UNIVERSIDADE FEREDAL DE SANTA CATARINA
CENTRO DE COMUNICAÇÃO E EXPRESSÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM LITERATURA

Verônica Ribas Cúrcio

Palavras de Rosa: análise estilométrica da obra de João Guimarães Rosa

Tese

Florianópolis
2013

Verônica Ribas Cúrcio

Palavras de Rosa: análise estilométrica da obra de João Guimarães Rosa

Tese apresentada ao Programa de Pós-Graduação em Literatura da Universidade Federal de Santa Catarina como requisito para a obtenção do título de doutora em Teoria Literária.

Orientador: Prof. Dr. Alckmar Luiz dos Santos

Florianópolis
2013

Dedico este trabalho ao meu amigo poliglota Jessé Gabriel da Silva (*in memoriam*) e à minha mãe Vera Lúcia, professora dedicada que me ensinou a ler e a fazer continhas.

AGRADECIMENTOS

Agradeço a Deus. A todo o povo brasileiro que, por sua contribuição em impostos, concedeu-me a oportunidade de estudar e pesquisar em uma universidade pública e ser contemplada com bolsa de estudos durante toda a minha formação acadêmica.

À compreensão de minha família, por jamais duvidar de minhas vontades. Aos que de alguma forma contribuíram para que este trabalho ganhasse fôlego: professor, mestre e amigo Alckmar Luiz dos Santos, colegas e amigos que participam ativamente do NUPILL.

Ao programa de pós-graduação de Literatura da UFSC, em especial à professora Tânia Regina de Oliveira Ramos pelo apoio de sempre e à funcionária e amiga Elba Ribeiro que muito me incentivou para o estágio-sanduíche.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo financiamento e pela oportunidade de estágio de doutoramento no laboratório (BCL), da Universidade de Nice *Bases, Corpus, Langage* (França), cuja troca de conhecimentos favoreceu muito o desenvolvimento deste trabalho.

RESUMO

Nesta tese analisamos a obra literária de João Guimarães Rosa a partir da estilometria, com o objetivo de buscar as características do estilo rosiano que sejam possíveis de detectar por meio de ferramentas informatizadas de análise estatística. Nossa tese é verificar se esse tipo de estudo permite confirmar ou complementar intuições derivadas de leituras convencionais e, além disso, oferecer novos elementos textuais e estilísticos que nem sempre estão explícitos no texto. Para isso, privilegiaremos três estudos rosianos: *Signo e sentimento* (SPERBER, 1982) sobre a organização da linguagem de Rosa; *O insólito em Guimarães Rosa e Borges* (COVIZZI,1978), a qual propõe uma linha cronológica de expressão e explicação da obra rosiana; *João Guimarães Rosa: travessia literária*, (DANIEL, 1968), que afirma haver uma separação do léxico rosiano em duas fases: uma rural e outra urbana. A ferramenta estatística adotada foi o programa *Hyperbase*, de Étienne Brunet. Para os procedimentos de análises, trabalhamos com duas bases, uma com a cronologia de produção de escrita das obras e outra seguindo a cronologia de primeira publicação das obras, que foram respeitadas a fim de viabilizar a verificação da evolução e do crescimento do vocabulário do escritor. Levantamos muitas características do léxico rosiano, dentre elas verificamos que mais da metade de seu vocabulário não se repete; e que as obras de caráter sertanejo apresentam vocabulário menos diversificado. Por fim, veremos como Guimarães Rosa, ao final de sua carreira literária, tratou mais do seu material linguístico.

Palavras-chave: Guimarães Rosa. Estilometria. Estatística textual.

RÉSUMÉ

Cette thèse s'occupera de l'œuvre littéraire de João Guimarães Rosa à partir de la stylométrie, ayant pour objectif la recherche de caractéristiques du style de l'auteur qui puissent être identifiées par moyen d'outils informatisés d'analyse statistique. Notre thèse consiste à verifier si ce type d'étude permet de confirmer ou d'enrichir des intuitions résultantes de lectures conventionnelles et aussi à offrir de nouveaux éléments textuels et stylistiques qui ne sont pas toujours explicites dans le texte. Pour cela, on privilégiera trois études sur l'écrivain : *Signo e sentimento* (SPERBER, 1982), texte sur l'organisation du langage de Rosa ; *O insólito em Guimarães Rosa e Borges* (COVIZZI, 1978), qui propose une frise chronologique de l'expression et explication de l'œuvre de Rosa ; *João Guimarães Rosa, travessia literária*, (DANIEL, 1968), selon laquelle il y a deux phases concernant le lexique de l'écrivain : l'une rurale et l'autre urbaine. L'outil statistique adopté a été le logiciel *Hyperbase*, d'Étienne Brunet. Pour le procédés d'analyse, on a travaillé avec deux bases, à savoir, l'une qui contenait la chronologie de production d'écriture des œuvres ; l'autre, la chronologie de la première publication des œuvres, considérée dans ce travail afin de rendre possible une investigation effective de l'évolution et de l'expansion du lexique de l'auteur. Parmi la grande quantité de donnéés recuillies, on a pu constater que plus de la moitié de son lexique ne se répète pas. En outre, les œuvres "sertanejas" présentent moins de variations lexicales. Pour conclure, on démontrera comment Guimarães Rosa, à la fin de sa vie, s'est occupé davantage de son matériel linguistique.

Mots clés: Guimarães Rosa. Stylométrie. Statistique textuel.

SUMÁRIO

1 INÍCIO DA TRAVESSIA

Podemos ainda estudar a obra de João Guimarães Rosa (Cordisburgo, MG, 27/06/1908 – Rio de Janeiro, RJ, 19/11/1967) por meio de seu vocabulário e trazer algo de novo? Escrever sobre Guimarães Rosa não é tarefa fácil, a começar pela quantidade de críticas[1] já feitas sobre seus textos, incluídas aí reflexões filosóficas, abordagens de cunhos sociológico e linguístico, mapeamentos sociogeográficos, enfim, estudos que exploram possibilidades muito heterogêneas de leitura da obra rosiana.

A esse respeito, o pesquisador Willi Bolle comenta a quantidade de estudos sobre o texto mais estudado de Rosa, *Grande sertão: veredas*, e acrescenta:

> A fortuna crítica do romance, que já acumula mais de 1.500 títulos, confirma uma observação de Joseph Maistre sobre a recepção em geral: dois ou três críticos fixam inicialmente a opinião, e a maioria dos que vêm depois segue por essas mesmas trilhas. Assim, as marcas dos ensaios pioneiros, ambos publicados em 1957, respectivamente, por Antonio Candido [com o ensaio *O sertão e o mundo*] e M. Cavalcanti Proença [*Alguns aspectos formais de Grande sertão: veredas*]. (BOLLE, 2004, p. 19).

Partindo desses dois mestres, Bolle identifica três modos de abordagem metodológica do texto rosiano, com seus respectivos estudiosos:

> 1. Os estudos linguísticos e estilísticos, como os de Mary Lou Daniel (1968) e Teresinha Souto Ward (1984) [...] Nei Leandro de Castro (1970) e Nilce Sant'Anna Martins (2001). [...]
> 2. As análises de estrutura, composição e gênero, como as de Roberto Schwarz (1965a e 1965b), Eduardo Coutinho (1980, 1983, 1991 e 1993), Benedito Nunes (1985), Rosemary Arrojo (1985) e Davi Arrigucci Jr. (1994) [...]

[1] Buscamos pela combinação "Guimarães+Rosa" no banco de teses e dissertações da CAPES, e obtivemos 578 resultados relacionados à pesquisa nas áreas de Linguística, Letras e Artes. A pesquisa foi feita em 07 jun. 2012.

> 3. A crítica genética[2], com contribuições de Maria Célia Leonel (1985 e 1990), Lenira Covizzi e Maria Neuma Cavalcante (1990), Walnice Galvão (1990), Edna Maria dos Santos Nascimento (1990), Elizabeth Hazin (1991 e 2000), Cecília de Lara (1993, 1995 e 1998) e Ana Luiza Martins Costa (1997-8 e 2002) [...] (BOLLE, 2004, 19-20).

Podemos observar que das autoras que nos apoiamos para esta tese, Daniel (1968) se encontra na abordagem número 1 e Covizzi (1978) na abordagem 3, dessa forma, podemos afirmar que nossa pesquisa está localizada entre os estudos estilísticos e genéticos da obra rosiana.

Paradoxalmente, essa quantidade de pesquisas dificulta e facilita os atuais estudos. Dificulta, pela tarefa de mapear elementos ou perspectivas minimamente originais (pois essa é a maior dificuldade quando se trabalha com grandes autores); facilita, pelo acúmulo de elementos e características já repertoriados e que ajudam a guiar as leituras contemporâneas. Não deixar de referenciar mesmo as pesquisas mais relevantes impõe um exercício de escolha a todo momento, e tecer comentários sobre o léxico de Rosa requer o cuidado de não repetir o que até o momento já foi dito.

Por isso, a tarefa que propomos aqui, como caráter inédito, é a reunião abrangente e exaustiva do léxico de Guimarães Rosa, incluindo os resultados de desenvolvimento de vocabulário ao longo de sua produção literária e utilizando uma ferramenta informatizada. Faremos uma leitura cronológica dentro de análises da estatística de textos, conhecida também como estilometria, lexicometria e, ainda, textometria.

Um dos nossos objetivos é verificar qual o léxico básico, diferencial e preferencial de Rosa, além disso nossa tese é a de que esse tipo de análise permitirá confirmar intuições derivadas de leituras convencionais[3] (ou nas quais tais intuições se basearam), abrindo, a partir daí, um leque de novos elementos textuais que, nem sempre sendo reconhecíveis facilmente ou de maneira explícita no texto, ainda assim devem ser considerados como marcadores do estilo do escritor.

É claro que, quando falamos em *quantificar*, a análise *qualitativa* não pode nos escapar. Sendo assim, com a praticidade proporcionada pelas ferramentas informatizadas que servem de apoio para pesquisas na

---

2 Incluiríamos aqui os trabalhos de Suzi F. Sperber.

3 Vale salientar que aqui leitura convencional é aquela realizada tão somente entre o leitor e a obra, sem a interferência de alguma máquina.

área de Letras[4], realizaremos um estudo de estilo, a partir de um programa de tratamento estatístico de textos chamado *Hyperbase*[5], método que nunca foi antes aplicado à obra completa de Guimarães Rosa.

Segundo Andrew Olivier (1998, p. 480), os estudos lexicométricos estão fundados sobre a importância dos lexemas e tendem a uma leitura do texto em função das informações destes. Portanto, nossa leitura seguirá uma linha aproximada da metodologia pontuada como de número 1 na citação referenciada anteriormente por Bolle (2004), porém auxiliada pela matemática e pela nova tecnologia informatizada.

Ao estudar obras literárias empregando banco de dados textuais e hipertextuais, Nathalie Ferrand (1997, p. 15) afirma:

> [...] *l'hypertexte ou l'écriture non séquentielle sont des notions qui appartiennent depuis toujours à la littérature* […] *mais les technologies informatiques de mise en rapport ont le mérite de renouveler le regard sur l'objet littéraire.*

Esse mérito renovador, comentado por Ferrand, nos servirá de base para estudarmos e descrevermos, por meio do uso de ferramentas informatizadas, as características de contagem estatística sobre o vocabulário rosiano, bem como a análise de comportamento ao longo de sua produção ficcional.

Quando analisamos o comportamento do vocabulário de um escritor, a cronologia é fator imprescindível para tanto; por isso, há uma grande preocupação por parte do pesquisador em inserir os textos no programa de análise estatística sempre respeitando a ordem cronológica de produção da obra. Daí procurarmos, na medida do possível, o momento em que cada texto foi elaborado — o que não significa exatamente o momento da primeira publicação. Para o embasamento cronológico da produção da obra, utilizamos como parâmetro a tese da

---

[4] Tais ferramentas constituem ainda um método de pesquisa e de leitura pouco difundido no Brasil, no contexto literário. Ressaltamos que o trabalho aqui proposto, ao focar a prática investigativa por meio da estatística textual nesse âmbito, dá continuidade ao método de pesquisa feito anteriormente no período de nosso mestrado: *Sintaxe da frustração. Análises estatísticas de textos de Franz Kafka*. (CÚRCIO, 2007). Trata-se de um estudo que analisou o vocabulário kafkiano, na sua língua original, a partir de quatro obras: *O Processo*, *O Castelo*, *América ou O Desaparecido*, e *A Metamorfose*.

[5] O *Hyperbase* é um programa desenvolvido pelo professor Étienne Brunet e sua equipe de linguistas e programadores, no laboratório *Bases, Corpus, Langage* (BCL) da Universidade de Nice, na França.

antropóloga Ana Luiza Costa intitulada "Veredas de Viator" (2006), publicada nos *Cadernos de Literatura Brasileira*, do Instituto Moreira Salles.

De outro lado, como nossa tese trabalha com a ficção completa, seria exaustivo e inútil abordar igualmente toda a recepção da obra de Guimarães Rosa, em suas distintas vertentes. Escolhemos, coerentemente, apenas estudos que se direcionam ao vocabulário, aos estudos de criação lexical, enfim, às leituras críticas que investigam a obra no que diz respeito a seu material linguístico.

É assim que privilegiaremos, para nossa análise vocabular, três estudos rosianos. O primeiro deles é a pesquisa de Suzi Sperber (1982)[6], que estuda a produção literária da obra rosiana e foca o fenômeno da organização da linguagem pelo viés da prática da *Textkritik*. Nesse trabalho a pesquisadora verificou, pelos estudos de vocabulário, como, desde o primeiro rascunho até a versão definitiva, a obra completa foi criada, mapeando o amadurecimento da escrita do autor. Sobre sua pesquisa, explica a autora:

> [...] o método comparativo permitia a percepção clara de um detalhamento nos temas, da elaboração da linguagem, dentro de uma perspectiva diacrônica que permitiria mais tarde, depois desta primeira apreensão, uma visão sincrônica mais trabalhada. (SPERBER, 1982, p. 4).

Outra análise que escolhemos foi a de Lenira M. Covizzi (1978), cuja pesquisa propõe uma linha cronológica que se estende em dois períodos (inicialmente, de expressão, e posteriormente, de explicação) da obra de Guimarães Rosa. Do estudo de Covizzi, o que nos interessa mesmo como exercício de pesquisa é avaliar, por meio da estatística de texto, o percurso da criação ficcional, explorando dois caminhos propostos, de modo que possamos verificar quantitativamente se esse movimento acontece também no léxico.

Por fim, trabalharemos também com a hipótese de Mary Lou Daniel (1968), que afirma haver uma separação, em termos lexicais, de duas fases (rural e urbana) na obra do escritor brasileiro. Verificaremos se essa distinção entre o rural e o urbano se reflete no léxico. Para tanto,

---

6 SPERBER, Suzi F. *Guimarães Rosa*: *signo e sentimento*. São Paulo: Ática, 1982.

utilizaremos, aqui, uma ferramenta que mede a evolução e a distância lexicais[7] de um texto a outro.

A verificação informatizada que propomos para esses três estudos apenas demonstra alguns exercícios possíveis de levantamento estatístico de uma obra literária. A intenção deste estudo é, principalmente, apresentar as possibilidades de leitura que a ferramenta estatística oferece para análises literárias.

Vale ressaltar também que buscamos a estatística de textos literários não para dar o veredicto, nem o parecer final de uma obra, mas mostrar os múltiplos caminhos de leitura, haja vista a quantidade de percursos que foram abertos enquanto levantamos os dados. Nas palavras de Ferrand (1997):

> [...] *les résultats produits par les calculs automatisés de la machine sont là pour relancer l'interprétation en suscitant de nouvelles questions, et non pas pour la stopper en laissant le chercheur muet de béatitude devant ses graphiques*. (FERRAND, 1997, p. 11).

Nossa pesquisa está estruturada da seguinte maneira: no capítulo O estilo "pede roupa nova e escova", auxiliados por Pierre Guiraud (1970), José Maria Pozuelo Yvancos (1994) e René Wellek e Austin Warren (1966), esboçamos um percurso que resgata estudos sobre estilo, abarcando desde a arte retórica à estilística com a ajuda de algumas correntes teóricas: Leo Spitzer (1968), Amado Alonso (1966) e Charles Bally (1951).

O próximo capítulo intitulado "Este mundo é muito misturado": os números em prol das letras, aborda, ancorado em Pierre Guiraud (1959), Anthony Kenny (1982), Susan Hockey (2004), Valérie Beaudouin (2000), de uma forma histórica, os estudos na área de estatística textual que envolvem a linguística e a informática mais voltados para o campo da Estilometria. A importância desse capítulo está em apresentar como duas áreas de naturezas tão distintas, a literatura e a estatística, podem se aliar e resultar em trabalhos importantes na área das Letras. Trataremos das origens dos estudos estatísticos e da distinção entre as duas grandes escolas, a inglesa e a francesa, com a finalidade de justificar a escolha do nosso método.

---

7 Sobre o *Hyperbase* e suas ferramentas: <http://ancilla.unice.fr/~brunet/PUB/hyperwin/hypermenu.htm>. Acesso em: 30 dez. 2011.

Ainda, apoiado por pesquisadores como Charles Muller (1968), André Salem e L. Lebart (1994), Étienne Brunet (1983, 1988, 2003, 2011), apresentamos o programa que será usado para as análises, o *Hyperbase* (versão 5.4); explanando mais rigorosamente sobre o tratamento estatístico de textos literários, bem como o esclarecimento da nomenclatura utilizada para o estudo de linguística de *corpus* na área de literatura.

No capítulo *Corpus* de baile, apresentamos todas as obras que serão incluídas nas análises e sua contextualização, ou seja, devidamente datadas segundo a pesquisa que realizamos sobre a produção ficcional de Rosa, para que a inserção da mesma no programa respeite a ordem cronológica, auxilie na leitura dos gráficos e na compreensão da análise.

"A bordo da nave com os timoneiros": dados da crítica é o capítulo em que apresentamos alguns estudos da crítica rosiana que pensamos ser interessantes como ideias para o tipo de abordagem que utilizamos para ler o texto literário. Dessa forma, traremos à baila algumas resenhas que retratam o aspecto estilístico e linguístico da obra de Guimarães Rosa. Nesse capítulo, abordaremos as leituras de Nilce Sant'Anna Martins (2001) e Walnice Nogueira Galvão (1978). Ao final dele, ressaltamos as informações que a crítica apontou e comparamos com os dados estatísticos escolhidos para a análise. Aí estão Oswaldino Marques (1983), Eduardo Coutinho (1993), Suzi Frankl Sperber (1982), Lenira Marques Covizzi (1978), e Mary Lou Daniel (1968).

No capítulo "A Nhanina sabe as letras mas... não decora os números, de conta de se fazer...", encontram-se os dados, propriamente ditos, seguidos de suas análises e apresentações das características lexicais. São apresentados aspectos da estatística textual tais como: extensão do vocabulário, riqueza lexical, *hapax*, evolução do vocabulário, crescimento lexical, as altas e médias frequências, os gráficos em árvores, oriundos dos cálculos de análises fatoriais, a distribuição das frequências (distância lexical). É nesse capítulo que relacionamos as proposições de Covizzi (1978), Sperber (1982) e Daniel (1968) com os dados obtidos.

"... cada um o que quer aprova, o senhor sabe: pão ou pães, é questão de opiniães": o qualitativo e o quantitativo nos textos literários, trata-se de um capítulo que apresenta uma breve reflexão a respeito da importância da análise estatística para os textos literários.

E, por último, "...ao fim retomo emendo o que vinha contando": desdobramentos é o capítulo conclusivo, no qual encontram-se resumidamente os resultados e as comparações com as hipóteses da crítica estabelecida para este trabalho.

2 “O ESTILO PEDE SEMPRE ROUPA NOVA E ESCOVA”

Le style c’est l’homme même.
(Buffon)

Para chegarmos à estilometria, precisamos antes entender como surgiu a preocupação em definir o estilo literário, como ele foi explorado ao longo dos séculos e como houve o envolvimento de diferentes áreas em função da necessidade de se formar atualmente uma crítica do estilo. Para isso, faremos um sobrevoo por muitos séculos da cultura ocidental e sua preocupação em definir o estilo.

Os estudos linguísticos (fonologia, morfologia, sintaxe etc.) e a literatura caminham em alguns momentos juntos, esta servindo de base para aqueles, como exemplo, as obras Owl and the Nightingale (poema do século XII ou XIII) e Sir Gawain and the Green Night (romance versificado em inglês do séc. XIV) que serviram como fontes de documentos para estudar alguns dialetos do inglês (WELLEK e WARREN, 1966). Apenas no século XX, com Charles Bally (Suíça, 1865-1947) e Karl Vossler (Alemanha, 1872-1949) uma disciplina dos estudos linguísticos, a estilística, passa a se aliar com a literatura. Desta parceria entre uma área que se preocupa com as línguas correntes e a outra com a poética de uma língua, resulta o que tivemos de maior valor sobre análise literária pois fez alavancar os estudos literários. Quando partimos do princípio que o estilo é marcado por ser diferença, tratamos da estilística como uma vertente da linguística que estuda a língua de uma obra literária cuja inserção se dá em uma língua geral.

Pierre Guiraud (1970) descreve, dentro de uma perspectiva histórica, sobre a noção de estilo informando que teve início na Grécia antiga por meio de estudos sobre a retórica (aproximadamente no século V a. C, sendo o filósofo sofista Górgias um de seus principais precursores), passa pela a Idade Média e constitui ainda no século XIX parte integrante da educação ocidental. A retórica, considerada como uma teoria da eloquência artística, tinha como função auxiliar o retórico (locutor, declamador) na composição de discursos e escritos, por isso, ela abarcou a expressão linguística, as artes poéticas e a literatura. Teoricamente, ela se constituiria sobre as noções de gênero, estilo (tom) e figuras (meios de expressão).

Muitos formulários e tratados foram redigidos para ensinar a arte de bem escrever e falar; alguns exemplos trazidos por Guiraud (1970), por exemplo, A Retórica (Aristóteles), De Oratore e o Orator (Cícero) e De Institutione Oratoria (Quintiliano) – nos permitem verificar a

trajetória de séculos que a arte da retórica atravessou. Tais modelos ofereciam algumas etapas de elaboração, por exemplo, a organização de argumentos, temas e provas (*inventio*), a ordem para a apresentação dos argumentos (*dispositio*), a exposição e a forma dos argumentos (*elocutio*) - essas três primeiras etapas constituem a elaboração do enunciado -, a entonação, fluência, gesticulação e fisionomia (*actio*) – esta última orientava a atividade de enunciação do discurso. Os tratados ofereciam uma série de outras etapas, mas não é nosso objetivo aqui nos estendermos sobre eles.

A arte da oratória influenciou os diferentes modos da expressão literária que compunham a retórica, e assim, os gêneros épico, lírico e dramático surgiram por meio de adaptação; daí resultaram tratados, por exemplo, de Aristóteles (A Poética) e de Horácio (A Arte Poética). Guiraud (1970) nos relata que é possível encontrarmos diferença entre os gêneros prosa e verso desde o século IV, havendo maior dedicação ao verso, pois a partir deste último foram categorizados – pelas formas de versificação que continham, bem como pelo vocabulário e sintaxe - ainda outros subgêneros, tais como a prosódia, o ditirambo etc.

Segundo José M. P. Yvancos (1994), tanto a Poética quanto a Retórica sempre estiveram relacionadas, Cícero (106-43 a.C.) e Tácito (58-120 d.C.) se utilizaram das duas técnicas para enriquecer o discurso oratório, dentro de um processo evolutivo, da poética à retórica. Esta acabou sofrendo mudanças, por motivos políticos (pois deveria se prestar como prática para o discurso de persuasão na democracia grega) e pedagógicos (tendência da difusão escolar, a retórica se tornou uma prática pedagógica sobre as formas de ornamento verbal). Desse modo, a arte discursiva passa a ser arte elocutiva, tendo sua função cada vez mais voltada para a própria linguagem.

Para Guiraud (1970) a Retórica seria a estilística de tempos e de escritores (ou declamadores) da antiguidade, com desenvolvimento por mais de vinte séculos. A partir do século XVIII, com o romantismo, devido à queda de paradigmas, a retórica clássica sofre uma grande mudança. Pois o iluminismo alcançou propriedade sobre o discurso teórico e prático, e o romantismo sobre o domínio estético, mas apenas com o romantismo que a retórica foi extinguida da expressão da imaginação; o conceito de literatura alcançou praticamente toda a escrita; ela tornou-se imaginativa e autônoma. (WELLBERY, 1998). Dessa forma, surge uma decadência da retórica, pois a linguagem deixa de refletir uma forma exterior ao homem, tornando-se o meio pelo qual o homem exprime sua experiência:

> A linguagem constitui a expressão de uma situação concreta; exprime diretamente as ideias e os sentimentos dos indivíduos e se confunde com ele; e através deles expressa os de um determinado temperamento social, de uma coletividade e de seus costumes e instituições. Já não se trata de reconhecer uma forma linguística num catálogo de formas universais anteriores e exteriores à expressão; a vida e a linguagem passam a ser consideradas no que têm de único e insubstituível, e essa autenticidade do vivido é que fundamenta sua autoridade. (GUIRAUD, 1970, p. 49).

A função da literatura se desloca e, em vez de traduzir o mundo, expressa a experiência do homem. "O estilo é a expressão do gênio individual." (GUIRAUD, 1970, p. 51). Assim, a retórica proporcionou à teoria um novo horizonte para compreender o literário que, mais tarde, fora projetado em escolas de poética linguística. Com isso, fundamentaram-se duas ideias da retórica herdadas pelo ocidente: a teorização feita a partir do interior da linguagem literária sobre a própria linguagem – cujo paradigma teórico institui a oposição entre línguas literária e gramatical; e, a noção de *desvio* (YVANCOS, 1994).

Com essa noção de desvio se desenvolveram escolas, cujo fundamento de seus estudos estava na teoria do estilo, descrevendo o literário como fenômeno linguístico diferenciado, ou seja, que se desvie das normas comuns que regem o uso da linguagem corrente. Essas escolas partem da existência de estruturas, recursos, formas e procedimentos que figuram a língua literária como um tipo específico. Surgem a estilística idealista (ou genética) e a estilística gerativa; ambas concordantes que a língua literária deva se apresentar diferenciada da língua padrão.

Segundo Guiraud (1970), a linguagem foi retomada como meio passivo da coletividade e do ato criador do indivíduo, a partir de estudos de Wilhelm Wundt (1832-1920), o qual seguia uma espécie de teoria linguística universal, e Hugo Schuchardt (1842-1927), seguidos por Karl Vossler (1872-1949). Este último reconheceu a língua como criação e evolução. Posteriormente, Leo Spitzer (1887-1960), a partir desses preceitos, formou a escola idealista alemã que tinha como influências o intuicionismo de Henri Bérgson (1859-1952) e as doutrinas de Benedetto Croce (1866-1952), as quais lançavam o princípio de autonomia e liberdade individual. Surgem então, as escolas estilísticas

do início do século XX: a *estilística do indivíduo* (ou estilística genética) que opera nas relações da expressão com o indivíduo, a partir da coletividade que a cria ou a emprega; e a *estilística da expressão* (estilística descritiva), cujos estudos enfocam as relações da forma com o pensamento, relacionando-se com a elocução clássica.

A estilística idealista inaugurada por Vossler e Spitzer elaborou a crítica da obra no seu todo inclusive no seu contexto. Spitzer, influenciado por Vossler, estabeleceu uma crítica fundamentada em características da obra (a crítica é imanente à obra, toda obra constitui um todo, cada detalhe deve permitir acesso ao centro da obra), vale dizer, que para Spitzer, a consciência de um detalhe que chama a atenção do leitor, mais a convicção de que tal detalhe guarda uma relação importante para a obra como um todo deve ser considerado. (YVANCOS, 1994). Em outras palavras, o crítico faz a leitura da obra mediante uma intuição, portanto, tomando como ponto de partida um traço característico constituindo um desvio estilístico individual.

As ideias de Spitzer provocaram na época críticas que colocavam em xeque a legitimidade do seu método, pelo fato do mesmo se fundamentar na intuição, o que resultaria em julgamentos derivados da subjetividade. (GUIRAUD, 1970). A estilística idealista acreditava que os desvios se correspondiam e explicavam particularidades psíquicas, pois os estudos linguísticos desta linha andavam em parceria com os estudos da psicologia. Além de Spitzer, outros teóricos importantes dessa linha são Amado Alonso, Dámaso Alonso, Haimut Hatzfeld e Carlos Bousoño.

