Métodos de Clusterização para a Criação de Corpus para Rastreamento Ocular durante a Leitura de Parágrafos em Português

Sidney Evaldo Leal[1], **Sandra Maria Aluísio**[1],
Erica dos Santos Rodrigues[2],
João Marcos Munguba Vieira[3], **Elisângela Nogueira Teixeira**[3]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
[2] Departamento de Letras - Pontifícia Universidade Católica do Rio de Janeiro (PUC)
[3] Departamento de Letras Vernáculas - Universidade Federal do Ceará (UFC)

[1]sidleal@gmail.com, [1]sandra@icmc.usp.br, [2]ericasr@puc-rio.br
[3]joaomvieira@gmail.com, [3]elisteixeira@letras.ufc.br

Abstract. *This paper presents a method for automating the process of choosing a short passages subset of a large corpus to be used in psycholinguistic research that investigates reading using eye-tracking. To show the method effectiveness, a corpus with 100 short passages of 3 textual genres was used to choose a smaller corpus with 50 passages, using clustering methods and 58 metrics of several linguistic levels. The groups resulting from clustering were evaluated by similarity criteria and the method proved to be useful in supporting the selection of material to be used in psycholinguistic studies.*

Resumo. *Este trabalho apresenta um método para automatização do processo de escolha de um subconjunto de parágrafos de grandes corpora a ser utilizado em pesquisas psicolinguísticas que investigam a leitura usando rastreamento ocular. Para mostrar a efetividade do método, foi utilizado um corpus com 100 parágrafos de 3 gêneros textuais para a escolha de um corpus menor com 50 parágrafos, via métodos de clusterização, usando 58 métricas linguísticas. Os grupos resultantes da clusterização foram avaliados com base em critérios de similaridade e o método mostrou-se útil para apoiar a seleção de material para estudos psicolinguísticos.*

1. Introdução

Atualmente, corpora de rastreamento ocular são frequentemente utilizados no estudo de custos de processamento de estruturas linguísticas para, por exemplo, (i) avaliar modelos e métricas de dificuldade sintática [González-Garduño and Søgaard 2017], (ii) para melhorar ou avaliar modelos computacionais de simplificação via compressão sentencial [Klerke et al. 2016] e (iii) avaliar a qualidade da tradução automática com métricas objetivas [Klerke et al. 2015]. No entanto, existem poucos destes recursos, para um pequeno número de idiomas, por exemplo, inglês [Luke and Christianson 2017, Cop et al. 2016], inglês e francês [Kennedy et al. 2003], alemão [Kliegl et al. 2004] e russo [Laurinavichyute et al. 2018].

Para o português do Brasil, o rastreamento ocular já é utilizado há algum tempo nas pesquisas da área de Psicolinguística. Por exemplo, [Maia et al. 2007]

utilizaram para investigar o papel do processamento morfológico na identificação de palavras; [Leitão et al. 2012] utilizaram na investigação do processamento anafórico; [da Silva e Forster 2013] investigou o processamento incremental de orações relativas restritivas de objeto; e [Teixeira et al. 2014] para evidenciar o custo de resolução de pronomes nulos e plenos. Entretanto, não há nenhum grande corpus do português, publicamente disponível, com dados de rastreamento ocular de jovens adultos e com normas de previsibilidade para a tarefa de leitura silenciosa. Essa é uma grande lacuna que restringe as possibilidades de pesquisa nas áreas de Psicologia Cognitiva, Psicolinguística e Processamento de Línguas Naturais (PLN).

Pesquisas na área de Psicolinguística, especificamente de processamento da sentença, podem se beneficiar de corpora de textos autênticos linguisticamente anotados, que permitem fazer uma correlação dos tempos de leitura com fenômenos linguísticos, por exemplo os elencados abaixo:

1. complexidade estrutural do período (períodos simples vs. compostos);
2. transitividade verbal;
3. animacidade do sujeito e do objeto;
4. tipos de sentenças (ativas/passivas/relativas);
5. mecanismos de construção de relações de correferência, entre outros.

