A complexidade narratológica de textos escolares: avaliação por padrões gramaticais através do aprendizado de máquina

Cibele Ribeiro da Cunha Oliveira[1], Claudia Lage Rebello da Motta[1], Carlo Emmanoel Tolla de Oliveira[1]

[1]Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais.– Universidade Federal do Rio de Janeiro (UFRJ)
Caixa Postal 2324 – 21941-916– Rio de Janeiro – RJ – Brasil
{cibele,claudiam,carlo}@nce.ufrj.br

Abstract. *Linguistic framework NLTK based that translates written texts in grammatical symbols in order to find language patterns that are recursive in different narratives. The purpose is evaluative and it allows the analysis of writing in metalinguistic archetypes.*

Resumo. *Arcabouço linguístico baseado em NLTK que traduz obras escritas em símbolos gramaticais com o intuito de encontrar padrões da linguagem que se repetem em diferentes narrativas. O propósito é avaliativo e permite analisar a escrita em arquétipos metalinguísticos.*

1. Introdução

A tríade Informática, Educação e Sociedade surge no contexto que, segundo Fagundes (1988), representa a entrada da sociedade na era da informação e desde então soluções computacionais têm sido elaboradas com o intuito de apoiar e otimizar conteúdos e processos educacionais. O modelo proposto é um estudo integrante deste movimento, visto que abrange aspectos de cada um dos âmbitos formadores da tríade em questão, ou seja, o modelo usa a informática com ferramenta que torna possível a viabilidade da pesquisa, quanto a educação, é uma ferramenta de apoio ao avaliador de produções textuais que muitas vezes se depara com uma quantidade considerável de redações para corrigir sem nenhum suporte métrico para embasar suas ponderações, por fim o viés social está atrelado a descoberta de padrões gramaticais que podem ser utilizados para aprimorar a coesão textual e, consequentemente, em paralelo, o nível de letramento do usuário.

O presente artigo tem como objetivo apresentar o estudo realizado sobre a categorização de narrativas textuais em diferentes níveis com base na clusterização de padrões encontrados a partir da análise gramatical correspondente a metodologia Montessori (2017). Para tal, foi usado o Spekuloom – software em construção – para realizar a categorização textual. Este examinou os textos de 34 alunos entre os diferentes anos do segmento básico escolar do ensino fundamental 2 de uma escola no estado do Rio de Janeiro de Metodologia Montessori. Os textos utilizados foram retirados da redação de uma das suas avaliações escolares e vieram acompanhadas da correção respectiva do professor responsável pelo conteúdo.

2. Trabalhos Relacionados

A perspectiva da pesquisa, na qual etapas implementares são descritas no item 4, é, através da ferramenta computacional, encontrar, no escopo de narrativas, padrões textuais pertencentes a diferentes níveis do texto com influência da análise gramatical Montessori (descrita no item 4.2).

Um artigo recente sobre o uso de algoritmos dos Símbolos Gramaticais Montessori Sasi (2018) comprova que o uso dessa metodologia na classificação gramatical é uma prática relevante no ambiente acadêmico, quando é utilizada na criação de estratégias que tratam de simplificar uma abstração no âmbito da linguagem para que haja uma concretização desta, promovendo, dessa forma, um melhor entendimento do conteúdo.

Há também estudos na área de informática sobre a categorização de gêneros textuais através de “O Jogo do Jornal”, ferramenta que analisa o texto e o classifica em diferentes gêneros, otimizando, através da computação, atividades que para professores seriam enfadonhas se feitas apenas pela mão humana Barros (2009).

3. Metodologia

O estudo apresenta a dimensão da linguagem escrita como base do Modelo Dimensional. É a partir desta dimensão que os padrões gramaticais são analisados e distribuídos em três diferentes níveis escalares, são eles respectivamente: Básico, Intermediário e Transitório.

O processo de elaboração iniciou-se a partir da escolha do escopo. A narrativa foi eleita em função desta ser o tipo textual mais utilizado no ensino fundamental. Em seguida, houve uma pesquisa sobre qual seria a melhor fonte para servir de modelo no aprendizado da máquina. Constatou-se que os textos utilizados em grande escala nesse segmento de ensino são os livros de literatura que englobam todas as idades dos mais novos até os dezoito anos.