Falemos brevemente deles. Amado Alonso insistiu fortemente no caráter integrador e unitário da obra de arte, entendendo como recursos formais (desvios) partícipes de um sistema expressivo formado por elementos substanciais (psíquicos, temáticos ou expressivos) e materiais (recursos verbais):

> Estas intuiciones (la del creador y la del lector) literarias, artísticas, se diferencian de la científica (mucho más simple) en que movilizan, por decirlo así, la totalidad psíquica del hombre: la memoria, a la cual llamamos fantasía cuando – en un estado lúcido, que tiene sin embargo relación con el ensueño – entremezcla con libertad sus datos, al par que los actualiza (realidad ilusoria: se trata de una intuición fantástica); la voluntad, que matiza afectivamente la imagen, deseada o repelida (aunque con "querencia" no práctica, es decir, sin

> finalidad posesoria: se trata de una intuición afectiva); y en fin – en literatura – básicamente el entendimiento (se trata de una intuición intelectual). Científicamente, intuimos con sólo una veta de nuestra psique (la intuición científica no es fantástica, ni es afectiva). Estéticamente, intuimos con toda nuestra psique, puesta de modo automático en una especie de vía muerta, o ensueño, [...] Pero el lector sabe que sueña, sabe que sabe que juega. (ALONSO, 1966, p. 38-39).

Já a estilística da expressão é definida da seguinte forma por Guiraud:

> A estilística da expressão é, portanto, o estudo dos valores expressivos e impressivos próprios aos diferentes meios de expressão de que a língua dispõe. Esses valores se acham ligados à existência de variantes estilísticas, isto é, de diferentes formas para expressar a mesma idéia, de sinônimos que exprimem um aspecto particular da comunicação. (GUIRAUD, 1970, p. 73).

Essa definição podemos encontrar em *Traité de Stylistique Française* (1902), de Charles Bally, sucessor de Ferdinand Saussure, tratado que fundamenta as bases racionais à estilística da expressão. Nele, a estilística é considerada como um estudo sobre os fatos da expressão da linguagem do ponto de vista do seu conteúdo afetivo, isto é, carrega a expressão dos fatos da sensibilidade mediante a linguagem e a ação dos fatos da linguagem sobre a sensibilidade. (GUIRAUD, 1970). O objeto de estudo da estilística de Bally se fundamenta no conteúdo afetivo da linguagem, dividindo os *caracteres afetivos* em *efeitos* naturais, que podem ter valor pejorativo e *efeitos por* evocação, caracterizados por tonicidade, línguas de época, classes ou grupos sociais, a geografia, a biologia:

> *Définition de la stylistique: elle étudie la valeur affective des faits du langage organisé, et l'action réciproque des faits expressifs qui concourent à former le système des moyens d'expression d'une langue. La stylistique peut être, en principe, générale, collective ou individuelle; mais l'étude ne peut présentement se fonder que sur le langage d'un groupe social organisé; elle doit commencer*

> *par la langue maternelle et le langage parlé.* (BALLY, 1951, p. 1).

A linguagem constituiria um instrumento de comunicação, um sistema de sinais destinados para a transmissão do pensamento. Partindo disso, a escola saussuriana se negaria a estudar o estilo individual, porque o considera como ato livre, isolado, original, incomensurável, não servindo para análise ou classificação:

> A análise saussuriana retoma, ao mesmo tempo, a posição de Humboldt entre a linguagem criadora e livre do indivíduo e a linguagem fixa e normalizada da coletividade; essa distinção, que se tornou clássica, entre a palavra e a língua, apresenta novamente ao linguista o problema do estilo. (GUIRAUD, 1970, p. 60).

A noção de desvio no estruturalismo se difere da estilística idealista, pois o ponto de vista de estilo do criador é substituído pelo ponto de vista do "estilo funcional" da literatura. A linguagem poética passa a interessar apenas como fenômeno geral. A explicação da linguagem literária se desenvolve a partir dos estudos linguísticos e a poética estruturalista adota um viés sistemático. Desse modo, a língua literária de um criador não fará referências às particularidades psíquicas dele, mas às formas às quais se opõe ou das quais se desvia. (YVANCOS, 1994).

Após essa imersão retrospectiva, ousamos apresentar neste trabalho, a comparação da obra completa de Guimarães Rosa, resgatando o primeiro *modus operandi*[8] comentado por Wellek e Warren (1966), contudo, para além disso, tomaremos o método estatístico auxiliado pelo programa *Hyperbase* que nos facilitará alcançar uma análise sobre a obra completa do autor. A junção desses dois métodos de pesquisa, a estilística e a estatística, é o que chamamos de estilometria, nas palavras de Freitas (2007) encontramos uma definição mais resumida:

> A estilometria é área que busca os padrões de repetição de elementos que compõem o texto. São esses padrões que compõem em grande parte a

[8] Para recapitular o *modus operandi*, indicamos a leitura do segundo parágrafo deste capítulo.

> identidade de um autor. A eles se misturam os padrões da língua, as características próprias do gênero literário, do tema e as influências da época, o que torna o desafio maior ainda. (FREITAS, 2007, p. 49).

3 "ESTE MUNDO É MUITO MISTURADO": OS NÚMEROS EM PROL DAS LETRAS

Chaque fois qu'on se risque à dire d'un auteur qu'il aime, ou qu'il préfère une tournure, un thème, un effet de style, chaque fois qu'on utilise pour le caractériser les mots fréquent, rare, souvent, jamais, même, autre, tout, recherché, banal, commun, original, caractéristique, typique etc., on fait appel à une statistique implicite, à des fréquences attendues et observées et en fin de compte à la notion d'écart.
(Étienne Brunet)

Auxiliaire de la synthèse, l'ordinateur est un outil mental: après l'organum d'Aristote et le Novum Organum conçu par Bacon, n'est-il pas le Novius Organum, " l'outil le plus nouveau " ?
(Benzécri)

Os estudos de estatística textual são desenvolvidos em vários países, mas há uma forte tradição no contexto europeu. Tony B. Sardinha (2004) afirma que a Linguística de *Corpus* chegou nos Estados Unidos muitas décadas depois dos estudos europeus, e seu obstáculo maior estava na linguística de Noam Chomsky. Enquanto a Europa tinha projetos como o *Cobuild*, o BNC, o *Longman Corpus Network*, o *Cambrigde International Corpus*, os EUA não ganhavam financiamento para pesquisa de *corpus*. Contudo, o reconhecimento foi chegando aos poucos e, hoje em dia, um dos expoentes norte-americanos é Douglas Biber[9], com as análises multidimensionais voltadas para a sociolinguística.

Apesar do interesse maior das pesquisas girar em torno de análises do discurso no campo sócio-político, principalmente nos atuais estudos realizados na França, temos a cada dois anos a JADT[10]

[9] Atualmente, professor do programa e Linguística Aplicada da Universidade do Norte do Arizona (EUA).

[10] A versão 2012 que aconteceu em Liège, Bélgica, contemplará temas como: análise exploratória de dados textuais; linguística de *corpus* e linguística quantitativa; tratamento automático de linguagem natural; etiquetagem; lematização; enriquecimento linguístico; análise estatística de dados estruturais e não estruturais; classificação de textos; cartografia lexical e textual; pesquisa documentária; web semântica; estilometria e análise do discurso; programas para análise textual; aprendizagem automática para análise de dados textuais e

(*Journées d'Analyse Statistique des Donées Textuelles* cuja primeira edição foi em 1990) no qual se apresentam trabalhos interessantes nessa área e se aponta, cada vez mais, para uma diversificação nos temas de análise. No Brasil, mais especificamente nos cursos de Letras, os estudos que envolvem o auxílio do computador estão mais relacionados ao campo da linguística e aos estudos da tradução; haja vista o evento intitulado *Encontro de Linguística de Corpus* (ELC) que acontece anualmente em nosso país. Nos estudos literários, infelizmente ainda não temos nenhum encontro nacional periódico voltado para a discussão sobre as ferramentas estatísticas informatizadas para auxílio tanto da pesquisa em literatura[11], quanto do ensino da mesma. Por vezes, é exatamente em contextos de linguística de *corpus* que vamos encontrar algumas pesquisas voltadas à literatura. É o caso do artigo de Lourdes Bernardes Gonçalves intitulado "Linguística de *corpus* e análise literária: o que revelam as palavras-chave", único trabalho que envolvia literatura entre os estudos de linguística de *corpus* publicado em um volume organizado em 2008 pelo Programa de Pós-graduação em Estudos Linguísticos e Literários em Inglês da Faculdade de Filosofia, Letras e Ciências Humanas da USP[12]. A autora analisou a obra *Dubliners*, de James Joyce, comparando a um *corpus* paralelo composto por escritores como Katherine Mansfield, David Herbert Lawrence e Virginia Woolf. Ela utilizou duas ferramentas do programa *Word Smith Tools*, de Mike Scott: *KeyWord* (como o nome diz, busca palavras-chave) e *Concord* (organiza listas de concordância).

Em 2010, o ECL[13] teve uma sessão de literatura, e nela se apresentaram três trabalhos: "Medidas de complexidade textual entre

---

corpora multilíngues e *corpora* paralelos. Disponível em: <http://www.jadt2012.ulg.ac.be/>; acesso em 3 jan. 2012.

11 Temos alguns eventos organizados em nível nacional (na sua última edição, em 2012, internacional) pelo NUPILL (Núcleo de Pesquisa em Informática, Literatura e Linguística da UFSC) que tem se esforçado bastante para estabelecer parcerias com o BCL (*Base Corpus Langage* – laboratório de pesquisa que se encontra na Universidade de Nice, local onde fizemos nossa pesquisa-sanduíche em 2010), com o grupo espanhol LEETHI (*Literatura Españolas y Europeas: el Texto al Hipermedia*), com o NUPLID (Núcleo de Pesquisa em Literatura Digitalizada) da UFPI e muitos outros parceiros que podem ser verificados em: <www.nupill.org>. Acesso em: 20 jan. 2013.

12 Na USP também temos contribuições da professora Ieda Maria Alves na área de lexicologia e lexicografia. O NILC (Núcleo Institucional de Linguística Computacional) também é um núcleo de pesquisa da USP que desde 1997 tem contribuído com pesquisas na área de linguística computacional. Mais informações podem ser encontradas no sítio do projeto: <http://www.nilc.icmc.usp.br/nilc/>. Acesso em: 04 de abr. 2012.

13 Segundo a lista de seus anais, o ECL de 2012 não apresentou nenhum trabalho no âmbito literário. Disponível em: <http://143.107.232.109/elc-

traduções brasileiras e originais de literatura inglesa: um estudo piloto baseado em *corpus*" (PAQUALINI; FINATTO; EVERS, 2010); "Palavras-chave e *hapax legomena*: aliados na análise literária?" (GONÇALVES, 2010) e "*Corpora* e operações enunciativas: um estudo sobre as adversativas do português brasileiro" (CELLI, 2010). Contudo, as pesquisas são mais voltadas para a metodologia da linguística de *corpus*, ou seja, sem análises literárias.

Alguns programas de análises estatísticas de textos realizam contagem de palavras, medição da distância lexical, da evolução e crescimento de vocabulário e da riqueza lexical[14]. Dessas análises, podemos também extrair dados interessantes para estudos temáticos, identificar autoria de textos apócrifos, desenvolver estudos estilísticos que, nesse caso, enviam ao que chamamos estilometria. Na análise de textos dentro do viés da estatística, as ferramentas propiciam estudos comparatistas, pois esse é o princípio da estatística, exercer comparações. Assim, por exemplo, pode-se verificar se um grupo de autores de um determinado período se diferencia no uso de vocabulário; quais deles o utiliza de forma mais rebuscada etc. É o que afirmam os estudiosos desse campo Ludovic Lebart e André Salem (1994, p. 11), para eles, a estatística textual pretende resolver uma série de problemas levantados por pesquisadores de estudos de estilo, como medidas comparativas de vocabulário de diferentes autores, medidas da evolução do vocabulário de um mesmo autor ao longo de uma cronologia (nosso caso) etc.

Entre os exemplos de estudos que aplicaram o método de estatística de textos, podemos mencionar: Gustav Herdan (1941), linguista, estatístico e jurista, foi o primeiro a estudar a estilometria com análise fatorial[15] para descobrir a relação entre seis autores e para identificar quais deles se utilizavam de palavras mais complexas. O mesmo tipo de análise também serviu para estabelecer o ancestral

---

ebralc2012/index.php/pt/elc2012anais>. Acesso em: 7 jan. 2012. O evento de 2011 também não trouxe nenhum trabalho na área de literatura. Disponível em: <http://www.letras.ufmg.br/CMS/index.asp?pasta=linguisticacorpus2011>. Acesso em: 7 jan. 2012.

[14] No HYPERBASE: *Le programme de préparation, entre autres tâches, constitue le tableau de distribution des classes de fréquences, le relevé des hapax (ou mots employés une seule fois) et bien d'autres résultats qui intéressent la structure du vocabulaire* (BRUNET, E. *Hyperbase*. 1997) Disponível em: <*http://ancilla.unice.fr/~brunet/PUB/hyperwin/richesse.htm*>. Acessado em 12 ago. 2011.

[15] Explicaremos sobre o procedimento das análises fatoriais no capítulo 5 deste trabalho.

comum de um determinado número de línguas indoeuropeias (Johnson e Kotz, 1967). Roger Peng e Nicolas Hengartner (2002) se utilizaram da análise fatorial para examinar a contagem de palavras gramaticais de cada autor. David Mealand (1997) estabeleceu, por meio da mesma técnica, que amostras de diferentes partes do *Evangelho de São Marcos* variam no estilo e também identificou os marcadores de escrita mais usados nessas passagens (BAGAVANDAS, M.; MANIMANNAN, G., 2004, p. 72).

Da produção nacional[16], Deise Freitas (2007) em sua tese, discorre sobre a divisão estilística em relação aos contos de Machado de Assis, reforçando a hipótese de que a transformação do estilo do autor acontece gradualmente na sua composição, não no material linguístico. Há também nossa dissertação de mestrado em que analisamos quatro romances (em alemão) de Franz Kafka (*O Castelo*, *América ou o Desaparecido*, *O Processo* e *A Metamorfose*) em níveis sintáticos e semânticos utilizando dos aplicativos *Hyperbase* e *Lexico 3*.

A informática permitiu o surgimento de novas formas de análise de textos e vocabulários: por exemplo, pode-se afirmar com segurança qual termo é ausente no texto, ou ainda, verificar não somente o que o texto diz, mas também o que ele evita dizer. A vantagem dessas ferramentas está na maneira como se tornam maiores, mais objetivas e acessíveis as informações, dando condições de chegar com mais segurança aos resultados das análises de informações. Todavia, é importante ressaltar que esse tipo de exploração da obra literária não se restringe apenas aos aspectos quantitativos, pois estes devem ser empregados como suporte à leitura direta da obra, ou seja, o contato direto do leitor-crítico com a criação literária é imprescindível e nem pode ser substituído.

Apesar da incipiência dos estudos estatísticos no âmbito literário, há de se concordar que os estudos tanto linguísticos, quanto literários assistidos por computador têm um caráter interdisciplinar, pois especialistas de diferentes áreas se reúnem para elaborar as ferramentas, utilizá-las, buscando sempre aprimoramento e reflexão. Precisamos de estatísticos, matemáticos, linguistas, críticos literários, sociólogos *etc*. É um trabalho em equipe que marca uma nova comunidade científica unida pelas novas mídias e pelo tipo de pesquisa que elas proporcionam.

---

16 Como dissemos anteriormente, os estudos de estatística de textos literários voltados para a teoria literária ainda é muito recente, nas Letras há uma tradição desses estudos nos campos da tradução e da linguística, mas na literatura ainda estamos a abrir caminhos.

3.1 BREVE HISTÓRICO DOS ESTUDOS ESTATÍSTICOS PARA TEXTOS LITERÁRIOS

Trazer um apanhado histórico das pesquisas já feitas sempre auxilia na compreensão do objeto de estudo, pois assim, podemos perceber as problemáticas e as soluções que este tipo de trabalho gera. Mas a importância maior de apresentarmos um panorama histórico sobre o assunto em tese é a orientação para os futuros leitores e pesquisadores desta área de estudos estatísticos para textos literários, pois não há muita bibliografia a respeito do tema traduzido para a língua portuguesa.

O linguista francês Pierre Guiraud (1959) esclarece que a estatística linguística vem de uma tradição muito antiga, os gramáticos alexandrinos já haviam criado uma lista de *hapax legomena*[17] de Homero e de textos massoretas (escribas judeus), relacionando-os com todas as palavras da Bíblia. Mas, é a partir do século XIX, com a gramática histórica, que o estudo das línguas se ampliou apoiado sobre os inventários numéricos, e, na década de trinta do século XX, esses estudos começaram a ser acompanhados por análises estatísticas.

Para Guiraud (1959), os índices e as concordâncias constituem a primeira matéria do estatístico, o *index verborum* é um repertório alfabético de todas as palavras de um texto com a indicação das passagens (verso – se for o caso –, página, linha). É a partir desse índice que se pode extrair o número total de palavras (formas) de um texto, o número de palavras diferentes, o número palavras para cada categoria gramatical (substantivos, adjetivos, singular, tempo verbal etc.), o número de palavras para cada categoria semântica, enfim, a distribuição do léxico no texto.

A frequência das palavras é, desde os trabalhos de George Kingsley Zipf (1902-1950), um dos problemas maiores da estatística textual. Colocou-se desde sempre o problema geral da distribuição dessas frequências e de sua forma particular. Há muito, por exemplo, se observou que um pequeno número de palavras seguidas e repetidas constitui a maior parte de um texto. Segundo Guiraud, os pesquisadores Margaret Sinclair Ogden e Charles E. Palmer efetuaram vastas compilações que definiram o vocabulário mínimo do inglês, estabelecendo uma lista de um milhão de palavras que satisfaziam todas as necessidades de expressão.

---

17 Do grego ἁπάξ (uma só vez) e τό λεγόμενον (o que se diz), ou seja, palavra que ocorre apenas uma vez no *corpus* (frequência = 1).

Guiraud (1959, p. 16) parte do princípio de que "uma língua é um sistema de signos". Ele divide sua discussão sobre esse sistema em três níveis: o primeiro se identifica pelos sons ou fonemas; o segundo nível combina esses fonemas em formas portadoras de sentido que, em conjunto, constituem um léxico, e finalmente, no terceiro nível, as combinações dessas formas exprimiriam suas relações de sintaxe. Esses três tipos (fonéticos, lexicais e sintáticos) se identificam, por um lado, pela sua forma e em oposição às outras formas do sistema; e de outro, por seu sentido, quer dizer, pela imagem mental que elas evocam. Essas duas funções, diacríticas e semânticas, definem o signo qualitativamente. Já o terceiro nível seria caracterizado pela frequência do signo, que, segundo Guiraud, teria valor funcional e constituiria um de seus atributos.

Nesse sentido, existe aí, para ele, uma função estatística da linguagem (já que falamos de frequência do signo), função não menos importante, não menos objetivamente real que as funções diacrítica e semântica; não menos indispensável para uma compreensão do fato linguístico. Os signos (sons, palavras, marcas e construções gramaticais, elementos de estilo) se repetem com uma frequência fixa:

> *Ceci constitue le postulat sur lequel repose l'application de la méthode et sa légitimité, et plus qu'un postulat c'est un fait désormais si universellement observé et vérifié qu'on doit le considérer comme une loi du langage et quelle que soit l'attitude du lecteur à son égard, les doutes qu'il peut avoir sur sa valeur, la méfiance que peut lui inspirer son interprétation, la répugnance qu'il peut entretenir à l'utiliser, il doit au moins être certain qu'il est impossible de nier son existence* (GUIRAUD, 1959, p. 16).

Para o autor, a estabilidade de frequência de sons é um fato observado e reconhecido. A criptografia já estabeleceu desde o século XVI a frequência de letras e de suas combinações nas diferentes línguas. Essas frequências dependem do estilo (língua culta e popular, prosa e poesia) e são notavelmente estáveis em uma língua. A frequência de oclusivas sonoras, por exemplo, é aproximadamente a metade das surdas correspondentes e isso ocorre em um grande número de línguas (GUIRAUD, 1959, p. 17).

Assim define Guiraud (1959) a estatística como o método que permite estabelecer estruturas, apreciar desvios e decidir em que medida elas (as estruturas) são aleatórias, e, portanto, sem significação; ou pelo

contrário, se elas têm um valor determinado. Acrescentaríamos ainda ao raciocínio do autor, que no caso de um texto literário, as estruturas podem determinar um valor estilístico. O interesse do método estatístico é medir e analisar precisamente, constatando não somente as diferenças na estrutura numérica de dois textos, como também quantificando desvios e os colocando em fórmulas passíveis de transformações abstratas e de expressões gráficas de leitura e interpretação imediatas.

De todos os comportamentos humanos, ainda segundo Guiraud (1959), a expressão linguística aparece como a menos individualizada; nossa língua, em boa parte, não é uma criação pessoal, nós a recebemos de outrem. A palavra criada por um indivíduo toma seu valor apenas na medida em que ela é aceita, retomada, repetida e, finalmente, definida pela soma de seus empregos. Tais empregos, no conjunto, acabam refletindo sua situação linguística, as causas frequentemente complexas que determinam sua escolha e seu uso. Emprega-se a palavra pelo seu sentido, pelo seu valor linguístico, pela sua forma fonética, ou por essas relações com outras palavras, entre outros. A estatística, para Guiraud, é precisamente o método destinado a estabelecer e interpretar, no nível coletivo, esses acontecimentos que não são definíveis no nível individual. Essas relações são muito mais gerais e muito mais sutis entre as palavras e estabelecem relações de hábitos que modificam os sentidos. É pela diferença das palavras ou de outros elementos da linguagem que o escritor atua sobre o leitor e sobre a língua: "*un style est un écart qui se définit quantitativement par rapport à une norme*" (GUIRAUD, 1959, p. 19).

Certamente não se trata de substituir a apreciação qualitativa subjetiva por uma análise quantitativa objetiva; as duas são inseparáveis. O que a estatística propõe é introduzir rigor na apreciação e no emprego desse elemento quantitativo inerente a todo discurso. Para o autor, foi sobre o estudo quantitativo que se fundou a gramática comparativa, encontrando em diferentes línguas a presença de características comuns.

Como grande parte da bibliografia relacionada ao tema dos estudos estatísticos de textos está publicada em inglês ou francês, traçaremos agora o ponto de vista histórico de pesquisas das fontes que encontramos, de modo que possamos assim ter uma visão mais ampla das preferências investigativas.

Anthony Kenny e Susan Hockey, pesquisadores da Estatística Textual, afirmam que os estudos estatísticos na Inglaterra retomam o ano de 1851, com uma pesquisa de autoria que media o tamanho das palavras das epístolas de São Paulo. Augustus de Morgan (1806-1871), professor de matemática da Universidade de Londres responsável pela

pesquisa, chegou à conclusão de que as palavras usadas na *Epístola aos Hebreus* eram mais longas em relação às outras cartas escritas pelo mesmo apóstolo. Ainda no mesmo século, outro pesquisador retoma a mesma ideia de Morgan, Thomas Corwin Mendenhall, que publica o artigo *The characteristic curves of composition* (1887), testando a mesma hipótese. Segundo Kenny (1982), Mendenhall estudou também o *Novo* e o *Velho Testamento*, e as obras de autores como Dickens, Thackeray, Shakespeare, Bacon e Marlowe, seu método foi construir listas de frequências e comparar tamanhos de palavras (KENNY, 1982, p. 3). No século XX, o estudo que se destaca é do estatístico de Cambridge Udny Yule, que publicou *The statistical study of literary vocabulary*. Nesse estudo, o autor verificou o tamanho médio das frases de alguns escritores, entre eles, Bacon, Coleridge e Macaulay, destacando as suas diferenças estilísticas.

Kenny (1982) acrescenta que os estudos relacionados à estatística de textos se desenvolveram enormemente entre Mendenhall e Yule e que, com a invenção do computador, surgiram grandes pesquisas quantitativas em estudos de textos literários. Os estudos sobre o *Novo Testamento*, tendo como ponto de partida a questão da autoria das *Epístolas*, foram novamente retomados por William C. Wake e A. Q. Morton que estudaram o comprimento das frases nas *Epístolas* e fizeram extensas comparações com outros autores gregos.

Kenny retrata dois estudos de textos em língua inglesa, do início dos anos sessenta, que são considerados como modelos de pesquisa estatística de estilo. Um deles é o de Alvar Ellegård sobre *Junius Letters*[18], que apresentou uma lista de aproximadamente quinhentas palavras e expressões caracterizando um padrão de escrita de Junius, chegando à conclusão de que Sir Philip Francis era esse escritor. O outro trabalho de referência é o de Frederick Mosteller e David Wallace sobre os *Federalist Papers*, em que foram empregadas técnicas mais elaboradas que as de Ellegård, como os métodos estatísticos baseados nos teoremas de probabilidades de Thomas Bayes[19].

---

[18] *Junius Letters* é uma reunião de 69 cartas assinadas pelo pseudônimo "Junius" que foram publicadas entre os anos de 1769 e 1772 no jornal inglês *General Advertiser* (Londres).

[19] Teólogo inglês, autodidata em matemática, contribuiu com o cálculo de probabilidades por meio de um estudo intitulado "Ensaio voltado para a solução de um problema na Doutrina do Acaso", publicado postumamente pela *Royal Society* (1763). O ensaio virou lei fundamental da matemática por apresentar técnicas de estatística e estimativa e tornando-se conhecido como "Teorema de Bayes". O inovador na pesquisa de Bayes era o caráter subjetivo na previsão de um evento, considerando significativa a opinião do matemático

Apresentamos brevemente a escola inglesa e, assim, pudemos perceber que ela se caracteriza fortemente pelos estudos e análises de autoria. Trataremos agora, da escola francesa, cuja tradição das análises se identifica mais com a nossa pesquisa.

Segundo Freitas (2007, p. 47), nos estudos de estatísticas de textos realizados na França, a pesquisa que se destaca como pioneira data de 1642, sendo um trabalho que sistematizou uma codificação técnica que desenvolvia listas de concordâncias dos textos bíblicos. O responsável pelo desenvolvimento dessa empreitada chamava-se Hubert de Phalèse (monge beneditino de Afflighem, Bélgica), e, em homenagem a essa figura, fundou-se em 1989 um grupo de pesquisa na Universidade de Paris III[20], que desenvolve pesquisas em torno das tecnologias computacionais relacionadas à análise de textos.

Em artigo sobre análise distribucional, Valérie Beaudouin (2000) comenta que os estudos sobre estatística textual, no contexto da França do século XX, surgiram no final dos anos 50 entre Besançon e Estrasburgo, no *Centre d'Études du Vocabulaire Français de Besançon*. Tudo começou com uma contagem mecanizada da obra de Corneille feita pelo pesquisador francês Charles Muller. Diz o autor ainda, que foi em 1957, em conferência na cidade de Estrasburgo, que se deu a origem ao projeto *Trésor de la Langue Française*[21].

Ainda dentro dos estudos francófonos, Beaudouin (2000) faz distinção entre duas tendências: a estatística lexical e a estatística textual. A primeira é representada pelo pesquisador Charles Muller (que tomou a empreitada de contabilizar a obra de Corneille em análise lexicométrica informatizada) e a segunda, que trata de dados linguísticos ou textuais, é conduzida pelo pesquisador estatístico Jean-Paul Benzécri.

A estatística lexical compara os dados observados com os dados calculados partindo de um modelo teórico que aposta na ideia de que um texto analisado é uma amostra representativa de uma língua e que no

---

nos cálculos, contudo, essa subjetividade do pesquisador também foi motivo de polêmica para o teorema.

[20] Disponível em: < http://www.cavi.univ-paris3.fr/phalese/>. Acesso em: 7 jan. 2013.

[21] O *Trésor de la Langue Française* (TLF) é um grande projeto de dicionário que contém a língua francesa dos séculos XIX e XX (entre os anos 1789 a 1960), seu *corpus* possui 100.000 palavras, 270.000 definições e 430.000 exemplos com citações. Foi editado entre os anos de 1971 a 1994, e apenas no início dos anos 90 passou por um processo de informatização, sendo finalizado em 2001. Disponível em: <http://atilf.fr>. Acesso em: 18 fev. 2012.

estudo de um *corpus*, é possível então retirar informações sobre a língua que o compõe.

