Identificação de Padrões de Versificação em Texto em Prosa da Língua Portuguesa

Ricardo Sena Carvalho

Orientadores: Angelo Loula (UEFS), João Queiroz (UFJF)

ricardo.sys@gmail.com, angelocl@ecomp.uefs.br, queirozj@gmail.com

Nível: Mestrado

Pós-graduação em Computação Aplicada

Universidade de Estadual de Feira de Santana (UEFS), Feira de Santana –Bahia - Brasil

Ano/semestre de ingresso no programa: 2013.2

Época esperada de conclusão: 2015.2

Etapas já concluídas: Apresentação do projeto da dissertação 07/2014

Etapas futuras: Exame de Qualificação e Defesa de Dissertação em 2015.2

Resumo. *Historicamente baseada em padrões métricos de versificação, a poesia difere categorialmente da prosa, especialmente sob este aspecto. Mesmo com fronteiras demarcadas, diversos autores têm chamado atenção pro fato de que poesia e prosa compartilham diversas propriedades e recursos formais. Guilherme de Almeida em 1946 e Augusto dos Campos em 1996 revelaram diversos padrões heterométricos (versos de variadas extensões) na prosa de Euclides da Cunha, Os Sertões. O trabalho, manualmente executado, de Almeida e Campos, exige uma análise detalhada, baseada em um complexo sistema de decisões. Aqui é apresentada a proposta para análise e desenvolvimento de uma ferramenta capaz de identificar estruturas heterométricas de versificação em um texto escrito em prosa no português do Brasil.*

Abstract. *Historically based on metric standards of versification, poetry differs categorically of prose, especially in this respect. Even with delimited borders, several authors have called attention to the fact that poetry and prose share several properties and formal resources. Guilherme de Almeida in 1946 and Augusto dos Campos in 1996 revealed several heterometric patterns (verses in varying length) in Euclides da Cunha's prose, Os Sertões. The work, performed manually, by Almeida and Campos, requires a detailed analysis based on a complex decision system. Here it is presented the proposal for analysis and development of a tool to identify heterometric structures of versification in a text written in prose in Brazilian Portuguese.*

Palavras-chave*: Prosa, Poesia, Mineração de Textos, Linguística Computacional.*

1. Introdução

A história da literatura no ocidente exibe uma enorme variação morfológica de entidades estruturadas, linguísticas (verbais) e paralinguísticas (métrica, ritmo, prosódia, etc). Duas grandes categorias se distinguem, em termos classificatórios -- prosa e poesia. Sua distinção, nem sempre de clara demarcação, baseia-se na identificação de uma grande variedade de estruturas e processos. Os textos em prosa, ou discurso direto, de modo geral, podem ser considerados mais comuns, pois estão presentes em um universo maior de obras em diversas áreas por conta de sua natureza mais próxima à prática da comunicação verbal e referencial. A prosa pode ser encontrada, dentre outras fontes, em textos de ficção, tratados em geral, no teatro, livros escolares ou científicos, ofícios, livros de meditação religiosa, especulação política e filosófica. A poesia, por sua vez, lida, mais acentuada ou explicitamente, com paramorfismos fonológicos, como os paralelismos sonoros (rimas, trocadilhos e paronomásias), métricos e rítmicos, morfo-sintáticos e gramaticais (Jakobson 1985). Na poesia, "o ritmo segue uma modulação mais atenta às unidades melódicas - semântico - emotivas que a sintaxe, e muitas vezes os segmentos frásicos articulam-se com relativa simetria, que lembra a regularidade do verso" (MASSAUD,2002).

Em português o sistema de versificação é silábico-acentual – conta-se o número de sílabas de cada verso, e verifica-se a alternância entre as sílabas fortes, acentuadas (ver Spina, 2003). Tal alternância fixa um certo número de padrões que, combinado às cesuras, ou repetições posicionais da sílaba acentuada, cria segmentos internos, estabelecendo as regras de versificação, ou metrificação.