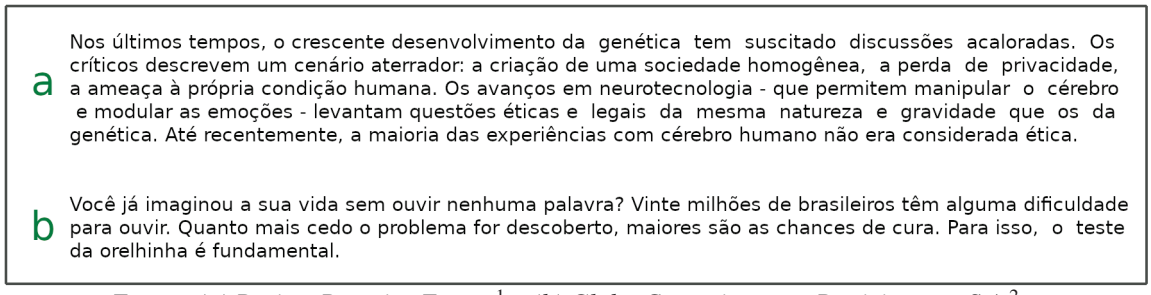
Nos experimentos psicolinguísticos, estímulos são construídos para examinar o efeito de fatores/variáveis independentes no comportamento do participante e, assim, poder investigar hipóteses de trabalho. Uma crítica muitas vezes feita a esses trabalhos diz respeito ao nível de naturalidade dos estímulos experimentais, com consequências em termos do grau de validade ecológica das pesquisas. Assim, projetos atuais utilizam textos autênticos, envolvendo diferentes gêneros textuais (jornalísticos, científicos, literários, etc.), para permitir uma avaliação da influência conjugada de um conjunto de fatores linguístico-textuais que podem afetar o processamento linguístico durante a leitura, em condições menos artificiais de realização da tarefa. Esses corpora são compilados para trazerem uma rica diversidade de fenômenos linguísticos, como, por exemplo, os cinco tipos de descrição das estruturas sintáticas elencados acima, para que se possa correlacionar a diversidade destes fenômenos com tempos de leitura, o comportamento durante a leitura e a avaliação de modelos complexos de controle dos movimentos dos olhos durante a leitura (por exemplo, o E-Z reader - modelo de processamento lexical serial [Reichle et al. 2006] - e o Swift - modelo de processamento lexical paralelo [Engbert et al. 2002]), implementados por simulações computacionais.

Entretanto, uma dificuldade para a compilação desses corpora é a anotação manual destes fenômenos, que idealmente deveria usar mais de um anotador para a avaliação do nível de concordância entre eles [Carletta 1996]. De posse deste corpus anotado com os fenômenos, se pode escolher aquele subconjunto com os atributos variados dos fenômenos linguísticos para a adequação do estudo. Por exemplo, os parágrafos da Figura 1 são do gênero de divulgação científica e jornalístico, respectivamente, e apresentam o mesmo número de sentenças, mas eles diferem em vários níveis linguísticos, por exemplo, na complexidade de seu léxico, na complexidade sintática e tamanho das sentenças, no nível de formalidade.

Dada a disponibilização pública de várias métricas automáticas para avaliação da coesão e coerência de textos escritos ou falados para a língua portuguesa

([Scarton and Aluísio 2010]; [Aluísio et al. 2016]), várias métricas, além do número de sentenças, poderiam ser analisadas para que a escolha dos parágrafos seja adequada para uma dada pesquisa em psicolinguística.

Figura 1. Parágrafos de gêneros diferentes, com mesmo número de sentenças



Fontes: (a) Revista Pesquisa Fapesp[1] *e (b) Globo Comunicação e Participações S.A.*[2]

Esta pesquisa apresenta um método para automatização do processo de escolha de um subconjunto, tomado de um grande corpus de parágrafos para pesquisas que utilizam rastreamento ocular durante a leitura destes parágrafos. Ela é parte integrante do projeto RastrOS[3].