Nesse sentido, compara-se a subfrequência observada de uma palavra em uma subpartição de um *corpus*, na qual a subfrequência teórica é calculada a partir da frequência da palavra no conjunto do *corpus*, medindo-se assim o desvio entre as duas (frequência e subfrequência). Desse modo, obtém-se a lista de palavras significativamente mais presentes ou ausentes em cada partição do *corpus*. Além de se trabalhar com uma referência interna (o conjunto do *corpus*), pode-se escolher uma referência externa, como por exemplo, a representada pela *Trésor de la Langue Française* (TLF) que seria a frequência de todas as palavras presentes na base textual FRANTEXT[22] ou, por exemplo, trazendo para o nosso contexto, os *corpora* desenvolvidos para a língua portuguesa[23]. Nesse tipo de abordagem probabilística, da lexicometria, o modelo é elaborado empiricamente a partir de dados, providos seja de um *corpus* em seu conjunto ou de um *corpus* externo.

Segundo Beaudouin (2000), as pesquisas de cunho estatístico, giram em torno de análises sobre a riqueza lexical, especificidades, crescimento e evolução cronológica do vocabulário. Ele afirma, ainda, que análises dessa natureza são bem empregadas por países anglo-saxões em estudos de estilo e de busca de autoria.

A outra modalidade de estatística para textos que Beaudouin (2000) aborda é a estatística textual, tendo como representante Jean-Paul Benzécri, considerado o precursor da análise de dados entre os franceses. Em 1964, Benzécri apresenta, na Faculdade de Ciências de Rennes, teorias e métodos de escalas multidimensionais, contribuindo principalmente na ordem dos dados, dispondo-os em um quadro, sob a forma de matrizes, para aplicação do método de análise que permite

---

22 *FRANTEXT* é uma base de dados de textos literários e filosóficos (científicos e técnicos – 10%) em língua francesa, desenvolvido e mantido pelo ATILF-CNRS (*Laboratoire Analyse et Traitement Informatique de la Langue Française*). A base foi criada nos anos 70 com o intuito de fornecer exemplos para o *Trésor de la langue française*. FRANTEXT. Disponível em: <http://www.frantext.fr/>. Acessado em: 15 set. 2011.

23 Para a língua portuguesa também temos muitos corpora para pesquisar, seguem algumas fontes para consulta:

Corpus Histórico do Português Tycho Brahe <http://www.tycho.iel.unicamp.br/~tycho/corpus/index.html>;

Corpus Brasileiro, disponível em: <http://corpusbrasileiro.pucsp.br/cb/Inicial.html>;

Projeto Linguateca, disponível em: <http://www.linguateca.pt/www_linguateca_pt.html>. Acesso em: 20 jan. 2013.

sintetizar a informação contida nessas matrizes. Benzécri tinha por ambição teórica abrir um novo campo de pesquisa nos estudos linguísticos, pois estes estavam dominados pelos estudos da linguística gerativista, se opondo à tese generalista de Chomsky, que nos anos 60, afirmava que não poderiam existir procedimentos sistemáticos para determinar a gramática de uma língua, ou de estruturas linguísticas a partir de um conjunto de dados (banco de textos).

É contra essa tese que Benzécri propõe um método indutivo de análise de dados linguísticos que efetua uma abstração quantitativa partindo de tabela de dados diversos. Ele se opõe, igualmente, às teorias idealistas que declaram a existência de um modelo e verificam a pertinência do mesmo por meio da observação.

O pesquisador Max Reinert, aluno de Benzécri, desenvolveu uma metodologia inspirada pelas análises de dados chamada ALCESTE (sigla para *Analyse des Lexèmes Cooccurrents dans les Énoncés Simples d'un Texte*): "*Il s'agit, non pas de comparer les distributions statistiques des 'mots' dans différents corpus, mais d'étudier la structure formelle de leurs cooccurrences dans les énoncés d'un corpus donné"* (REINERT, 1993, p. 9 *apud* BEAUDOUIN, 2000). Reinert considera um *corpus* como uma sequência de enunciados elementares produzidos por um sujeito-enunciador. Desse modo, o texto é modelado em uma tabela composta por linhas de enunciados, que trazem a marca do sujeito-enunciador, e por colunas preenchidas por palavras ou lexemas que reenviam aos objetos do mundo. Sua hipótese considera o vocabulário de um enunciado particular como um traço pertinente de um ponto de vista, de um lugar referencial e de uma atividade coerente do sujeito-enunciador.

Os procedimentos estatísticos que aproximam enunciados empregando o mesmo tipo de léxico permitem identificar diferentes redes lexicais (ou como prefere chamar Reinert, mundos lexicais) que podem revelar a visão de mundo de um sujeito-enunciador. Em um trabalho sobre a obra *Aurelia* (1855), de Gérard de Nerval (1808-1855), por exemplo, Reinert (1990 *apud* BEAUDOUIN, 2000) identifica três tipos de mundos lexicais classificando por enunciados: o imaginário, o real e o simbólico. A metodologia ALCESTE possibilitou, nesse caso, identificar o universo do discurso, as classes dos enunciados, que devem ser o objeto de uma interpretação específica em função da natureza do *corpus* e dos objetivos de uma análise.

Em termos gerais, as duas áreas da estatística de textos apresentadas por Beaudouin (a lexical e a textual) são desenvolvidas paralelamente, cada uma com suas próprias publicações (em menor

número aqui no Brasil). Nossa pesquisa pretende mediar essas duas áreas, pois acreditamos que a complementação de ambas enriquecerá a pesquisa: as análises fatoriais e os cálculos de especificidades[24].

3.2 O PROGRAMA HYPERBASE E SUAS FERRAMENTAS

A ferramenta utilizada para esta tese foi o aplicativo *Hyperbase*[25]. Desenvolvido por Étienne Brunet, (BCL, Universidade de Nice-FR), é utilizado em análises de unidades ou macroestruturas, por exemplo, podemos analisar uma peça de teatro ou um conjunto de peças. Também é possível com o aplicativo determinar os termos mais específicos de uma obra ou de um conjunto de obras em relação a outro *corpus*. O *Hyperbase* traz outras funções tais como lista de concordâncias, riqueza e evolução de vocabulário e análise fatorial que representa graficamente a distância lexical entre os textos que compõem o *corpus*.

Dentro das operações de funções estatísticas e documentárias, podemos definir:

- o dicionário de frequências do vocabulário de um autor a partir do *corpus*;
- o vocabulário específico de cada texto ou de todo o *corpus* através de uma lista de formas significativamente excedentes ou deficitárias no texto em análise;
- o desenvolvimento temático de uma palavra ou de um grupo de palavras;
- a correlação cronológica, ou seja, a frequência teórica de cada palavra é avaliada para ser identificada a progressão ou a regressão das formas (também chamada de evolução do vocabulário);
- o efetivo dos vocábulos e das palavras empregadas apenas uma vez no *corpus* inteiro (*hapax*);
- a conexão lexical que ilustra a distância de cada texto de todos e aos pares, com os outros que compõem o *corpus* por meio do léxico comum ou exclusivo;
- a riqueza lexical.

---

[24] Ambos serão explicados e exemplificados no capítulo 5.

[25] O *Hyperbase* não é um software livre. Mas existe uma versão demo de 2011 para baixar em: <http://ancilla.unice.fr/>, acesso em 3 jan. 2012.

Portanto, utilizando o *Hyperbase*, além de reunir todo o vocabulário rosiano de forma categorizada, definiremos suas características lexicais, quais sejam as altas ou baixas frequências de palavras, *hapax legomena*, coocorrências e outras ferramentas que apontem os traços mais fortes do estilo de Rosa. Verificaremos, também, pela cronologia de sua produção literária, por meio da distância lexical, como se estabeleceu o seu vocabulário, ou seja, como se deu o percurso evolutivo da obra. Por meio de todos esses dados que desenvolveremos estratégias de leitura as quais devem legitimar a estatística textual como percurso plausível para se interpretar um texto literário.

O *Hyperbase*, como o próprio nome diz, é um aplicativo que reúne vários programas de diferentes funções. Ele apresenta um menu principal com ferramentas que garantem uma exploração metódica da documentação. As duas funções básicas de característica documentária são o *concordance* e *contexte*, que obedecem aos mesmos princípios e se distinguem apenas pela apresentação de seus resultados. No *contexte*, cada ocorrência que se busca é situada e mostrada em seu contexto no parágrafo. O *concordance* também apresenta uma contextualização, contudo mais restrita, apresenta a ocorrência da palavra centrada na lista de linhas (não mais de parágrafo), acompanhada de um número mínimo de palavras para a direita e para a esquerda de sua posição no enunciado, ou seja, parte de seu contexto, ressaltando desse modo, os sintagmas repetidos que por vezes podem revelar restrições sintáticas, mas também, tendências fraseológicas do autor.

No capítulo 4, privilegiamos alguns trabalhos que nos auxiliaram a encontrar os percursos para nosso estudo, mas das leituras da crítica, principalmente, os trabalhos de Suzi Sperber (1982), Lenira Marques Covizzi (1978) e Mary Lou Daniel (1968), no que diz respeito ao vocabulário, foram nossos principais guias para pensar o objeto desta tese em termos quantitativos. Essas leituras não estão diretamente ligadas à evolução do vocabulário — trabalho que será feito por nós —, mas serviram certamente como reflexões a nortear nossas análises.

Inicialmente, o procedimento foi reunir a obra completa de Guimarães Rosa em formato digital, ou seja, os textos todos foram escaneados por um programa de reconhecimento óptico de caracteres e depois revisados[26]. Utilizamos a edição da *Nova Aguilar* (1994) como

[26] Atualizamos o *corpus* segundo as normas do novo acordo ortográfico da língua portuguesa, mas optamos em não alterar as palavras hifenizadas, pois tal tarefa alteraria bastante o estilo do autor.

referência para a escanerização e, também, para o procedimento seguinte, a revisão dos textos, embora outras edições também tenham sido consultadas[27].

Após a revisão, balizamos os textos conforme a necessidade do aplicativo de análise estatística, ou seja, padronizamos os caracteres para que o *corpus* fosse reconhecido e delimitado. Balizamento de textos é um procedimento que pode variar de acordo com cada programa, sua característica é o emprego de alguns códigos para que o aplicativo reconheça, por exemplo, onde começa e termina um texto, onde começa e termina um parágrafo etc. Por exemplo, para o *Hyperbase*, fundamentalmente, e dependendo da versão, devemos grafar com a sequência de caracteres *&&& TÍTULO&&&* toda vez que se inicia um texto no documento de extensão **.txt* que será utilizado no programa. Outro software que conhecemos, mas que não aplicamos neste trabalho é o *Lexico3*[28], nele, o procedimento para que se reconheça o início de um texto é <*texto=título*>. É importante salientar que dentro dos textos que serão inseridos no programa não existam tais símbolos em uma mesma sequência, pois eles podem impedir ou interferir na leitura dos programas.

3.3 TERMINOLOGIA DE CORPUS ESTATÍSTICO E DAS FERRAMENTAS

O método estatístico tem como alicerce medidas e contagens realizadas a partir de objetos que se deseja comparar. Esse tipo de procedimento requer que identifiquemos tais objetos de maneira ordenada. Geralmente os programas de análise estatística textual utilizam uma nomenclatura que não costuma variar muito. Traremos aqui uma sucinta explicação dos termos empregados pelo aplicativo *Hyperbase*. Os termos são originalmente em francês e estão traduzidos para o português, a fim de facilitar o diálogo entre os resultados e a leitura desta tese.

Na prática, empregam-se alguns princípios gerais que devem ser seguidos para que ocorra a comunicação entre o usuário do programa e a exposição dos dados. Há também alguns limites a respeitar. Por

---

[27] As outras edições utilizadas para conferência na revisão dos textos estão listas nas referências bibliográficas.

[28] O aplicativo LEXICO3 pode ser baixado gratuitamente em: <http://www.tal.univ-paris3.fr/lexico/lexico3.htm>. Acesso: 2 ago. 2012.

exemplo, o tamanho dos textos a serem comparados não deve apresentar diferenças exorbitantes; mesmo existindo cálculos de ponderação no programa para que haja um equilíbrio, o melhor é não trabalhar com textos muito desiguais. Segundo Brunet (2011, p. 18), deve-se evitar ultrapassar o limite de 500.000 palavras por texto. Na versão atual do *Hyperbase* pode-se trabalhar com 75 textos ao mesmo tempo. Contudo, vale a pena ressaltar que o limite também existe para possibilitar a leitura dos gráficos, pois uma análise de 250 textos ao mesmo tempo, inviabilizaria a leitura na tela. Neste caso, o limite existe para possibilitar a leitura.

Inicialmente, quando inserimos os textos (em formato **.txt*) no aplicativo, há uma operação que recorta o texto, ou o conjunto de textos, em *unidades mínimas* (unidades que serão decompostas mais adiante no procedimento estatístico); é a fase chamada de *segmentação* do texto. Após esse recorte, sucede a etapa de *identificação*, quando ocorre um reagrupamento, a partir do texto, das unidades idênticas. Para que se concretize uma segmentação automática do texto em ocorrências de formas gráficas, é necessário apontar, no início do processo, um conjunto de caracteres que são conhecidos como *delimitadores de texto* (LEBART; SALEM, 1994, p. 36).

Os caracteres considerados delimitadores pelo *Hyperbase* são:

, . ; : ? ! " ' ( ) < > - — + / = { } [ ] …

Sem a delimitação, o aplicativo pode considerar, por exemplo, "pois," (com vírgula) como uma *forma* e "pois" (sem vírgula) como outra *forma*, (sendo que é uma só!) a vírgula é um caractere que está ali na sequência das *formas* como qualquer outro, porém, o aplicativo não faz diferenciação entre caracteres, a não ser que o pesquisador etiquete e escolha quais caracteres são os *delimitadores* de texto.

A distinção de caracteres do texto em *caracteres delimitadores* e *não-delimitadores* permite definir uma série de descritores relativos às formas simples. Para abordar a descrição dos *segmentos compostos* de várias *formas* e *segmentos repetidos* (duas formas que aparecem juntas mais de uma vez) em um *corpus*, é necessário especificar o estatuto de cada um dos *caracteres delimitadores*. Por isso, sempre que se cria uma base, o programa apresenta uma lista desses caracteres que pode ser alterada a critério do pesquisador.

Para os procedimentos formais que comumente se elaboram, nos contentaremos em dar a alguns sinais de pontuação (o ponto final, o ponto de exclamação e o ponto de interrogação) o estatuto de separador

forte ou separador de frase. Entre esses caracteres delimitadores nós escolheremos igualmente um subconjunto correspondente às pontuações fracas e fortes (em geral: a vírgula, o ponto e vírgula, os dois pontos, as aspas e os parênteses) e chamaremos o conjunto de *delimitadores de sequência*. A continuação, então, das ocorrências situadas entre dois *delimitadores* é considerada como uma *sequência*.

A operação feita, primeiramente, pelos programas específicos em análise de estatística textual é um corte das *sequências de caracteres* em *formas*, de onde se retiram as *ocorrências*. Por exemplo, se uma forma aparece uma vez em um *corpus*, dizemos que ela tem uma *ocorrência*, e que, portanto, essa é a sua *frequência*. Os aplicativos, geralmente, distinguem e fornecem listas de todas as formas do texto a ser trabalhado, essa lista pode ser lida em ordem alfabética ou em ordem decrescente de frequências (conhecida também como dicionário de frequências, alfabético ou hierárquico). Na maioria dos casos, as formas mais frequentes que surgem no topo da listagem são as *palavras gramaticais*, como os artigos, os dêiticos[29], as preposições, os pronomes, as conjunções etc.

O *conjunto de formas* de um texto constitui o seu *vocabulário*. Uma segmentação assim permite considerar o texto como uma sequência de ocorrências separadas entre elas por um ou mais caracteres delimitadores. O número total de ocorrências contidas em um texto é o seu *tamanho* ou seu *comprimento*.

Para os pesquisadores Lebart e Salem (1994, p. 36), do ponto de vista lexicométrico, o *corpus* deve ser submetido a uma *lematização*[30], ou seja, ser submetido a regras que identifiquem as mesmas formas gráficas correspondentes às diferentes flexões de um mesmo lema, tais como, levar as formas verbais ao infinitivo, os substantivos ao singular, os adjetivos ao masculino singular etc. A lematização atua como uma espécie de filtro que evita, em partes, a ambiguidade dos vocábulos homógrafos, de modo que impeça a sua repetição, opera nas diferentes conjugações de um mesmo verbo, tornando as flexões todas em infinitivo e inferindo também tanto no número como no gênero das formas. A lematização garimpa o texto, de modo que as formas sejam

---

29 Há um artigo sobre atribuição de autoria por meio de comparação das *Cartas Chilenas* que utilizou o programa *Lexico 3* para a geração dos dados. Através de um levantamento dos dêiticos e palavras gramaticais, o artigo afirma que esses léxicos são usados de forma mais independentes por partes dos autores durante a elaboração de um texto (BRANDÃO, 2006).

30 Lematizar é dar regras de identificação que permitem reagrupar nas mesmas unidades as formas gráficas que correspondam às diferentes flexões de um mesmo lema.

contabilizadas mais estritamente. Contudo, o processo de lematização, a nosso ver, não é um método obrigatório para os tratamentos estatísticos de texto, pois dependendo da maneira que se quer analisar um texto, pode limitar a leitura. Por exemplo, uma análise estilística a partir dos tempos verbais mais usados por um dado autor será inviabilizada pela lematização, pois todas as conjugações serão substituídas pelos seus infinitivos[31].

Com exceção dos textos lematizados, os verbos aparecem com maior frequência porque suas formas sofrem flexões, diferente dos substantivos, que variam menos. Geralmente, para uma análise temática, os substantivos são considerados os elementos mais significativos do texto. Contudo, devemos estar atentos, pois nem todo tema é expresso pelos significantes, a leitura ocorre também através do contexto. Vejamos o clássico exemplo do romance *Dom Casmurro*, de Machado de Assis, que traz como temática a traição; porém, em nenhum momento do texto aparece a forma "traição". Por isso, devemos ter cautela com os resultados, pois os números podem ser ilusórios, ocasionando leituras contraditórias. Dessa forma — salientamos uma vez mais —, é necessário o domínio crítico do texto literário que se pretende analisar. Concordamos, igualmente, com FREITAS (2007, p. 66) que, na busca de rigor, pesquisadores das ciências humanas[32] adotem métodos de outras áreas, tal como a estatística; contudo, é necessário prudência para não se deixar levar pelo excesso de objetividade no que diz respeito à quantificação, evitando paradoxalmente os impressionismos que os resultados também podem trazer.

É possível realizar outros tipos de organizações das formas e das ocorrências do texto: as ocorrências de uma mesma forma se encontrarão agrupadas em uma mesma direção, acompanhadas de um pequeno fragmento do *contexto* imediato, no qual se fixará o

---

31 Por motivos como esse, escolhemos não lematizar nosso *corpus*.

32 Mas ainda assim, existe uma dificuldade a ser superada nos cursos de Letras, em se tratando de estudos estatísticos. Maria Tereza de Almeida Camargo, em artigo intitulado *Estatística Linguística* (1967), comenta que a maioria dos linguistas recua diante de tratados como o de Gustav Herdan que analisa sob um método de cálculo linguístico as *Geórgicas* de Virgílio. Diz a autora: "Por outro lado, os matemáticos e estatísticos que têm se dedicado à Linguística Matemática não têm suficiente formação linguística para equacionarem devidamente os problemas linguísticos dentro do universo estatístico [...] perdem-se em labirintos matemáticos tratando de problemas que não interessam à linguística moderna, ou utilizam conceitos linguísticos ultrapassados. [...] seria desejável que os estudantes inclinados aos estudos linguísticos tivessem uma formação estatística elementar, durante os anos de licença universitária, como já acontece em outros domínios das Ciências Humanas – Sociologia, Psicologia". (CAMARGO, 1967, p. 118).

comprimento em função de necessidades particulares. A forma fixa que reagrupa os contextos chamamos de *forma-polo*. Esse tipo de reorganização permite estudar de modo mais fácil as relações existentes entre os diferentes contextos de uma mesma forma, tal procedimento também é conhecido como *concordância*. Desse modo, podemos verificar o emprego de uma forma desejada e quais as outras formas que se encontram vinculadas a ela em maior ou menor grau; há também a possibilidade de buscarmos as *coocorrências*, ou seja, palavras que surgem normalmente na mesma sentença, no mesmo parágrafo, ou em um mesmo contexto.

Existe uma maneira de verificar a frequência das ocorrências e suas localizações, esse modo de verificação geralmente é designado como *índice* (sistema de organização, em que se apresentam todas as formas). Dependendo do programa, as formas e suas ocorrências podem vir acompanhadas de sua frequência e localização no *corpus*. Os índices classificam as formas segundo critérios diferentes: *índice alfabético* (classificação segundo a ordem lexicográfica, ou seja, a ordem corrente dos dicionários) e *índice hierárquico* (as formas são posicionadas em ordem decrescente, segundo as suas frequências).

Em um *corpus* de tamanho grande, o *crescimento do vocabulário* tende a sofrer uma dupla influência, cada vez que uma nova ocorrência é apreendida, o número total de formas de um *corpus* também aumenta (mais ocorrências, mais formas distintas), porém, quando o tamanho do *corpus* aumenta, a taxa de formas novas trazidas para cada crescimento do número de ocorrências tende a diminuir.

Charles Muller, em *Initiation à la statistique linguistique* (1968), explica que o crescimento do vocabulário é, primeiramente, feito a partir de uma contagem das palavras que compõem um *corpus*, obtendo assim um valor numérico **N** (número total de palavras), esse número é a exata medida de extensão do texto. Desse modo, o programa associa cada uma dessas palavras a um *vocábulo*[33] (*forma*), para obter-se um segundo valor numérico, **V** (número de vocábulos que têm ao menos uma ocorrência no texto). **V** está em função de **N**, ou seja, **V** tende a crescer com **N**, mas é evidente que **V** cresça de modo mais lento que **N**, pois, cada palavra que aumenta o *corpus* pode ser um vocábulo que já estava presente nele. Desse modo, diminui a proporção entre as quantidades **V**

33 Charles Muller apresenta a seguinte distinção entre vocábulo e palavra: «Le **vocable** est une **unité de lexique**, le **mot** une **unité de texte** ; on a lu un mot dans le texte, mais c'est un vocable que l'on trouvera dans le dictionnaire» (MULLER, 1968, p. 133). (grifo nosso).

e **N**, ou seja, entre o número de palavras não repetidas e a totalidade delas. Ao iniciar uma contagem de um texto, Muller repara que **V** é igual a **N** até a primeira repetição de um vocábulo qualquer.

Muller dá outro exemplo: partindo de um texto qualquer considerado homogêneo, ele extrai dois fragmentos de comprimento distintos. Deve-se prever que o mais longo terá um vocabulário de extensão superior em relação ao mais curto. Porém, a extensão do vocabulário é função também do estilo, ou seja, ele é determinado, no mínimo, pelo léxico do autor na situação estilística em que ele se encontra. Se recolhermos dois textos de estilos muito distintos e de comprimento igual, observaremos um *desvio* entre a extensão do vocabulário dos dois textos, e esse desvio é uma característica estilística de primeira importância. Admitiremos assim que, o vocabulário mais extenso, significa também um léxico mais extenso; porém, essa extensão não significa *riqueza de vocabulário* (MULLER, 1968, p. 156-7).

Étienne Brunet, autor do *Hyperbase*, em seu trabalho sobre o vocabulário de Proust, afirma que o crescimento de vocabulário é uma noção relativa e dinâmica (BRUNET, 1983, p. 20), ao contrário da riqueza lexical, que se apresenta como uma medida absoluta, independente da ordem dos textos considerados. Ao dispor os sete romances proustianos componentes de *À la recherche du temps perdu*: *Du côté de chez Swann* (1913), *À l'ombre des jeunes filles en fleurs* (1918), *Le côté de Guermantes* (1921-1922), *Sodome et Gomorrhe* (1922-1923), *La prisonnière* (1923), *Albertine disparue* (1925), *Le temps retrouvé* (1927), todos dispostos em ordem cronológica, Brunet percebe que, na sequência deles, as entradas de palavras novas eram cada vez mais raras conforme a contagem chegava ao final da obra. Isso confirma o juízo de que quanto maior for o tamanho do *corpus* investigado, menor será o número de formas novas que ele irá apresentar.

A apresentação dos métodos e nomenclaturas dos trabalhos aqui mencionados demonstra uma pequena parcela do que pode ser produzido com o auxílio de ferramentas informatizadas. Por isso, durante a utilização das análises estatísticas, traremos reflexões teóricas sobre estudos de estilo e de vocabulário de nosso *corpus*, unindo os aspectos lexicais e formais da obra para, assim, detectarmos o comportamento do vocabulário rosiano, ou seja, para compreendermos de que modo ocorre uma evolução e como ela se manifesta durante toda a obra.

No próximo capítulo apresentaremos o *corpus* de nossa análise, ou seja, os textos que irão compor as bases de nossas análises

estatísticas, bem como sua contextualização cronológica, pois essa é de grande importância para o tratamento estatístico a ordem de entrada dos textos no programa de estatística textual, quando se quer investigar o movimento progressivo e evolutivo do léxico de um autor.

4 *CORPUS* DE BAILE

Neste capítulo apresentaremos o *corpus* de maneira mais detalhada no que diz respeito aos dados cronológicos de produção e publicação de João Guimarães Rosa. Como nossa proposta de trabalho é estudar a evolução do vocabulário e suas especificidades, faz-se necessário inserir os textos em ordem cronológica no programa, por isso, o critério de organização do *corpus* foi esse.

Elaboramos duas bases distintas[34] para a extração dos dados, todas respeitando o critério cronológico compreendem o vocabulário ficcional de Guimarães Rosa que se estende entre os anos de 1929 a 1967. Numa das bases, a base A, encontram-se todos os textos, e a divisão deles se apresenta, primeiramente, por data aproximada de elaboração que resgatamos por meio da biografia elaborada por Ana Luiza Martins Costa[35], ou por ordem de data de primeira publicação (pois um texto nem sempre é publicado no mesmo ano em que é escrito). Desse modo, conseguimos estimar os possíveis anos de produção ou de primeira publicação de cada texto. Por consequência dessa organização, os textos de *Primeiras estórias*, *Estas estórias*, *Tutameia*, e *Ave, palavra* se encontram separados e distribuídos segundo a cronologia da primeira publicação que encontramos nos periódicos, ou seja, diferente da maneira como a editora estipulara em primeira publicação reunida. Ainda é preciso advertir que, com relação a *Sagarana*, *Grande sertão: veredas* e *Corpo de baile*, por motivos de elaboração mais extensa, ocupando muitos anos de trabalho, decidimos distribuí-los pela base em uma cronologia aproximada, simulada, pois eles não têm data específica. Para a exploração dos dados, nessa base, os contos de 1929 e 1930 estão reunidos num só conjunto[36].

A outra base que estabelecemos para análise, a base B, obedece à entrada dos textos no *Hyperbase* seguindo a organização das datas de primeira publicação em obra reunida, tanto em vida quanto, nos caso de alguns textos, póstuma (*Ave, palavra*, por exemplo). Os primeiros contos escritos durante os anos de 29 e 30 foram dispostos separadamente nesta segunda base, bem como os grandes textos,

---

34 Nas legendas dos gráficos é possível reconhecer qual base está sendo analisada.

35 COSTA, Ana Luiza Martins. *Veredas de Viator*. Cadernos de Literatura Brasileira. Instituto Moreira Salles. 2006.

36 Para que o programa identifique quando começa e quando termina um texto, devemos inserir no arquivo algumas balizas de reconhecimento para tal informação, por isso, o programa vai considerar o limite que dermos. Sobre isso tratamos no capítulo 3, no item 3.3 "Terminologia de *corpus* estatístico e das ferramentas".

Grande sertão: veredas ficou divido em três partes de acordo com a distribuição das páginas, e *Corpo de baile* fora divido segundo os contos.

As bases apresentam o vocabulário de Rosa que se estende dos anos de 1929 a 1967. A apresentação que segue contextualiza os textos com algumas informações a respeito das publicações e das elaborações. Começaremos pelos primeiros contos escritos entre os anos 1929 e 1930. Vejamos, então, os textos.

