Do contato entre a Literatura, a Linguística de *Corpus* e o Processamento de Língua Natural: o caso dos anagramáticos de Guimarães Rosa

From the contact between Literature, Corpus Linguistics and Natural Language Processing: the case of the anagrammatics of Guimarães Rosa

Átila Augusto Soares Vital *[1]

[1] *Universidade Federal de Minas Gerais, Faculdade de Letras, Belo Horizonte, MG, Brasil.*

Resumo

Da tentativa de realizar a cooperação entre a Linguística de *Corpus* e o Processamento de Língua Natural (PLN), foram alcançados importantes frutos, como a possibilidade de processamento de grandes dados linguísticos e o desenvolvimento de tecnologias que se utilizam de dados da língua. A relação entre essas duas áreas e os Estudos Literários, no entanto, tem sido pouco explorada, o que abre espaços para o presente trabalho, que tem por objetivo fazer uma análise exploratória da construção dos poemas atribuídos a anagramáticos de João Guimarães Rosa, em *Ave, Palavra*, obra de 1970. Para isso, foram utilizadas, em conjunto, abordagens da Linguística de *Corpus* e do PLN, associadas aos trabalhos de Rossi (2007), Brito (2012) e Vital (2021), acerca da obra rosiana. Com o processamento computacional do *corpus*, pudemos extrair: a) o número de palavras; b) a razão type-token; c) o número de estrofes e de versos e d) as palavras mais frequentes para cada um dos anagramáticos. Os dados foram dispostos em gráficos e nuvens de palavras (*wordclouds*). Desses resultados, foi observado que existem, de fato, diferenças quantitativas e qualitativas presentes no nível poético, reafirmando, por meio de observações das epígrafes de cada anagramático, a complexidade envolvida na criação da metapoeticidade de suas máscaras.

Palavras-chave: Linguística de *Corpus*. Processamento de Língua Natural. Guimarães Rosa.

Abstract

From the attempt to achieve the cooperation between *Corpus* Linguistics and the Natural Language Processing (NLP), important products have been created, as the possibility of processing lots of linguistic data and developing technologies that use language. The relationship between those areas and the Literary Studies, however, has been less studied, opening spaces for this study, which has the objective of carrying out an exploratory analysis of the poems assigned to the anagrammatics of João Guimarães Rosa, in *Ave, Palavra*, from 1970. In order to do so, approaches of *Corpus* Linguistics and NLP were used together, associated with the works of Rossi (2007), Brito (2012) and Vital (2021), about the rosian oeuvre. Using computational processing, we extracted the following data from the *corpus*: a) the number of words; b) type-token ratio; c) the number of stanzas; d) the most frequent words for each anagrammatics. The data were displayed in the form of graphics and word clouds. From the results, we observed that there are quantitative and qualitative differences for each poet, reinforcing, through observations of the epigraphs of each author, the complexity involved in the metapoeticity of anagrammatic masks.

Keywords: *Corpus* Linguistics. Natural Language Processing. Guimarães Rosa.

Textolinne
Linguagem e Tecnologia

DOI: 10.35699/1983-3652.2022.39316

Seção:
Artigos

Autor Correspondente:
Átila Augusto Soares Vital

Editor de seção:
Daniervelin Pereira
Editor de layout:
Leonado Araújo

Recebido em:
20 de abril de 2022
Aceito em:
3 de setembro de 2022
Publicado em:
14 de setembro de 2022

1 Introdução

Nas últimas décadas, com o desenvolvimento acelerado da Linguística de *Corpus* (LC) e da possibilidade do processamento de grandes volumes de dados em computadores domésticos, foram estabelecidos novos paradigmas nas Ciências da Linguagem. Através de uma ótica diferente das proposições lógico-dedutivas a respeito da natureza da língua, a linguística de base empírico-indutiva ganhou seu lugar de destaque dentro dos Estudos Linguísticos, já que é cada vez mais fácil processar dados – sejam eles escritos, falados ou, até mesmo, multimodais (que mesclam diferentes modos de uso da

*Email: guto.soares.victal@hotmail.com

língua) – para análises segundo diferentes modelos teóricos. No Brasil, temos exemplos de importantes corpora utilizados para a descrição sincrônica ou diacrônica do português brasileiro, tais como o C-ORAL-BRASIL I[1] (RASO; MELLO, 2012), que compilou dados de fala em contextos informais, e o *Corpus* Brasileiro[2], que tem se valido de diferentes modalidades. Na esteira do crescente desenvolvimento das teorias *corpus-driven*, isto é, induzidas através do estudo de corpora, há também o Processamento de Língua Natural, doravante PLN, cujas técnicas são recorrentemente aplicadas em programas de reconhecimento de voz, tradução automática, extração de informações relevantes, dentre outras aplicações presentes em nosso dia a dia. Importantes reflexões históricas sobre esse assunto são abordadas por Berber Sardinha (2000), que salienta que a descrição de um *corpus* está ligada à representatividade e, portanto, ao caráter probabilístico do uso linguístico.

