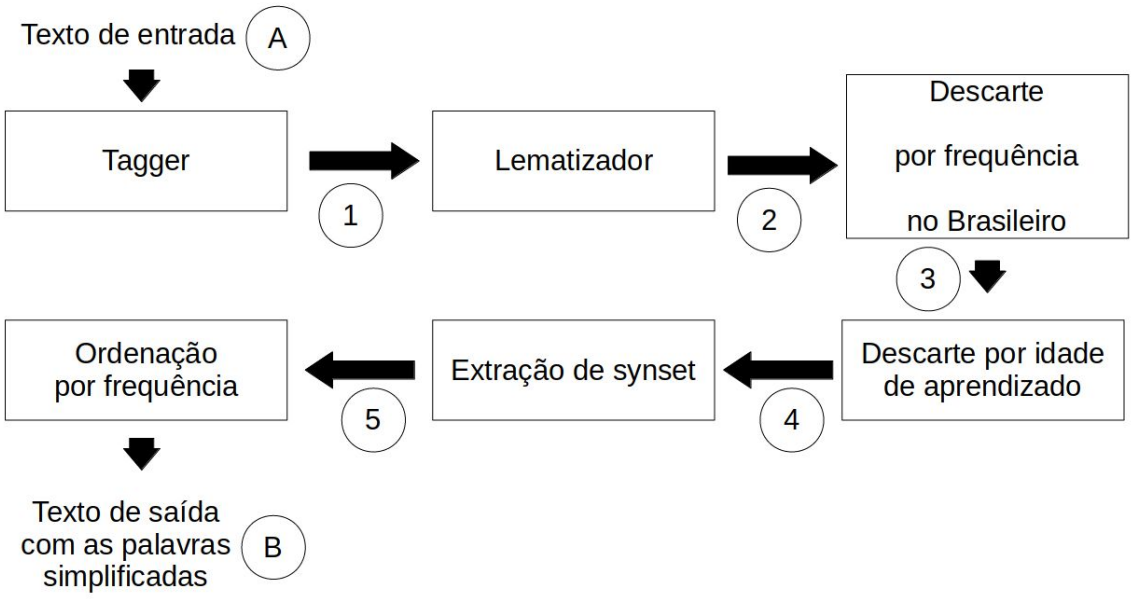


Figura 4: os dados de entrada e saída são chamados de A ou B, com várias etapas de processamento para a mudança; os dados intermediários estão descritos pelos números de 1 a 5 (que são internos ao programa).

O *pipeline* começa com a classificação morfossintática do texto de entrada, treinado com o MacMorpho para que o processo continue para os adjetivos, advérbios e substantivos. Essas palavras são o conjunto de dados (1). A etapa de lematização então torna-se necessária para descobrir o *synset* equivalente a cada palavra (na verdade, todos os *synsets* em que a palavra aparece). As palavras lematizadas estão em (2).

Então as etapas seguintes são para eliminar do processo de simplificação palavras que não sejam raras e portanto, possivelmente, simples. A primeira eliminação por frequência retira de (2) palavras que estejam acima de um limiar de frequência no Córpus Brasileiro após lematização deste (e soma das frequências). Saem então as palavras restantes em (3) para novo corte por frequência de acordo

com uma métrica de idade de aprendizado das palavras. A métrica de idade de aquisição das palavras é baseada em Santos et al (2017), estudo que, entre outras coisas, utilizou *word embeddings* e trabalhos realizados para o português europeu para estimar automaticamente a idade de aquisição de 26.875 palavras.

Dessa etapa saem as palavras a serem simplificadas (4). Então mapeia-se (conceito de chave-valor, como em banco de dados) cada uma destas palavras para um conjunto com todos os *synsets* que a contenham. Assim, restam em (5) pares de palavra a ser simplificadas e um conjunto de todas as simplificações possíveis.

Na etapa de ordenação, para cada palavra “difícil” ordenam-se as possíveis simplificações por frequência de ocorrência no Córpus Brasileiro após lematização. Mantêm-se apenas as 3 palavras mais frequentes, ou o máximo de palavras encontradas. Por fim, para produzir a saída B, são eliminadas as palavras para as quais não foram encontradas substituições possíveis.