O mistério de Highmore Hall – conto enviado para concurso literário promovido pela revista *O Cruzeiro* (n. 57, RJ), escrito enquanto Rosa ainda era estudante de Medicina. Foi publicado em 07/12/1929, com ilustrações de Carlos Chambelland.

Makiné – conto publicado no suplemento dominical "De tudo um pouco" em *O Jornal* no dia 9/02/1930 com ilustrações de Chambelland.

ΧΡΟΝΟΣ χαι ΑΝΑΓΚΗ (Tempo e Destino) – conto publicado em *O Cruzeiro* (RJ) em 21/06/1930, com ilustrações de Chambelland.

Caçadores de camurças – conto publicado n'*O Cruzeiro* (RJ) em 12/07/1930, com ilustrações de H. Cavalleiro.

Magma – primeiro e único livro de poesias, vencendor em 22/11/1936 do *Concurso Literário da Academia Brasileira de Letras*, conquistando o primeiro lugar, entre 24 inscritos. Recebeu elogio de Guilherme de Almeida em parecer da comissão julgadora. Publicado apenas em 1997.

Sagarana – livro produzido entre os anos de 1936 a 1945 e publicado em abril de 1946, ganhando o *Prêmio Felipe d'Oliveira*. Concorreu ao *Prêmio Humberto de Campos* (em 1936), da Livraria José Olympio Editora. Graciliano Ramos era membro do júri, votando contra, e o livro ficou em segundo lugar. Composto por 9 contos: *O burrinho pedrês*; *A volta do marido pródigo*; *Sarapalha*; *Duelo*; *Minha gente*; *São Marcos*; *Corpo fechado*; *Conversa de bois* e *A hora e vez de Augusto Matraga*.

Sobre a elaboração de *Sagarana*, Rosa diz em carta ao editor João Condé:

> Assim, pois, em 1937 — um dia, outro dia, outro dia... — quando chegou a hora de o "Sagarana" ter de ser escrito, pensei muito [...] O livro foi escrito — quase todo na cama, a lápis, em cadernos de 100 folhas — em sete meses de exaltação, de deslumbramento. (Depois, repousou

> durante sete anos; e, em 1945 foi "retrabalhado", em cinco meses, cinco meses de reflexão e lucidez).
> [...] Lá por novembro, contratei com uma datilógrafa a passagem a limpo. E, a 31 de dezembro de 1937, entreguei o original, às 5 e meia da tarde, na Livraria José Olímpio. O título escolhido era "Sezão"; mas, para melhor resguardar o anonimato, pespeguei no cartapácio, à última hora, este rótulo simples: "Contos (título provisório, a ser substituído) por Viator". Porque eu ia ter de começar longas viagens, logo após (ROSA, 1993, p. 7-9).

Sagarana foi, primeiramente, apresentado com um total de 12 contos, os três retirados pelo próprio autor foram: *Questões de família* "História fraca, sincera demais, meio autobiográfica, mal realizada. Foi expelida do livro e definitivamente destruída" (ROSA, 1993, p. 10); *Uma estória de amor* "Um belo tema, que não consegui desenvolver razoavelmente. Teve o mesmo destino da novela anterior." (ROSA, 1993, p. 10) e *Bicho Mau* "Deixou de figurar no 'Sagarana', porque não tem parentesco profundo com as nove histórias deste, com as quais se amadrinhara, apenas por pertencer à mesma época e à mesma zona. Seu sentido é outro. Ficou guardada para outro livro de novelas, já concebido, e que, daqui a alguns anos, talvez seja escrito" (ROSA, 1993, p. 10). *Sagarana* foi sucessivamente revisado pelo autor ao longo de 21 anos, até a sua 5ª edição, em 1958. Como foi uma obra retrabalhada durante dez anos e depois seguiu com reedições revisadas pelo autor até o ano de 58, preferimos deixá-la na organização do *corpus* no ano de 1946.

Com o vaqueiro Mariano - 1947 (*Correio da Manhã*, RJ) publica a primeira parte, em 1948 publica as outras duas partes no mesmo jornal. No final do ano de 1952 foi publicado em versão integral com ilustrações de Daniel Valença Lins, pela editora Hipocampo. Foi o seu segundo livro publicado. Para seguir com a cronologia nas análises, retiramos o conto "Entremeio com o vaqueiro Mariano" da reunião de *Estas Estórias*, e o dispomos como obra publicada separada, já que esse conto teve elaboração e publicação na década de 40 (sendo, na verdade, o seu segundo livro publicado e pouco comentado pela crítica nessa qualidade). *Estas Estórias* reúne textos produzidos na década de 60.

Corpo de baile – produzido entre os anos de 1952 e 1955. Publicado em janeiro de 1956, em dois volumes, contando com 7 novelas e ilustrações de Poty. A partir da terceira edicão (1964) o seu

conteúdo foi subdividido pelo próprio autor nas seguintes partes que foram publicadas separadamente:

- *Manuelzão e Minguilim*: Campo geral; Uma estória de amor.
- *No Urubuquaquá, no Pinhém*: O recado do morro; Cara-de-bronze; A história de Lélio e Lina.
- *Noites do Sertão*: Dão Lalalão; Buriti.

Grande sertão: veredas – produzido na primeira metade da década de 50, publicado em maio de 1956, mesmo ano de *Corpo de baile*. Igualmente ilustrado por Poty. Ganhou três prêmios: Machado de Assis (INL); Carmen Dolores Barbosa (SP) e Paula Brito (RJ).

Primeiras estórias - foi escrito entre os anos de 1960 a 62 aproximadamente, e publicado em 1962, pela Editora José Olympio, com ilustrações de Luís Jardim, num total de 21 contos (15 deles foram publicados entre os meses de janeiro a agosto 18/03/1961 no jornal *O Globo* e na revista *Senhor*, depois incluído em *Primeiras Estórias* e 6 inéditos). Em 1963 foi premiado pelo Pen club Brasileiro.

Tutaméia (Terceiras estórias) – três contos foram publicados anteriormente no jornal *O Globo* em 1961 e a sua grande maioria publicada em 1967 no periódico *Pulso*. Em 1966, publicou mais 24 contos em *Pulso*. A coletânea foi publicada em julho de 1967 com capa de Luís Jardim, pela José Olympio, num total de 40 contos. Ganhou mais 4 contos depois na republicação: *Nós, os temulentos*, *Melim-meloso*, *Hipotrélico* (*O Globo*, 1961) e *Aletria e Hermenêutica*, em versão reduzida sob o título de *Risada e meia* (Letras e Artes, 1954).

Estas estórias – produzido durante a década de 60, com publicações esparsas, primeiramente na revista *Senhor* e depois reunidas para publicação em 1969. A editora José Olympio lançou esse volume com desenhos de Poty, 8 contos e republicou aí "Com o vaqueiro Mariano".

Ave, palavra – trata-se de uma coletânea de textos publicados anteriormente em periódicos e suplementos como *Correio da Manhã*, *Folha da Manhã*, *A Manhã*; *Diário de Minas*; *Jornal de Letras*; *O Globo* (19 contos); *Pulso* (6 deles em 1967) durante os anos de 1947 a 1967, muitos deles escritos durante as estadias de Rosa na Europa. Postumamente reunido e publicado, pela editora José Olympio, em 1970, com capa de Gian. Reúne 54 textos de natureza variada, bem como se explica na edição da Nova Aguilar, em nota (ROSA, 1994, p. 916):

> O volume, preparado pelo autor, reúne notas de viagem, diários, poesias, contos, flagrantes, reportagens poéticas e meditações, tudo o que constituirá sua colaboração de vinte anos, descontínua e esporádica, em jornais e revistas brasileiros, durante o período de 1947 a 1967.

O mistério dos MMM – romance policial elaborado em conjunto e constituído por dez capítulos, dos quais, o sétimo fora desenvolvido por Rosa. As ilustrações são de Percy Deane. Publicado entre outubro e dezembro de 1961. Esse romance foi coordenado por João Condé e composto por Viriato Corrêa, Dinah Silveira de Queiroz, Lúcio Cardoso, Herberto Sales, Jorge Amado, José Condé, João Guimarães Rosa, Antônio Callado, Orígenes Lessa, Rachel de Queiroz.

Finalizada a apresentação e contextualização da obra de Rosa, partiremos desses dados e comporemos uma cronologia com base nas supostas épocas de produção e nas primeiras publicações. Os textos serão inseridos no *Hyperbase* nesta ordem:

- 1929-1930: O mistério de Highmore Hall; Makiné; Tempo e destino; e Caçadores de camurça;
- 1936: Magma (publicado em 1997);
- 1937: Sezão ou Contos (Sagarana) - publicado, depois de reformulado, em 1946.
- 1941: Zoo (*Hagenbecks Tierpark*); Aquário - publicados depois no periódico *Pulso* 18/02/67; Mau humor de Wotan - publicado depois em *Correio*, RJ, 29/02/48); *A senhora dos segredos* (publicado depois em *Correio*, RJ, 6/12/52); *Homem, intentada viagem* (*O Globo*, 18/02/1967); *A velha* (*O Globo*, 03/06/1961);
- 1942-44: em Bogotá, vivenciou a história *Páramo* que foi finalizada em 1967;
- 1945: *Sanga Puytã* (publicado depois em *Correio*, RJ 17/08/47); *Cipango* (relato de uma visita a uma colônia japonesa no Pantanal, publicado na *Folha da Manhã*, SP, 17/02/53); *Ao Pantanal* (*Diário de Minas*, BH, 05/04/53); *Uns índios – sua fala* (*Letras e Artes*, RJ, 25/05/1954);
- 1946: Sagarana;

- 1947-48[37]: *Com o vaqueiro Mariano* (*Correio da Manhã*, RJ, 1947); *Meu tio Iauaretê*; *Em Cidade* (*Correio da Manhã*, RJ, 15/02/48);
- 1950: visita os lugares: *Jardin de Plantes* e Zoológico de *Vincennes* de onde publica os textos: *Do diário de Paris*; *Zoo (Parc Zoologique du Bois de Vincenne)*; *Áquario* (Nápoles); *O burro e o boi no presépio* e *Zoo* (Whipsnade Park, Londres);
- 1951: *O lago do Itamaraty* (pulicado depois pela *Seleções Digest Readers*);
- 1952: escreve *Mensagem da ordem do vaqueio: pé-duro, chapéu-de-couro* (*O jornal*, RJ, 28/12);
- 1952-1954: anos de elaboração de Corpo de baile e Grande sertão: veredas;
- 1953: *Teatrinho* (*Folha de SP*, 17/02/53); *No diário de Paris* (17/05/53); *O homem de Santa-Helena* (3/05/53); *Terrae vis* (10/05/53); *Fantasmas dos vivos* (*Diário de Minas*, BH, 24/01/53);
- 1954: publica em *Letras e Artes*: *Subles* (6/4/54); *Risada e meia* (4/5/54, publicado posteriormente como o prefácio de Tutameia, *Aletria e hermenêutica*, não foi incluído no *corpus* no ano de 1954 porque sofreu muitas alterações até Tutameia);
- 1958: *O último dos Maçaricos* (tradução)[38];
- 1960: no mês de abril na revista *Senhor* (n. 22) é publicado *A simples e exata estória do burrinho do comandante* (republicado em Estas estórias, em 1969);
- 1961: entre os meses de janeiro e agosto publica 34 contos em *O Globo*, 12 delas republicadas em **Primeiras estórias**, 1962): *Sorôco, sua mãe, sua filha* (18/03); *A terceira margem do rio* (15/04); *Menina de lá* (06/05); *Sequência* (13/05); *Os irmãos Dagobé* (10/06); *As margens da alegria* (01/07); *O cavalo que bebia cerveja*

---

37 De 1948 a 1951: mora em Paris, elabora *Náutikon* (escrito em seu diário de viagens 4/11/48 a 18/02/1951, contudo, ainda não publicado).

38 Do original *The last of the Curlews*, de Fred Bodsworth, um dos diretores da Fundação de Naturalistas de Ontário (Canadá). Consideramos acrescentar a tradução desse texto, pois segundo a crítica, essa tradução está bem elaborada e traz muito do estilo rosiano. Nas palavras de Manuel Bandeira: "Eu sabia que era assim com Rosa. Sabia do que se passou com ele quando foi convidado a traduzir para *Seleções* um romance condensado. Era a história de um pássaro. Rosa mandou vir dos Estados Unidos o romance completo. Mandou vir também tratados de Ornitologia. Fez a tradução, reescreveu-a cinco vezes. No fim saiu obra perfeita, coisa que não era no original." (BANDEIRA, 1986, p. 320).

(8/07); *Um moço muito branco* (29/7); *A benfazeja* (05/08); *Tarantão, meu patrão* (12/08). Depois saíram mais 3 publicações que foram republicadas em **Tutameia** (1967): *Hipotrélico* (O Globo, 14/01); *Nós, os temulentos* (28/01); *Melim-meloso* (22/04) e um conto não republicado: *O inverso afastamento* (15/07). Mais a diante 19 contos foram publicados no mesmo jornal, *O Globo*, e tiveram republicação em **Ave, palavra** (1970): *De stella et advendtu magorum* (07/01); *Além da amendoeira* (21/01); *Uns inhos engenheiros* (04/02); *O grande samba disperso* (11/02); *Homem, intentada viagem* (18/02); *Às Coisas de poesia* (25/02); *O riachinho Sirimim* (04/03); *Circo do Miudinho* (25/03); *Outras coisas de poesia* (01/04); *Novas coisas de poesia* (20/05); *Jardim fechado* (27/05); *A velha* (03/06); *A caça à lua* (17/06); *Sempre coisas de poesia* (22 ou 29/07); *Recados do Sirimim* (19/08); *Evanira!* (26/08); *Alguns bichos* (Brasil, 12/1961 e 01/1962); publica o sétimo capítulo no romance *O mistério dos MMM* na *Revista O Cruzeiro* (out. e dez.);

- 1962: publica no periódico *Senhor* de março a agosto: *A estória do homem do Pinguelo* (n. 37, março, novela republicada em **Estas estórias**); *Substância* (n. 38, abril); *Partida do audaz navegante* (n. 39, maio); *Nenhum, nenhuma* (n. 42, agosto); *Pirlimpsiquice* (em *Comentário*, RJ, n.11, publicado posteriormente em **Primeiras estórias**);
- 1963: *Maior meu Sirimim* (Diário Carioca, RJ, 21/07, republicado em **Ave, palavra**).
- 1964: *Fita verde no cabelo* (suplemento literário, de *O Estado de São Paulo*, 08/02); *As garças* (22/02, republicado em **Ave, palavra**); *Os chapéus transeuntes* (sobre a soberba, o primeiro capítulo do livro *Os sete pecados capitais* (RJ, Civilização Brasileira, a pedido de Enio Silveira, republicado em **Estas estórias**, 1969);
- 1965: em maio começa a publicar pequenos contos no jornal de medicina *Pulso*, RJ, de maio a dezembro, dos 17 contos publicados, 14 deles irão parar em **Tutameia** (1968): *A escova e a dúvida* (15/05); *Desenredo* (29/05); *Orientação* (26/06); *Tapiiraiauara* (10/07); *Uai, eu?*

(07/08); *João Porém, o criador de perus* (21/8); *Tresaventura* (4/9); *Azo de almirante* (18/09); *Hiato* (02/10); *O outro ou o outro* (16/10); *No prosseguir* (13/11); *Como ataca a sucuri* (27/11); *A vela ao diabo* (11/12); *Presepe* (25/12). Três contos serão republicados em **Ave, palavra** (1970): *O porco e seu espírito* (12/06); *Sem tangência* (24/07); *Quemadmodum* (30/10);

- 1966: publica entre janeiro e dezembro 26 contos no jornal *Pulso* (RJ), desses, 24 estarão em **Tutameia**: *Antiperipleia* (22/01); *Umas formas* (19/02); *Se eu seria personagem* (05/03); *Sota e Barla* (19/03); *Grande Gedeão* (02/04); *Reminisção* (16/04); *Intruje-se* (30/04); *Lá, nas campinas* (14/05); *Barra da vaca* (28/05); *Retrato de cavalo* (11/06); *Estoriinha* (25/06); *Curtamão* (09/07); *Rebimba, o bom* (23/08); *Esses Lopes* (30/09); *Estória n. 3* (17/09); *Sinhá Secada* (01/10); *Os três homens e o boi dos três homens que inventaram um boi* (15/10); *Zingaresca* (29/10); *Vida ensinada* (12/11); *Faraó e a água do rio* (26/11); *Droenha* (10/12). E dois outros foram republicados em **Ave, palavra**: *Cartas na mesa* (08/01); *Nascimento* (24/12);
- 1967: entre janeiro e julho publica 13 pequenos contos em *Pulso*, RJ (seis deles estarão em **Ave, palavra**): *Zoo* (Whipsnade Park, Londres, 07/01, mas elaborado em seu diário de viagem em 1950); *Aquário* (18/02, também escrito durante sua estadia na Europa pelos anos 50); *Zoo* (Rio, Quinta da Boa Vista, 01/04); *Os abismos e os astros* (27/05); *Redolbra* (10/06). Seis serão de **Tutameia**: *Mechéu* (21/01); *Palhaço da boca verde* (04/02); *Sobre os planaltos* (4/03); *Caderno do Zito* (18/03); *Inteireza/incessância* (15/4); *Trastempo* (22/4) e um conto que não foi republicado *Rogo e Aceno* (29/07)[39]. *Poemas de Natal* (*Das Pastorinhas* e *Quatro poemas sobre o burro e o boi no presépio*, Revista Realidade, SP, está no *corpus* no ano de 1950). Quatro

[39] Tal texto não consta em nosso *corpus*. Os textos *O verbo e o logos* (publicado logo depois de sua morte, em jornais do Rio de Janeiro (*Correio da Manhã*, 25/11)); *Carta ao Cônsul Cabral* (Jornal da Tarde, SP, 25/11); *Viver é muito perigoso...* (Suplemento Literário de Minas, BH, 25/11); *Oração aos novos* (Diário de Notícias, RJ, 26/11) não foram resgatados e portanto não se encontram no *corpus*.

> contos inéditos datilografados, faltando apenas a última revisão do autor: *Páramo*; *Bicho mau* (que pertenceu à *Sagarana* antes mesmo da primeira publicação); *Retábulo de São Nunca*; *O dar das pedras brilhantes* (todos publicados em **Estas estórias**).

Conforme o levantamento que fizemos da obra com base na biografia elaborada por Ana Luiza Martins Costa, ainda nos faltou a localização temporal de alguns textos que decidimos remanejá-los desta maneira, respeitando a organização das edições:

- Publicados em **Primeiras estórias**: Famigerado, Fatalidade, O espelho, Nada e a nossa condição, Luas-de-mel, Darandina, Os cimos. Publicados postumamente **em Ave, palavra**: Histórias de fadas, Grande louvação pastoril, Quando coisas de poesia, Reboldra, Ainda coisas de poesia, Dois soldadinhos mineiros, Minas Gerais e Mais meu Sirimim. Esses textos serão ordenados no corpus junto aos textos de 1962;
- Publicados em **Estas estórias**: Bicho mau, Páramo, Retábulo de São Nunca, O dar das pedras brilhantes. Esses estarão junto aos textos de 1964;
- Publicados em **Tutameia**: Aletria e hermenêutica, Arroio-das-Antas, Quadrinho de estória, Ripuária. Esses serão inseridos juntos aos textos de 1965-66.

Já os textos: *Sobre os planaltos; Caderno do Zito; Inteireza/incessância; Trastempo*, que foram publicados separadamente, encontram-se diluídos em *Tutaméia*. Tal opção tem como fundamento a dissertação em crítica textual de Sandra Regina Paro:

> Entre janeiro e dezembro de 1966 publicou mais 26 pequenos contos no *Pulso*: 24 deles, republicados em *Tutaméia*. E, finalmente, entre janeiro e julho de 1967, publicou 13 contos no *Pulso*, dos quais, seis deles aparecerão mais tarde organizados em *Tutaméia*. Desses seis, quatro contos aparecem incorporados no prefácio "Sobre a escova e a dúvida", são eles: "Sobre os Planaltos" (04.03); "Caderno de Zito" (18.03), ambos incorporados no item VII do dito prefácio; "Inteireza/incessância" (15.04), incorporado no item II do prefácio e "Transtempo" (22.04), incorporado ao item III do prefácio. (PARO, 2008, p. 52-53).

Toda essa preocupação com a ordenação dos textos é para suprir a análise de evolução de vocabulário — lembrando que evolução aqui é apenas um termo da análise estatística de textos e não implica nenhum juízo de valor, a evolução requer tempo, ou seja, evolução do vocabulário é a análise sobre o comportamento do mesmo dentro de uma evolução cronológica — os textos precisam ser inseridos no *Hyperbase* em ordem cronológica e por isso a nossa preocupação em ordená-los de acordo com a produção de escrita. Desse modo, inserimos os arquivos de textos na ordem e na nomenclatura a seguir, vale reparar que os textos de dimensões maiores (*Grande sertão: veredas* e *Corpo de baile*) foram segmentados para não haver muita discrepância na geração dos dados:

Ordem	**Nomenclatura**	**Ordem**	**Nomenclatura**
1	1929-1930	13	LINA
2	1936_MAGMA	14	LALALÃO
3	1941	15	BURITI
4	1945	16	GSV
5	1946_SAGARANA	17	GSV1
6	1947-1948	18	GSV2
7	1950	19	1953-1954
8	1951-1952	20	1958
9	CAMPO	21	1960-1961
10	ESTAMOR	22	1962
11	RECADO	23	1963-1964
12	BRONZE	24	1965-1966-1967

Quadro 1: Ordenação e nomenclatura da base A das obras para inserção no *Hypebase.*

Como as obras *Corpo de baile* e *Grande sertão: veredas* são de grande fôlego e não encontramos data (ano) exata de elaboração das mesmas, decidimos separar a primeira em 7 partes (respeitando a ordem dos contos no livro) e a segunda em 3 partes. Para dar sequência com a lógica temporal, dispomos as partes entre os anos iniciais da década de 50.

No capítulo a seguir reuniremos os dados da crítica rosiana que tratam do material linguístico de Rosa e que nos servirão de apoio e reflexão para esta pesquisa. Cabe lembrar que esse capítulo apresentará também um dos nossos objetivos da tese, que será verificar algumas

intuições derivadas de leituras tradicionais (a partir de Sperber, Covizzi e Daniel) e buscar novos elementos textuais não reconhecíveis tão facilmente no texto, mas que podem ser considerados como marcadores estilísticos de Rosa.

5 "A BORDO DA NAVE COM OS TIMONEIROS": DADOS DA CRÍTICA

A crítica literária, que deveria ser uma parte da literatura,
só tem razão de ser quando aspira a contemplar, a preencher,
em suma a permitir o acesso à obra.
(Rosa em entrevista com Gunter Lorenz, 1973)

Se vi mais longe foi por estar de pé sobre ombros de gigantes.
(Isaac Newton)

A poeticidade e a complexidade da obra rosiana comporta um amplo espectro, e dentro dele percebemos criação de vocábulos, camuflagem de provérbios, ricas manipulações do ritmo da prosa etc. Por consequência disso, os estudos sobre a obra rosiana se tornam, igualmente, variados e bastante ricos, podemos citar aqui, entre muitos outros trabalhos, uma publicação que interessa a esta pesquisa, pois congrega o vocabulário de toda a produção de Rosa: *O léxico de Guimarães Rosa*[40]. É o resultado de um estudo de Nilce Sant'Anna Martins, apresentado em formato de dicionário, contém explicações hipotéticas sobre o acervo de palavras empregadas, classificando-as como dicionarizadas ou como inventadas pelo autor.

Contudo, o trabalho de Martins, apesar de imenso, nos traz apenas um recorte do vocabulário que considera de valor estilístico maior, segundo sua perspectiva:

> Procurei selecionar, de preferência, os vocábulos empregados com algum valor estilístico mais acentuado, vocábulos com alguma expressividade particular, como neologismos, arcaísmos ou vocábulos arcaizantes, empréstimos, onomatopéias, palavras populares, regionais ou eruditas. Assim sendo, não foram incluídos

40 MARTINS, Nilce Sant´Anna. *O Léxico de Guimarães Rosa*. São Paulo: EDUSP, 2001. Trata-se de uma pesquisa (20 anos de duração) de valor imenso para os que pesquisam Guimarães Rosa, pois reúne cerca de 7 mil verbetes, dos quais 2.700 não são dicionarizados. É, sem dúvida, nossa importante fonte de consultas.

> vocábulos do léxico básico da língua, aqueles que todos conhecem e usam, a não ser que seu emprego ultrapasse o puramente referencial, estando enriquecidos de uma conotação especial. (MARTINS, 2001, p. xiii).

Nossa pesquisa, além de reunir a totalidade do léxico, traz como diferencial a análise estatística e estilística da evolução do vocabulário de Rosa a partir da cronologia de sua produção.

Outra característica muito explorada nos estudos rosianos é a criação de neologismos. Declarações do próprio autor nos instigam a desvendar esse léxico tão amplo, criativo e trabalhado:

> Mas o mais importante, sempre, é fugirmos das formas estáticas, cediças, inertes, estereotipadas, lugares comuns etc. Meus livros são feitos, ou querem ser pelo menos, à base de uma dinâmica ousada, que se não for atendida, o resultado será pobre e ineficaz. Não procuro uma linguagem transparente. Ao contrário, o leitor tem de ser chocado, despertado de sua inércia mental, da preguiça e dos hábitos. Tem de tomar consciência viva do escrito, a todo momento. Tem quase de aprender novas maneiras de sentir e de pensar. Não o disciplinado – mas a força elementar; selvagem. Não a clareza – mas a poesia, a obscuridade do mistério, que é o mundo. E é nos detalhes, aparentemente sem importância, que estes efeitos se obtêm. A maneira-de-dizer tem de funcionar, a mais, por si. O ritmo, a rima, as aliterações ou assonâncias, a música subjacente ao sentido – valem para maior expressividade (in MARTINS, 2001, p. ix).

> Os sertanejos de Minas Gerais, isolados entre as montanhas, no imo de um Estado central, conservador por excelência, mantiveram quase intacto um idioma clássico-arcaico, que foi o meu, de infância, e que me seduz. Tomando-o por base, de certo modo, instintivamente tendo a desenvolver suas tendências evolutivas, ainda embrionárias, como caminhos que uso (ROSA in DANIEL, 1969, p. 91).

Ou ainda: “Aprendi algumas línguas estrangeiras apenas para enriquecer a minha própria [...]” (ROSA *in* COUTINHO, 1991, p. 87). Relatos como esses de Rosa nos estimulam a verificar se houve algum momento menos ou mais produtivo, mais repetitivo ou mais diversificado em termos lexicais no seu projeto literário, ou ainda, um momento em que se deu um maior amadurecimento.

Em *Mínima mímica*, Walnice Nogueira Galvão, no capítulo intitulado “As listas de palavras”, discorre sobre as folhas avulsas (manuscritas ou datilografadas) do Arquivo Guimarães Rosa que está sob a guarda do Instituto de Estudos Brasileiros da Universidade de São Paulo. No trecho abaixo, quando analisa o método de trabalho de Rosa, Galvão expõe uma dúvida muito pertinente ao nosso trabalho:

> Ora, as que podem [as folhas avulsas] ser rastreadas até textos definitivos colocam-se tardiamente na cronologia da obra: não há folhas avulsas relacionadas aos primeiros e mais volumosos livros, inclusive o único romance. Deve-se atribuir este fato ao nascimento de uma consciência da importância dos prototextos e paratextos, devido à fama, ou deve-se pensar que o escritor sofreu uma evolução em seus métodos de trabalho? (GALVÃO, 2008, 155-56).

A autora descreve as folhas avulsas em grupo, respeitando o critério da mais simples à mais complexa, ou seja, por listas de palavras, locuções de diferentes tipos, elencos de títulos, notas de leitura, lembretes, e observações de pesquisa de campo. No primeiro e no segundo grupos de documentos analisados em sua pesquisa no Arquivo, Galvão define quatro frentes de trabalho de elaboração por parte de Rosa: seleção no eixo paradigmático, dando privilégio às palavras mais raras em relação ao léxico da língua portuguesa[41]; intervenção no eixo sintagmático, por meio de resumos de verbetes; diferentes ocorrências de vocábulos adaptados ou apropriados de forma diferenciada ao usual dicionarizado e, por fim, criação.