A relação entre a Linguística de *Corpus* e o PLN foi delineada por Finatto, Lopes e Silva (2015), que sustentou a utilidade da cooperação entre as duas áreas dos saberes, explorando a noção do conexionismo, presente em versões recentes da Inteligência Artificial (I.A.), e que sugere a possibilidade de integração entre diferentes níveis de análise. Após processarem dois corpora de textos científicos de áreas diferentes – de um lado, textos sobre Pneumopatias Ocupacionais, e, de outro, o Curso de Linguística Geral, de F. de Saussure – o objetivo dos autores foi de caracterizar cada uma das obras, formalizando suas diferenças a partir dos resultados obtidos na fase de processamento.

É precisamente no centro desses caminhos abertos pelas novas tecnologias que as pesquisas em Literatura têm as possibilidades de se reinventar, aliando a utilidade dos processamentos computacionais dos textos em línguas naturais às críticas e às análises literárias, já consagradas nos estudos das Artes. Nesse sentido, para além de uma possível aversão ao racionalismo extremo e à linguagem limitada das máquinas, consideramos, para este trabalho, com base nas vantagens já abordadas por Finatto, Lopes e Silva (2015), que as ferramentas introduzidas pelas técnicas de PLN e pela Linguística de *Corpus* poderão, se usadas corretamente, ser importantes para os avanços nos Estudos Literários, garantindo a análise crítica e o respaldo da natureza complexa do campo das Humanidades. Neste mesmo contexto, curiosamente, muito se fala da real interdisciplinaridade entre as chamadas Ciências Humanas e as Ciências Exatas, que, até então, é pouco explorada no ambiente escolar e acadêmico brasileiro, ficando a cargo da simples menção em documentos oficiais e de tímidas abordagens por parte dos grandes vestibulares. Vislumbramos, por isso, mostrar a possibilidade de uma aproximação verdadeira entre os dois campos, garantindo a formalização e a reflexão crítica a respeito dos objetos gerados pelo contato entre eles. Em proposta parecida, Kauffmann (2020), em sua tese – de posse dos recursos da LC – apresenta uma análise multidimensional do *Corpus* Literário de Machado de Assis (CLIMA) e do *Corpus* Literário Congênere (CLIC), entrecruzando os dados linguísticos com discussões estilísticas machadianas a partir da frequência de palavras, lemas e suas coocorrências.

A obra *Ave, Palavra*, de 1970, reúne uma coletânea de textos diversos de João Guimarães Rosa, um dos maiores expoentes do modernismo literário brasileiro, movimento que completa 100 anos em 2022. Publicado após a morte do escritor, o livro em questão é caracterizado pelos próprios paratextos editoriais como uma obra miscelânica e, mesmo no campo acadêmico, possui menos visibilidade do que Grande Sertão: Veredas e Primeiras Estórias, escritos majoritariamente em prosa. Assim, poucos foram os trabalhos publicados a respeito de *Ave, Palavra* e de seu lugar na literatura brasileira. A saber, temos a dissertação de Rossi (2007), que aponta as qualidades principais das poesias da coletânea e os artigos de Brito (2012) e Vital (2021), que analisam a criação de máscaras anagramáticas nos poemas da obra. Nos textos, Rosa (1985), num movimento que se aproxima das heteronímias de Fernando Pessoa, cria autores fictícios, cujos nomes são diferentes combinações das letras do seu próprio nome. Assim, em *Ave, Palavra*, somos apresentados a Soares Guiamar, Meuriss Aragão, Sá Araújo Ségrim e Romaguari Sães, os autores-anagrama, aos quais Guimarães Rosa atribui uma coletânea de 26 poemas, seguidos de breves considerações a respeito das características de cada autor-anagrama. Sobre esse aspecto, Brito (2012) salienta que a escolha de manter as letras do nome do autor empírico reforça sua identidade – ainda que de forma embaralhada.

Com vistas a um movimento de prospecção inicial para a aplicação de recursos computacionais

1 http://www.c-oral-brasil.org/
2 http://corpusbrasileiro.pucsp.br/cb/Acesso.html

nos textos literários, nosso principal objetivo neste trabalho é fazer uma análise de cunho quantitativo e qualitativo de poemas ainda pouco conhecidos de Guimarães Rosa sob o ferramental metodológico da Linguística de *Corpus*, evidenciado as marcas textuais que podem ser utilizadas nos poemas. Como objetivos complementares, pretendemos (i) identificar cada um dos anagramáticos, através das palavras mais frequentes, número de versos e estrofes, e (ii) tecer possíveis discussões linguístico-literárias com as descrições de cada um, dadas pelo próprio autor empírico. Além disso, o *script* criado para o processamento do *corpus* gera nuvens de palavras, exibindo graficamente os resultados qualitativos da análise do *corpus*.