4.2. Simplificação gramatical

A complexidade das sentenças utilizadas por Machado de Assis, quando comparada à de textos cultos contemporâneos, não se deve aos pontos geralmente associados a complexidade pela bibliografia especializada (ALUÍSIO et al, 2008). As sentenças do autor, considerando as obras disponíveis no *Corpus Machado*, têm uma média de 15,56 palavras por frase. Comparado a textos jornalísticos contemporâneos, trata-se de um valor baixo: o *corpus* Mac Morpho tem uma média de 22,76 palavras por frase, enquanto o Floresta Sintá(c)tica apresenta o valor semelhante de 22,86. Há 0,17 aposto por sentença nos cinco romances utilizados, e 0,04 estruturas passivas. A média de verbo por sentença é de 2,62, enquanto conjunções subordinativas surgem em 0,22 das sentenças. Os resultados de palavras por sentença são comparáveis àqueles encontrados para textos simplificados em ALUÍSIO et al (2008), e os outros não são superiores aos encontrados para textos jornalísticos, gênero que um aluno de ensino médio, público-alvo da aplicação pretendida, já deve ser capaz de ler. Teóricos apontam que a literatura oitocentista teve o jornal (à época bastante popular) como suporte de divulgação majoritário, difundida em uma sociedade de maioria analfabeta,

sendo identificável "a presença de um certo traço de oralidade nos textos brasileiros" (GUIMARÃES, 2001, p. 23). Assim, a simplificação linguística de suas estruturas envolve somente a identificação de fenômenos que caíram em desuso. Esta seção discute os fenômenos gramaticais relevantes e sua modelagem. Todos eles estão ligados à colocação dos clíticos. Lidamos com o fenômeno utilizando expressões regulares.

De acordo com as gramáticas normativas, a ordem normal de ordenamento dos pronomes oblíquos no português é a ênclise, sendo a próclise permitida em contextos de palavras atrativas. O uso da mesóclise é obrigatório no emprego do futuro, tanto indicativo quanto subjuntivo (BECHARA, 2009). Estudos baseados em *corpora* de produção escolar mostram que que, embora prefiram formas proclíticas, alunos de ensino fundamental e médio utilizam a ênclise em redações (VIEIRA, 2012). A mesóclise, no entanto, é extremamente rara mesmo em dissertações e teses (VILELA, 2005), sendo que suas poucas ocorrências se dão em textos de ciências humanas. Manuais de estilo de grandes jornais recomendam que a forma seja evitada (MARTINS, 1997, p. 69).

Nas ocorrências de mesóclise em verbo plenos, optamos por transformar os futuros mesócliticos na estrutura perifrástica com "ir" (exemplo: "vão nos dizer"), solução baseada no Manual de Redação e Estilo do Estado de São Paulo (MARTINS, 1997). A solução foi escolhida por dois motivos. Primeiro, porque consideramos o texto jornalístico como representativo do português culto atual. Segundo, porque, em alguns contextos, o uso do futuro sintético forçaria o apagamento do pronome. Não é permitido pela gramática normativa iniciar sentença com clítico ("o darei"), nem é permitido pela gramática de qualquer falante nativo pospor o clítico no futuro sintético ("*darei-o, *daria-o"), o que dificulta a implementação do futuro sintético, solução que melhor agradaria a quem vê o uso da perífrase com "ir" como anglicismo. No entanto, de acordo com Oliveira (2006), algumas gramáticas normativas mencionam (ainda que com pouca ênfase) a existência dessa forma que, segundo a autora, possui frequência crescente na língua escrita, e se mantém como a forma mais frequente de futuro na língua falada.

A mesóclise em verbos auxiliares recebe tratamento distinto. Observa-se que o auxiliar "ir" não soa natural quando combinado a outro auxiliar ("vai ter o dado", "ia

ir o comprar", "vai poder o trazer"), nem mantém sua semântica em estruturas interrogativas ("Tê-lo-ia dado ao irmão?" é diferente de "ia o ter dado ao irmão?"). Afixando o clítico ao verbo pleno, opta-se aqui pelo uso do futuro sintético e posposição ("Teria o dado ao irmão?").

Quanto às contrações entre pronome acusativo e dativo, BAKKEJORD (2008, p. 51) aponta que, nos contextos em que a contração seria regra no português europeu, há tendência a omitir o objeto direto e utilizar o clítico de objeto indireto. Para este trabalho, tendo em vista o caráter didático da aplicação pretendida, optamos por não recorrer à supressão (que seria mais natural em português corrente), mas sim à separação dos pronomes, mostrando mais explicitamente ao aprendiz a dinâmica das contrações pronominais. Assim como ocorre com o fenômeno anterior, tratamos as contrações pronominais por meio de três tipos de processo que se aplicam a diferentes casos: a transformação do objeto indireto em sintagma preposicional com pronome dativo seguido ao verbo; a mesma estratégia, mas seguida a um segundo verbo; e a transformação do clítico dativo em pronome possessivo.