A partir da classificação dos livros utilizados, surgiu a proposta de classificar os textos em diferentes níveis que correspondessem à Classificação Decimal Universal Almeida e Santos (2005) dos livros, sendo textos básicos aqueles que possuem padrões similares aos do livro para crianças até seis anos, textos intermediários aqueles que possuem padrões similares aos da literatura infantil e textos transitórios aqueles que possuem padrões similares aos da literatura juvenil.

A quantidade de níveis do texto pode ser ampliada se abranger outras classificações de livros infantis. Há, também, a possibilidade de escolher outro escopo que não livros infantis, criando inúmeras possibilidades. Para o presente trabalho, a análise está dividida nos três níveis mencionados.

Transformar esses livros em dados não é uma tarefa trivial. Foram feitas tentativas através de fotos, digitação e leitura para aplicativos que transformassem em textos. Porém, a busca de material de Domínio Público em arquivos no formato .pdf foi a forma que otimizou a alimentação e limpeza de dados.

O método utilizado nesta pesquisa foi a experimentação com base na teoria PPAE - Planejamento, Projeto e Análise de Experimento - Montgomery (1991),

ocorrendo os três princípios básicos por ele descrito. O tratamento experimental foi replicado algumas vezes para aumentar sua precisão diminuindo erros associados à amostra. A aleatorização ocorreu tanto na escolha de livros para ensinar a máquina quanto na escolha de textos escritos pelos alunos. Finalmente, a blocagem ocorre quando há a clusterização de níveis textuais a partir de padrões gramaticais sequenciais que mais se repetem em cada nível. Trata-se de uma pesquisa quantitativa, pois mensura a recorrência de padrões de classes gramaticais em um dado texto.

O instrumento utilizado foi criado a partir do uso da NLTK, uma plataforma de programação em Python, nela se pode trabalhar com dados da língua humana Bird (2009). Por conseguinte, é possível utilizar suas bibliotecas para classificar, simbolizar, etiquetar, entre outras funções. Por ser capaz de programar o processamento de linguagem, é de grande valia para futuras contribuições relacionadas ao letramento através da linguística computacional.

4. Estratégia de Aplicação do PPAE no Spekuloom

Com base na estratégia de Hockman e Berengut (1995), o desenvolvimento do Spekuloom está descrito em etapas de implementação de modo a caracterizar a pesquisa experimental do mesmo modo que explicita e descreve os diferentes estágios do estudo.

4.1. Coleta de informações

Esse estágio caracterizou-se pela busca de informações referente à classificação textual no que tange a diferenciação entre textos de um mesmo tipo e gênero textual, porém com níveis diferentes de elaboração, ou seja, níveis diferentes de letramento que, segundo Soares (2005), é a aptidão de usar o sistema da leitura e escrita para exercer uma prática social em que a sua utilização é necessária.

A partir desse pressuposto, a atenção foi voltada para como se classifica o texto no caso da leitura e deparou-se com a Classificação Decimal Universal, utilizada em bibliotecas, desta foram retiradas as seguintes classificações: 82-92 Livros para crianças até aos seis anos e 82-93 literatura infantil (seis aos doze anos) e juvenil (doze aos dezoito anos); que serviram de parâmetro para classificar também o sistema da escrita.

4.2. Definição de Objetivos

Com as classificações do sistema escrito amparado sobre as do sistema da leitura, iniciou-se a etapa de definição de objetivos representados pela busca de como encontrar padrões que caracterizassem cada categoria de classificação em um texto escrito. Neste estágio, surgiu a Psicogrammatica Montessori (2017) que através dos símbolos gramaticais, representados pela figura 1, faz a análise morfológica de um texto.

Figura 1. Símbolos Gramaticais

Em seguida, observou-se que Montessori aglutinava classes gramaticais em padrões de três palavras sequências, o que ela denominou família do nome e família do verbo, representados respectivamente pelas figuras 2 e 3. Esta investigação, conduziu a pesquisa a seu próximo patamar com o objetivo já traçado, usar aglutinações sequenciais de palavras com base em suas classificações gramaticais para encontrar padrões que configurem cada classificação: Livros para crianças até aos 6 anos, literatura infantil (seis aos doze anos) e literatura juvenil (doze aos dezoito anos); que passaram a se chamar, a partir desta etapa, respectivamente: nível básico, nível transitório e nível intermediário. É importante apontar que os símbolos utilizados para este estudo se aproximaram ao máximo possível aos da simbologia Montessori, porém houveram algumas adaptações devido às formas disponíveis para uso na programação.