Para mostrar a efetividade do método, utilizamos, como exemplo, um corpus com 100 parágrafos de três gêneros (jornalístico, divulgação científica e literário) (Seção 3), para a escolha de um subcorpus que traga 50 parágrafos, sendo 35 dos gêneros jornalístico e literário e 15 de divulgação científica, via métodos de clusterização, detalhados na Seção 2. O método proposto faz uso de um grande conjunto de métricas de vários níveis linguísticos (Seção 4), disponíveis publicamente na Plataforma Simpligo (https://simpligo.sidle.al/nilcmetrixdoc). Particularmente, foram escolhidas 58 métricas, agrupadas em quatro conjuntos; três destes conjuntos – tipos de sentenças (7 métricas), complexidade da estrutura sintática (22 métricas) e análise de correferência (8 métricas) foram escolhidos para modelar diretamente os três estudos de comparação dos tempos de leitura abaixo: (i) complexidade estrutural do período (períodos simples vs. compostos); (ii) tipos de sentenças (ativas/passivas/relativas); (iii) mecanismos de construção de relações de correferência, entre outros. E o conjunto denominado morfossintaxe (21 métricas) foi escolhido para modelar indiretamente os estudos sobre transitividade verbal e animacidade do sujeito e do objeto. Finalmente, a Seção 5 mostra o conjunto de agrupamentos resultante, juntamente com métodos para avaliar sua qualidade.

2. Aprendizado de Máquina e Métodos de Clusterização

Inicialmente, a área de Inteligência Artificial (IA) era considerada uma área teórica, mas nas últimas décadas com o crescimento do volume de dados e complexidade de problemas que necessitam de tratamento computacional, as técnicas de Aprendizagem de Máquina (AM) começaram a se destacar [Faceli et al. 2011]. Elas são boas ferramentas na criação

[1] https://revistapesquisa.fapesp.br/2002/07/01/manipuladores-de-cerebros/

[2] https://g1.globo.com/bemestar/noticia/mais-de-20-milhoes-de-brasileiros\-tem-alguma-dificuldade-para-escutar.ghtml

[3] Um grande corpus com medidas de RASTReamento Ocular e normas de previsibilidade durante a leitura de estudantes do ensino Superior no Brasil - http://www.nilc.icmc.usp.br/nilc/index.php/rastros

de hipóteses (ou funções) a partir da experiência passada, para predizer respostas ou descrever dados dos problemas que se deseja tratar. Hoje são utilizadas em tarefas tão diversas quanto reconhecimento de fala, detecção de fraudes financeiras, condução autônoma de automóveis, diagnóstico de doenças, dentre outras.

Dentro da AM, existem algoritmos que procuram identificar padrões ou tendências relevantes em conjuntos de dados sem necessidade de um elemento externo servindo de guia do aprendizado. Essas técnicas são chamadas de aprendizagem não supervisionada. Destas, as de clusterização (ou agrupamento) são de especial interesse deste trabalho, pois permitem analisar um grande número de métricas e dados, gerando sugestões de grupos por afinidade.

Os algoritmos dessas técnicas geralmente são classificados em [Faceli et al. 2011]:

- **Baseados em centróides:** Otimizam o critério de agrupamento de forma iterativa, procurando minimizar o erro quadrático ou variação dentro do *cluster*.
- **Hierárquicos:** Geram uma sequência de partições aninhadas a partir de uma matriz de proximidade. Podem ser do tipo **aglomerativo**, que começa com um grupo para cada objeto e vai combinando, ou **divisivo**, que começa com um único grupo e vai dividindo sucessivamente.
- **Baseados em densidade**: Assumem que cada *cluster* é uma região de alta densidade de objetos, separada das demais por regiões de baixa densidade.
- **Baseados em grafos:** Os dados são representados em um grafo de proximidade, no qual cada nó representa um objeto e as arestas, a similaridade ou distância.
- **Baseados em redes neurais:** Sistemas paralelos compostos de unidades simples de processamento; por exemplo o algoritmo SOM (*Self-Organizing Map*).
- **Baseados em *Grid*:** Define um *grid* (reticulado) para o espaço de dados. Muito eficiente para grandes conjuntos de objetos.