2 Metodologia

Para formação do *corpus*, foram transcritos os poemas de cada um dos anagramáticos, na ordem em que eles foram publicados em *Ave, Palavra*, para arquivos .txt, compatíveis com a linguagem de programação Python, através da qual foi desenvolvido o *script*. Para contabilização posterior, ao final de cada estrofe, foram adicionadas "//". Essa foi a única anotação adicionada durante o processo de transcrição. Foi criado um arquivo de texto para cada anagramático, nomeados com o nome do poeta e as páginas do livro em que se encontram as poesias (no seguinte padrão: Nome_Sobrenome_vx-yz), além de um outro arquivo com todos os textos compilados. O *corpus* está disponível para *download* neste *link*: https://bityli.com/wOGlv. A seguir, temos um trecho dos textos do *corpus*, como exemplo do poeta anagramático Meuriss Aragão, cujas poesias se encontram entras as páginas 90 e 92 da edição de 1985, da Editora Nova Fronteira:

Exemplo 1 (Meuriss_Aragao_90-92)

ele entranha e em torno e erra
o milagre monótono
//
intacto em colméias;
nem e sempre outro adeus
me não-usa, gasta o
fim não fim:
repete antecipadamente
meu único momento?

Além disso, percebemos que os poemas de Sá Araújo Ségrim se encontram organizados em dois momentos diferentes do livro, com uma coletânea entras as páginas 112-114, e outra entre páginas 184-186. Pelo fato de nossa análise considerar uma classificação com base em cada anagramático, optamos por unir, em um único arquivo, todos os poemas atribuídos a Ségrim, mesmo que, em alguma medida, haja diferenças pontuais entre suas duas aparições poéticas. Esta decisão foi tomada considerando, portanto, critérios puramente metodológicos.

Em seguida, os arquivos foram processados pelo *script*, que contabilizou os seguintes parâmetros para cada um dos anagramáticos:

a. Total de palavras (tokens);
b. Total de palavras sem repetição (types);
c. Total de palavras (após a retirada de *stopwords*);
d. Número de estrofes;
e. Número de versos escritos;
f. Dez palavras mais frequentes.

Como a transcrição dos poemas para os arquivos .txt não levou em consideração nenhuma alteração linguística significativa, como retirada de pontuações, normalização de letras maiúsculas e etc., foi necessário realizar o pré-processamento automático dos dados do *corpus*, uma vez que esses elementos poderiam interferir na contagem total de palavras e caracteres. Portanto, no pré-processamento, foram retiradas todas as pontuações, e os caracteres foram transformados em minúsculos, garantindo que palavras com iniciais maiúsculas (como "Alma") ou escritas em caixa alta (como "ALMA") não fossem contabilizadas como entidades diferentes.

Após esses comandos, foi contabilizado o número total de palavras para cada anagramático. Em seguida, foram retiradas as chamadas *stopwords*, isto é, palavras funcionais e pouco informativas que compõem boa parte dos enunciados das línguas humanas. Para isso, foi utilizada a lista padrão de *stopwords* para o português da biblioteca *Natural Language Toolkit*, NLTK (BIRD; KLEIN; LOPER, 2009).

Além da biblioteca NLTK, foi utilizada a biblioteca Pandas (MCKINNEY, 2010) para geração das nuvens de palavras. O *script* foi feito por meio da interface do Google Colaboratory, ambiente virtual para programação em Python e que dispensa muitas das instalações prévias, estando disponível neste *link*: https://bityli.com/rYEOS.

Como métricas utilizadas, foram calculados os números brutos de palavras por anagramáticos, número de estrofes e de versos. Para o número de estrofes, em virtude da diferença de tamanho e do número e de textos para cada autor, optamos por fazer uma normalização em relação ao número de palavras. Nesse sentido, evitando vieses, os números de estrofes foram calculados em relação ao número de palavras para cada autor-anagrama, como é comum em trabalhos de Linguística de *Corpus*.