Figura 2. Artigo, substantivo e adjetivo

Figura 3. Pronome, verbo e advérbio

4.3. Projeto do experimento

O software Spekuloom foi idealizado, nesta etapa, como ferramenta que através da linguagem de programação Python usa o framework NLTK e a Análise Bayesiana para validar a classificação textual, ensinar a máquina a fazer a análise automática e, finalmente, avaliar diferentes textos submetidos ao software.

Em um primeiro momento, utiliza-se o framework NLTK para etiquetar as classes gramaticais com os símbolos gramaticais Montessori adaptados. Em seguida, busca-se por padrões sequenciais de palavras com base nas classes gramaticais em diferentes textos. Ao encontrar os padrões, analisam-se quais são os que mais se repetem em cada texto. Depois, alimentam-se os dados com textos das diferentes classificações da CDU usa-se a estatística Bayesiana para prover prognósticos das diversas situações exemplo, ou seja, usa a equação da Análise Bayesiana, apresentada a seguir, para converter os dados e calcular sobre a recursividade de padrões e sua

frequência de probabilidade nos diferentes níveis textuais sendo caracterizado pela a maior porcentagem que qualifica a aprendizagem da máquina. Por fim, textos escritos são submetidos ao programa para serem analisados dentro dos níveis previamente propostos.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|x) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n|c) \times P(c)$$

4.4. Rodar o experimento

O primeiro experimento, após a alimentação de dados para a aprendizagem da máquina com quarenta livros classificados nas diferentes categorias da CDU, foi realizado em dez textos que continham entre 700 e 1800 caracteres.

4.5. Análise do experimento

O software Spekuloom foi utilizado para realizar a análise do texto dentro do critério de padrões simbólicos Montessori. Como resultado, apresentou sete diferentes padrões pertinentes a cada um dos dez textos e sua recursão em cada um deles entre colchetes representados pela figura 4

Figura 4. Análise de padrões gramaticais de dez textos submetidos.

4.6. Interpretação dos resultados

Ao observar apenas sete padrões encontrados, em comparação a análise dos livros – trinta e dois padrões – logo se observou a necessidade de incluir novos textos para que novos padrões surgissem e pudessem ser analisados de forma mais apurada.

4.7. Adequação do experimento

Foi necessário analisar trinta e quatro textos para que a quantidade de padrões encontrados aumentasse para vinte, representado pela figura 5, número significativo e propício para que haja a avaliação piloto de textos com base no aprendizado feito pela máquina.

Figura 5. Análise de padrões gramaticais de trinta e quatro textos submetidos

4.8. Rodadas de confirmação

De posse dos resultados, realizamos a rodada de confirmação para verificar se o modelo estatístico previsto é confirmado. Ao comitar os trinta e quatro textos, observa-se que todos com a exceção de um, figura 6, são considerados de nível básico de acordo com a avaliação, revelando, assim, duas possibilidades.

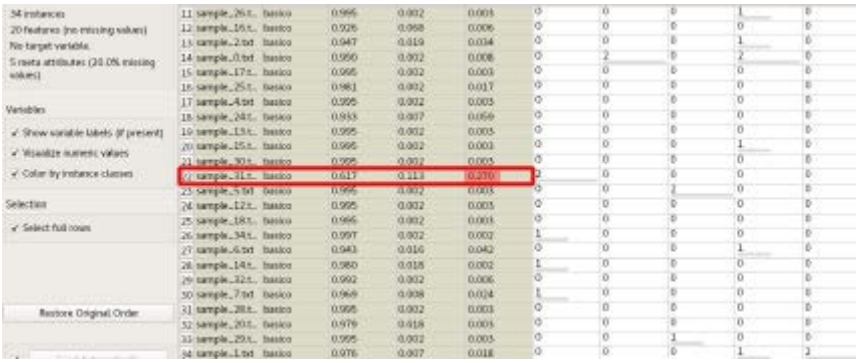

Figura 6. Avaliação Textual em Básico, Transitório ou Intermediário

A primeira, e mais provável, indica que há algum erro na classificação textual, provavelmente devido a quantidade de caracteres utilizados para a busca de padrões nos livros, aproximadamente 8000, em comparação a quantidade de caracteres utilizados na busca de padrões nos textos, de 700 – 1800. A segunda indica que o nível de produção textual dos alunos selecionados ainda encontra-se em estágio inicial, necessitando haver uma atenção maior ao ensino de estratégias da escrita.