A primeira estratégia se aplica à maioria dos casos, nos quais o clítico dativo instancia o objeto indireto de um verbo pleno. Como a regência varia, o verbo é lematizado e tem sua preposição checada em uma lista baseada em argumentos dativos de verbos bitransititvos ("arg2") encontrados no PropBank.Br. Quando o verbo não possui objeto indireto, é assumido que se trata de uma construção benefactiva, sendo-lhe atribuída a preposição "para", como no exemplo abaixo.

"(...) e nós, quando voltávamos à noite para a Glória, vínhamos suspirando as nossas invejas, e pedindo mentalmente ao Céu que no-las matasse..."

no-las matasse -> as matasse para nós

A segunda estratégia é utilizada em caso de verbo auxiliar. A sintaxe do português brasileiro corrente pede que o objeto indireto (ou benefactivo) venha associado ao verbo principal. Novamente, por meio de etiquetação morfossintática, procura-se empareamento de verbos na estrutura-alvo.

"Já meus padrinhos mo haviam dito, e eu reconheço que diziam a verdade".

mo haviam dito -> o haviam me dito

Os outros dois fenômenos cobertos são a simplificação da inversão da negativa (“me não” se transforma em “não me”) e determinadas formas enclíticas

pouco frequentes, especialmente com o verbo “fazer” (“fi-la” se transforma em “eu a fiz”, por exemplo).

4.3 Aplicação

Para exemplificação dos resultados de forma visual e para simular como os resultados da simplificação foram criadas páginas web para as cinco obras de Machado de Assis. Essas páginas estão anexadas ao projeto e há uma descrição de como usá-las junto a elas.

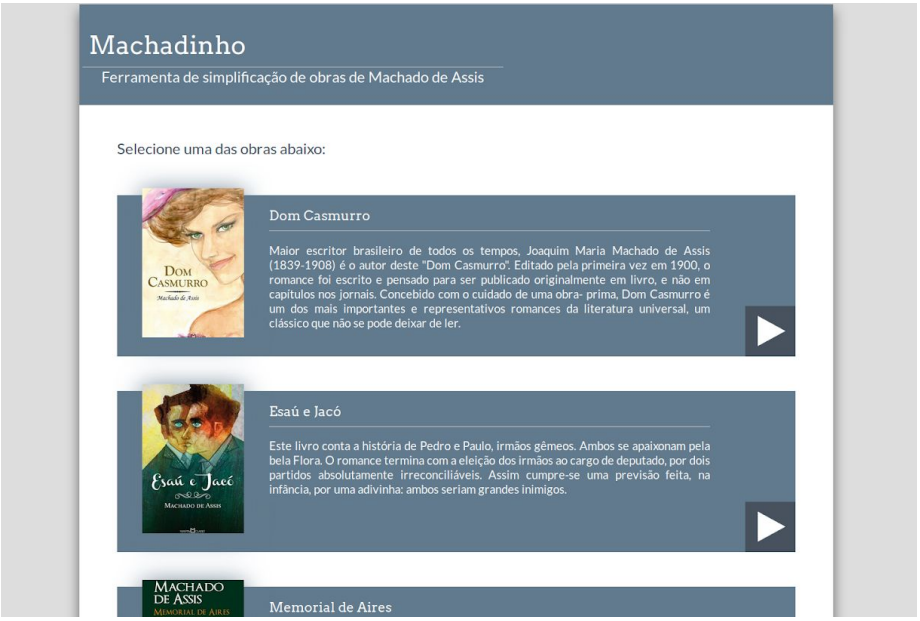


Figura 5: página inicial do site com as simplificações das obras.