Para a análise de dados, nos baseamos nos referenciais propostos nos trabalhos de Rossi (2007), Finatto, Lopes e Silva (2015), Berber Sardinha (2000) e Vital (2021). Uma parte da proposta metodológica descrita no trabalho também está presente na obra *Text Analysis with R for Students of Literature*, de Jockers (2014), que, apesar se apresentar como uma introdução a métodos computacionais para o tratamento de textos literários através da linguagem R, possui seções metodológicas próximas daquelas usadas em linguagem Python. Ao final, foram tecidas breves considerações a respeito das palavras mais frequentes e de seu papel na composição dos poemas, com o objetivo de demonstrar a possibilidade de se conciliar os dados de processamento com discussões linguístico-literárias e com as epígrafes presentes em cada capítulo de apresentação dos autores fictícios.

3 Resultados

Após o processamento de cada um dos arquivos de texto, pudemos ter uma noção do tamanho do *corpus*, que conta, ao todo, com 1.596 palavras e 746 types, distribuídas entre os 26 poemas, segundo a Tabela 1. A razão type-token, que relaciona a quantidade de tipos de palavras em relação ao número total de palavras no *corpus* é de 0,4674. Os autores-anagrama Soares Guiamar e Sá Araújo Ségrim possuem, juntos, um total de palavras correspondente a cerca de 69,6% do *corpus*, enquanto que Meuriss Aragão e Romaguari Sães contam, respectivamente, com 13,4% e 17%. Tais resultados podem ser visualizados na Figura 1, que mostra a quantidade de palavras com e sem as *stopwords* para cada um dos anagramáticos. Nessa mesma figura, é possível que percebamos com clareza a relevância quantitativa das *stopwords* para a língua em uso, reforçando a necessidade do pré-processamento para, em seguida, a análise das palavras mais frequentes.

Tabela 1. Dados após o processamento do *corpus*.

	Total de Palavras (com *stopwords*)	Total de Palavras (sem *stopwords*)	N° de Estrofes	N° de Versos
Soares Guiamar	527	306	31	138
Meuriss Aragão	214	137	7	78
Sá Araújo Ségrim	583	333	16	152
Romaguari Sães	272	171	24	84
TOTAL	1596	947	78	452

Fonte: Elaborada pelo autor.

Na Tabela 1, com exceção dos termos mais frequentes, analisados a seguir, através das nuvens de palavras, encontramos todas as informações relativas ao processamento do *corpus*.

Embora Sá Araújo Ségrim seja o anagramático com o maior número de palavras – e, portanto, o escritor da obra mais longa dentre os quatro – a avaliação do número de estrofes demonstrou que, após Soares Guiamar, com 31 estrofes, Romaguari Sães foi o poeta que mais se destacou, tendo escrito

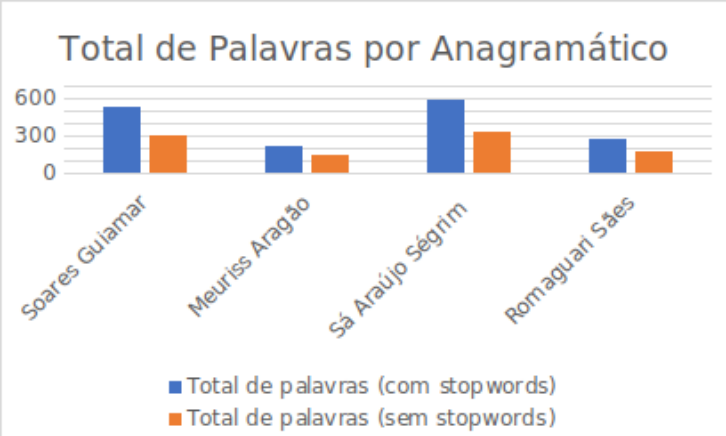

Figura 1. Total de palavras com e sem *stopwords*.
Fonte: elaborada pelo autor.

24 estrofes com apenas 272 palavras, conforme a Figura 2, que mostra a relação entre o número de estrofes e o número de palavras, sugerindo uma marca estilística única dentre os anagramáticos.

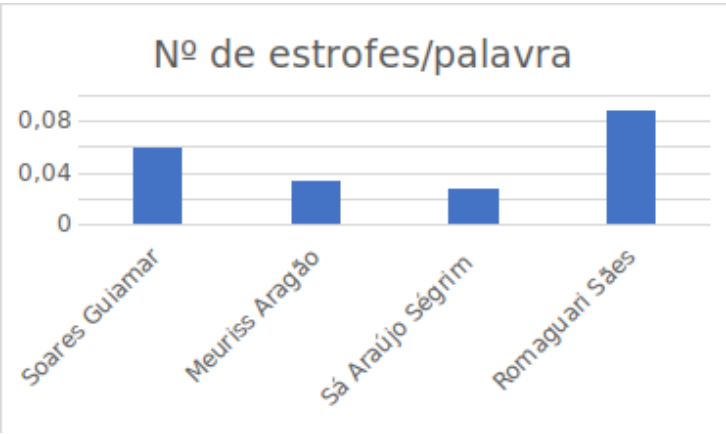

Figura 2. Gráfico com o número de estrofes por anagramático.
Fonte: elaborada pelo autor.