O algoritmo mais simples e mais utilizado é o K-Means[4] que utiliza técnica baseada em centróides. Nesta pesquisa, além do K-means, também foram avaliados dois outros algoritmos – o AgglomerativeClustering[5], do tipo hierárquico e o DBScan[6], baseado em densidade. Este trabalho utilizou a implementação deles na biblioteca scikit-learn em python. O DBScan não teve bons resultados no nosso cenário devido ao tamanho e distribuição do conjunto de dados.

3. Conjunto de dados separados por gêneros de texto

Foram selecionados manualmente 100 parágrafos de três gêneros e várias fontes, procurando incluir uma boa amostra para abranger o máximo dos fenômenos do português brasileiro escrito. Os parágrafos do gênero jornalístico foram obtidos de portais de notícias bem conhecidos como G1, Metro, BBC, Reuters, Terra, Estadão, Folha de São Paulo, Jornal da USP, dentre outros. Os parágrafos do gênero literário vieram de romances em domínio público. Os parágrafos de divulgação científica vieram das fontes: Revista Pesquisa Fapesp, Galileu, Aventuras na História, Época, Exame, Isto é, caderno ciência e

[4]https://scikit-learn.org/stable/modules/clustering.html#k-means
[5]https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering
[6]https://scikit-learn.org/stable/modules/clustering.html#dbscan

tecnologia do Jornal do Brasil, Mente e Cérebro, National Geographic Brasil, Piauí, Scientific American Brasil, dentre outras.

A distribuição dos parágrafos pode ser vista na Tabela 1. O objetivo deste trabalho foi selecionar, dentre os 100 parágrafos, um subconjunto, com 50 parágrafos, que mantivesse a maior variância possível dos fenômenos da língua, relacionados com os cinco estudos de comparação dos tempos de leitura, descritos na Seção 1.

Tabela 1. Distribuição dos parágrafos por gênero

Gênero	Quantidade disponível	Alvo da seleção
Jornalístico	43	35
Literário	9	
Divulgação científica	48	15
Total	100	50

4. Métricas Selecionadas

Para representar cada parágrafo do conjunto de dados, foram escolhidas 58 métricas calculadas com o apoio da ferramenta NILC-Metrix[7]. Essas métricas foram agrupadas em 4 conjuntos, resultando em 22 sobre complexidade estrutural/sintática (e.g. períodos simples vs compostos), 7 com tipos de orações (e.g.ativas/passivas, relativas), 8 com mecanismos de construção de relações de correferência e 21 relacionadas com a morfossintaxe (e.g. categorias gramaticais e flexão de substantivos e verbos); a lista completa pode ser vista na Tabela 2.

5. Método para Escolha de Subconjuntos via Clusterização e Avaliação

Após selecionar as métricas, os três conjuntos de parágrafos foram processados e foram executados diversos experimentos, buscando a melhor divisão de grupos, dentro de cada conjunto. Os melhores resultados foram obtidos utilizando a técnica chamada "Método do Cotovelo"[8] (do inglês *Elbow Method*) para encontrar o número ideal de agrupamentos. Esta técnica simula diversas divisões em número crescente de grupos e calcula as variâncias internas de cada grupo, buscando o ponto de equilíbrio [Dangeti 2017]. O gráfico com o cálculo do "cotovelo" para o gênero jornalístico pode ser visto na Figura 2, com o título "Cotovelo Kmeans", no exemplo ele indica 7 grupos ótimos, que foram plotados no gráfico com título "Grupos", com números de 0 a 6.