O terceiro grupo de exemplares apresentado pela autora se compõe de locuções acompanhadas ou não por verbete explicativo, não se trata mais apenas de vocábulos, mas de algumas expressões idiomáticas do português como, por exemplo, “sacudir o sono – acordar” (GALVÃO, 2008, p. 158). O quarto grupo traz exemplos de

41 Estudaremos as palavras raras de Rosa em relação ao *corpus* como índices de *hapax* no capítulo 6.

locuções destacadas como se fossem citações ou notas de leitura e não apresentam um verbete que resuma a ideia. O quinto grupo se caracteriza por abordar listas de palavras ou locuções, mas trata de um único e peculiar assunto: a fabricação artesanal do polvilho, que será aproveitado no conto "Substância". Para a autora, é possível verificar nesse grupo de exemplares um caráter de prototexto mais complexo por se revelar próximo ao que se dará na narrativa final. Os dois últimos grupos analisados trazem orações bem lapidadas (como matéria-prima) e prontas para inserção em alguma narrativa. Resumindo, o trabalho de Walnice N. Galvão apresenta um método básico de criação adotado por Rosa: coleção de palavras isoladas ou sintagmas e unidades frásicas para uso imediato ou em lista de espera para futuro emprego, praticamente, um banco de dados textuais.

Partindo do trabalho de Galvão e dos próximos que apresentaremos, percebemos que comentar a obra de Guimarães Rosa é quase inevitável sistematizar, listar ou contar palavras (mesmo em números pequenos e sem o auxílio de computador). Veremos isso, nos próximos parágrafos, pois a crítica rosiana não escapa desse processo de garimpar as palavras de Guimarães Rosa.

Por isso o trabalho do estatístico textual também tem muito a contribuir com as leituras já realizadas, tanto para reafirmá-las, complementá-las ou refutá-las. Muitas dessas leituras que trataremos têm caráter quantitativo, o que viabiliza o diálogo com os programas de análise estatística. Afirmações a respeito da prosa rosiana, tais como "profusão desnorteante do seu vocabulário, dos mais ricos que já manejou um prosador de língua portuguesa" (MARQUES, 1983, p. 101) animam leitores e pesquisadores em literatura e linguística a se aproximar dos textos rosianos de modo a verificar que profusão é essa e como ela se dá. Uma das virtualidades do texto rosiano é a criação de palavras novas a partir de prefixos e sufixos, é o que veremos no item a seguir.

5.1 A QUESTÃO DOS PREFIXOS E SUFIXOS

Selecionamos alguns exemplos de pesquisas que podem propiciar essa parceria entre crítica literária e estatística textual, contudo, nem todas que citamos aqui serão analisadas no *Hyperbase*, elas são trazidas apenas para ilustrar a potencialidade desse tipo de estudo. A começar pelo estudo de Oswaldino Marques (1983) que aponta alguns recursos de verbalização mais frequentes nos textos rosianos, explorando alguns processos de construção dos termos forjados por Rosa, sempre os

relacionando a critérios semânticos. Vejamos alguns exemplos selecionados por Marques:

Prefixo "des" ou "de": desterrestre, desviveu, descrespo, desmoverem-se, despaga etc.; retroação (desavança, degressivo); restabelecimento de situação anterior (desescorregar); caracterização de estado, com tênue intensificação (desvago).

Sufixação: abundância, plenitude (almado); tendência, inclinação, abandono (sonhosa); sugestão de esguicho, com uma intensa componente sonora: (escorrijo); pomosas, manhanil, velhorro, soproso, crispim, ruivim, herculesco, ninhagem, boólatras, feeril, solsim etc (MARQUES, 1983, p. 102-6).

Para Marques (1983), Rosa alcança alta originalidade na maneira como maneja os prefixos consagrados da língua portuguesa, com algumas exceções como *van, ja, ber*, alcançando efeitos estilísticos ao se utilizar de outras modalidades de verbalização que importam superposições e cruzamentos semânticos como: *obluz, dismenso, admugem* e *desterrestres*.

Na sua tese, Mary Lou Daniel (1968) afirma que a prosa rosiana é mais rica em léxico coloquial que erudito e aposta, de modo intuitivo, em dados quantitativos:

> É interessante notar que existe maior porcentagem de palavras brasileiras e regionais em *Sagarana* do que nas outras obras do autor, e que decai sucessivamente esta porcentagem até atingir o seu ponto mais baixo nas Primeiras estórias. A escassez de tais elementos nesta obra está em relação direta com o caráter e o assunto das *estórias* do volume, sendo estas as mais internacionais e menos regionais de toda a obra rosiana, em vivo contraste com grande parte de sua prosa anterior (DANIEL, 1968, p. 28).

Os processos específicos empregados no léxico por Rosa para a elaboração de neologismos, segundo Daniel, são de natureza analítica (afixação, prefixação —, elementos preposicionais de origem grega ou latina, sugerindo uma significação para cada uma delas). Ela os distingue em dois tipos de prefixos que resumimos deste modo:

Prefixos direcionais: *para-* (indecisão ou vagueza no movimento), *per-* (inclinação irregular ou sem direção fixada, que vai além do significado normal do termo), *por-* (movimento pelo meio ou através), *pro-* (sentido geral do movimento), *so-* (conotação de movimento para baixo ou de tipo sub-reptício) e *trans-* ou *tra-* (ideia

mista de *através* e *por*) —, mudança interna e abreviação de palavras, e derivativa (mistura, influência ou criação interparadigmática e de formação popular).

Prefixos de intento: *em-* ou *en-* (sugere participação comum objetiva); *es-* (elaboração ou resultado de um processo ou capacidade intensificadora); *con-* (conotação de participação simultânea ou subjetiva na ação); *pes-* (intensivo); *tres-* (intensificador absoluto). A autora indica ainda os prefixos mais usados por Rosa: *re-* (como função intensificadora ocorre[42] "umas dezoito vezes na obra rosiana" (DANIEL, 1968, p.37)); *de-* (empregado "mais de vinte vezes" como neologismo, mas não é utilizado em *Sagarana* nem em *Primeiras estórias*, mas em *Grande sertão: veredas* e *Corpo de baile* desempenha outras funções); *a-* ("ocorre mais de 50 vezes" (p. 37)) tem função direcional, derivada do seu emprego etimológico *ad*, mas na maioria dos casos tem intento de ênfase pleonástica); e *des-* ("prefixo negativo de maior frequência na obra rosiana [...] o qual ocorre umas quarenta vezes" (p. 39), esse prefixo, por vezes, denota inversão ou substituindo o prefixo *in-*, por exemplo, 'desfeliz' ou desempenha função intensiva). A autora ainda afirma que:

> É significativo o fato de serem estes os prefixos prediletos do autor, pois são os mais inerentes à língua portuguesa e os mais empregados e abusados na expressão coloquial. Guimarães Rosa escolhe como os recursos primários do seu artesanato, não os elementos exóticos e eruditos, senão os normais e familiares, para lhes dotar de vida nova nas suas novas funções. (DANIEL, 1968, p. 37).

Não é o nosso caso tratar dos sufixos e prefixos neste trabalho, mas vale ressaltar que há possibilidades de busca de ocorrências por prefixação ou sufixação em programas de análise estatística de textos. Dessa forma, o pesquisador consegue extrair todas as palavras e listá-las de modo a ter garantido todas as ocorrências reunidas e, por consequência, analisar quais as tendências estilísticas (palavras raras ou repetidas), substituindo a leitura intuitiva e manual pela leitura informatizada, que garante maior abrangência e precisão.

---

42 Não há, em estudos de vocabulário desse teor, como escapar das contagens, Daniel no final da década de sessenta já estava fazendo, de certo modo, seu levantamento estatístico textual, em várias passagens do texto, ela indica quantidades de ocorrências.

5.2 A QUESTÃO DO REGIONALISMO

Outro tema muito desenvolvido pela crítica rosiana é o regionalismo. Para tratarmos dessa questão, começaremos com a *História concisa da literatura brasileira* de Alfredo Bosi (1994). Considerando os sertanistas que partiram do filão de José de Alencar, o crítico indica os seguintes nomes: Bernardo Guimarães (primeiramente com *O Ermitão de Muquém*, de 1858), Alfredo D'Escragnolle Taunay (com *Inocência*, de 1872) e Franklin Távora (com *O Cabeleira*, de 1876). Ressalta o primeiro pela mistura de elementos da narrativa oral com uma dose de idealização; o segundo, pela demonstração da cultura e do temperamento mais sóbrio ao regionalismo romântico; e o terceiro, em formato de manifesto e reivindicação, sendo mais vigoroso com o critério da verossimilhança. O crítico afirma ser o *sertanismo* variado, apresentando característica dentro do estilo romântico, naturalista, acadêmico ou até mesmo modernista (BOSI, 1994, p. 141). Sabemos que essas diferentes formas nasceram do contato da cultura urbana e letrada com a matéria bruta brasileira, ou seja, o arcaico, o rural, o provinciano. Esse *sertanismo* resultou quase sempre em uma prosa híbrida, cujo prosador não consegue fundir, artisticamente, seus métodos ideológicos e estéticos com a vida rural elegida para o retrato. Por tais características, Bosi conclui que o regionalismo está fadado a dividir-se em dois extremos: a busca pelo registro puro da fala regional, considerada por ele como uma "concepção ingênua de realismo" (Visconde de Taunay, com alguns trechos de *Inocência*, é exemplificado para esse caso, acompanhado de Valdomiro Silveira e Simões Lopes Neto); e o resgate de formas que atuam na expressão da vida rústica e que sofrem uma "reelaboração" para o entendimento do leitor culto. Dessa reelaboração, Bosi cita o trabalho empenhado por Rosa, que ele chama de "invenção revolucionária", afirmando que Rosa "conseguiu universalizar mensagens e formas de pensar do sertanejo através de uma sondagem no âmago dos significantes" (BOSI, 1994, p. 141). Para Bosi (1994), tal trabalho apresenta características experimentais da arte moderna, exigindo maior nível de abstração por parte do leitor.

No prefácio da obra completa de Rosa publicada pela editora Nova Aguilar, Eduardo Coutinho recorda que, quando *Sagarana* foi lançado, uma das ficções que predominava no Brasil era a do romance do Nordeste marcado fortemente pelo caráter de protesto e calcado numa linguagem descritiva, voltada para o convencional. Ciente disso, segundo o autor, Rosa tomou como sua principal atividade literária a revitalização da linguagem; e fez isso violando regras estabelecidas,

atingindo o leitor por meio de sua originalidade que foge do lugar-comum, abrindo mão de formas e estruturas fossilizadas. Para Rosa, a linguagem corrente se desgastou no uso, e, consequentemente, não expressava ideias, apenas clichês. Sua missão foi então buscar o novo na escrita e a renovação poética para Rosa está em fazer o leitor estranhar e refletir a todo o momento, assumindo um papel importante de participação ativa na leitura.

A respeito da prosa rosiana sobre o caráter regionalista e sua linguagem, Antonio Candido (et al. 2011, p. 20-21) diz:

> [...] a linguagem dele não era propriamente documentária, o que acontece no regionalismo. A impressão que se tinha é que ele estava criando uma linguagem. Eu não tinha formação linguística para saber até que ponto, mas senti que ele estava inventando uma linguagem que ao mesmo tempo era plantada na região, mas estava ligada, por exemplo, ao passado da língua portuguesa – que a região tem, aliás, um certo arcaísmo – e a uma criação dele, uma criação de palavras, uma invenção, uma coisa que acontece muito na língua alemã, a pessoa pode fundir meias palavras, palavras, para fazer uma nova, e ele fazia muito isso.

Não temos dúvida que o teor regional é traço estilístico do autor, para Daniel (1968) o ruralismo é um fator importante para o viés de sua leitura e portanto, imprescindível para a nossa análise, pois ele é peça determinante nas pesquisas das estudiosas em que nos apoiamos neste trabalho.

5.3 A QUESTÃO DA REVITALIZAÇÃO DA LINGUAGEM

Na visão de Coutinho, o processo de revitalização da linguagem executado por Rosa pode ser verificado em dois âmbitos, o vocabular e o sintagmático. No primeiro âmbito encontram-se as afixações e aglutinações como "sozinhozinho[43]" (ao perceber a inexpressividade do vocábulo, Rosa reaviva o seu significado originário, servindo-se do mesmo processo que acreditava ter sido utilizado um dia: repetiu o sufixo diminutivo no final e criou a forma "sozinhozinho"). Para o procedimento de aglutinação, comentado na citação anterior de

[43] Exemplos retirados de *Grande sertão: veredas* por Coutinho.

Candido, Coutinho exemplifica com o caso de palavras como: “nenhão” (nenhum+não); “fechabrir”; “prostitutriz”; “adormorrer”. O autor repara que nesses casos exemplificados se observa alteração ou recriação de significante, mas nunca a invenção de significantes totalmente novos que estejam dissociados das formas existentes da língua, assim como o exemplo comentado por Haroldo de Campos (*et al* 2011, p. 54):

> [...] a expressão “Num nú” [referindo-se à frase ‘Num nú, os adversários se engalfinharam’] é uma tradução que ele faz direta do alemão, onde existe a frase, *im Nu*, significando num átimo, num momento, e traduz isso diretamente para o português, e nós pensamos na palavra nu como nudez, e aquela coisa de “Num nú” qualquer pessoa entende, sem explicação de dicionário, que aquilo significa num repente, num momento, num átimo.

No âmbito sintagmático, explicado por Coutinho, por vezes surgem sentenças inteiras de clichês[44] que ganham outra expressividade, tais como: “nu da cintura para os queixos” (nu da cintura para cima); “não sabiam de nada coisíssima” (não sabiam coisíssima nenhuma).

Outros procedimentos apontados por Coutinho:

> [...] enumeração de palavras pertencentes à mesma classe gramatical e ao mesmo campo semântico, que introduz uma ruptura na estrutura sintagmática do discurso, e contribui para uma espécie de neutralização da oposição entre prosa e poesia; a inversão da ordem tradicional dos vocábulos e sintagma na oração, que constitui talvez o traço mais erudito do estilo do autor e o responsável, em grande parte, pelo rótulo que diversos críticos quiseram emprestar-lhe de neo-barroco; e o uso de orações justapostas e construções elípticas, típicas da linguagem oral, que revelam uma preferência acentuada pela coordenação sobre a subordinação e por um tipo de estilo fluido, linear e direto (COUTINHO in ROSA, 1993, p. 16).

---

44 Aqui podemos fazer uma referência ao ensaio de Walnice Nogueira Galvão sobre as listas de palavras que Rosa costumava colecionar, que comentamos no capítulo 5 deste trabalho.

Em *Caos e Cosmos* (1976), ao tratar das leituras filosóficas de João Guimarães Rosa, Suzi Frankl Sperber mostrou, por meio de recursos comparativos, que a somatória de leituras permite compreender o processo de crescimento do autor e que isso se reflete na estruturação da obra (do texto) e no estilo de forma mais saliente do que na temática. Segundo Sperber (1982), as leituras e preocupações espirituais de Rosa atuavam significativamente na técnica, no estilo e na linguagem de seu fazer literário. Mas é em *Signo e Sentimento* que sua análise parte para um viés que se relaciona mais com a ideia deste trabalho. Sperber (1982) inicia por *Sagarana*:

> Pretendo apresentar a formação de visão literária e dos modismos criadores de Guimarães Rosa, a partir de *Sagarana*, em função do que vimos em *Caos e Cosmos* como material inspirador e em função dos arranjos que o Autor achou para ordenar os temas centrais de sua obra, que incluem a linguagem por ele forjada. Pareceu-me que só a partir deste estudo seria possível entender a contribuição de Rosa na literatura brasileira. Mas como fazer o estudo da linguagem forjada? (SPERBER, 1982, p. 3).

Ao longo de sua análise, a autora questiona se é na diferença que se encontram os dados suficientes para uma compreensão das formas de combinação e organização de um texto. Para ela, as opções do autor são mais facilmente notadas na "diferencialidade". Partindo de comparações entre a obra rosiana e as possíveis fontes que influenciaram Rosa em sua ficção, Sperber (1982) estuda o fenômeno da organização da linguagem de Rosa e indica que, ao longo da evolução de um mesmo texto, catálises e distaxias são descobertas nos sintagmas, fenômeno que ela nomeou de palavras-instrumento[45]. Segundo a autora, "o efeito de distorção dos elementos da narrativa corresponde à distorção dos elementos do sintagma" (SPERBER, 1982, p. 6) e essa distorção (daí a diferencialidade, pois ao buscar estratégias de organização da ficção, Rosa inseria uma diferença ou suprimia os traços herdados de uma cultura a respeito de um sentido) ocasiona dificuldade na leitura, a ação

[45] Aprofundaremos mais esse estudo de Sperber numa abordagem estatística no capítulo 6. Esses procedimentos surgiram de acordo com o amadurecimento de escrita do autor, isso se relaciona diretamente com os estudos de evolução lexical que faremos no próximo capítulo. Contudo, salientado a diferença de que Sperber tratou da evolução da versão manuscrita à publicada de uma mesma obra. Nossa análise parte da evolução da primeira obra publicada à última.

da narrativa é disfarçada, camuflada por palavras-instrumento que propiciam exigência de uma leitura muito mais atenta, haja vista a própria definição da linguagem dada por Rosa: "Não procuro uma linguagem transparente. Ao contrário, o leitor tem de ser chocado, despertado de sua inércia mental, da preguiça e dos hábitos. Tem de tomar consciência viva do escrito, a todo momento." (in MARTINS, 2007, p. ix).

Comparando os contos de *Sagarana*, a versão primitiva (*Sezão*, primeira elaboração) e a publicada, Sperber (1982) repara em uma economia da escrita que se dá inicialmente por meio de eliminação das palavras estrangeiras ou rebuscadas. A autora considera tal procedimento como um "amadurecimento de consciência de nacionalidade da linguagem" (SPERBER, 1982, p. 30). *Signo e Sentimento* entra em nossa discussão por apresentar processos de escolhas (a economia da escrita comentada anteriormente) que Rosa fez durante as reelaborações de *Sagarana*. Questionamo-nos, então, se esse amadurecimento ocorreu com todos os textos, no seu léxico propriamente dito, ou seja, se esse processo de refinamento de vocabulário aconteceu também ao longo da produção da obra. Para detectar de maneira efetiva esse processo, vamos utilizar ferramentas estatísticas que nos apresentem dados sobre crescimento e a evolução de vocabulário[46].

Para o conto *A hora e a vez de Augusto Matraga*, Sperber (1982) ensaia uma contagem, bastante simples, sobre as conjunções aditivas e adversativas: *e*, *ou*, *mas*, *porque*, *que*, *porém* e *pelo que*. Pelo resultado obtido, ela afirma que a ênfase dessas conjunções traduz um sentido de causalidade como processo fundamental do relato (do conto específico). Diz a autora:

> Nas cinco primeiras páginas do conto temos um total de 57 "e", 18 "ou", "mas" e "porém" e 8 "porque" ou equivalentes causais. Comparando-se com outros contos de Sagarana, com "O burrinho pedrês", por exemplo, notamos uma diferença considerável. Nas cinco primeiras páginas de "O burrinho pedrês" há um total de 53 "e", 13 "ou", "mas", e "porém" e 6 "porque" ou equivalentes causais. Porém, "porque", propriamente dito, há 3 em "O burrinho pedrês" e 6 em "A hora e a vez de Augusto Matraga". A

46 Os dados e seus resultados podem ser observados no capítulo 6.

> ênfase ao sentido da causalidade, propiciada por este processo é fundamental para o relato como todo. (SPERBER, 1982, p. 42).

O diferencial, apontado por Sperber (1982) acerca do emprego das conjunções coordenativas aditivas e adversativas, está na posição onde a conjunção se encontra na frase e sua frequência em tal posição. Ambos os contos (*A hora e a vez de Augusto Matraga* e *O burrinho pedrês*), segundo a autora, mostraram o emprego de conjunções coordenativas aditivas e adversativas no início de períodos, elemento de escrita que não pode ser ignorado quando a intenção é radiografar os traços estilísticos significativos do autor.

No terceiro capítulo de *Signo e sentimento*, Sperber analisa *Corpo de baile*, ainda sob o viés da articulação entre signos e sintagmas, e dá atenção especial a *Campo geral* tratando da conjunção coordenativa *mas*. Ela percebe, igualmente, maior incidência da conjunção no início de frases:

> O ‘mas’ revela-se restritivo. Esta restrição pretende coordenar sintagmas que em princípio dispensavam esta articulação. Ora, a articulação forçada quer explicar o que não se explica. Deste modo, buscamos, nós, leitores, compreender o sintagma anterior através do seguinte. Consequentemente, o sentido do sintagma anterior, em si completo, é minimizado, é limitado pela restrição da conjunção coordenativa adverbial inicial (SPERBER, 1982, p. 72).

Sobre a comparação entre o esboço e a finalização de *Grande sertão: veredas*, Sperber informa que Guimarães Rosa renova a linguagem pelo uso que faz de um léxico arcaico (do sul da Bahia e norte de Minas), criando neologismos que, segundo a autora, parece ser uma prática do próprio povo dessa região (SPERBER, 1982, p. 72). Por fim, questiona: “Qual o sentido da evolução estilística e narratívica de Guimarães Rosa?” (SPERBER, 1982, p. 93). Essa é justamente a questão que retomaremos ao longo dessa pesquisa, pois, se existe uma evolução na escrita, é preciso determinar quais elementos fazem parte desse processo e de que forma ela transparece no material linguístico.

5.4 COVIZZI E A LINHA TEMPORAL DA PRODUÇÃO LITERÁRIA DE ROSA

Do percurso que fizemos a partir da leitura de Sperber (1982), tentamos ressaltar os comentários e relacioná-los ao crescimento e à evolução de vocabulário. Para que possamos analisar essa evolução em termos estatísticos, é necessário, como já dissemos anteriormente, inserir os textos nos aplicativos de modo a respeitar uma cronologia. Sabemos que Guimarães Rosa teve muito apreço e dedicação a cada vocábulo de sua criação literária, haja vista o caso de *Sagarana* que foi lapidado pelo autor, praticamente durante dez anos seguidos depois de sua primeira publicação. Então nos perguntamos: qual seria a melhor maneira de dispor os textos nos aplicativos para uma análise de evolução de vocabulário? A versão da primeira publicação? A versão da última publicação em vida do autor? A versão da primeira elaboração do texto? Para a nossa pesquisa, decidimos pela última edição publicada em vida pelo autor.

Como dissemos na introdução, pensar na cronologia da produção literária de Rosa nos levou a buscar apoio nos estudos de Sperber (1976), de Covizzi (1978) e de Daniel (1968), pois todas elas retratam de alguma maneira o aspecto cronológico da obra rosiana. A primeira propõe leituras diacrônica e sincrônica da obra; a segunda apresenta uma linha do tempo que se caracteriza inicialmente por uma forte expressividade que finaliza apresentando um caráter explicativo da obra (que logo veremos, teria inicialmente uma grande repercussão no âmbito da expressão, ou seja, com alto índice de elaboração de novos vocábulos), e por último, Daniel, que separa a obra de Rosa em duas fases: rural e urbana - por esse motivo, pontuamos a questão do regionalismo no item anterior (4.2).

Suzi Sperber (1976), em sua grande empreitada de fichar mais de 1.000 livros da biblioteca-espólio de Guimarães Rosa, toma por base o trabalho de Benedito Nunes em *O dorso do tigre*, publicado originalmente em 1969. A partir disso, ela faz um inventário de comparações entre as obras filosóficas, indicadas pelas leituras de Nunes, nas entrevistas com Guimarães Rosa e na própria obra deste. Diz ela que:

> a obra em si deve ser vista não só sincrônica, como também diacronicamente, uma vez que não é homogênea do primeiro ao último livro publicado [...] o inventário lexemático-filosófico

> deve incluir a si (como lexema básico filosófico) e às atualizações na obra, de modo a podermos avaliar as articulações dentro de cada livro (sincronia) e de livro a livro (diacronia) (SPERBER, 1976, p. 16).

O outro estudo que nos serve de base é o de Lenira Marques Covizzi (1978), pois trata de dois escritores latino-americanos comparativamente sob a ótica do insólito. Ela determina alguns aspectos "estranhos" que surgem na ficção de Guimarães Rosa e Jorge Luis Borges, no que diz respeito à significação. Por insólito, a autora compreende: "a inadequação irrefutável da realidade perceptual e a sua representação artística [...] por se utilizarem de expressões não realistas" (COVIZZI, 1978, p. 46).

Contudo, o interessante para a nossa pesquisa é que Covizzi (1978) cria uma linha de curso da narrativa, remetendo ao caráter metalinguístico da obra de Guimarães Rosa. Ela afirma que a obra rosiana segue uma linha, um percurso que vai da expressão (arte superior) à explicação (comentário dessa arte). Na apresentação do estudo de Covizzi (1978), João Alexandre Barbosa (1978) comenta que não somente a partir de *Primeiras Estórias* Guimarães Rosa se apropria de uma metalinguagem, ela já existiria antes, de maneira implícita; depois, surge quase como "uma obsessão ou motivo de sua atividade criadora" (in COVIZZI, 1978, p. 18).

Ainda sobre essa linha de percurso da narrativa, Covizzi (1978) afirma que as duas obras (*Grande sertão: veredas* e *Corpo de baile*) abordam universos muito elaborados (do ponto de vista da linguagem e do ambiente rural) que ganharam densidade a partir de *Sagarana* até chegar aos textos de 1956. A partir de então, houve um processo de diluição do teor regional ou rural como em *Primeiras Estórias* e um aumento considerável na liberdade e na violentação da linguagem, diz a autora:

> *Tutaméia*, na mesma linha de produção das *Primeiras Estórias*, é composta por quarenta pequenas estórias [...] onde a ambientação rural já parece funcionar apenas como ilustração das preocupações que determinam o universo rosiano e como pretexto para exercer com mais liberdade e violentação da linguagem, que encontra menor verossimilhança se situada na cidade, porque seu público leitor é urbano. [...] Nessa linha ainda deve ser enquadrada *Estas Estórias*, mais

> próximas das *Primeiras Estórias* [...]. (COVIZZI, 1978, p. 59-61).

Para Covizzi (1978), o processo explicativo — o comentário da expressão da narrativa rosiana — se inicia a partir de *Primeiras Estórias*. A estreia de *Sagarana* foi ainda seguida de outras grandes expressões como em *Grande Sertão: veredas* e *Corpo de baile*, e depois desse procedimento criativo, Rosa revelaria a sua ficção, passaria então ao processo explicativo da obra:

> Uma ficção que sempre quis ser busca, procura, se esclarece a si mesma, atribuindo-se o direito de responder ao processo que a gerou. É uma ficção que se volta sobre si mesma, logo, de caráter metalinguístico, que ocorre de duas maneiras. A primeira é explícita, manifestando-se na referência discursiva ou alusiva de caráter crítico que já ocorre desde *Sagarana* [...] E há uma outra menos evidente, que se desvelou para nós quando procedemos à análise comparativa de alguns relatos das *Primeiras Estórias* (COVIZZI, 1978, p. 61-62).

Seria, em termos ilustrativos, aproximadamente assim:

Sagarana

Primeiras Estórias/Terceiras Estórias

linha ficcional da obra de Guimarães Rosa

EXPRESSÃO EXPLICAÇÃO

Barbosa (1978) questiona por que não pensar nas últimas obras como um outro período de experimentação de Guimarães Rosa com relação àquilo que ele produziu inicialmente (BARBOSA in COVIZZI, 1978, p. 18) e é, do mesmo modo, o que a própria Covizzi (1978) afirma:

> [...] a partir de 1962 Guimarães Rosa inicia a elaboração do reverso da moeda que é a sua produção anterior. Ou seja, inicia, ainda através da ficção, a explicação de sua criação anterior.

> *Primeiras Estórias* é o início da explicação, do avesso de sua ficção no seu caráter metalinguístico, de produto da imaginação e no de sua particular visão do mundo.
> *Terceiras Estórias* continua nesse esquema, enfatizando o caráter metalinguístico.
> *Estas Estórias* continua ainda no mesmo esquema, enfatizando sua perspectiva mística de compreensão do mundo. (COVIZZI, 1978, p. 62).