Curiosamente, com essa relação entre o número de estrofes e palavras, é possível percebermos uma particularidade formal dos escritos de Romaguari, já que seus poemas são construídos por um grande número de estrofes, mas de tamanhos reduzidos, fazendo jus ao número de palavras desse autor-anagrama. Essa característica pode ser facilmente observada no poema "Cândida", transcrito abaixo de Rosa (1985, p. 235), composto por estrofes de dois versos e um esquema métrico de sete sílabas poéticas:

> CÂNDIDA
> (Marjolininha)
>
> Candinha sonha comigo
> no sonho sou seu amigo.
>
> Eu que nunca vi Candinha
> Reconheço-a na poesia.
>
> Sonho que Candinha dorme
> sonho que Candinha sonha
>
> neste mundo certo e enorme
> nesta vida não tristonha.
>
> Candinha sonha um abrigo
> no futuro - no conforme.

Que da simples alegria
o seu sonho se componha.

Candinha? Um sonho se sonha.

Conforme vimos no gráfico da Figura 1, Sá Araújo Ségrim é o autor com o maior número de palavras, além de também apresentar um modelo formal característico, com menos divisão estrófica (Figura 2). Um exemplo representativo da estética de Ségrim pode ser o poema "Distância" (ROSA, 1985, p. 112), já explorado numa análise de Rossi (2007), e com duas de suas estrofes transcritas a seguir. Como podemos ver, há uma diferença clara entre o estilo de Romaguari e o de Ségrim, denotada não apenas nos assuntos de que tratam os poetas, mas, sobretudo, na relação entre o número de palavras e o número de estrofes, detendo, este último, estrofes densas e sem esquema métrico definido, e, aquele, estrofes menores em número de palavras e estrato fônico marcado, nos termos de Ramos (2011). Soares Guiamar e Meuriss Aragão, por outro lado, mantém a quantidade de estrofes proporcional à quantidade de palavras escritas. Quanto ao primeiro, análises como a de Vital (2021) e a de Rossi (2007) denotam que sua poesia, mesmo que não seja metricamente articulada, esconde importantes referências metapoéticas e exploração do estrato óptico do poema, isto é, a maneira como as palavras se distribuem na página.

DISTÂNCIA

Um cavaleiro e um cachorro
viajam para a paisagem.
Conseguiram que esse morro
não lhes barrasse a passagem.
Conseguiram um riacho
com seus goles, com sua margem.
Conseguiram boa sede.
Constataram:
cai a tarde.

Sobre a tarde, cai a noite,
sobre a noite a madrugada.
Imagino o cavaleiro
esta orvalhada e estrelada.
O pensar do cavaleiro
talvez o amar, ou nem nada.
Imagino o cachorrinho
imaginário na estrada.
Caía a tarde.

Como recurso a ser utilizado em análises qualitativas dos assuntos recorrentes nos poemas dos anagramáticos, propusemos a criação de nuvens de palavras (*wordclouds*), gráficos que possuem o objetivo de representar visualmente, com o auxílio de diferentes tamanhos de fontes e cores, a frequência de palavras em determinado conjunto de textos (JAFAR; BABB; DANA, 2012). Para cada um dos autores, foi criada uma nuvem de palavras, procurando sintetizar os temas recorrentes em seus estilos poéticos (Figura 3, Figura 4, Figura 5, Figura 6). No Apêndice A, há a lista das 10 palavras mais frequentes para cada anagramático.

Após o processamento de todos os arquivos, chegamos às seguintes palavras mais frequentes nas obras: "onde", "vida", "mim", "sempre", "tarde", "rio", "mar", "triste". Neste ponto, salientamos que, sem os recursos computacionais da Linguística de *Corpus* e suas interfaces com os Estudos Literários, tais resultados se tornariam exponencialmente trabalhosos, uma vez que a análise das palavras se dá *token* a *token*, isto é, uma a uma, e é realizada automaticamente pelo algoritmo criado. Na seção seguinte, são apontados breves caminhos de análise que conciliam a frequência das palavras com as epígrafes de cada autor-anagrama.

Figura 3. Nuvem de palavras dos poemas de Soares Guiamar.
Fonte: elaborada pelo autor.