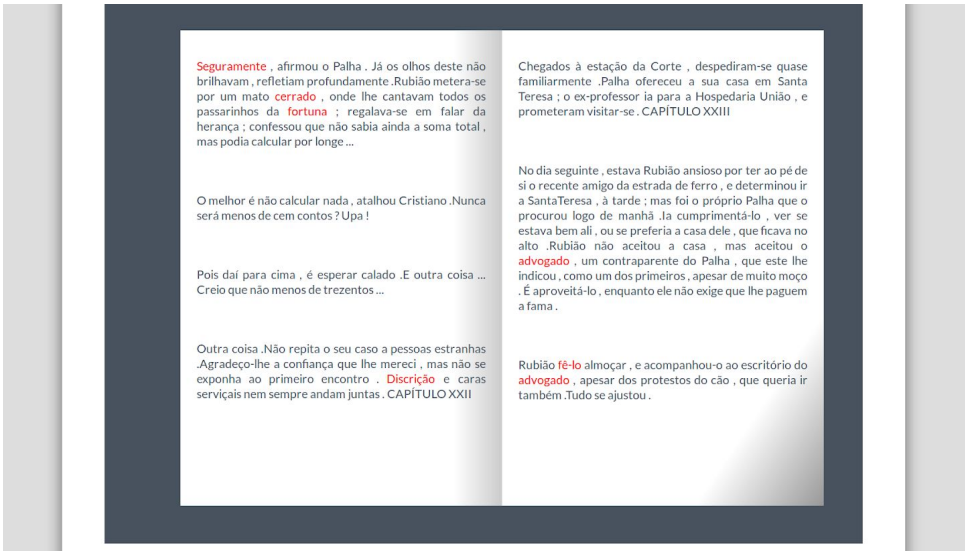


Figura 6: exemplo de livro com palavras difíceis em vermelho, ao passar o cursor, a simplificação aparece em uma nota.