Em 1946, Guilherme de Almeida publicou um estudo sobre aspectos poéticos (métricos e fonológicos) presentes em *Os Sertões* de Euclides da Cunha, publicada em 1902. A obra aborda o confronto armado ocorrido no nordeste brasileiro entre 1896 e 1897, conhecido como Guerra de Canudos. Mais tarde, precisamente em 1996, foi publicado um estudo realizado por Augusto dos Campos, e Guilherme de Almeida, revelando os aspectos poéticos da linguagem euclidiana. Neste trabalho, Almeida e Campos revelam a presença de estruturas heterométricas de versificação, padrões característicos do português. Os decassílados e alexandrinos estão entre os metros mais usados, em combinações e posições variadas. A variedade de padrões encontrados, desprezada "a metrificação estrita" e admitida "maior liberdade rítmica"(Campos, 2010, p. 29), cria surpreendentes zonas de tensão, "áreas pregnantes de poesia em trechos significativos de sua prosa"(Campos, 2010, p. 18), especialmente no início e final dos períodos, como assinalam Almeida e Campos (2010, p. 32).

Feitas manualmente, as análises realizadas por Almeida e Campos exigem, dependendo do tamanho da obra, horas, dias ou até meses de trabalho. Felizmente, hoje, a utilização do computador, com a aplicação de técnicas específicas, pode permitir que análises, antes feitas manualmente e, em muitos casos, durante um largo período de tempo, possam ser realizadas de forma automatizada reduzindo, desta forma, os custos e o tempo destinado a tarefa

Dentre outros motivos, a Internet como a conhecemos hoje fez aumentar a demanda pela análise de grandes volumes de dados, e em função disso, técnicas computacionais

foram desenvolvidas com o intuito de permitir a descoberta automatizada de conhecimento. Técnicas de mineração de textos, aliadas a outros recursos e procedimentos, possibilitam que determinados padrões de comportamento e estrutura possam ser identificados em diferentes tipos de documentos. Segundo Silva, Barros e Prudêncio (2005) a maior parte da informação armazenada nos repositórios digitais encontram-se na forma de documentos textuais. São livros, revistas, artigos científicos, relatórios técnicos, comentários em redes sociais, prontuários médicos, e-mails, propagandas, mídias sociais, dentro outros tipos, construídos e armazenados em formato semi-estruturado.

Este artigo apresenta uma proposta preliminar de pesquisa cujo objetivo é desenvolver e avaliar uma ferramenta computacional capaz de identificar padrões heterométricos de versificação em textos de prosa, em português.

2. Fundamentação Teórica

Ferramentas de avaliação de versos e estrofes e tipos de rima formam um universo de esforços para criar artefatos de softwares capazes de auxiliar iniciantes e experientes na escrita e avaliação de poemas.

Baseado na elaboração de módulos de software e conexão destes, Araújo e Mamede (2002) propuseram um esquema para um classificador de poemas para o português europeu. A proposta dos autores é baseada em uma arquitetura capaz de auxiliar escritores, então poetas, na classificação de estruturas escritas com base nos conceitos de estrofe, verso, sílaba, e rima. Por se tratar de um classificador de poemas, a estrutura do texto deve estar organizada em estrofes. Desta forma, o classificador é capaz de fornecer resultados como a quantidade de linhas, versos, estrofes, sílabas por verso e o tipo de rima.

Plamondon (2014) discute a identificação computacional de alguns padrões, como ritmo e rima, em textos poéticos de língua inglesa. O objetivo da identificação do ritmo não é produzir uma escansão métrica de um poema, mas identificar medidas dominantes com um certo grau de confiança. O trabalho é baseado na possibilidade de que técnicas de computação sejam capazes de identificar o número de sílabas e acentos dominantes em versos importantes, para análise de poemas. O objetivo final é viabilizar uma ferramenta capaz de analisar a estrutura do poema e, a partir desta análise, fornecer dados relevantes sobre tal estrutura ao seu utilizador.

Gervás (2000) propõe a aplicação de regras de lógica de programação para identificar estruturas métricas aplicadas a poesias em espanhol. A técnica apresentada pelo autor visa a identificação e classificação do verso e da rima. O objetivo é fornecer uma ferramenta pedagógica para auxiliar no ensino de poesia em espanhol. O autor realiza o processo por meio da extensão da lógica de programação para técnicas de processamento de linguagem natural. Segundo o autor, a ferramenta foi testada em uma base de sonetos em espanhol e apresentou bons resultados.