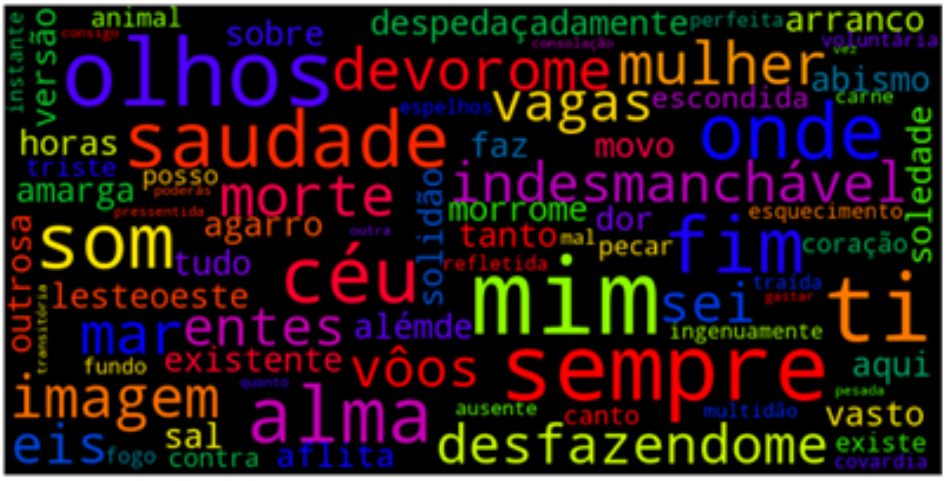

Figura 4. Nuvem de palavras dos poemas de Meuriss Aragão.
Fonte: elaborada pelo autor.

Figura 5. Nuvem de palavras dos poemas de Sá Araújo Ségrim.
Fonte: elaborada pelo autor.

4 Discussão linguístico-literária

Como dissemos acima, cada anagramático, em *Ave, Palavra*, é introduzido por uma epígrafe que, em muitos casos, dialoga com o leitor a respeito do autor e dos poemas que se encontrarão nas

Figura 6. Nuvem de palavras dos poemas de Romaguari Sães.
Fonte: elaborada pelo autor.

próximas páginas. Essa pequena introdução se repete a cada momento em que Rosa nos apresenta a novos autores-anagrama. Em Brito (2012), há a informação de que os poemas da obra de 1970 foram publicados, originalmente, em edições do jornal O Globo, em 1961, o que, ao nosso ver, nos proporciona uma explicação a respeito da natureza das epígrafes, uma vez que o autor empírico, como que apresentando suas máscaras anagramáticas aos leitores d'O Globo, adiciona breves parágrafos sobre informações de seus "poetas de bolso". A título de exemplo, a epígrafe de apresentação de Soares Guiamar, o primeiro dos anagramáticos, está transcrita, na íntegra, a seguir, e, num gesto ligeiro, acaba também por apresentar a ideia dos anagramas: "De Soares Guiamar – despercebido, impresso, inédito, fora de moda – que queria livro, o "Anagramas", e disse palpites: Ser poeta é já estar em experimentada sorte de velhice. Toda poesia é também uma espécie de pedido de perdão" (ROSA, 1985, p. 58).

Como vimos, os poemas de Guiamar representam boa parte da obra anagramática de Guimarães Rosa, uma vez que ele é o segundo poeta com o maior número de palavras e o primeiro no número de estrofes. Na sequência da obra, temos Meuriss Aragão, que é introduzido por uma epígrafe curiosa, pois é nela que Rosa relembra Soares Guiamar, restando poucas palavras para caracterizar Aragão:

> Perguntam-me por mais versos de Soares Guiamar. Não são possíveis. Ele agora para **longe**, certo à beira do **Riachinho Sirimim**, **lugar** de se querer **bem**. Tenho, porém, outro poeta de bolso: Meuriss Aragão. Jovem, sem jeito, em sua primeira fase, provavelmente extinta. Vejam, se serve. (ROSA, 1985, p. 90, grifos nossos).

Nessa apresentação, por sua vez, nos chamou a atenção a existência de palavras que se aproximam ou, em alguns casos, fazem parte daquelas mais frequentes na estética do primeiro autor-anagrama. Exemplos como "bem", "lugar", "Riachinho Sirimim" e "longe" dialogam diretamente com os termos mais frequentes resultados do processamento de seus poemas, que conta com termos como "bem", "onde", "mar" e "rio" como sendo bons representantes das 10 palavras mais frequentes (ver Figura 3). Ao mesmo tempo, Guimarães reserva pouco espaço para a entrada do próximo anagramático, Meuriss Aragão, que, inclusive, possui o menor número de palavras e estrofes de todos os quatro.

Neste ponto, caso não desejemos enquadrar as explicações anteriores apenas no terreno das coincidências literárias, é absolutamente relevante considerarmos as epígrafes e os trechos paratextuais como, juntamente dos poemas, veiculadores do sentido das obras. Revelar estes interstícios da forma literária, neste caso, só foi possível com a contagem de palavras e estrofes, e, desse modo, com os resultados do PLN.