5 Avaliação da simplificação

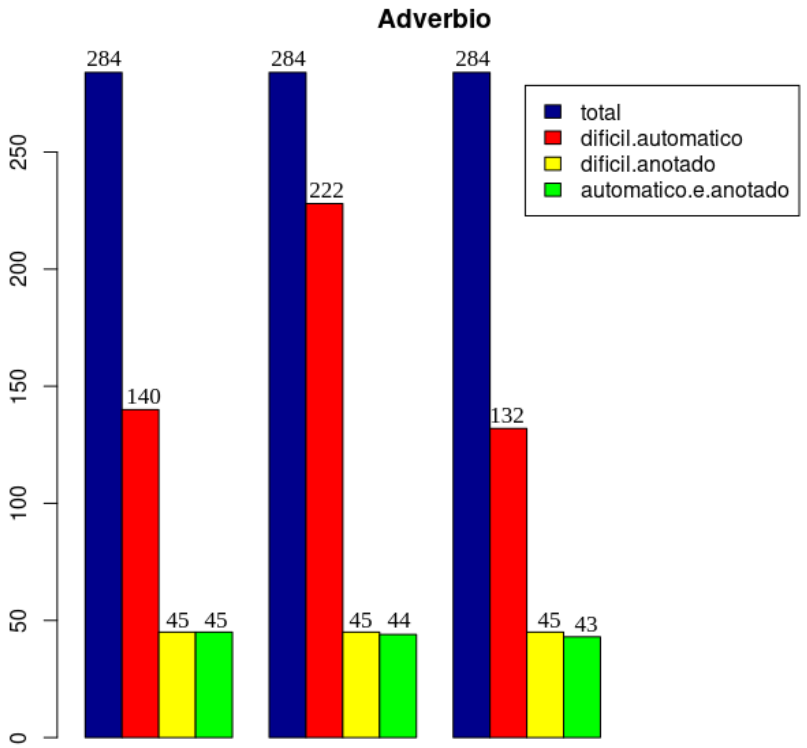
5.1 Parte lexical

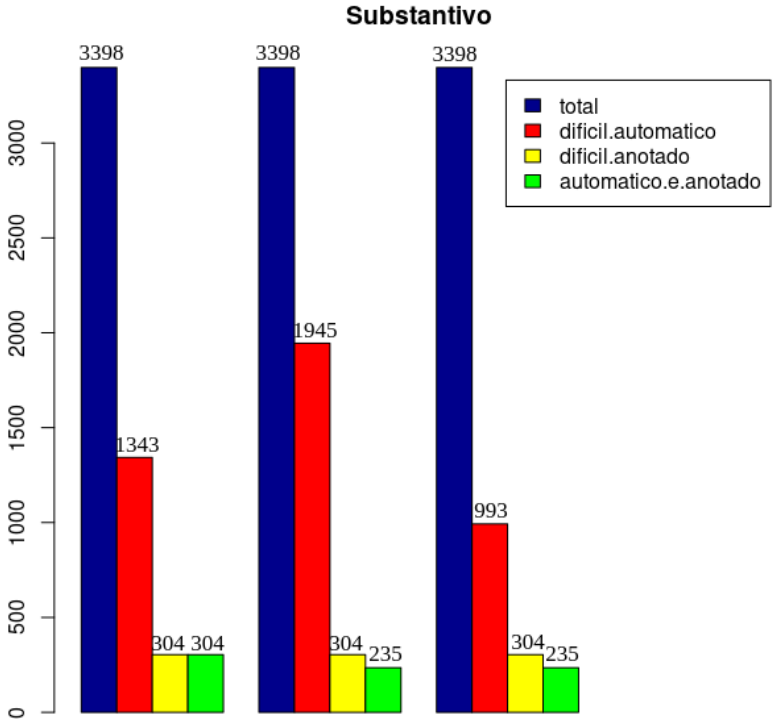
Como citado nas seções acima, a principal utilidade do córpus criado foi para a realização das medidas de avaliações sobre o nosso sistema. Com o objetivo de identificar as melhores técnicas para ser utilizado na etapa final, implementamos 3 destas e calculamos o quanto de discordância elas tinham em relação ao nosso corpus.

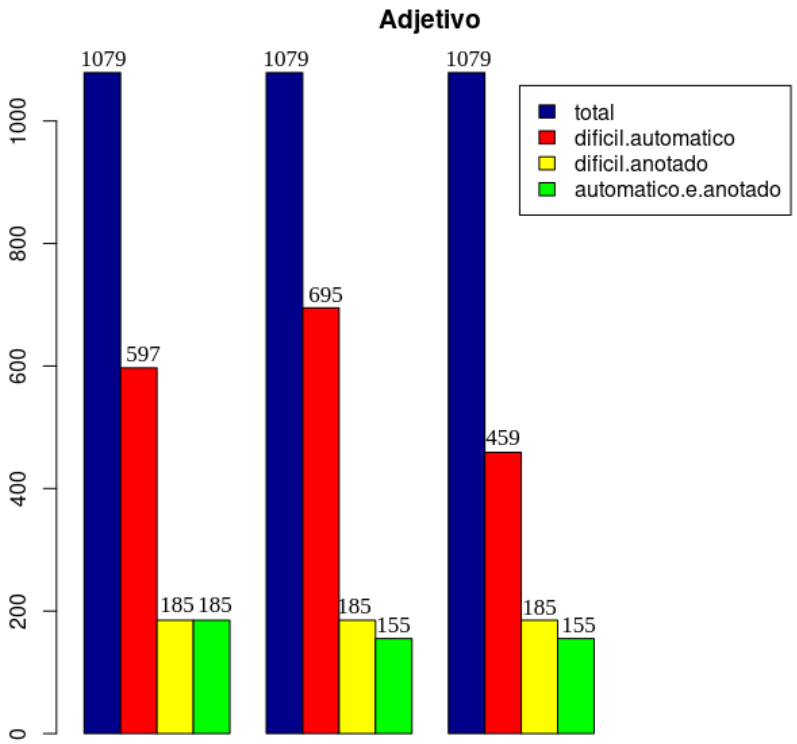
A primeira técnica consiste em utilizar somente as frequências das palavras no corpus Brasileiro, com base nas suas classes morfossintáticas e em um limiar definidos empiricamente por nós, sendo estes também descritos em seções anteriores. Desta forma, as palavras que possuíssem a classe morfossintática substantivo e que possuíam uma frequência de ocorrência inferior ao seu limiar - 5000 neste caso - eram consideradas difíceis, e portanto deveriam passar pelos processos que se seguiam, como obtenção dos sinônimos mais prováveis, substituição destes nos textos, etc.

A segunda técnica consistia em utilizar somente o léxico fornecido no trabalho Psycholinguistic Properties of Brazilian Portuguese (SANTOS, 2017), que consistia de um conjunto de 26875 palavras, sendo que cada uma continha o atributo AoA, que indica a idade média de aquisição da mesma da palavra. Utilizando o nosso córpus, calculamos a média do AoA para as palavras dele, cujo número forneceu um bom estimador para o limiar do AoA que devíamos considerar como difíceis; dessa forma, calculamos que palavras com AoA abaixo de 6.00 poderiam ser consideradas difíceis. Da mesma forma que a técnica anterior, as palavras consideradas difíceis continuavam o processo de simplificação.