Conforme exemplificado nos três últimos parágrafos, o desenvolvimento de soluções voltadas para identificação e classificação de estruturas poéticas pode ser encontrado para diferentes idiomas. Porém, para o português do Brasil não foram encontradas fontes que revelassem que este tipo de pesquisa já tenha apresentado resultados

consistentes, capazes de comparar com outros resultados apresentados para ferramentas de apoio a leitura de poesia em outros idiomas, como o espanhol e o inglês. Ao contrário dos trabalhos encontrados, que avaliam poemas com estruturas de versificação gráfica e explicitamente demarcadas, facilitando o processamento das informações e avaliação dos resultados, este trabalho propõe a automatização da procura de tais estruturas, em prosa, trabalho ainda não realizado para o português ou qualquer outra língua, segundo nossa revisão.

Embora não tenham sido encontrados trabalhos relacionados diretamente com esta pesquisa, no aspecto técnico e metodológico a pesquisa tem uma forte ligação com o processo de mineração de textos por incluir dentre suas etapas processos como o pré-processamento do texto e redução de dimensionalidade.

Para realizar a identificação de padrões heterométricos de versificação, um processo de separação e encadeamento de sílabas poéticas, ou sílabas não-gramaticais, precisa ser feito como uma das etapas do processo de mineração; este processo de separação e encadeamento de sílabas poéticas é chamado de escansão. Por fazer uso de características fonológicas, outra área de conhecimento, a Linguística Computacional, está sendo estudada. Isto decorre do tratamento dado aos textos por algoritmos de síntese de voz, sílabas tônicas e átonas que precisam ser identificadas pelo algoritmo para que sejam emitidos sinais de sonoridade compreensível.

Tal característica dos sintetizadores de voz permite que este trabalho utilize técnicas destas ferramentas porque elas consideram propriedades fonéticas importantes para emissão de um sinal de voz. Silva (2011) apresenta em sua tese de doutorado um trabalho intitulado "Algoritmos de Processamento da Linguagem e Síntese de Voz com Emoções Aplicados a um Conversor Texto-Fala Baseado Em HMM" dois algoritmos, o Algoritmo de Determinação de Tonicidade e o Algoritmo para Separação de Sílabas, que estão sendo implementados como parte do processo de mineração proposto.

3. Metodologia

Este trabalho envolve a aplicação de técnicas computacionais capazes de automatizar a procura por padrões de versificação, heterométricos, em textos de prosa. Para este trabalho, serão aplicadas técnicas de Mineração de Textos e de Linguística Computacional por se tratar de áreas da computação que já possuem resultados reconhecidos pela comunidade de computação.

Foram planejadas diversas etapas de processamento de um texto em prosa, para obter indicações sobre a distribuição de padrões heterométricos de versificação ao longo do texto. Na primeira etapa (seleção dos dados) será utilizado apenas um documento, ou seja, apenas uma obra literária, ou diversos trechos desta, onde serão descobertos os diversos padrões.

Na etapa seguinte ocorre a preparação dos dados. Esta atividade envolverá a separação das sílabas, identificação das tônicas, etiquetagem da sílaba tônica e demarcação da oração para redução da dimensionalidade (demarcação do início e fim da frase ou verso). A separação silábica consiste na separação silábica propriamente dita, utilizando

regras de separação silábica poética. Já o processo de identificação das tônicas se traduz pela procura de tônicas naturais ou artificiais. Este processo se faz necessário por conta da necessidade de identificação da posição da tônica na sentença, importante para demarcar os padrões de versificação estabelecidos pelo português e necessários para atingir os objetivos do projeto. A etiquetagem é o processo de inserção de marcações nas sílabas tônicas e em pontos que determinem o início e fim da frase para diminuição do espaço de avaliação

Após esse pré-processamento, o texto deverá passar para a fase de mineração de dados propriamente dita através da classificação baseada em padrões heterométricos determinados. Após a efetivação destas etapas, devem ser catalogados os trechos de acordo com sua classificação criando, desta forma, um índice.

Na penúltima etapa, será apresentado ao usuário um mecanismo de visualização dos resultados, uma interface que permitirá visualizar a presença dos diferentes padrões métricos, ao longo do texto. Esta interface permitirá, por exemplo, destacar a presença, frequência e distribuição dos versos em todo o texto facilitando a execução da etapa seguinte.

Na última fase, a análise será conduzida por um especialista de domínio, para verificar se os padrões encontrados e demarcados pelas fases anteriores realmente podem ser seguramente definidos como padrões de versificação, utilizando os resultados apresentados pela interface.