Da mesma forma, Rosa nos apresenta o terceiro dos autores-anagrama: Sá Araújo Ségrim, com a seguinte epígrafe:

> Poeta comprido – é outro dos anagramáticos, de que hoje disponho. Se bem talvez um

> tanto discípulo de Soares Guiamar, sob leves aspectos, sofre só e sozinho verseja. Sei que pensa em breve publicar livro: o 'Segredeiro', e do supracitado é, às vezes, o que prefiro. Será que conosco concordam? (ROSA, 1985, p. 112).

Em primeiro lugar, nos chama a atenção o adjetivo "comprido", utilizado para qualificar o poeta em questão. Tal qualidade é respaldada, por vezes, pelos resultados de nosso processamento, já que os poemas de Ségrim representam o maior número de palavras, sendo, dos quatro, o poeta mais comprido: 583 das palavras totais, isto é, 36,5% do *corpus*. Nessa mesma linha, a epígrafe salienta que Sá Araújo Ségrim é discípulo de Guiamar, outro dos anagramáticos, o qual possui o segundo maior número de palavras escritas: 527. Esta relação mestre-discípulo, além de se mostrar através da similaridade estrutural entre os dois autores – como é possível percebermos no número de palavras – há, na poesia de ambos, alta frequência dos termos "vida" e "rio", que estão entre as 10 palavras mais frequentes para os dois anagramáticos.

Num segundo momento, há a apresentação do segundo conjunto de poemas atribuídos a Ségrim. Na abertura, Rosa (1985) faz as seguintes explicações a respeito da volta do anagramático, que, como sabemos, é seu preferido:

> Se lhe não firo a modéstia, direi, aqui, depressa, que Sá Araújo Ségrim, em geral, agradou. Por isso mesmo, volta, hoje, com novos poemas, que só não sei se escolhemos bem. Sendo coisas mui sentidas. Sendo o que ele não sabe da vida. Digam-me, o mais, amanhã. Leiam-no, porém (ROSA, 1985, p. 184).

Segundo Rossi (2007), os poemas de Ségrim comportam uma grande variedade tanto no nível formal quanto semântico. Nesse sentido, a divisão de suas poesias em dois agrupamentos situados em partes diferentes do livro fortalece a ideia da diversidade.

O último anagramático é Romaguari Sães, que foi publicado apenas em *Ave, Palavra*, não tendo feito parte das publicações em jornal, como foi o caso dos outros autores-anagrama. De acordo com Rossi (2007), Sães é o autor que mais se diferencia dos outros, impondo um estilo único e que se coloca em contato com trovas e cantigas medievais. Reforçando este ponto, a autora acrescenta que a escolha de palavras que se aproximam do campo semântico dessas produções, como "amigo", "prado" e "bailai", sugerem ainda mais a relação com a cultura medieval. Em sua epígrafe, lemos os seguintes trechos:

> Outro anagramático é Romaguari Sães, o 'embevecido', escondedor de poemas. No grupo, é considerado como um tanto diferente. Tem outra música. Tem um amor mais leve, originário, avançado. Disse, uma vez, em entrevista, que a poesia devia ser um meio de 'restituir o mundo ao seu estado de fluidez, anterior, exempta'. Aprovam-no? (ROSA, 1985, p. 184).

Como já apresentamos na seção de resultados, Romaguari conta com uma construção estrófica única em relação aos seus pares, produzindo um grande número de estrofes pequenas, em geral, metrificadas. Dentre as 10 palavras mais frequentes, citamos "prado", também indicada por Rossi (2007) como um termo que aproxima sua estética da poesia medieval.

Com isso, complementando os trabalhos a respeito das infindáveis obras de Guimarães Rosa, temos um panorama inicial da cooperação que alinha a Linguística de *Corpus*, o PLN e o campo dos Estudos Literários, de modo a corroborar hipóteses interpretativas e lançar luz a recursos ainda pouco explorados na análise de obras literárias. Não obstante, a comparação realizada entre as máscaras anagramáticas de Guimarães Rosa se materializou através de tabela, gráficos e nuvem de palavras, recursos que poderão ser utilizados para estudos futuros do *corpus* poético de *Ave, Palavra*. Na tentativa de apontar os rumos interdisciplinares do contato inteligente entre as Ciências Exatas e as Humanidades, o presente estudo serve de base teórico-metodológica para o tratamento de textos literários, inclusive, em sala de aula, despontando interesses a partir de diversas disciplinas, como a Língua Portuguesa, a Literatura, a Programação e os conhecimentos lógicos. Nesse sentido, como denotado por Finatto, Lopes e Silva (2015), os conhecimentos empregados na compilação, anotação

e processamento de dados de corpora podem ser complementados por estas áreas, além da associação acertada entre Linguística de *Corpus* e PLN.