5.1. Redução de Dimensionalidade via Análise de Componentes Principais

Outra técnica utilizada para melhorar os resultados dos experimentos foi a Análise de Componentes Principais ou PCA (do inglês *Principal Component Analysis*). PCA é um procedimento matemático que cria novas métricas (ou variáveis) que são uma combinação linear das métricas originais e é utilizado para reduzir a dimensionalidade dos dados, sendo aplicado como uma etapa de pré-processamento antes de métodos de clusterização, como os apresentados na Seção 2. Ele permite visualizar os dados no espaço (cf. na Figura 2, os gráficos com títulos "Gênero Jornalístico" e "Textos - por índice") e também melhorar a generalização dos algoritmos [Dangeti 2017].

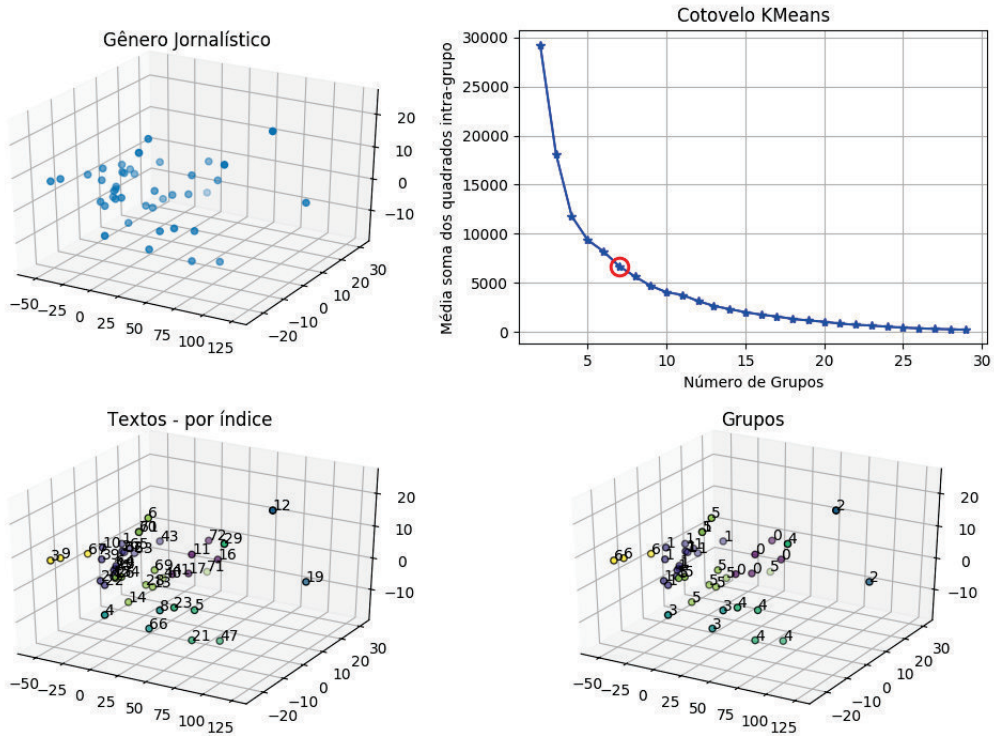
[7]https://simpligo.sidle.al/nilcmetrix

[8]A denominação vem do fato que se o gráfico relembra um braço, então o "cotovelo" (ponto de inflexão da curva) é uma boa indicação de que o modelo subjacente se encaixa melhor naquele ponto.

Tabela 2. Lista de todas as métricas utilizadas.