4.9. Aplicar Resultados

Mesmo a análise de resultados ainda não completamente calibrada, é possível aplicar o resultado nos textos pois ao avaliar o exemplo de nível intermediário este apresenta uma construção mais elaborada do uso da língua portuguesa em relação aos outros textos, indicando que a hipótese de encontrar padrões recorrentes e pertinentes a cada nível textual, figura 7, está sendo comprovada aproximando-se assim do objetivo do estudo de usar esses padrões para classificar os textos.

0	▲▲◡▲	11	72	43
1	▲◡▻▲	4	13	10
2	◡▲◡▲	16	31	29
3	▲◡▲▲	3	19	11
4	▻▲◡▲	6	15	11
5	▲▲▲●	9	1	6
6	▲▲▻●	3	16	10
7	●▲▲◡	13	30	37

Figura 7. Exemplo de padrões e sua recursividade em cada nível: básico, intermediário e trnsitório

Observe o contraste entre os trechos básico e intermediário respectivamente a seguir:

> Na minha opinião morar nas grandes cidades, eu acho não tem muitos problemas sociais, econômicos ou civis, mas isso não significam que não exita crimes nas grandes cidades, em todo lugar existe crime, mesmo que aparente não ter, uma vez inclusive quando estava voltando do trabalho,[...]

> Enxergar as grandes cidades como florestas, onde as pessoas são animais perdidos e abatidos e os prédios são grandes e misteriosas árvores. A indiferença e o medo estampam os rostos pela rua e a necessidade e os excessos se contrastam.

5. Arquétipo metalinguístico

A observação da existência de padrões diretamente relacionados às classes gramaticais das palavras indica que é possível matematizar a classificação textual em diferentes níveis, ou seja, surge uma nova dimensão arquetípica Kinox (2005) – estrutura fundamentadora – de cunho metalinguístico, pois trata da classificação textual em outro patamar, analisando a linguagem do texto em seu âmbito gramatical.

6. Considerações finais e aplicações futuras

Apesar de ser o resultado do estudo da aplicação piloto da experimentação é possível pontuar a que implicação pedagógica principal observada é: se há uma recorrência de padrões de classes gramaticais sequenciais em diferentes textos, é possível utilizá-los para viabilizar uma estratégia observável da evolução da escrita.

Ao calibrar de forma adequada, será possível identificar os padrões de contagem associados aos diferentes gêneros textuais do tipo narrativa e poderá ser usado para uma ferramenta de tutoração automática. Textos de diversos níveis de letramento poderão ser fornecidos a um processo de aprendizagem de máquina que poderá monitorar o avanço do estudante e fornecer suporte através da sugestão de padrões para o aprendizado da língua.

Referências

Almeida, A. C., e Santos,M.(2005), “eds.CDU: Classificação Decimal Universal: tabela de autoridade.” Biblioteca Nacional Portugal.

Barros, D.R., Marques, C.V.,Oliveira, C.E. e Vrabl, S. (2009), “O Jogo do Jornal: construindo novas estratégias de Letramento.” SBIE.

Bird, Steven, Edward Loper and Ewan Klein (2009), “Natural Language Processing with Python.” O’Reilly Media Inc.

Fagundes, L. (1988), “Informática e educação”, Rio de Janeiro: UFRJ/NCE.

Hockman,K.K. and Berengut, D. (1995), “Design fo experiments.” Chemical engeneering, November, pages 142-147.

Knox,J. (2005), “Archetype, attachment, analysis: Jungian psychology and the emergent mind.” London and New York: Routledge.

Montessori, M., Tornar, C. and Fresco, G. H. (2017), “Psicogrammatica”, FrancoAngeli.

Montgomery, D.C.(1991), “Designe and Analysis of Expiriments” 3rd Ed. New York: Jhon Wiley and Sons.

Sasi, S. (2018). “Creating Algorithmic Symbols to Enhance Learning English Grammar. “International Journal of Research in English Education”, pages 69-93.

Soares, M.B. e Batista, A. A. G.(2005) “Alfabetização e letramento: Caderno do Professor”. Belo Horizonte: Ceale/FaE/UFMG. (64 p.) Coleção Alfabetização e Letramento.

Wagner, W., Bird, S., Klein, E., and Loper, E (2010). “Natural language processing with python, analyzing text with the natural language toolkit”. Language Resources and Evaluation, v. 44, n. 4, (p. 421-424).