Em relação aos poemas processados, é de se considerar, ainda mais, o rigor (meta)poético empregado, a complexidade das características associadas a cada autor-anagrama e seus correspondentes formais, tanto na elaboração dos poemas quanto na criação das epígrafes. Tais fatos caminham na direção já discutida por Rossi (2007) e Vital (2021), que reconhecem a criação de máscaras anagramáticas na literatura rosiana como um objeto rico – embora pouco discutido pela crítica – mas que merece ser estudado em profundidade.

Referências

BERBER SARDINHA, T. Lingüística de Corpus: histórico e problemática. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, v. 16, p. 323–367, 2000. ISSN 0102-4450, 1678-460X. DOI: 10.1590/S0102-44502000000200005. Disponível em: http://www.scielo.br/j/delta/a/vGknQkZQGsGYbrQfKmTZY4s/?lang=pt. Acesso em: 11 set. 2022.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. 1ª edição. Beijing ; Cambridge Mass.: O'Reilly Media, jun. 2009.

BRITO, F. M. M. de. Aspectos metaficcionais na poética de Rosa e Pessoa: o artifício das máscaras heteronímicas e anagramáticas. *Letras de Hoje*, v. 47, n. 4, p. 425–429, dez. 2012. ISSN 1984-7726. Disponível em: https://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/view/10619. Acesso em: 11 set. 2022.

FINATTO, M. J. B.; LOPES, L.; SILVA, A. C. Processamento de linguagem natural, linguística de corpus e estudos linguísticos : uma parceria bem-sucedida. *Domínios de Linguagem*, v. 9, n. 5, p. 41–59, 2015. ISSN 1980-5799. Disponível em: https://lume.ufrgs.br/handle/10183/169398. Acesso em: 11 set. 2022.

JAFAR, M.; BABB, J.; DANA, K. A Framework for an Interactive Word-Cloud Approach for Visual Analysis of Digital Text using the Natural Language Toolkit. In: PROCEEDINGS of the Conference on Information Systems Applied Research. New Orleans: [s.n.], 2012. v. 5, n. 2240, p. 1–10.

JOCKERS, M. L. *Text Analysis with R for Students of Literature*. Switzerland: Springer, 2014. Disponível em: https://link.springer.com/book/10.1007/978-3-319-03164-4. Acesso em: 11 set. 2022.

KAUFFMANN, C. H. *Linguística de corpus e estilo: análises multidimensional e canônica na ficção de Machado de Assis*. 2020. Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem) – Pontifícia Universidade Católica de São Paulo, São Paulo.

MCKINNEY, W. Data structures for statistical computing in python. In: PROCEEDINGS of the 9th Python in Science Conference. [S.l.: s.n.], 2010. p. 56–61. Disponível em: https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf. Acesso em: 11 set. 2022.

RAMOS, M. L. *Fenomenologia da Obra Literária*. 4ª edição. Belo Horizonte: Editora UFMG, jan. 2011.

RASO, T.; MELLO, H. (Ed.). *C-oral - Brasil I: Corpus de Referência do Português Brasileiro Falado Informal*. 1ª edição. Belo Horizonte: Editora UFMG, jan. 2012.

ROSA, J. G. *Ave, Palavra*. Rio de Janeiro, RJ: Nova Fronteira, 1985.

ROSSI, E. A. *As poesias de Guimarães Rosa em Ave, Palavra: um caminho de leitura*. 2007. Dissertação (Mestrado em Estudos Literários) – Universidade Estadual Paulista, Faculdade de Ciências e Letras de Araraquara, Araraquara, SP.

VITAL, A. A. S. Uma possibilidade de leitura do poema "Alongo-me", de Soares Guiamar, anagramático de Guimarães Rosa. *Miguilim – Revista Eletrônica do Netlli*, v. 10, n. 3, p. 980–989, 2021.

A Apêndice

Tabela 2. Ranking de palavras significativas mais frequentes para cada autor-anagrama.

Ranking	Soares Guiamar	Meuriss Aragão	Sá Araújo Ségrim	Romaguari Sães	Corpus
1	bem	mim	tarde	candinha	onde
2	vida	olhos	cavaleiro	maria	vida
3	três	ti	noite	quero	mim
4	moça	sempre	cachorro	sonho	sempre
5	moço	som	conseguiram	meninas	tarde
6	rio	céu	vida	sonha	rio
7	então	onde	sol	prado	mar
8	mar	saudade	rio	majolininha	triste
9	onde	alma	distância	quanto	quero
10	triste	fim	morro	vou	candinha

Fonte: Elaborada pelo autor.