A terceira e última técnica consistia em obter as vantagens de ambas as anteriores. Primeiramente era realizado um corte de palavras utilizando as frequências com que aparecem no córpus Brasileiro, e posteriormente era utilizado o léxico do AoA. Os gráficos abaixo mostram os resultados obtidos pelas 3 técnicas com base nas classes morfossintáticas advérbio, substantivos e adjetivo.


Adverbio
284
284
284
140
222
132
45
45
45
44
45
43
0
50
100
150
200
250
total
dificil.automatico
dificil.anotado
automatico.e.anotado


Substantivo
3398
3398
3398
1343
1945
993
304
304
304
235
304
235
0
500
1000
1500
2000
2500
3000
total
dificil.automatico
dificil.anotado
automatico.e.anotado



Cada um dos 3 gráficos se referem a uma classe morfossintática, e para cada classe foi aplicado as três técnicas descritas acima, respectivamente. As colunas azul e amarela permanecem constantes em ambas as técnicas, pois a azul representa o número total de adjetivos obtidos, com base nos 5 textos do córpus, e a amarela representa o número total de adjetivos difíceis anotados. A coluna vermelha mostra o número de adjetivos que permaneceram após a realização dos cortes aplicados pelas técnicas, sendo aqui chamada de número de palavras classificadas; a coluna verde mostra a intersecção das palavras contidas na vermelha e amarela, indicando o número de palavras que estão no córpus anotado que também foram obtidas pelo corte das técnicas, sendo aqui chamada de acertos.

Um dos objetivos do nosso trabalho é diminuir as palavras classificadas e aumentar o número de acertos. As palavras classificadas mostram os números de palavras que permanecem no processo de simplificação, mesmo após os cortes; elas devem ser diminuídas pois não queremos que palavras que sejam fáceis sejam classificadas como difíceis pelo nosso algoritmo. O número de acertos mostra o número de palavras que foram classificadas como difíceis pelo nosso algoritmo e que também foram foram anotadas como tal, o que é o esperado.

Para melhorar as nossas técnicas, poderíamos diminuir o número de palavras classificadas através da diminuição dos valores dos limiares utilizados pelas técnicas, mas isso também diminuiria o número de acertos. Poderíamos aumentar o número de acertos pelo processo inverso, mas as palavras classificadas poderiam aumentar, o que é indesejável também. Esse conflito entre os atributos nos levaram a escolher a 3ª estratégia a ser utilizada na versão final, pois dentre as estratégias ela é a que forneceu o menor número de palavras classificadas, mantendo quase constante o número de acertos.

5.2 Parte gramatical

Para avaliar a simplificação gramatical, foram verificados os resultados dos algoritmos baseados em regras, sendo eles classificados segundo sua adequação ao contexto. Como os problemas encontrados se devem a erros de processamento ou situações de uso não previstas das estruturas-alvo bastante transparentes, foi possível fazer uma classificação bastante precisa dos problemas. Nas subseções abaixo, encontram-se tabelas registrando a frequência de cada problema.

Simplificação da mesóclise

Grande parte das ocorrências de mesóclise foram simplificadas adequadamente segundo os critérios estabelecidos na discussão teórica.

O principal problema foi a ocorrência das expressões opacas "dir-se-ia/á" e "dar-se-ia/á", bastante frequentes e que não podem ter seu significado explicitado por mera inversão. Seus significados aproximados são "parecia" (ou "parece") e "será", respectivamente. Exemplificamos abaixo.

Dir-se-ia que o próprio feto repercutiu a sensação e abençoou o pai.

* Dir-se-ia -> ia se dizer

Dar-se-á que a não ter carregado nada na meninice devo eu o aspecto de "moço" que as primeiras me acharam agora?

* Dar-se-á -> vai se dar

A regra de posposição do clítico, que seria utilizada nas situações minoritárias em que houvesse verbo auxiliar, acabou não surgindo nas sentenças de saída devido a erros de etiquetagem morfossintática.

Ter-me-iam espreitado?

* Ter-me-iam -> iam me ter

Uma ocorrência desse fenômeno não foi adequadamente transformada devido a uma ordenação imprevista para essa construção (auxiliar, sujeito, particípio, objeto). Para lidar com esse tipo de caso, seria indispensável um *parser* sintático eficiente em lidar com estruturas invertidas, algo que não pudemos encontrar. A simplificação correta aqui é "ela os teria abraçados".