Nos interessa avaliar esse percurso da criação ficcional, explorando os dois caminhos que Covizzi sugere, para verificarmos quantitativamente esse movimento em termos lexicais. Covizzi (1978, p. 83) também não escapa das listagens de palavras que ela considera como novidades de linguagem (na maioria, retiradas de *Primeiras Estórias*): descrevivendo-as; bis-viu; beladormeceu; deligentil; grimpava-a; muralhavaz; fixibilidade; frondosura; versão voxpopular; inacionais hinos; engenhingonça; excelentriste; abusufruto; milmaravilhoso; ultramuito; desnascer; mãos de enxadachim; redesimportância; já requiescante; capisquei; voz tonifluente; psiquiartista; psiquiatrista; esmarte; artimanhoso.

Para Covizzi (1978), é em *Terceiras Estórias* que encontramos um caminho que se volta para a ficção de Guimarães Rosa. Para ela, o autor se utiliza de recursos tais como redução e crítica da própria obra. O caráter explicativo, para ela, se dá no nível da linguagem após a criação exercida em *Grande sertão: veredas*. Como se essa obra retratasse um ideal de linguagem, uma 'característica fundante' e que a partir dela, toda a obra não pudesse retroceder tal nível de linguagem:

> comparações, definições, explicações redundantes, acúmulo de máximas, provérbios populares, teorização através de prefácios e da própria ficção, auto-avaliação implícita nos prefácios, reticências, síncopes, insistências obsessiva nas fugas às convenções através da violentação do pensamento lógico e da violentação de palavras e expressões estabelecidas, obsessão interrogativa, reiterações temáticas tais como a da loucura e cegueira, conflitos entre sentimentos absolutos tais como ódio/amor, presença do humor etc., numa evidente tentativa de alisar as arestas de sua matéria ao enfatizá-las. (COVIZZI, 1978, p. 84).

Para a tarefa de análise estatística sobre os dados apontados por Covizzi, faremos, inicialmente, um levantamento de vocabulário por ocorrências de *hapax* do *corpus*, para, em seguida, verificarmos se as obras iniciais representariam mais essa expressividade criativa; por outro lado, se, na fase final da produção, o índice de *hapax* for menor, consequentemente houve uma redução na criação de vocábulos novos. Acreditamos que se houver esse excedente de vocábulos, de *hapax legomena* e riqueza lexical na primeira fase em relação à final, isso já seria um indício[47] de que aquela possa ter sido bastante "expressiva" em termos lexicais.

Passemos agora para outro estudo em que também nos apoiaremos para a composição desta tese. Trata-se do já mencionado trabalho de Mary L. Daniel (1968) intitulado *João Guimarães Rosa: Travessia Literária*, um estudo sistemático do estilo linguístico, que se dá por três métodos de análise descritiva: sobre aspectos do léxico, do nível sintático-gramatical e dos elementos poético-religiosos. Daniel (1968) aponta duas vertentes na obra de Rosa, uma de caráter predominantemente rural (*Sagarana*, *Corpo de baile*, *Grande sertão: veredas*) e outra de caráter urbano (*Primeiras Estórias*), sendo essa menos oral. Ela estabelece uma progressão qualitativa, que, de obra em obra, anuncia o aumento do potencial comunicativo e indica *Primeiras Estórias* como a obra mais madura do autor. Faz-se necessário lembrar aqui que a tese da autora foi contemporânea à produção de Rosa, resultando em uma análise somente até essa obra, ou seja, o estudo não abarcou sua obra completa[48].

Para o exercício quantitativo sobre a análise de Daniel (1968), temos como premissa verificar se essa distinção entre o rural e o urbano se reflete no léxico. Para isso, utilizaremos uma ferramenta que mede a evolução e a distância lexicais de um texto a outro, pois ela detecta essa diferenciação, caso exista e reflita no vocabulário.

A distância lexical é uma das análises feitas pelo programa *Hyperbase* que considera o vocabulário completo de cada um dos textos do *corpus*, não se preocupando mais com a frequência dos vocábulos, mas apenas com a presença ou a ausência de uma determinada palavra

---

47 Contudo, resta-nos ainda pensar em como verificar o momento "explicativo" da obra, pois se o mesmo for também "expressivo" em seu léxico, o procedimento de análise deverá ser repensado.

48 Após a publicação de *Tutaméia*, Daniel (1968) publica em nova edição de sua pesquisa um anexo no qual ela explora alguns aspectos da obra, dando continuidade ao método empregado nas obras anteriores.

no texto em análise. Busca, igualmente, a conexão entre dois textos ou mais, dependendo do tamanho do *corpus*, por meio de uma palavra, pois ela pode contribuir com a aproximação desses textos (se ela for comum aos dois) ou aumentar a distância (caso essa palavra seja exclusiva de um texto apenas).

Em suma, neste breve capítulo, procuramos trabalhar com alguns estudos que colaboraram para nossa reflexão voltada à análise estilística sobre a matéria-prima da escrita, da linguagem do ficcionista Guimarães Rosa, especulações que têm muito a nos oferecer em termos de estudos estatísticos. Para que áreas tão distintas como literatura e estatística comecem a fazer mais sentido para o nosso trabalho, apresentaremos as afinidades que ambas podem ter — lembrando que seus resultados mais explorados, encontram-se no capítulo a seguir. Dispostos os exemplos da crítica literária, veremos os gráficos, os resultados bem como as análises também no próximo capítulo. Nele trataremos dos aspectos quantitativos gerais e específicos da obra completa de Rosa e, além de praticar o exercício da estatística textual sobre os trabalhos da crítica supracitados, ou seja, nos estudos de Sperber (1982), Covizzi (1978) e Daniel (1968). Discutiremos os aspectos do vocabulário rosiano e os possíveis problemas e soluções de leitura que podemos indicar por meio da estatística textual sobre esse *corpus* literário.

6 "A NHANINA SABE AS LETRAS MAS... NÃO DECORA OS NÚMEROS, DE CONTA DE SE FAZER..."

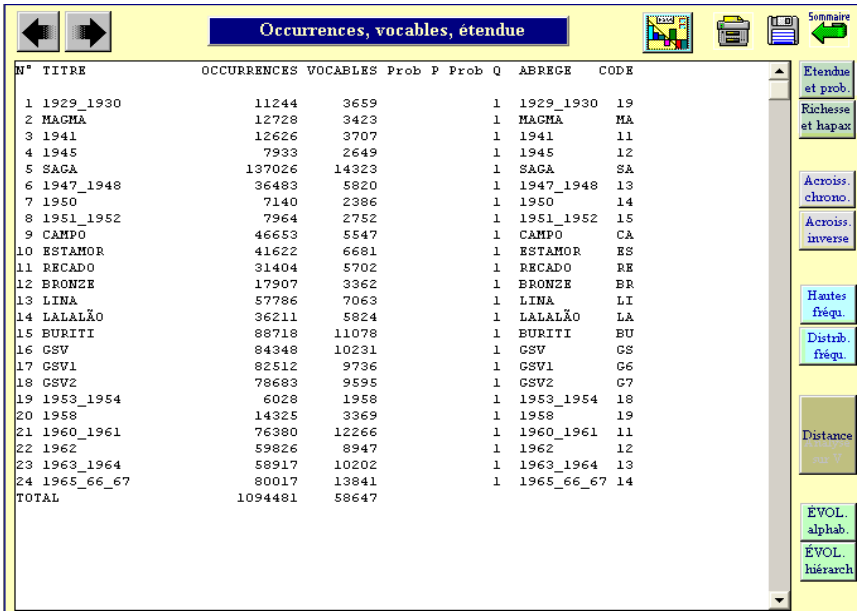
Milhões, bis, tris, lá sei, haja números para o Infinito.
Sobre a escova e a dúvida

Neste capítulo discorreremos, pelo viés estatístico, sobre a estrutura do vocabulário da obra ficcional de Guimarães Rosa. Cada subitem aborda uma ferramenta estatística do programa *Hyperbase* que oferece os dados necessários para o levantamento estatístico e documentário: extensão do vocabulário; riqueza do vocabulário e *hapax*; crescimento lexical; altas frequências; distribuição de frequências; distância lexical e evolução do vocabulário. Durante a descrição dos resultados, faremos relações sobre as hipóteses extraídas dos estudos de Sperber (1982), Covizzi (1978) e Daniel (1968).

6.1 CARACTERÍSTICAS GERAIS: EXTENSÃO DO VOCABULÁRIO

Para uma primeira apreciação, observaremos as informações básicas dos números gerados pelo aplicativo. Essa é a visualização que temos a respeito da extensão de vocabulário do *corpus*, ou seja, a quantidade total de ocorrências (N) e de vocábulos (V). No quadro a seguir, à direita, encontram-se os botões que acionam algumas ferramentas estatísticas que o *Hyperbase* possui e, acima no quadro, temos os botões de navegação do programa, bem como geração e impressão de gráficos. Na primeira coluna, lendo da esquerda para a direita, encontramos a lista de disposição das obras (24 textos no *corpus*); na segunda coluna, temos os totais de ocorrências de cada conjunto de textos; na terceira coluna, os totais dos vocábulos de cada conjunto; na quarta, temos os índices de probabilidades P e Q[49]; na quinta, apresentam-se os títulos abreviados dos conjuntos de textos; por fim, na sexta coluna, o código da cada conjunto de obras (uma forma mais abreviada de apresentação dos conjuntos). Vejamos então as primeiras características:

[49] Tratam-se de termos da distribuição binomial onde p = probabilidade de sucessos e q = probabilidade de falha que opera sobre um número x de eventos.



N°	TITRE	OCCURRENCES	VOCABLES	Prob P	Prob Q	ABREGE	CODE
1	1929_1930	11244	3659		1	1929_1930	19
2	MAGMA	12728	3423		1	MAGMA	MA
3	1941	12626	3707		1	1941	11
4	1945	7933	2649		1	1945	12
5	SAGA	137026	14323		1	SAGA	SA
6	1947_1948	36483	5820		1	1947_1948	13
7	1950	7140	2386		1	1950	14
8	1951_1952	7964	2752		1	1951_1952	15
9	CAMPO	46653	5547		1	CAMPO	CA
10	ESTAMOR	41622	6681		1	ESTAMOR	ES
11	RECADO	31404	5702		1	RECADO	RE
12	BRONZE	17907	3362		1	BRONZE	BR
13	LINA	57786	7063		1	LINA	LI
14	LALALÃO	36211	5824		1	LALALÃO	LA
15	BURITI	88718	11078		1	BURITI	BU
16	GSV	84348	10231		1	GSV	GS
17	GSV1	82512	9736		1	GSV1	G6
18	GSV2	78683	9595		1	GSV2	G7
19	1953_1954	6028	1958		1	1953_1954	18
20	1958	14325	3369		1	1958	19
21	1960_1961	76380	12266		1	1960_1961	11
22	1962	59826	8947		1	1962	12
23	1963_1964	58917	10202		1	1963_1964	13
24	1965_66_67	80017	13841		1	1965_66_67	14
TOTAL		1094481	58647				

Quadro 2: Dados de ocorrências, vocábulos e extensão.
Fonte: *Hyperbase* ©, versão 5.4.

Nossa base apresenta um *corpus*[50] de vinte quatro (24) divisões, (pela sua dimensão GSV foi divido em 3 partes) e um total de 1.094.481 ocorrências[51] e, segundo a contagem do aplicativo, podemos afirmar que o vocabulário de Guimarães Rosa se constitui de 58.647 vocábulos. Somando as partes de *Corpo de baile*, verificamos que ele é o texto de maior extensão em termos de vocabulário[52] possuindo 45.257 vocábulos. Na sequência temos *Grande sertão: veredas* com 29.562 vocábulos e *Sagarana* com 14.323.

50 Possui 4.889.367 caracteres.

51 Para uma melhor compreensão ou revisão da terminologia, indicamos a releitura do item "2.3 Terminologia de *corpus* estatístico e das ferramentas" que se encontra no capítulo 2 desta tese.

52 Dentro do total de ocorrências apresentado estão todas as repetições de vocábulos, e inclusive a pontuação, pois essa versão do *Hyperbase* considera como vocábulo cada caractere separado por espaço, isso indica que temos aí inclusos os números também e os sinais gráficos. Os outros textos que não tiveram destaque acima de 10.000 vocábulos não daremos importância nesse parágrafo.

6.2 RIQUEZA LEXICAL

A medida estatística que se baseia na relação do número de palavras repetidas e diferentes de um mesmo texto e o número total de palavras que o compõem é o que chamamos de riqueza lexical. Trata-se da razão entre número de palavras diferentes (vocábulos ou formas)[53] e o número total de palavras (ocorrências); sendo assim, podemos deduzir que quanto maior o número de vocábulos novos, maior será a riqueza e a variedade do vocabulário a ser estudado, caso contrário, mais repetitivo e restrito será o texto.

Maciel nos explica que a riqueza lexical:

> [...] *est un élément de structure du texte et est en rapport avec le thème de même qu'elle traduit des changements intervenus dans le style des oeuvres littéraires. Un passage narratif ou descriptif n'aura donc pas [...] la même structure lexicale qu'un passage dialogué ; la longueur de la phrase et la chronologie interviennent aussi et conditionnent à leur renouvellement*. (MACIEL, 1986, p. 85).

Do cálculo estatístico para a riqueza do vocabulário, Brunet (1988, p. 27) observa que:

> [...] *on mesure la part du vocabulaire théoriquement absent (et par la suite celle du vocabulaire théoriquement présent) dans chacun des textes. Cet effectif attendu est comparé à celui qu'on observe en réalité, et la distance est appréciée par un écart réduit."*

Para sermos mais didáticos, segue uma pequena análise de um trecho de "Campo geral" que exemplifica melhor a noção de riqueza lexical, a qual empregamos na estatística textual:

A mãe, quando ouvisse essa certeza, havia de se alegrar, ficava consolada. Era um presente; e a ideia de poder trazê-lo desse jeito de cor, como uma salvação, deixava-o febril até nas pernas. (ROSA, 1995, p. 465).

---

53 Nos estudos de linguística de *corpus* também é possível encontrar os termos *types* (vocábulos) e *tokens* (ocorrências).

Desse trecho, temos então:

- total de formas: 41
- total de formas diferentes: 33
Riqueza lexical (%) = 33 vocábulos / 41 ocorrências
33 x 100 / 41 = 80,48

Desse simples cálculo, concluímos que o trecho apresentado possui um percentual de variedade em seu vocabulário de 80,48%, ou seja, apenas 19,52% das palavras são repetidas.

Carlos Maciel (1986, p. 75) explica que a noção de riqueza lexical independe da presença ou ausência de algum vocábulo considerado raro numa obra literária. A riqueza contabilizada não deve ser confundida com nenhum juízo de valor, pois se trata de um elemento da estrutura do texto, correspondendo única e exclusivamente a dados quantitativos, e que carrega em si traços estilísticos. Um elemento que se relaciona diretamente com o conceito de riqueza lexical é o *hapax legomena*, ou seja, as palavras de um *corpus* que têm apenas uma ocorrência influenciam no resultado da riqueza lexical, e a proporção desses vocábulos não depende apenas de características estilísticas ou linguísticas diretamente, mas também do comprimento dos textos. "*En general, on peut toutefois affirmer que les textes les plus riches sont aussi ceux dont les phrases sont en moyenne plus longues et comportent plus de mots-outils* [...]" (MACIEL, 1982, p. 92).

No caso do programa que empregamos, o *Hyperbase*, podemos obter por meio de cálculos com base em distribuições de frequências e na extensão relativa dos textos, a riqueza do léxico de um autor em duas possibilidades, pelo vocabulário geral ou pelo índice de *hapax*[54].

O resultado geral que obtivemos traz os valores real (efetivo) e teórico, o desvio (*écart)*, desvio reduzido[55] (*écart reduit*), *hapax* e *hapax*

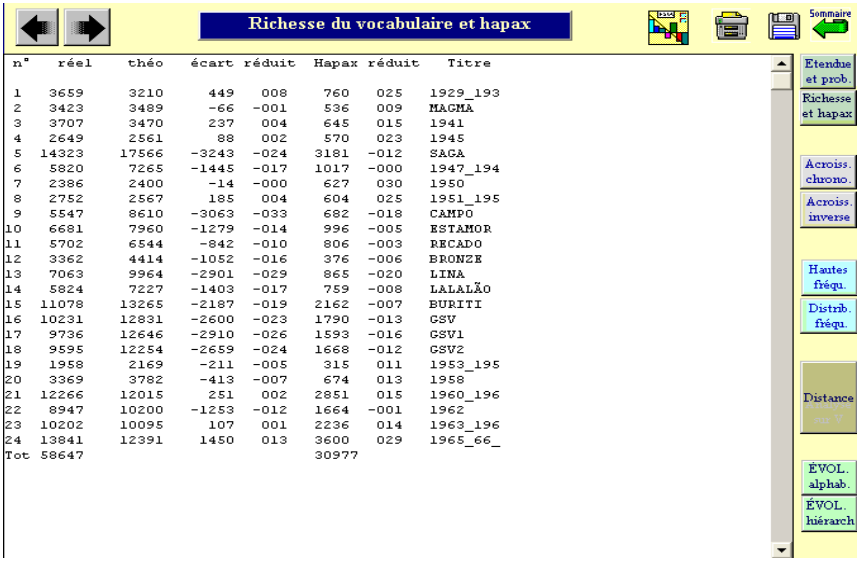
[54] Para uma explicação mais técnica, Brunet esclarece que o cálculo aplicado aos *hapax* está relacionado à distribuição normal: "*La methode est ici plus simple et se rattache à la loi normale. On aboutit pareillement à des écarts réduits qui servent d'ordonnées au programme de courbe*." (BRUNET, 2011, p. 57).

[55] O desvio reduzido é um índice que permite estimar a importância dos desvios e de os comparar; ele é igual ao quociente do desvio absoluto pelo "desvio-tipo" (GUIRAUD, 1959, p. 41). Sabe-se que a distribuição de uma palavra é raramente regular em um *corpus* (BRUNET, 2011, p. 37); o *écart reduit* se estabelece com uma simples regra de três (frequência teórica de uma palavra num texto = frequência de uma palavra no *corpus* ponderado pela probabilidade "p" ou parte do texto no *corpus*). Cabe ainda citarmos aqui a explicação de Ferreira (2005) para o desvio reduzido: O desvio reduzido, como o próprio nome

reduit. Para facilitar a leitura do quadro a seguir sobre a riqueza do vocabulário e *hapax*, basta, por exemplo, subtrair um valor teórico pelo valor real que assim teremos o desvio. Para um maior entendimento, explicaremos o caso da inserção da obra MAGMA no *corpus*:

valor real de formas (3.423) – valor teórico (3.489) = seu *écart* será – 66.

Vejamos os demais resultados da riqueza do vocabulário e *hapax* a seguir:



n°	réel	théo	écart	réduit	Hapax	réduit	Titre
1	3659	3210	449	008	760	025	1929_193
2	3423	3489	-66	-001	536	009	MAGMA
3	3707	3470	237	004	645	015	1941
4	2649	2561	88	002	570	023	1945
5	14323	17566	-3243	-024	3181	-012	SAGA
6	5820	7265	-1445	-017	1017	-000	1947_194
7	2386	2400	-14	-000	627	030	1950
8	2752	2567	185	004	604	025	1951_195
9	5547	8610	-3063	-033	682	-018	CAMPO
10	6681	7960	-1279	-014	996	-005	ESTAMOR
11	5702	6544	-842	-010	806	-003	RECADO
12	3362	4414	-1052	-016	376	-006	BRONZE
13	7063	9964	-2901	-029	865	-020	LINA
14	5824	7227	-1403	-017	759	-008	LALALÃO
15	11078	13265	-2187	-019	2162	-007	BURITI
16	10231	12831	-2600	-023	1790	-013	GSV
17	9736	12646	-2910	-026	1593	-016	GSV1
18	9595	12254	-2659	-024	1668	-012	GSV2
19	1958	2169	-211	-005	315	011	1953_195
20	3369	3782	-413	-007	674	013	1958
21	12266	12015	251	002	2851	015	1960_196
22	8947	10200	-1253	-012	1664	-001	1962
23	10202	10095	107	001	2236	014	1963_196
24	13841	12391	1450	013	3600	029	1965_66_
Tot	58647				30977		

Quadro 3: Riqueza do vocabulário e *hapax*.
Fonte: *Hyperbase* ©, versão 5.4.

A partir desses resultados verificamos que do vocabulário total da ficção (58.647 vocábulos) de João Guimarães Rosa, 30.977 são *hapax*.

---

diz, consiste em reduzir a zero os desvios relativos de todas as unidades lexicais. Estabelecemos, dessa forma, um centro de gravidade ao redor do qual orbita todo o léxico do *corpus*. Qual o motivo de estabelecermos esse centro? Simples: é a partir dele, de sua exploração e observação minuciosas, que podemos medir, com a mesma medida, todas as unidades lexicais, para então confrontá-las, independentemente de seus tamanhos, seus traços, e, assim, bem determinarmos quais são aquelas pertencentes, ou não, ao eixo normal de utilização pelo autor em seu discurso. (FERREIRA, 2005, p. 253).

Ou seja, mais da metade (aproximadamente 52,81%) do vocabulário de Rosa não se repete, o que em termos estilísticos demonstra muita habilidade de escrita do autor.

A seguir temos o gráfico que ilustra a riqueza lexical de Rosa:

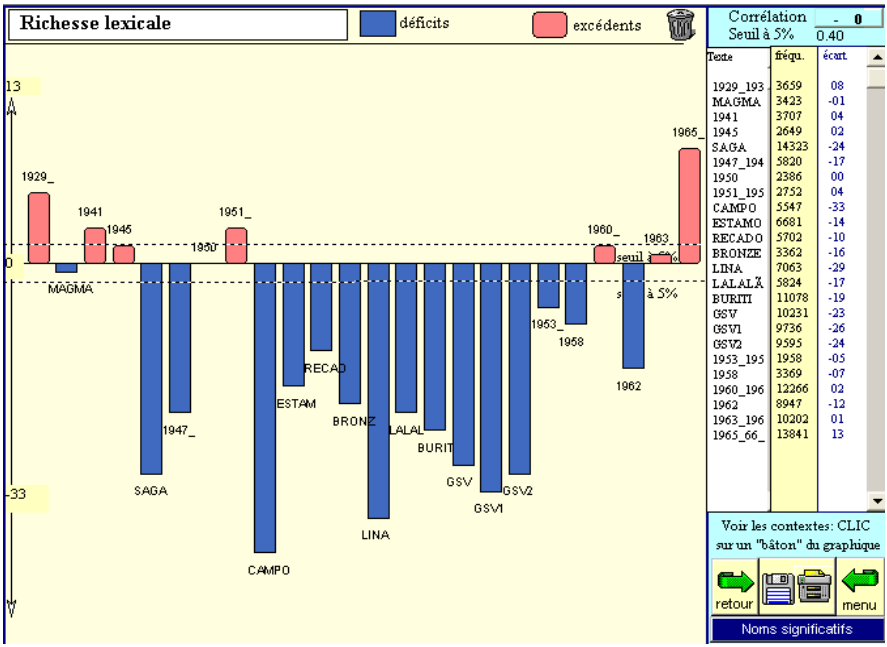


Gráfico 1: Riqueza lexical. *Hyperbase* ©, versão 5.4.

Existe uma margem de 5%[56] de tolerância (margem de erro), o dado que ultrapassar essa margem (as barras em vermelho) para cima é considerado como excedente. No gráfico 1 vemos que, na cronologia relacionada à riqueza de vocabulário de Rosa, o início apresenta um período positivo, pois são contos (textos curtos e que, portanto, trazem um vocabulário condensado sem a possibilidade de muita repetição), logo indica leve decréscimo em *Magma* e novamente apresenta saldo positivo (textos de 1941 e 1945 que figurarão em *Ave, palavra*) e a partir de *Sagarana* o vocabulário sofre um grande período deficitário

56 O termo "seul à 5%" designa a expressão numérica de um critério e constitui um tipo de base localizada em uma escala ordenada de resultados. (EDUMETRIE, 2011).

que vai durar até o surgimento dos textos[57] de 1965 a 1967 onde figuram *Estas Estórias*, *Tutameia* e *Ave, palavra*, apresentando o auge da riqueza lexical e deixando clara a tendência de que quanto menor o texto maior a riqueza vocabular. Uma outra característica que podemos ressaltar sobre os textos deficitários, é que eles contêm maior presença de diálogos, e, ao contrário dos textos narrativos curtos e descritivos, apresentam um vocabulário menos diversificado, como afirma Maciel:

> *Il est en effet connu que le lexique qui s'actualise dans le dialogue est plus réduit alors que les passages descriptifs et narratifs mettent en jeu un vocabulaire plus diversifié. Par ailleurs, en langue portuguaise, en raison de la structure de la réponse, on doit s'attendre à ce que les verbes aient des effectif plus importants dans les textes oú le dialogue prédomine.* (MACIEL, 1982, p. 107).

Com base nos resultados da riqueza lexical, concluímos que a maioria das obras rosianas tem caráter mais deficitário que excedente, ou seja, apresentam índice lexical muito baixo, não havendo muita variedade de vocabulário. As características gerais dessas obras são: ambientação fundamentada no sertão (*Sagarana*, *GSV* e *Corpo de baile*) e grande incidência de diálogos, salvo as obras que aparecem também em déficit ao final da linha do tempo: as de 1953 que estarão em *Ave, palavra*; a tradução *O último dos maçaricos* e as de 1962 que estarão reunidas em *Estas estórias* e em *Primeiras estórias*.

Outro aspecto bastante importante para a análise de riqueza lexical é a contagem de *hapax*. A seguir, apresentamos o histograma resultante das ocorrências de *hapax* na obra rosiana, que ilustra em quais obras o processo de renovação de vocabulário se desenvolveu mais intensamente:

---

[57] Para legenda do gráfico indicamos a leitura da listagem que expusemos no capítulo 4 desta tese.

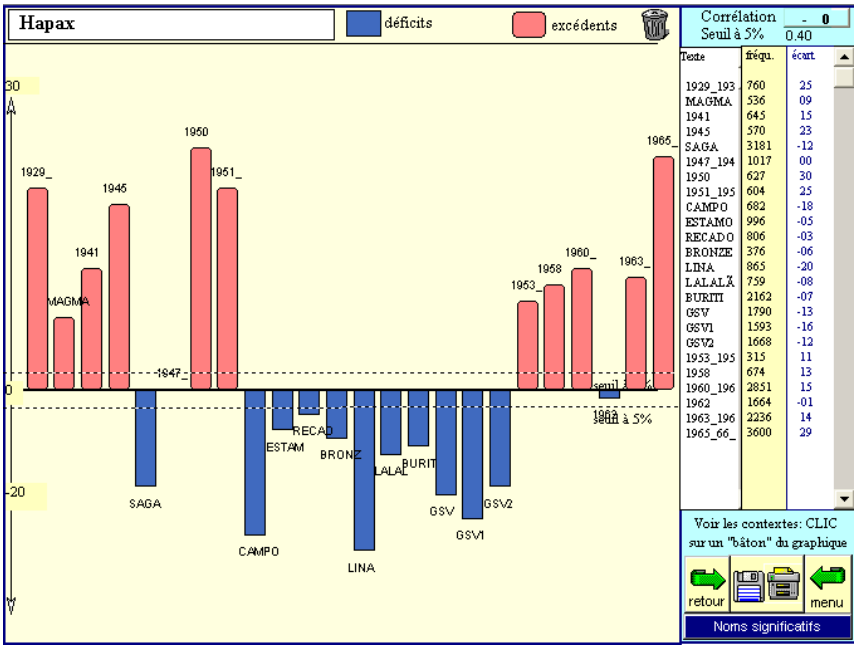


Gráfico 2: Ocorrências de *Hapax*. *Hyperbase* ©, versão 5.4.