Nome	Descrição
	Complexidade Estrutural
words_per_sentence	Média de palavras por sentença
sentences	Quantidade de sentenças no parágrafo
words	Quantidade de palavras no parágrafo
sentence_length_max	Quantidade máxima de palavras por sentença
sentence_length_min	Quantidade mínima de palavras por sentença
sentence_length_std	Desvio padrão da quantidade de palavras por sentença
yngve	Complexidade sintática de Yngve (árvores sintáticas fora do padrão de ramificação à direita)
frazier	Complexidade sintática de Frazier (baseada na profundidade das árvores sintáticas)
dep_distance	Distância na árvore de dependência
words_before_main_verb	Quantidade média de palavras antes dos verbos principais das orações principais das sentenças
clauses_per_sentence	Quantidade média de orações por sentença
sentences_with_zero_clause	Proporção de sentenças sem verbos em relação a todas as sentenças do parágrafo
sentences_with_one_clause	Proporção de sentenças com uma oração em relação a todas as sentenças do parágrafo
sentences_with_two_clauses	Proporção de sentenças com duas orações em relação a todas as sentenças do parágrafo
sentences_with_three_clauses	Proporção de sentenças com três orações em relação a todas as sentenças do parágrafo
sentences_with_four_clauses	Proporção de sentenças com quatro orações em relação a todas as sentenças do parágrafo
sentences_with_five_clauses	Proporção de sentenças com cinco orações em relação a todas as sentenças do parágrafo
sentences_with_six_clauses	Proporção de sentenças com seis orações em relação a todas as sentenças do parágrafo
sentences_with_7+_clauses	Proporção de sentenças com sete ou mais orações em relação a todas as sentenças do parágrafo
punctuation_diversity	Proporção de *types* de pontuações em relação à quantidade de *tokens* de pontuações no parágrafo
punctuation_ratio	Proporção de sinais de pontuação em relação à quantidade de palavras do parágrafo
non_svo_ratio	Proporção de orações que não estão no formato SVO (sujeito-verbo-objeto) em relação a todas as orações
	Tipos de orações
passive_ratio	Proporção de orações na voz passiva analítica em relação à quantidade de orações do parágrafo
relative_clauses	Proporção de orações relativas em relação à quantidade de orações do parágrafo
relative_pronouns_div_ratio	Proporção de *types* de pronomes relativos em relação à quantidade de *tokens* de pronomes relativos
subordinate_clauses	Proporção de orações subordinadas pela quantidade de orações do parágrafo
infinite_subordinate_clauses	Proporção de orações subordinadas reduzidas pela quantidade de orações do texto
coordinate_conj_per_clauses	Proporção de conjunções coordenativas em relação a todas as orações do texto
apposition_per_clause	Quantidade média de apostos por oração do texto
	Correferência
adjacent_refs	Média das proporções de candidatos a referentes na sentença anterior em relação aos pronomes pessoais do caso reto nas sentenças
anaphoric_refs	Média das proporções de candidatos a referentes nas cinco sentenças anteriores em relação aos pronomes anafóricos das sentenças
arg_ovl	Quantidade média de referentes que se repetem nos pares de sentenças do texto
adj_arg_ovl	Quantidade média de referentes que se repetem nos pares de sentenças adjacentes
stem_ovl	Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças
adj_stem_ovl	Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças adjacentes
adj_cw_ovl	Quantidade média de palavras de conteúdo que se repetem nos pares de sentenças adjacentes
coreference_pronoun_ratio	Média de candidatos a referente, na sentença anterior, por pronome anafórico do caso reto
	Morfossintáticas
verbs	Proporção de verbos em relação à quantidade de palavras do parágrafo
verbs_max	Proporção máxima de verbos por palavras em relação à quantidade de palavras das sentenças
verbs_min	Proporção mínima de verbos por palavras em relação à quantidade de palavras das sentenças
verbs_standard_deviation	Desvio padrão das proporções entre verbos e a quantidade de palavras das sentenças
verbal_time_moods_diversity	Quantidade de diferentes tempos-modos verbais que ocorrem no texto
adverbs	Proporção de advérbios em relação à quantidade de palavras do texto
adverbs_max	Proporção máxima de advérbios em relação à quantidade de palavras das sentenças
adverbs_min	Proporção mínima de advérbios em relação à quantidade de palavras das sentenças
adverbs_standard_deviation	Desvio padrão das proporções entre advérbios e a quantidade de palavras das sentenças
noun_ratio	Proporção de substantivos em relação à quantidade de palavras do parágrafo
nouns_max	Proporção máxima de substantivos em relação à quantidade de palavras das sentenças
nouns_min	Proporção mínima de substantivos em relação à quantidade de palavras das sentenças
nouns_standard_deviation	Desvio padrão das proporções entre substantivos e a quantidade de palavras das sentenças
pronoun_ratio	Proporção de pronomes em relação à quantidade de palavras do parágrafo
pronouns_max	Proporção máxima de pronomes em relação à quantidade de palavras das sentenças
pronouns_min	Proporção mínima de pronomes em relação à quantidade de palavras das sentenças
pronouns_standard_deviation	Desvio padrão das proporções entre pronomes e a quantidade de palavras das sentenças
adjective_ratio	Proporção de adjetivos em relação à quantidade de palavras do parágrafo
adjectives_standard_deviation	Desvio padrão das proporções entre adjetivos e a quantidade de palavras das sentenças
preposition_diversity	Proporção de *types* de preposições em relação à quantidade de *tokens* de preposições
syllables_per_content_word	Quantidade média de sílabas por palavra no parágrafo