Para validação da ferramenta, serão feitos testes de classificação e usabilidade; o texto base para testes utiliza as descrições e análises já realizadas por Almeida (1946) e Campos (2010), nos *Sertões*, de Euclides da Cunha. Deve-se enfatizar que Campos informa-nos que chegou a mais de 500 decassílabos no livro, entre sáficos e heróicos, e mais de duas centenas de dodecassílabos (Campos, 2010, p. 14). Os resultados serão tabulados, tendo como base a eficiência do método e o percentual de identificação apresentados em gráficos, para que o usuário possa visualizar o nível de relevância e conduzir um estudo mais detalhado do objeto alvo de análise.

4. Resultados Esperados

O principal resultado é uma ferramenta de suporte a pesquisas em crítica, história e teoria literária, de um lado, linguística estrutural e ciências cognitivas, de outro, especialmente seu ramo dedicado ao raciocínio diagramático (*diagrammatic reasoning*) em processos cognitivos simbólico-verbais. Pretende-se desenvolver mecanismos de análise, classificação e visualização de padrões heterométricos de versificação identificados inicialmente na prosa de Euclides da Cunha, a partir de estruturas silábico-acentuais normativizadas em língua portuguesa (decassílabos, alexandrinos, etc). Seu impacto e implicações são diversos, incluindo distribuição estatística de estruturas mais normativizadas, ou cristalizadas, em diversos momentos da prosa, e sua relação com padrões estruturais menos normativizados, incluindo uma comparação de frequências e densidades relativas destes padrões ao longo do texto. Interfaces serão desenvolvidas com o intuito de disponibilizar o máximo de informações sobre a localização dos padrões encontrados no texto.

5. Estado atual da pesquisa

A identificação de padrões de versificação envolve um processo chamado de escansão. Esta tarefa é caracterizada pela contagem de sílabas poéticas e identificação da posição das tônicas na sentença. A alternância entre as tônicas fixa um certo número de padrões que, combinado às cesuras, ou repetições posicionais da sílaba acentuada, cria segmentos internos, estabelecendo padrões de versificação. A criação de um algoritmo capaz de identificar a posição da tônica e a correta separação de sílabas poéticas encontra-se em desenvolvimento.

6. Etapas futuras do projeto

O desenvolvimento da solução passará pela criação de módulos com o objetivo de cumprir as etapas descritas na sessão 3. Após a conclusão do módulo de pré-processamento, será realizada sua avaliação e validação. Em seguida, será realizado a especificação e o desenvolvimento do módulo de reconhecimento de padrões de versificação e sua posterior avaliação. Por fim, será realizado um estudo, especificação e desenvolvimento da interface capaz de representar dados estatísticos (definição de medidas) e formas de visualização dos resultados de análise. Além das etapas de construção dos módulos uma fase será dedicada a aplicação da solução em textos literários diversos.

7. Referências

Almeida, G. "A poesia d'Os Sertões". Diário de São Paulo. 18 de agosto de 1946, 1946.

Araújo, P. & Mamede, N.–"Classificador de Poemas", CCTE 2002, Lisboa, Portugal, Maio 2002.

Campos, A.; Almeida, G. "Poética de Os Sertões". São Paulo: AnnaBlume, 2010.

Gervás, Pablo. "A logic programming application for the analysis of Spanish verse". In: *Computational Logic—CL 2000. Springer Berlin Heidelberg*, 2000. p. 1330-1344.

Jakobson, R.; Pomorska, K. "Diálogos". São Paulo: Cultrix, 1985.

Massaud, M. "Dicionário de Termos Literários(em português)". 11 ed. São Paulo: Cultrix, 2002. 520 p.ISBN9788531601309 Página visitada em 12 de janeiro de 2013.

Plamondon, Marc - R. "Virtual verse analysis: Analysing patterns" *in poetry.Literary and linguistic computing*, v. 21, n. suppl 1, p. 127-141, 2006.

Silva, D. C. "Algoritmos de processamento da linguagem e síntese de voz com emoções aplicados a um conversor texto-fala baseado em HMM". Tese de Doutorado, COPPE/UFRJ, Rio de Janeiro, 2011.

Silva, E.; Barros, F. A.; Prudêncio, R. BC. "Uma abordagem de aprendizagem híbrida para extração de informação em textos semi-estruturados." In: *XXV Congresso da Sociedade Brasileira de Computação* (Julho 2005). 2005. p. 504-513

Spina, S. "Manual de Versificação de Românica Medieval." São Paulo: Atelier Editorial, 2003.