Unidos os dois aqui , amados aqui, tê-los-ia ela abraçados ao próprio peito, e eles a ajudariam a morrer.

* Tê-los-ia -> ia os ter

Por fim, uma ocorrência não pôde ter seu significado claramente estabelecido.

Deputado , senador, ministro, vê-lo-iam tudo, com olhos tortos e espantados.

? Vê-lo-iam -> iam o ver

Categoria	**Ocorrências**	**Porcentagem**
Simplificação adequada	56	86%
Erro de etiquetagem morfossintática	2	3,2%
Expressões opacas	5	7,6%
Ordenação imprevista	1	1,6%
Desconhecido	1	1,6%
Total	65	100%

Tabela 1: desempenho da simplificação de mesóclise

Simplificação das contrações pronominais

Assim como o processo anterior, a simplificação das contrações pronominais também obteve uma taxa de sucesso alta, seguindo o modelo estabelecido nas

discussões teóricas, e os erros foram suficientemente transparentes para ser classificados. Os dois erros mais frequentes foram os de regência no sintagma preposicional, que ocorrem geralmente pela ausência do lema no dicionário de regência que elaboramos, e o de tratamento inadequado dos dativos co-referentes a substantivos parte do corpo, questão que explicamos abaixo.

O primeiro ocorreu principalmente com instâncias do verbo "agradecer" (correspondente a 1/3 dos problemas com preposição), que foi combinado à preposição "para" em vez de "a". Exemplificamos com um erro envolvendo outro verbo, cuja preposição só poderia ser estabelecida por análise de contexto, mostrando que a solução escolhida pode envolver tarefas bastante complexas. Tanto "furtar para" quanto "furtar de" são possíveis, mas selecionam papéis semânticos distintos. Na frase em questão, Quincas Borba furta o relógio de Brás Cubas, fazendo com que a simplificação correta seja a segunda.

O Borba furtara-mo no abraço.

* Furtara-mo -> Furtara-o para mim

O segundo problema já havia sido identificado durante a elaboração do algoritmo, e existia o projeto de desenvolver um análise alternativa que desse conta desses casos. Ela foi postergada devido à complexidade envolvida na tarefa, assumindo que, por hora, o algoritmo seria incapaz de lidar com isso. A questão é que algumas contrações pronominais fazem referência a uma entidade e aquilo que poderíamos chamar "nome parte do corpo" (Npc). Em geral, o Npc corresponde a merônimos. No entanto, aqui, qualquer objeto de uma frase com verbo "ter" pode ser chamada de Npc. Isso vale tanto para construções de posse ("João tem um computador. Pedro lho quebrou") quanto com verbo leve ("Pedro já tinha raiva de Maria, e a conversa lha aumentou "). Não foi possível chegar a um tratamento uno para esse fenômeno, que demandaria resolução de anáfora. Damos abaixo um exemplo do *corpus*.

Estendeu-lhe a mão; Camacho segurou-lha ao de leve, e tornou ao papel.

* Segurou-lha -> segurou-a para ele(a)

Assim como ocorreu com as mesóclises, a análise morfossintática não foi capaz de estabelecer os verbos auxiliares, fazendo com que todas as ocorrência de contração com auxiliar resultassem em estruturas pouco fluentes. Uma boa simplificação da frase abaixo seria "Deus é que os há de pagar para ele(a)".

Deus é que lhos há de pagar.

* Lhos há -> os há para ele(a)

Por fim, em dois casos, o módulo em questão teve dificuldade em lidar com a saída dos outros módulos de simplificação gramatical (nomeadamente, o da mesóclise e inversão da negativa), resultando na não alteração da sentença original.

Categoria	**Ocorrências**	**Porcentagem**
Simplificação adequada	113	75,33%

Erro de etiquetagem morfossintática	6	4%
Erro de regência	15	10%
Substantivo parte do corpo	14	9,33%
Sobreposição de módulos	2	1,33%
Total	150	100%

Tabela 2: Desempenho da simplificação de contrações pronominais

Inversão da negativa e ênclises raras

Como essas duas tarefas possuem complexidade muito baixa, os resultados foram todos adequados àquilo que nos propusemos a realizar. Os únicos problemas, no caso da inversão de negativa, são a sua integração com os outros módulos (não plenamente concretizada) e o não reconhecimento de clíticos ambíguos, como o já mencionado "mas".