O resultado que temos levanta novamente a questão do gênero literário, pois o primeiro bastão da esquerda representa a série de contos curtos — que Rosa publicou em jornais ainda na sua carreira de médico iniciante — com vocabulário muito mais diversificado que em *Sagarana*, por exemplo. Podemos afirmar também que a temática do sertão é mais restrita em seu vocabulário e, portanto, mais repetitiva, pois ela aparece em praticamente todas as obras ilustradas pelos bastões em azul. Das considerações sobre a quantidade de *hapax*, salientamos que a obra menos repetitiva, isto é, mais diversificada e renovada em termos lexicais é *Ave, palavra* (que inclusive é uma miscelânea de gêneros literários e anotações de viagens), e as obras[58] que se

[58] As obras *Sagarana*, *Corpo de baile* e *Grande sertão: veredas* estão dispostas de modo diferenciado (sem data definida) por apresentarem essa dificuldade de pontuar uma data específica para a produção das mesmas. *Sagarana* traz quase 10 anos de reedições e alterações; *Grande sertão: veredas* e *Corpo de baile* são obras muito maiores e que, supostamente, foram elaboradas na primeira metade da década de 50, por esse motivo, a disposição das obras se encontram dessa maneira no histograma.

apresentam deficitários, de acordo com as análises do histograma são: *Sagarana*, *Corpo de baile* e *Grande sertão: veredas*. Ao lado de *Sagarana* se encontram em déficit minoritário os textos de 1947 e 1948 que são: *Com o vaqueiro Mariano*; *Meu tio Iauaretê*; *Em Cidade* e os textos de 1962 *A estória do homem do Pinguelo*; *Substância*; *Partida do audaz navegante*; *Nenhum, nenhuma* e, por fim, *Pirlimpsiquice*.

Sobre os resultados que ora analisamos, concluímos que a riqueza lexical de um texto é avaliada em relação a alguns critérios determinantes, tais como, o gênero literário, o estilo e a temática. Se retomarmos, o que vimos com Sperber no capítulo anterior, em *Caos e Cosmos* (1976), lembraremos que a autora mostrou o processo de estruturação da obra rosiana afirmando que o estilo de Rosa se sobressai à temática. Podemos justificar o resultado do gráfico apontando para uma saída relacionada diretamente com a forma, com o estilo. Porém, é importante considerar que o resultado de variação de vocabulário mostra déficit nas obras de temática fortemente marcada pelo sertão e que nas outras obras, em vermelho no gráfico, o que há é uma alta diversificação do léxico. Portanto, a temática é um fator também ligado ao estilo, porém, no caso de criação vocabular de Rosa, o tema parece não contribuir com a diversificação do léxico.

Partiremos agora para a análise de Covizzi (1978), aproveitando o resultado do mesmo gráfico, quando a autora afirma haver uma fase expressiva de *Sagarana* até *Primeiras estórias* e que, a partir dela, o processo de escrita de Rosa volta-se a si mesmo, em tentativa de explicar a própria atividade literária. Diz Covizzi (1978, p. 88):

> Dosando um pouco a intenção artística e outro tanto a de surpreender pelo uso de palavras e expressões nada comuns, os prefácios às *Terceiras Estórias* de Guimarães Rosa seriam textos exemplares e eficientes como introdução aos estudos de teoria literária.

Ora, podemos suscitar aí o caráter didático e explicativo da obra de Rosa. Por meio do levantamento de vocabulário por ocorrências de *hapax*, podemos verificar se as obras iniciais representariam mais essa expressividade criativa (até mesmo Covizzi (1978 p. 88) afirma ser o neologismo um recurso expressivo) por conta da elaboração de vocábulos novos. Consequentemente, teríamos na fase após *Primeiras estórias* uma diminuição considerável de entrada de novos vocábulos.

Sob essa ótica de Covizzi (1978), visualizamos no gráfico 2 que a fase expressiva de Rosa, cujo ápice, segundo a autora, se encontra em

Grande sertão: veredas, parece não ter sido em termos lexicais. Caso essa expressividade tenha se dado por meio de muitos vocábulos novos, esse fenômeno teria refletido de forma inversa no gráfico. Porém, o que podemos afirmar, seguindo a perspectiva de Covizzi (1978), é que a fase explicativa se utilizou de uma diversidade lexical muito maior que a própria fase considerada expressiva e que, da mesma forma, a expressão não está relacionada diretamente a inserção de palavras inéditas no texto.

6.3 EVOLUÇÃO DO VOCABULÁRIO DE ROSA

A evolução de um vocabulário é, *grosso modo*, a soma de palavras novas empregadas até um determinado ponto de um *corpus* estudado, e que são acrescidas ao seu efetivo desde o primeiro texto do *corpus* até o último, sempre respeitando a ordem cronológica (de publicação ou de produção). Estudar a evolução de um vocabulário nos permite verificar a dinâmica lexical do autor do texto a ser analisado, bem como, a sua tendência (ou preferência) de escrita levando em consideração a renovação do vocabulário ou, ao contrário, a repetição e estagnação de formas já utilizadas pelo escritor. Sabemos (haja vista o estudo do item anterior sobre a riqueza do léxico) que a frequência de uma forma pode revelar as temáticas mais ou menos abrangidas no *corpus* e, a partir disso, é possível questionar, por exemplo, se o tema que emerge é uma característica das ideias ou do movimento literário do período (se comparado com outros autores da mesma época, mas esse não é o nosso caso) em que a obra foi elaborada, ou ainda, se o é exclusividade lexical do autor em estudo.

A evolução de vocabulário se relaciona com o crescimento lexical, e sobre isso, Charles Muller explica:

> [...] *qui s'établit entre N, l'étendue d'un texte, et V, l' étendue de son vocabulaire;* [...] *en considérant non point N et V comme des valeurs fixes, et leur rapport comme une relation statique, mais en essayant de déterminer comment V évolue quand on fait croître N régulièrement, c'est-à-dire quand on lit le texte en notant les vocables nouveaux. Il faut donc se figurer un lecteur doué d'une attention et d'une mémoire surhumaines, qui serait capable, tout au long, de compter à la fois les* ***mots*** *(qui*

> *constituent N) et les **vocables** (qui forment V), en remarquant chaque vocable nouveau qui vient accroître V d'une unité*. (MULLER, 1968, p. 183).

O coeficiente de correlação mede a evolução de uma forma (ou de um lema) e permite visualizar quais formas surgiram progressivamente no léxico de um autor e quais desapareceram ao longo de sua escrita. Para cada palavra, esse coeficiente estabelece uma relação entre o *ranking* (classificação) das formas e os valores do desvio (*écart reduit*). Obtém-se assim, por meio de uma base previamente definida, uma lista de termos que crescem ou que, pelo contrário, são progressivamente deixados de uso. O diagnóstico se exprime por um índice, positivo e negativo, segundo a forma, e a significância é proporcional ao valor absoluto desse índice. Vejamos a seguir, os resultados sobre o vocabulário do autor mineiro.

1 Evolução do vocabulário rosiano — coeficiente positivo

O resultado que segue está relacionado a um índice[59] hierárquico de frequência de palavras em progressão, ou seja, palavras (ou sinais de pontuação) que foram sendo cada vez mais utilizadas ao longo da produção literária de Rosa. Para a leitura dos resultados, vale ressaltar que a primeira informação visível é o coeficiente positivo, a segunda se refere à frequência da forma no *corpus* e a terceira indica a forma (que no caso do *Hyperbase* pode ser uma palavra ou uma pontuação) em si. Nessa listagem, visualizamos as palavras que Guimarães Rosa mais empregou até o final da produção ficcional (de 1929 a 1967), são elas:

Coefic.	**freq.**	**forma**
+ 0.001	116617	,
+ 0.001	50861	.
+ 0.001	14426	se
+ 0.001	8190	em
+ 0.001	2303	ou
+ 0.001	1874	à
+ 0.001	1239	nos
+ 0.001	844	às
+ 0.001	570	amor
+ 0.001	535	menos

[59] O *Hyperbase* também oferece índice em ordem alfabética, mas apresentamos aqui a listagem hierárquica porque queremos visualizar as palavras ordenadas pelas suas frequências.

+ 0.001	443	porém
+ 0.001	308	apenas
+ 0.001	288	contra
+ 0.001	265	seja
+ 0.001	248	sob
+ 0.001	241	seria
+ 0.001	230	talvez
+ 0.001	223	alma
+ 0.001	219	paz
+ 0.001	218	quanto
+ 0.001	201	azul
+ 0.001	206	esta
+ 0.001	177	decerto
+ 0.001	150	fato
+ 0.001	139	espírito
+ 0.001	134	justo
+ 0.001	128	espaço
+ 0.001	127	real
+ 0.001	114	segundo
+ 0.001	113	neste
+ 0.001	110	senão
+ 0.001	109	exemplo
+ 0.001	108	haver
+ 0.001	108	geral
+ 0.001	103	presença
+ 0.001	102	forma
+ 0.001	102	enfim
+ 0.001	99	fio
+ 0.001	96	estrelas
+ 0.001	96	amarelo
+ 0.001	90	própria
+ 0.001	90	par
+ 0.001	90	maneira
+ 0.001	82	perdido
+ 0.001	82	ante
+ 0.001	81	súbito
+ 0.001	81	seco
+ 0.001	79	sentido
+ 0.001	78	joaquim
+ 0.001	75	papel
+ 0.001	75	nesta

+ 0.001	75	mente
+ 0.001	75	estes
+ 0.001	75	contudo
+ 0.001	74	natureza
+ 0.001	73	uso
+ 0.001	73	grave
+ 0.001	73	centro
+ 0.001	72	livro
+ 0.001	69	memória
+ 0.001	68	vêm
+ 0.001	68	algo
+ 0.001	67	efeito
+ 0.001	65	entanto
+ 0.001	65	acaso
+ 0.001	64	minas
+ 0.001	63	fatos
+ 0.001	62	seguir
+ 0.001	61	aberto
+ 0.001	60	crer
+ 0.001	60	ação
+ 0.001	57	repouso
+ 0.001	57	posto
+ 0.001	55	ato
+ 0.001	54	novas
+ 0.001	54	engano
+ 0.001	53	ninho
+ 0.001	53	matéria
+ 0.001	53	jardim
+ 0.001	53	almas
+ 0.001	51	movimentos
+ 0.001	50	raro
+ 0.001	49	relógio
+ 0.001	48	mesmos
+ 0.001	48	longa
+ 0.001	47	pão
+ 0.001	45	navio
+ 0.001	45	muda
+ 0.001	45	futuro
+ 0.001	45	dita
+ 0.001	44	eis
+ 0.001	44	amar

+ 0.001	44	aliás
+ 0.001	43	dom
+ 0.001	42	sr
+ 0.001	42	penso
+ 0.001	41	plano
+ 0.001	40	pura
+ 0.001	40	praça
+ 0.001	40	humana
+ 0.001	40	botou
+ 0.001	39	várzea
+ 0.001	39	única
+ 0.001	39	liberdade
+ 0.001	38	salvo
+ 0.001	38	alva
+ 0.001	37	antiga
+ 0.001	36	ermo
+ 0.001	36	enterro
+ 0.001	35	situação
+ 0.001	35	puro
+ 0.001	35	porquanto
+ 0.001	34	quantas
+ 0.001	34	mediante
+ 0.001	34	ilha
+ 0.001	34	cujas
+ 0.001	34	amores
+ 0.001	33	toque
+ 0.001	33	falecido
+ 0.001	33	consciência
+ 0.001	32	alheio
+ 0.001	31	semi
+ 0.001	31	poesia
+ 0.001	31	face
+ 0.001	31	espaços
+ 0.001	30	vero
+ 0.001	30	profundo
+ 0.001	30	perfeito
+ 0.001	30	ondas
+ 0.001	30	escrito
+ 0.001	30	cova
+ 0.001	30	apontou

Lista de coeficiente positivo da obra de Guimarães Rosa. *Hyperbase* ©, versão 5.4

Segundo os resultados obtidos da lista de coeficiente positivo sobre a evolução do léxico[60], podemos verificar que as formas cujas ocorrências ultrapassam o número de 10.000 e que tiveram um crescimento em seu uso na cronologia da obra de Rosa foram: a vírgula (116.617 ocorrências), o ponto-final (50.861) e o vocábulo "se" (14.420, que pode ser tanto pronome ou conjunção, para fazer a distinção basta gerar uma lista de concordâncias e verificar ocorrência por ocorrência; não faremos tal exercício por não ser de interesse aos nossos objetivos). Em termos de pontuação, ao final de sua jornada literária, Rosa se apossou mais da vírgula e do ponto-final e abriu mão da exclamação – como veremos mais adiante na listagem correspondente ao coeficiente negativo — cujo emprego diminuiu significativamente ao longo da produção.

Para detectarmos em quais textos essa pontuação esteve mais presente, faremos uso da ferramenta que busca vocábulos, e por meio dela, geraremos uma lista com os dados de identificação da forma que se quer analisar no *corpus* (informações como: em quais obras o vocábulo se encontra, quantas vezes ele aparece em cada uma etc.). Após a obtenção desses dados, é possível desenvolver um histograma que apresente o comportamento da frequência do vocábulo, tornando viável a leitura dos dados e a identificação do auge de um vocábulo que se quer analisar. Vejamos, então, da lista dos coeficientes positivos, a frequência da vírgula:

60 O *Hyperbase* gera uma lista de evolução de uma palavra com frequência mínima de 30.

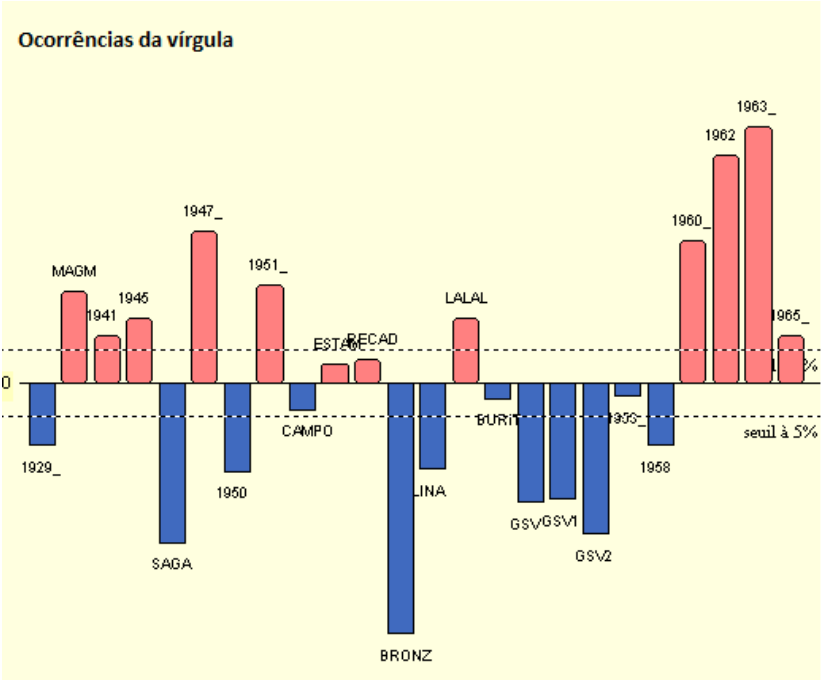


Gráfico 3: Ocorrências da vírgula na obra de Guimarães Rosa. *Hyperbase* ©, versão 5.4.

Nesse gráfico vemos a ascensão da vírgula em *Estas estórias* e déficit nos textos representados pelos bastões em azul. Uma pequena diferença sobressai na última parte de *Corpo de baile*, em *Noites do sertão*. A presença da vírgula também é motivo de observação para Covizzi (1978), e que também pode ser explorado em nossos estudos, pois segundo a autora, a pontuação em geral ganha maior presença na fase explicativa, com excessos de pausa, exigindo inclusive maior atenção para o que se lê.

Vejamos agora o ponto-final:

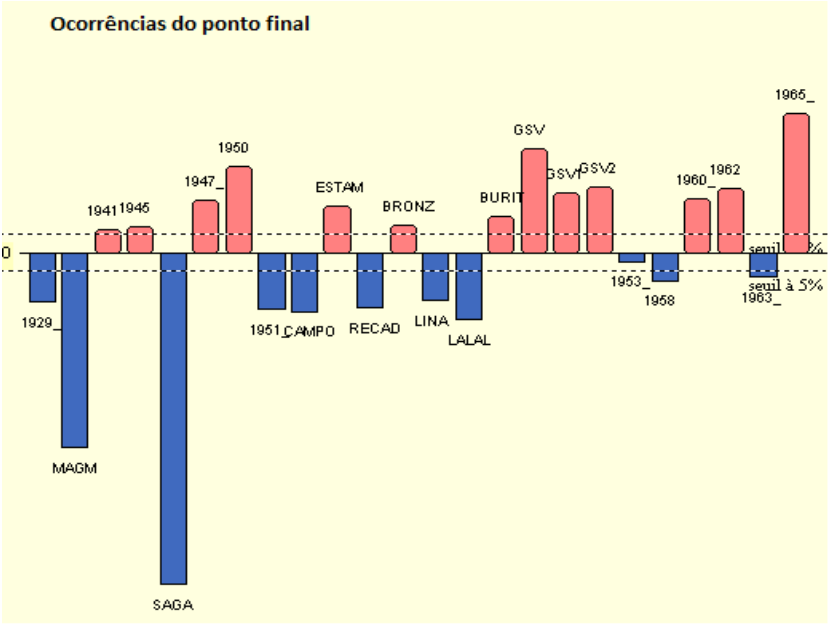


Gráfico 4: Ocorrências do ponto-final na obra de Guimarães Rosa. *Hyperbase* ©, versão 5.4.

Percebemos que o ponto-final apresenta uma grande variedade em seu uso. Ele é mais empregado nas obras dos anos de 1950 e a partir de *Grande sertão: veredas* ganha maior presença, com uma diferença pouco significativa de duas obras de natureza distinta: a tradução e o capítulo de uma novela policial. Teriam as obras em azul as frases mais longas, já que inclusive a incidência maior da vírgula também recai sobre elas?

Vamos tentar responder essa questão. Para isso, precisamos chegar ao comprimento das frases dividindo o número total de palavras pelo número total de frases de cada texto, desse modo, alcançamos a média do tamanho da frase. Com base nisso, chegamos aos resultados[61] a seguir:

61 Não incluímos *Magma*, o capítulo de MMM e a tradução por serem obras de gêneros muito distintos.

TEXTOS	PALAVRAS/FRASE
Contos de 29-30	12,4
Sagarana	5,8
Corpo de baile	12,66
GSV	11,9
Primeiras estórias	9,2
Tutaméia	9,2
Estas estórias	8,5
Ave, palavra	9,46

Quadro 4[62]: Média de palavras por frase.

Pelos dados brutos podemos verificar que os contos de 29-30, *Corpo de Baile* e GSV são os que apresentam frases mais extensas. Ao contrário do que pensamos, *Sagarana* apresentou a média menor de 5,8 palavras por frase, ou seja, não apresentou longos períodos em relação à média das outras obras. Percebemos que nos dois gráficos (sobre vírgula e sobre o ponto-final) *Sagarana* é deficitária. Contudo, o uso da exclamação – que veremos no gráfico 5 a seguir – dá-se quase que exclusivamente nessa obra. É importante salientar o equilíbrio existente entre as obras de *Primeiras estórias* (produzida entre 1960 e 1962) e *Tutaméia* (entre 1966 e 1967) que apresentam o mesmo tamanho e *Ave, palavra* (entre 1965 e 1967) que se aproxima em medida. O movimento que percebemos então é que ao final da vida como escritor, entre os anos de 1960 a 1967, Rosa atingiu um ritmo frasal que foi mantido por quase uma década de produção. Vale lembrar que a literatura modernista apresentou uma grande liberdade na linguagem, incluindo experimentos lexicais, sintáticos e semânticos, Rosa como escritor modernista, experimentou dessa liberdade com muita astúcia, e os dados obtidos refletem essa atividade oscilatória de criação.

Podemos analisar um pouco mais o tamanho da frase rosiana nos voltando para o ponto de vista das palavras dentro da obra completa. Antes decompomos o tamanho da frase pela pontuação de cada obra, agora vamos às palavras de modo geral. No *corpus* temos 1.094.481

62 Vale observar que neste quadro a organização está fundamentada pela cronologia da primeira publicação.

ocorrências; se somarmos a pontuação forte e dividirmos por esse total, chegaremos à média de palavras por frase. Assim, nos aproximamos do tamanho da frase, e com isso, do ritmo de frase da prosa rosiana como um todo. Contudo, faz-se necessário descontar alguns números desse total, pois como dissemos, o programa considera todo caractere separado por espaços como ocorrência. Então descontaremos os sinais gráficos:

Caractere	Frequência
'	43
-	34902
!	5660
(	371
)	375
*	374
,	116617
.	50861
/	17
:	8677
;	457
?	6522
‘	37
’	1104
“	3905
”	7914
...	11575

Lista de caracteres

Vamos ao cálculo, que é muito simples:

1.094.481 (total de ocorrências, incluindo os sinais gráficos) – 249.411 (total de sinais gráficos) = 845.070 (palavras apenas[63])

63 Não excluímos números e datas pois eles não afetam na pontuação, são componentes de uma frase e carregam importância semântica.

Na sequência, somamos os sinais fortes, para totalizar as frases:

6522 (?) + 50861 (.) + 5660 (!) + 11575 (...) = 74.618 frases

E por fim, dividimos o total de palavras pelo total de frases, para obter a média de palavras por frase:

845.070 / 74.618 = 11, 32 (palavras/frase)

Dessa forma percebemos que *Grande sertão: veredas* (segundo os resultados do Quadro 3) é a obra que mais representa a média de comprimento de frase rosiana. Considerando alguns estudos já realizados a respeito do tamanho de frase de alguns autores, trazemos aqui para comparação a média proustiana estabelecida por Brunet (1983) com a extensão de aproximadamente 31 palavras por frase. O autor apresenta uma tabela que relaciona alguns escritores de língua francesa, e a partir dela podemos observar as seguintes características:

Escritores	**Palavras**	**Frases**	**Média**
Émile Rousseau	257154	9280	27,71
Chateaubriand	1398984	62919	22,23
Giraudoux (romances)	412268	19971	20,64
Corpus XIX – XX	70273552	4611432	15,24
Prosa literária de 1893 a 1926	12216571	914130	13,36

Quadro 5: Média de frases de escritores franceses.
Fonte: Adaptado de BRUNET (1983, p. 124).

Brunet conclui que Proust não seguiu a tendência dos escritores de sua época, pois o comprimento de sua frase alcança o dobro da dimensão dos outros prosadores. Para a presente pesquisa, o ideal seria um levantamento parecido como esse, em língua portuguesa e com escritores brasileiros contemporâneos a Rosa. Porém, comparando o tamanho da frase rosiana com a produção francesa, podemos concluir que Rosa parece ter dado continuidade à tendência da prosa literária do século XX. Não temos pesquisa que demonstre resultado equivalente na prosa brasileira, portanto, vamos trabalhar com um estudo sobre o conto machadiano[64] que apresenta uma média variante entre 14 a 18 palavras

64 FREITAS, (2007).

por frase (um pouco além da média rosiana que concluímos). A partir desses dados, já identificamos uma diferença entre gênero literário, pois as frases de *Grande sertão: veredas* e *Corpo de baile* parecem ser mais longas que as frases dos contos rosianos (com exceção dos primeiros contos publicados do autor). Podemos pensar que o gênero textual influencia no tamanho da frase? Segundo Brunet (1983, p. 124), a complexidade do discurso influencia diretamente o tamanho da frase.

Nos resultados sobre o emprego do ponto de exclamação em Rosa, vimos que há um grande declínio ao longo da produção literária, pois foi o primeiro sinal de pontuação a surgir na lista de coeficiente negativo (que será apresentada logo a seguir). Vejamos o gráfico que ilustra o movimento do sinal gráfico:

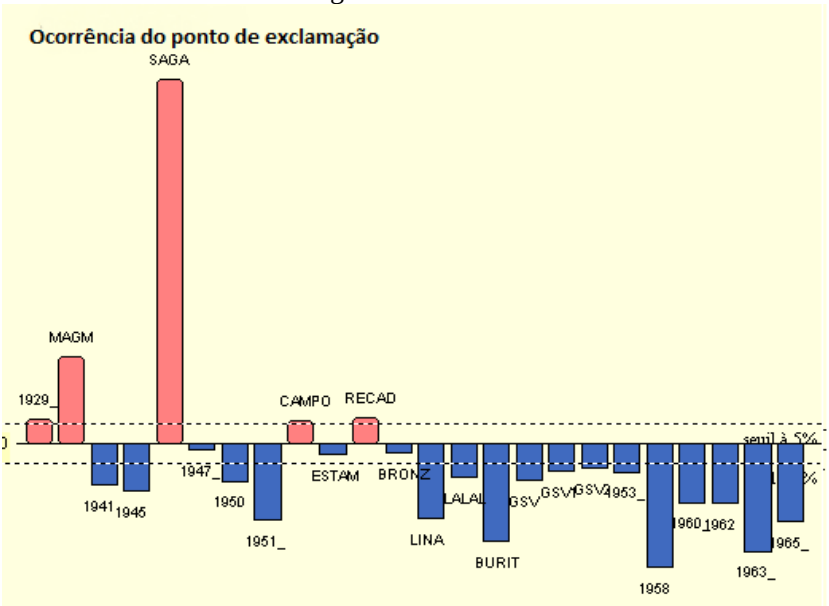


Gráfico 5: Ocorrências do ponto de exclamação na obra de Guimarães Rosa. *Hyperbase* ©, versão 5.4.

Com a ilustração do gráfico, é possível verificar que a maior incidência da exclamação está em *Sagarana* e que antes ainda de *Corpo de baile*, já encontramos o declínio dessa pontuação, bem como um processo deficitário até o final da produção literária. Vejamos agora a incidência do ponto de interrogação:

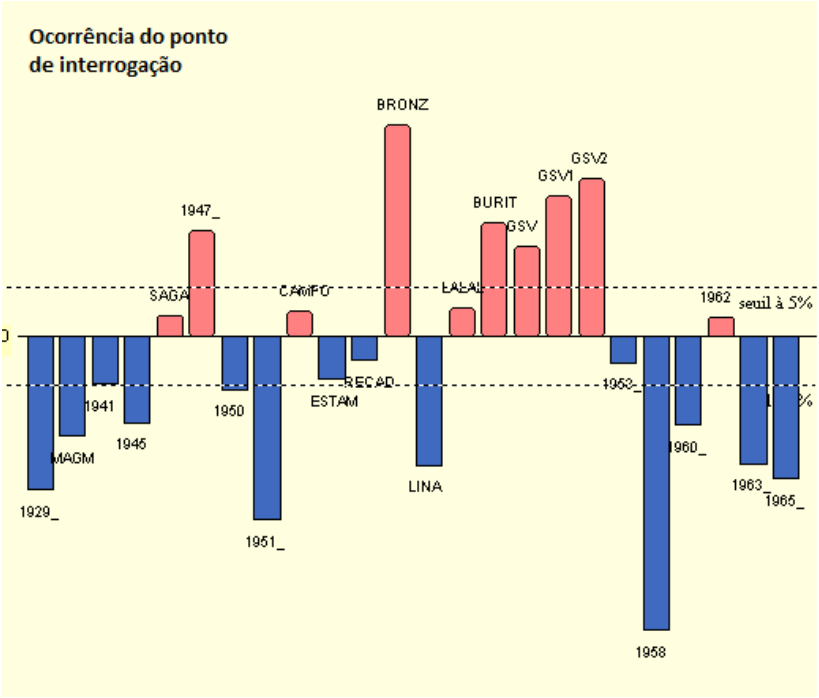


Gráfico 6: Evolução do emprego do ponto de interrogação na obra de Guimarães Rosa. *Hyperbase* ©, versão 5.4.

Do gráfico 6 percebemos que a obra de Rosa se inicia e finaliza com um déficit relacionado ao uso da interrogação. Temos a interrogação em *Corpo de baile*, mas é em *Grande sertão: veredas* que encontramos com força maior a presença da interrogação.

Tratemos agora da evolução dos substantivos. Partindo do coeficiente positivo, o primeiro substantivo que ganhou uma progressão cronológica foi "amor" (570), que aparece entre as dez primeiras (entre as palavras gramaticais inclusive, que são sempre as mais frequentes no *ranking*, o que é notável), seguido de "alma" (223), "paz" (219), "azul" (207), "fato" (150), "espírito" (139), "espaço" (128), "presença" (103) e "forma" (102); algumas delas, se reunidas, configuram um grupo particular, é o caso de *amor*, *alma*, *paz*, *azul*, *espírito*.

Há um estudo de Maria Célia Leonel (2000), em que a autora trata sobre *Magma* e ressalta um grupo de sete poemas[65], no qual cada um deles é intitulado com uma cor do espectro da luz. Dentro das cores, Leonel discorre sobre um subgrupo de poemas que abordam a temática da morte: os poemas "Vermelho", e "Roxo"; são respectivamente, o primeiro e o último do grupo: "O arco-íris rosiano, portanto, abre e fecha com a morte". (2000, p. 127).