Figura 2. Visualização dos parágrafos e grupos do gênero jornalístico.



5.2. Avaliação do Método

Neste trabalho, os agrupamentos foram gerados utilizando o algoritmo K-Means e AgglomerativeClustering, em seguida foram calculadas as medidas de silhueta (*Silhouette*) para os grupos e *V-Measure* [Rosenberg and Hirschberg 2007] para medir a concordância entre os dois algoritmos. Os resultados podem ser vistos na Tabela 3. A silhueta mede o quão similar é um objeto em seu grupo, em comparação com os demais grupos, e varia de -1 a +1. No nosso cenário, o valor médio 0,38 pode ser considerado bom, tendo em vista que os parágrafos já possuem certa similaridade pela seleção prévia (parágrafos curtos). Já a *V-Measure* obtida reforça que os algoritmos concordam com a divisão dos objetos nos grupos em mais de 90%. A Homogeneidade (*Homogeneity*) avalia se cada grupo contém somente membros de uma única classe, a Completude (*Completeness*) avalia se todos os membros de uma classe estão no mesmo grupo, sendo a *V-Measure* a média harmônica entre elas duas.

Tabela 3. Resultados

Gênero	Número de Grupos	Itens por grupo Med (Min-Max)	K-Means Silhouette	Agglomerative Silhouette	Homogeneity	Completeness	V-Measure
Jornalístico	7	7 (2-14)	0,38	0,38	0,93	0,92	0,92
Literário	4	2 (1-4)	0,39	0,39	1,00	1,00	1,00
Divulgação científica	7	5 (2-15)	0,38	0,35	0,82	0,79	0,81
Média	6	11 (1,6-4,6)	0,38	0,37	0,92	0,90	0,91

A Tabela 1 mostra os alvos de seleção para montar o corpus de 50 parágrafos a partir do corpus inicial de 100 parágrafos (35 parágrafos dos gêneros jornalísticos e literários

e 15 do gênero de divulgação). O experimento realizado selecionou 7, 4 e 7 grupos (cf. Tabela 3). Assim, o trabalho final para montar o corpus de pesquisa pode ser realizado pela escolha manual, apoiada por algum critério importante para a pesquisa, como o tamanho dos parágrafos. No exemplo deste artigo, de 11 grupos serão selecionados 35 parágrafos e de 7 grupos de divulgação, os 15 parágrafos finais.