6 Considerações finais

Os principais problemas pelo quais o nosso sistema passou foi associado a maneira a qual usamos os recursos obtidos, como o Unitex, o Tep, dentre outros. Para realizar a lematização das palavras, nós transformamos o unitex do formato txt para o formato json, sendo utilizado na forma de dicionário. O principal problema associado a ele, que descobrimos posteriormente, foi que durante o processo de transformação do unitex, deixamos de considerar a classe morfossintática da palavra, o que resultou em lematizações não previstas, como nos exemplos abaixo:

- obras: obrar
- advogados: advocar

Esse problema influenciou nos possíveis sinônimos que poderiam ser obtidos para as palavras, resultando em substituições indesejadas, e no pior caso, não havia sinônimos para as palavras, o que ocorreu em 26.8% das vezes. Com base nos sinônimos contidos em nosso corpus, que foram obtidos através da anotação de 156 frases com palavras difíceis, verificamos que os sinônimos obtidos automaticamente correspondem 41.7% das vezes com os sinônimos anotados por nós, o que está diretamente associado com os problemas aos quais citamos nesta seção. Abaixo estão alguns exemplos de sinônimos indesejáveis:

- poeta: versar, rimar, poetizar
- folhas: ler, manusear, folhear
- palestra: falar, tratar, conversar

Outro fator que contribui para alguns erros de substituição foi associado ao pré-processamento do córpus Tep. Durante a transformação do Tep do formato txt para o formato json, deixamos de levar em consideração os diferentes synsets aos quais as palavras pertenciam, o que resultou que no processo de obtenção dos sinônimos, somente palavras com mais frequentes fossem escolhidas. Isto é um problema, pois se levássemos em consideração os diferentes synsets, poderíamos obter palavras de conjuntos diferentes, aumento a probabilidade de que pelo menos um sinônimo pertencesse ao synset correto.

Para futuras versões do mesmo trabalho, realizaríamos as melhorias citadas acima, além de fornecer novas funcionalidades em nosso sistema, dentre eles, hospedar o nosso site na web, de forma que fosse acessível a todos. Além disso, disponibilizaríamos a possibilidade do usuário informar palavras que consideram difíceis e/ou informar os sinônimos adequados para tais palavras. Por fim, uma última funcionalidade seria permitir que o usuário fornecesse um texto que gostaria que fosse simplificado, sendo retornado a sua simplificação.

7 Referências

SANTOS, Leandro Borges dos. **A Lightweight Regression Method to Infer Psycholinguistic Properties for Brazilian Portuguese.** 2017. 9 f. University Of São Paulo, Institute Of Mathematics And Computer Sciences, São Carlos, 2017.

BAKKEJORD, K. **Técnicas de substituição e supressão dos clíticos no português do Brasil**. Dissertação, 2008.

BECHARA, E. **Moderna gramática portuguesa**, ed. 37. Rio de Janeiro : Nova Fronteira, 2009.

GUIMARÃES, H. **Os leitores de Machado de Assis: o romance machadiano e o público de literatura no século 19**. Tese. Campinas, 2001.

MARTINS, E. **Manual de Redação e Estilo de O Estado de S. Paulo**, ed. 3. São Paulo: O Estado de S. Paulo, 1997.

JUNIOR, A et al. Simplifica: a simplified texts web authoring system. **Proceedings of the XV Brazilian Symposium on Multimedia and the Web**. ACM, p. 46, 2009.

OLIVEIRA, J. **O futuro da língua portuguesa ontem e hoje.** Tese. Faculdade de Letras - UFRJ. Rio de Janeiro, 2006.

SANTOS, L et al. A Lightweight Regression Method to Infer Psycholinguistic Properties for Brazilian Portuguese. **Text, Speech, and Dialogue: 20th International Conference, TSD 2017**, p. 27-31, Praga, 2017.

SARDINHA, T. et al. O Corpus Brasileiro, **VII Encontro de Lingüística de Corpus**, Unesp, São José do Rio Preto, SP, 2008.

VIEIRA, S. Variação estilística na escrita escolar monitorada. **Revista do Gelne**, v. 14, p. 213 - 237, Natal, 2012.

VILELA, A. Mesóclise em textos acadêmicos. **SIGNUM**, n.8/2, p. 149-163, Londrina, 2005.