6. Considerações finais

A possibilidade de selecionar textos com características linguísticas específicas pode ser muito relevante em estudos de natureza experimental na área de psicolinguística. A contribuição desta pesquisa com um método de clusterização que atende esse propósito se mostrou bastante eficiente, pois o conjunto de métricas automáticas ajudou a agrupar parágrafos com características semelhantes, realizando uma anotação dirigida ao agrupamento. Com a lista dos parágrafos agrupados, a tarefa de selecionar manualmente a amostra final tornou-se bem mais simples e informada. É possível selecionar um número de itens de cada grupo, de forma aleatória, ou com alguma forma de ranqueamento (maiores ou menores parágrafos, por exemplo).

Acreditamos que a utilização dos recursos de PLN e Aprendizagem de Máquina não supervisionada podem ajudar bastante em tarefas trabalhosas como a anotação manual dos fenômenos da língua em um corpus. Como continuação deste trabalho, os autores pretendem disponibilizar uma ferramenta web com o método apresentado, para automatizar a análise e permitir que outros pesquisadores consigam replicar o experimento sem esforço de codificação.

7. Agradecimentos

À Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP, processo número 2019/09807-0, pelo apoio financeiro.

Referências

Aluísio, S. M., Cunha, A., Toledo, C., and Scarton, C. (2016). Computational tool for automated language production analysis aimed at dementia diagnosis. In *International Conference on Computational Processing of the Portuguese Language, Demonstration Session*.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, 22(2):249–254.

Cop, U., Dirix, N., Drieghe, D., and Duyck, W. (2016). Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49:602–615.

da Silva e Forster, R. A. M. (2013). *Aspectos do Processamento de Orações Relativas: Antecipação de Referentes e Integração de Informação Contextual*. PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio).

Dangeti, P. (2017). *Statistics for Machine Learning*. Packt Publishing, E-Book.

Engbert, R., Longtin, A., and Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, 42(5):621 – 636.

Faceli, K., Lorena, A. C., Gama, J., and de Carvalho, A. C. P. L. F. (2011). *Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina*. LTC - Livros Técnicos e Científicos, Rio de Janeiro.

González-Garduño, A. V. and Søgaard, A. (2017). Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443, Copenhagen, Denmark. Association for Computational Linguistics.

Kennedy, A., Hill, R., and Pynte, J. (2003). The dundee corpus. *Proceedings of the 12th European conference on eye movement*.

Klerke, S., Castilho, S., Barrett, M., and Søgaard, A. (2015). Reading metrics for estimating task efficiency with mt output. In *Conference on Empirical Methods in Natural Language Processing*, pages 6–13. Association for Computational Linguistics.

Klerke, S., Goldberg, Y., and Søgaard, A. (2016). Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. Association for Computational Linguistics.

Kliegl, R., Grabner, E., Rolfs, M., and Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16*, pages 262–284.

Laurinavichyute, A. K., Sekerina, I. A., Alexeeva, S., Bagdasaryan, K., and Klieg, R. (2018). Russian sentence corpus: Benchmark measures of eye movements in reading in russian. *Behavior Research Methods*, pages 1–18.

Leitão, M. M., Ribeiro, A. J. C., and Maia, M. (2012). Penalidade do nome repetido e rastreamento ocular em português brasileiro. *Revista LinguíStica*, v8 n2.

Luke, S. G. and Christianson, K. (2017). The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*.

Maia, M., Lemle, M., and França, A. I. (2007). Efeito stroop e rastreamento ocular no processamento de palavras. *Ciências e Cognição 2007*, 12:02–17.

Reichle, E. D., Pollatsek, A., and Rayner, K. (2006). E-z reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cogn. Syst. Res.*, 7(1):4–22.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CoNLL)*, pages 410–420.

Scarton, C. E. and Aluísio, S. M. (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.

Teixeira, E. N., Fonseca, M. C. M., and Soares, M. E. (2014). Resolução do pronome nulo em português brasileiro: Evidência de movimentação ocular. *VEREDAS: Sintaxe das Línguas Brasileiras*, 18.