Em se tratando de cores, a incidência da forma "azul" nos chamou a atenção, decidimos explorar o conteúdo do poema cujo título é homônimo. Inicialmente, transcrevemos o poema "Azul" da coleção de cores de *Magma*:

> Uma vanessa tropical travou na campânula
> de uma ipoméia
> o vôo oscilatório e helicoidal.
> Dobra o quimono de franjas sinuosas,
> marchetado e hachureado
> com minérios de cobre:
> aréolas, anéis, jóias concêntricas,
> olhos de íris elétrica e de pupila enorme,
> ocelos de um leque de pavão.
> Sinto o perfume da flor nova,
> com mais dois estames, buliçosos,
> e quatro pétalas, de uma esmalte raro,
> molhadas nas tintas de céus fundos,
> e cromadas com a faiança das lagoas...
> (ROSA, 1997, p. 57).

No poema, percebemos o voo de um lepidóptero[66] sobre uma flor nova de pétalas azuis. A imagem da borboleta compartilha dois temas entre muitas civilizações: a vida e a morte. Na civilização asteca, a borboleta está relacionada ao sopro vital que sai da boca do agonizante, é o símbolo da alma (CHEVALIER, 1990). A simbologia da borboleta está relacionada tanto à vida quanto à morte; entre os mexicanos é

---

65 Para esse grupo de poemas, a autora direciona uma temática filosófica, contudo, alerta, que são apenas nuanças temáticas, nada muito aprofundado no campo da metafísica. Aproveitamos aqui para acrescentar que nossa leitura sobre este grupo de palavras, também não será investigado minuciosamente.

66 Vanessa é o gênero de uma espécie de borboletas chamada Vanessa Atalanta, cujo nome é referência de uma personagem da mitologia grega, filha do arcádio Íaso, que só queria filhos do sexo masculino. Quando Atalanta nasceu, seu pai a abandonou no monte Pártenon, e lá, Atalanta foi alimentada por uma ursa e recolhida por caçadores que a criaram, teve uma criação parecida com a de Ártemis, foi caçadora e como tal jurou resistir às tentações do amor. (HACQUARD, 1996).

símbolo do sol negro, que atravessa os mundos subterrâneos durante o seu curso noturno. (CHEVALIER, 1990). Da mitologia grega, nos afrescos de Pompeia, Psiquê é representada como uma criança com asas de borboleta; a figura alegórica da alma[67] (interessante perceber que a incidência da forma "alma" também é ressaltada na lista) na cultura helenística era difundida como imagem de uma borboleta (BRUNEL, 2000). Vale ainda acrescentar o comentário de Câmara Cascudo (2000, p. 179) a respeito da crença popular brasileira de que a borboleta: "significa para o povo uma mensageira. Anuncia alma dos mortos ou presságios agoureiros".

Toda essa descrição sobre a imagem da borboleta nos provoca questionamento: estaria em *Magma* o âmago temático que nortearia a obra de Rosa? Assim como o próprio título do livro de poemas já denota um termo da geologia, essa massa mineral de alta temperatura que se encontra em grande profundidade no núcleo terrestre, e que por vezes é expelida à superfície em erupções. Seriam os poemas de *Magma* pequenas erupções, amostras de uma grande obra? Leonel já nos adiantou, em *Guimarães Rosa*: *Magma e a gênese da obra* (2000), quando a autora percorre o movimento da poesia brasileira modernista contemporânea de *Magma*. Amparada nas teorias do texto de Gérard Genette e na crítica genética, ela estabelece relações intertextuais sobre o único livro de poesias de Rosa e *Sagarana.* Além das temáticas de princípios religiosos e filosóficos que abarcam mitos e lendas, Leonel identifica procedimentos rosianos que se repetem entre as diferentes narrativas. À luz de *Palimpsestes*, a pesquisadora considera alguns poemas de *Magma* como hipotexto e alguns contos de *Sagarana* como hipertexto, pois os segundos derivariam dos primeiros: *Sarapalha* viria de *Maleita*; *São Marcos*, de *Reza brava*; *O burrinho pedrês*, de *Boiada* e *Chuva*; *Gruta do Makiné* resultaria em *O recado do morro* de *Makiné*; *A hora e vez de Augusto Matraga* derivaria de *Boiada*. Diz a autora:

> Magma é matéria, substância, massa que está na origem de procedimentos inventivos de Guimarães Rosa, mas é também forma. São temas, modos de compor o andamento, o ritmo, de construir palavras, de criar textos, muitos deles retomados. Há procedimentos e reiterações visíveis e há cristalitos, que precisam ser procurados com mais acuidade. (2000, p. 170)

---

67 Em grego, *psyqué*.

Nesse sentido, ao verificarmos as incidências do coeficiente positivo sobre o vocabulário de Rosa, concluímos que se encontram em *Magma* não apenas as ideias de *Sagarana*, mas também a reunião de vocábulos que culminariam ao final da obra.

Retomando a ideia do vocábulo "azul" – pois foi o que mais se distinguiu do conjunto particular de vocábulos destacados no coeficiente positivo da evolução do léxico rosiano – podemos afirmar que o azul "destoa" das outras palavras que surgiram (alma, paz, espírito e amor) pois carrega mais materialidade. Esse grupo de palavras nos faz resgatar a tese de Walnice Nogueira Galvão, *As formas do falso* (1972). Nessa obra, Galvão trabalha o conceito da ambiguidade em *Grande sertão: veredas* partindo de dois motes "tudo é e não é" e "a coisa dentro da outra" (1972, p. 13). Por meio da história de Maria-Mutema que se encontra no meio do romance, ela destaca as seguintes situações: o conto no meio do romance, o diálogo dentro do monólogo, a personagem dentro do narrador, o letrado dentro do jagunço, a mulher dentro do homem, o diabo dentro de Deus. Seguindo essa lógica da "coisa dentro da outra", compartilhamos da ideia ao retomarmos alguns pontos já costurados aqui neste trabalho: "Azul" é o nome de um poema que está em *Magma*; *Magma* é uma amostra do porvir rosiano. "Azul" é o menos abstrato dos signos (alma, paz, espírito e amor) que mais aparecem ao final da produção de Rosa. Disso podemos afirmar que, ao final da produção ficcional do escritor mineiro, a figura do imanente dentro do transcendente tomou lugar. Por mais reduzida que possa aparentar essa conclusão, não podemos esquecer que a estilística idealista de Spitzer, tomava a consciência de um detalhe chamativo à leitura, e desse detalhe, esmiuçava de modo singular a obra como um todo.

2 Evolução do vocabulário rosiano — coeficiente negativo

Voltemos à lista de coeficiente, mas, ao contrário da lista anterior, apresentaremos a lista hierárquica de frequência de palavras em regressão (de coeficiente negativo), isto é, palavras ou pontuação que perderam o seu uso ao longo da produção ficcional. Comentaremos apenas as formas com frequência acima de 100:

Coef.	**Freq.**	**forma**
- 0.001	8369	com

- 0.001	5660	!
- 0.001	5473	mas
- 0.001	3952	na
- 0.001	2857	tinha
- 0.001	2637	mesmo
- 0.001	1990	estava
- 0.001	1805	muito
- 0.001	1767	bem
- 0.001	1650	até
- 0.001	1597	quando
- 0.001	1533	todos
- 0.001	1447	agora
- 0.001	1411	então
- 0.001	1193	quem
- 0.001	1154	você
- 0.001	1104	’
- 0.001	1099	depois
- 0.001	1016	isso
- 0.001	1004	ter
- 0.001	907	vai
- 0.001	863	coisa
- 0.001	854	nas
- 0.001	594	ir
- 0.001	531	fazer
- 0.001	497	tinham
- 0.001	497	corpo
- 0.001	454	melhor
- 0.001	422	estavam
- 0.001	409	cima
- 0.001	400	boca
- 0.001	341	ficar
- 0.001	339	falar
- 0.001	280	gostava
- 0.001	278	teve
- 0.001	271	tivesse
- 0.001	256	filho
- 0.001	221	causa
- 0.001	214	direito
- 0.001	204	querendo
- 0.001	202	tristeza
- 0.001	200	vontade
- 0.001	200	pobre
- 0.001	199	passar
- 0.001	184	tomar
- 0.001	182	pena
- 0.001	182	conversa
- 0.001	179	fala
- 0.001	177	ruim
- 0.001	174	compadre
- 0.001	173	tanta
- 0.001	172	daqui

- 0.001	172	chegar
- 0.001	170	sete
- 0.001	161	contar
- 0.001	159	deixar
- 0.001	157	morro
- 0.001	153	águas
- 0.001	147	pensando
- 0.001	142	resto
- 0.001	139	verdes
- 0.001	133	ria
- 0.001	132	alegre
- 0.001	126	levar
- 0.001	126	depressa
- 0.001	119	terras
- 0.001	118	pedras
- 0.001	115	entrar
- 0.001	114	vergonha
- 0.001	111	gostar
- 0.001	110	sela
- 0.001	108	chegando
- 0.001	102	companheiro
- 0.001	101	passando
- 0.001	99	pôde
- 0.001	96	mandar
- 0.001	93	companhia
- 0.001	92	esperando
- 0.001	90	pasto
- 0.001	90	escuta
- 0.001	89	mandava
- 0.001	87	riso
- 0.001	85	mole
- 0.001	85	conversar
- 0.001	82	rezar
- 0.001	82	pastos
- 0.001	81	pretos
- 0.001	81	pedir
- 0.001	81	lagoa
- 0.001	81	córrego
- 0.001	80	milho
- 0.001	80	falado
- 0.001	80	comprido
- 0.001	76	perguntar
- 0.001	74	conseguia
- 0.001	74	bravo
- 0.001	73	cantar
- 0.001	72	visto
- 0.001	72	reza
- 0.001	71	olhando
- 0.001	70	esteve
- 0.001	69	pano
- 0.001	68	saco

- 0.001	65	vermelha
- 0.001	65	grito
- 0.001	62	chama
- 0.001	61	escondido
- 0.001	60	contou
- 0.001	58	mudar
- 0.001	58	aprender
- 0.001	57	levando
- 0.001	57	cantando
- 0.001	56	vender
- 0.001	55	gostando
- 0.001	54	cantiga
- 0.001	53	confiança
- 0.001	52	conhecido
- 0.001	51	bateu
- 0.001	50	casamento
- 0.001	50	ara
- 0.001	49	servia
- 0.001	49	grota
- 0.001	49	algibeira
- 0.001	46	viola
- 0.001	46	rindo
- 0.001	44	sentindo
- 0.001	44	nesses
- 0.001	43	uai
- 0.001	43	escutando
- 0.001	43	achando
- 0.001	42	sentiu
- 0.001	42	perdeu
- 0.001	42	cascos
- 0.001	41	engraçado
- 0.001	40	zebu
- 0.001	40	réis
- 0.001	40	jogar
- 0.001	39	monte
- 0.001	38	rego
- 0.001	37	parecendo
- 0.001	37	abraçou
- 0.001	36	conversando
- 0.001	36	chifre
- 0.001	35	morrido
- 0.001	35	compridas
- 0.001	35	bonitas
- 0.001	34	estreito
- 0.001	33	porcos
- 0.001	32	rezando
- 0.001	32	jogou
- 0.001	32	brigar
- 0.001	31	varas
- 0.001	31	tocando
- 0.001	31	encosta

- 0.001	31	benção
- 0.001	31	apanhar
- 0.001	30	fechados

Lista de coeficiente negativo da obra de Guimarães Rosa – *Hyperbase* ©, versão 5.4

Não comentaremos aqui o ponto de exclamação, pois o mesmo já foi apresentado no item anterior sobre a pontuação. Vamos agora aos verbos. Na lista de coeficiente negativo, observamos que há menor emprego por parte do autor[68] de verbos nas conjugações do infinitivo[69] "ter" (1004), "ir" (594), "fazer" (531), "ficar" (341), "falar" (339), "passar" (199), "tomar" (184), "chegar" (172), "contar" (161), "deixar" (159), "levar" (129), "entrar" (115), "gostar" (111). Antes ainda da presença de algum infinitivo, o imperfeito surge na lista regressiva: "tinha" (2857), "estava" (1990), "tinham" (497), "estavam" (422), "gostava" (280). E, ainda, outras formas de passado: "teve" (278), "tivesse" (271) e do gerúndio: "querendo" (204), "pensando" (147), "chegando" (108), "passando" (101).

Ao compararmos os resultados dos dois coeficientes que temos dos verbos, perceberemos que as formas "seja"/"seria" (coef. +) e "tinha"/"estava" (coef. -) têm maior representatividade de tempos verbais positivos e negativos, respectivamente, na produção. Dessa forma, verificamos o subjuntivo[70] do presente e o futuro do pretérito como tempos que ganharam mais uso na ficção e, por outro lado, o pretérito imperfeito teve diminuição no emprego.

De acordo com o estudo de Harald Weinrich (1968) sobre o tempo na narrativa, o pretérito imperfeito pertence ao mundo narrado (*erzählen*) e de acordo também com um estudo de Carlos Maciel (2005, p. 435) o campo[71] do verbo "ter" é pertencente ao passado mais na forma "tinha", atraindo também para o mesmo campo a terceira pessoa do singular. Se compararmos o resultado negativo do verbo "ter" em sua flexão "tinha" na listagem de evolução do vocabulário, veremos que Rosa se desvia do tempo linguístico tradicionalmente empregado na

68 Salientamos os verbos com frequência ≥ 100.

69 Lembrando que o corpus não está lematizado.

70 Sobre o emprego do subjuntivo, temos o trabalho de Ivana Versiani que trata da questão do subjuntivo em *Grande sertão: veredas* no ensaio *Para a sintaxe Grande sertão: veredas.* Valores do subjuntivo. In: COELHO, Nelly Novaes, VERSIANI, Ivana. *Guimarães Rosa*. São Paulo: Quíron, 1975.

71 No item sobre as altas frequências trazemos o gráfico gerado por Maciel que retrata essa questão dos verbos relacionados com o tempo da narrativa.

literatura, segundo o levantamento de Maciel (?, p. 435), abarcando mais o subjuntivo do presente e o futuro do pretérito.

No ensaio *O tempo na narrativa*, Benedito Nunes discorre:

> O *pretérito perfeito*, o *imperfeito* e o *mais-que-perfeito* indicam, pelo distanciamento e pelo curso livre que imprimem à linguagem, que estamos contando ou narrando. Configuram, por conseguinte, uma situação de *locução narrativa*, ao contrário do *presente*, do *passado* composto e do *futuro*, que configuram uma situação de *locução discursiva*, de comentário. (NUNES, 1988, p. 39-40).

Ainda Paul Ricoeur (1995), comentando as análises de Weinrich (1964) afirma que as situações de locução correspondem a dois tipos distintos de tempos verbais, para o mundo comentado: o presente, o pretérito perfeito composto e o futuro; e para o mundo contado: o pretérito perfeito simples, o imperfeito, o mais-que-perfeito e o condicional.

Teria Rosa comentado mais que narrado ao final de sua obra? Nesse momento, retomamos a proposição de Covizzi quando afirma ter uma linha distinta na produção de Rosa que se inicia com um processo de expressividade e se finaliza com um momento de explicação. Os indícios dos tempos verbais que chegamos por meio dos coeficientes nos fazem acordar com a autora, que Guimarães Rosa, ao final de sua carreira literária tratou mais do seu material linguístico compondo locuções discursivas, e comentários, instituindo um narrador-comentador de seu mundo comentado, alguém com um olhar mais crítico para com o seu texto ao invés de um narrador de seu mundo narrado. Luiz Valente (1988), do mesmo modo, argumenta sobre os prefácios de *Tutaméia*:

> *The four prefaces of Tutaméia [...] are of paramount importance for the understanding of his aesthetics, for they present in compact from virtually all of his major ideas about literary creation: the role of imagination, the magical power of literary language […] the partnership between the writer and the reader in the creative process.* (VALENTE, 1988, p. 349).

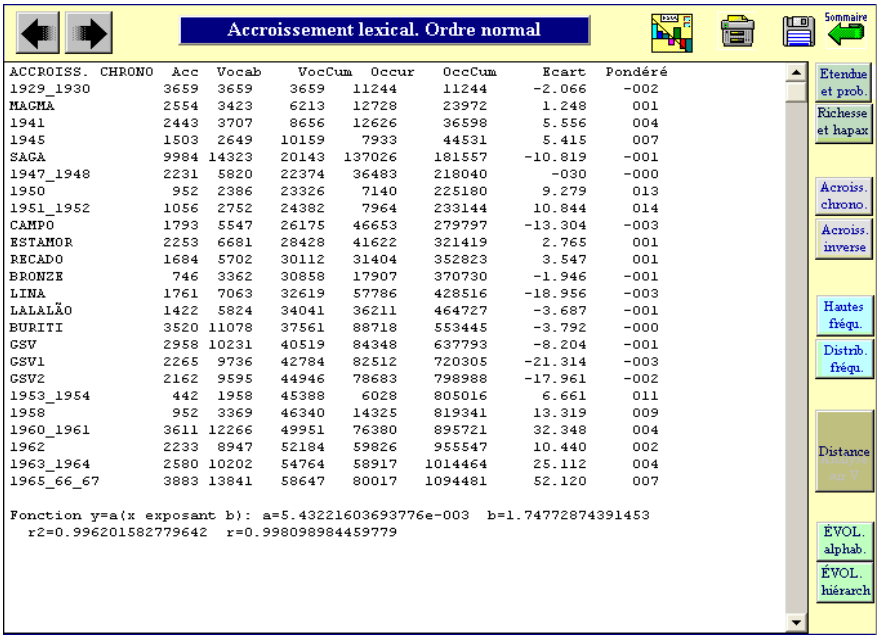
Ainda pensando sobre os prefácios de Tutameia, discorre Paulo Rónai:

> Estórias à primeira vista, num segundo relance os prefácios hão de revelar uma mensagem. Juntos compõem ao mesmo tempo uma profissão de fé e uma arte poética em que o escritor, através de rodeios, voltas e perífrases, por meio de alegorias e parábolas, analisa o seu gênero, o seu instrumento de expressão, a natureza da sua inspiração, a finalidade da sua arte, de toda arte. (RÓNAI in ROSA, 1976, p. 195).

6.4 CRESCIMENTO LEXICAL

Diferente da riqueza lexical, o crescimento lexical estuda diacronicamente a evolução cronológica do vocabulário. É importante aqui diferenciar ou reforçar as diferenças entre a análise de "evolução de vocabulário" e a análise de "crescimento de vocabulário". A primeira trabalha com coeficientes positivos e negativos e organiza as palavras de acordo com a escala de seu emprego, ou seja, analisa a quantidade de vezes que uma palavra X foi usada ou menos usada ao longo da produção; a segunda verifica as inserções de novas palavras que surgem a partir de um conjunto já estabelecido a partir da primeira obra inserida no aplicativo. O tratamento informático dado pelo programa é representado numa tabela de onde podemos acompanhar o desenvolvimento da obra. Eis como aparecem os dados obtidos:

Tabela 1: Crescimento lexical cronológico da obra de Guimarães Rosa.



ACCROISS. CHRONO	Acc	Vocab	VocCum	Occur	OccCum	Ecart	Pondéré
1929_1930	3659	3659	3659	11244	11244	-2.066	-002
MAGMA	2554	3423	6213	12728	23972	1.248	001
1941	2443	3707	8656	12626	36598	5.556	004
1945	1503	2649	10159	7933	44531	5.415	007
SAGA	9984	14323	20143	137026	181557	-10.819	-001
1947_1948	2231	5820	22374	36483	218040	-030	-000
1950	952	2386	23326	7140	225180	9.279	013
1951_1952	1056	2752	24382	7964	233144	10.844	014
CAMPO	1793	5547	26175	46653	279797	-13.304	-003
ESTAMOR	2253	6681	28428	41622	321419	2.765	001
RECADO	1684	5702	30112	31404	352823	3.547	001
BRONZE	746	3362	30858	17907	370730	-1.946	-001
LINA	1761	7063	32619	57786	428516	-18.956	-003
LALALÃO	1422	5824	34041	36211	464727	-3.687	-001
BURITI	3520	11078	37561	88718	553445	-3.792	-000
GSV	2958	10231	40519	84348	637793	-8.204	-001
GSV1	2265	9736	42784	82512	720305	-21.314	-003
GSV2	2162	9595	44946	78683	798988	-17.961	-002
1953_1954	442	1958	45388	6028	805016	6.661	011
1958	952	3369	46340	14325	819341	13.319	009
1960_1961	3611	12266	49951	76380	895721	32.348	004
1962	2233	8947	52184	59826	955547	10.440	002
1963_1964	2580	10202	54764	58917	1014464	25.112	004
1965_66_67	3883	13841	58647	80017	1094481	52.120	007

Fonction y=a(x exposant b): a=5.43221603693776e-003 b=1.74772874391453
r2=0.996201582779642 r=0.998098984459779

Legenda das abreviações: Acc (crescimento); Vocab (vocabulário); VocCum (vocabulário acumulado); Occur (ocorrências); OccCum (ocorrências acumuladas).
Fonte: *Hyperbase* ©, versão 5.4.

O crescimento é então calculado a partir da primeira obra (conjunto 1), ou seja, o número de formas da primeira obra é apenas o efetivo real dela (podemos verificar que, para a entrada de "1929_1930", seus valores são todos iguais: 3659). Na sequência, os cálculos estatísticos e o tratamento informático acrescentam a esse efetivo todas as formas novas encontradas em cada obra (conjunto 2, conjunto 3 etc.) que se seguem no *corpus*; desse modo, o número de formas inéditas, como vimos em item anterior, dependerá tanto da extensão da obra quanto do tema tratado nele.

Para que a leitura da tabela anterior seja facilitada, esclareceremos tomando como exemplo os resultados de *Sagarana*:

SAGA apresenta um crescimento de 9.984 formas, o seu vocabulário é de 14.323 vocábulos, seu vocabulário acumulado de 20.143, tem um total de ocorrências de 137.026 e seu acúmulo de ocorrências é de 181.557.

A partir dos dados brutos gerados na tabela citada anteriormente, podemos visualizar os seguintes resultados do crescimento lexical:

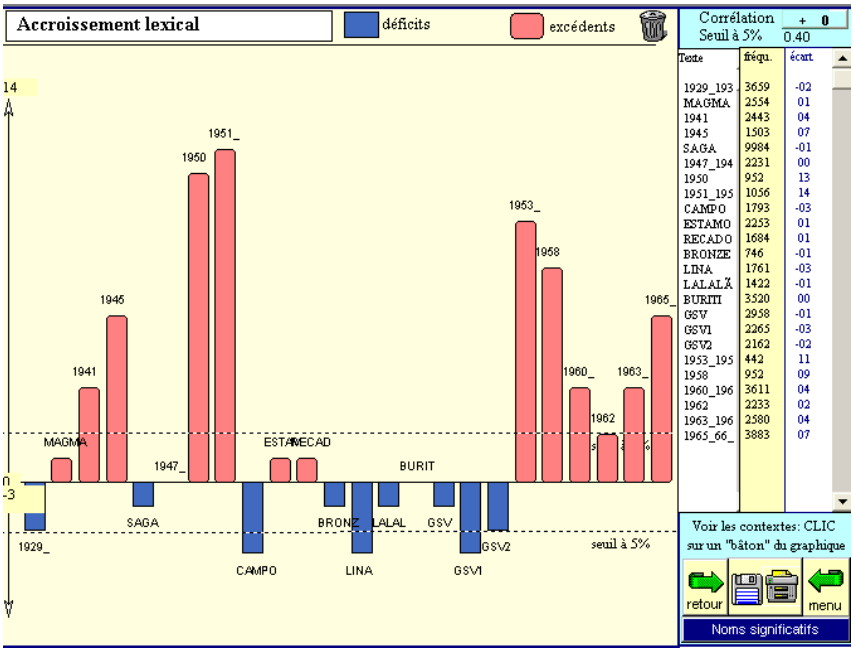


Gráfico 7: Crescimento lexical da obra de Guimarães Rosa. *Hyperbase* ©, versão 5.4.

Da leitura desse gráfico nos questionamos por que *Corpo de baile* e *Grande sertão: veredas*, juntamente a *Sagarana* e aos contos dos anos de 1929 e 1930 se encontram tão díspares do resto do *corpus*? É justamente aí que percebemos que as obras mais voltadas para o tema sertanejo — num procedimento menos diluído como afirma Covizzi e em termos de vocabulário como procuramos mostrar — são obras deficitárias (tão deficitárias que apenas algumas chegam a ultrapassar a margem de 5%) no que diz respeito ao crescimento de vocabulário. Podemos concluir aqui que além do gênero literário, a temática também é fator determinante para o enriquecimento ou o empobrecimento de um vocabulário, mas não o tamanho do texto.

Essa divisão pode soar estranha (rural/não-rural ou rural/urbana), pois tanto *Primeiras estórias* e como em *Tutaméia* são obras repletas de histórias rurais, contudo é bom lembrar que essa divisão existe por causa

do vocabulário empregado, logo, podemos afirmar que o autor, em alguns momentos, se utiliza de um léxico mais impregnado da característica sertaneja e em outros momentos apresenta-se como um vocabulário mais genérico. Galvão (1972), em sua tese sobre *Grande sertão: veredas*, "As formas do falso", afirma que "Guimarães Rosa tem, portanto, um pé na linguagem do sertão e o outro pé no mundo" (GALVÃO, 1972, p. 74).

Ainda sobre *Primeiras estórias* vale lembrar que o conto que inicia e o conto que finaliza ("Os cimos") a obra são uma história só, a de um menino que passa uns dias com os tios em uma cidade. Extraímos, para exemplificar, o primeiro parágrafo de "As margens da alegria", conto que inicia a obra:

> Esta é a estória. Ia um menino, com os tios, passar dias no lugar onde se construía a **grande cidade**. Era uma viagem inventada no feliz; para ele, produzia-se em caso de sonho. Saíam ainda no escuro, o ar fino de cheiros desconhecidos. A mãe e o pai vinham trazê-lo ao **aeroporto**. A tia e o tio tomavam conta dele, justinhamente. Sorria-se, saudava-se, todos se ouviam e falavam. O **avião** era da companhia, especial, de quatro lugares. Respondiam-lhe a todas as perguntas, até o **piloto** conversou com ele. O voo ia ser pouco mais de duas horas. [...] (ROSA, 1974, p. 21) (grifos nossos).

Destacamos algumas palavras para mostrar que já na metade dos parágrafos que iniciam e finalizam *Primeiras estórias*, podemos identificar quatro vocábulos que compõem um quadro urbano (cidade, aeroporto, avião e piloto). É a partir de vocábulos como esses, que, se acrescentados ao léxico geral, definem novas entradas e compõem o que chamamos de obras de caráter não-rural (as que apareceram em vermelho no gráfico 7).

6.5 AS ALTAS FREQUÊNCIAS

Pelas altas frequências podemos ter acesso ao tipo de palavra mais empregada pelo autor: quais verbos, conjugações, substantivos, enfim, quais categorias gramaticais dizem respeito ao traço estilístico de uma escrita. O *corpus* não foi lematizado – a intenção foi essa mesma, pois gostaríamos de acessar as incidências das flexões verbais – por

conta disso, encontraremos formas que podem ser consideradas repetidas quanto ao valor sintático, por exemplo, a preposição “em” e suas flexões (no, na, nos, nas, num).

A seguir, temos a lista das frequências mais altas até a 100ª posição, o que já oferece um panorama interessante sobre a natureza do vocabulário mais empregado por Rosa, podemos assim, ter uma noção tanto da temática quanto do ritmo do texto (ao observarmos a pontuação, pois se percebemos que a vírgula incide mais que o ponto-final, isso é característica de frases longas). A primeira informação na lista indica a posição no *ranking*, a segunda indica a frequência da forma, ou seja, a quantidade de vezes que ela aparece no *corpus*, e a terceira informação é a forma em si:

Ranking	Frequência	Forma
1	117857	,
2	50812	.
3	36379	de
4	35129	-
5	29856	o
6	23767	e
7	23693	que
8	22046	a
9	14590	se
10	12835	não
11	11961	...
12	9281	um
13	9009	do
14	8613	:
15	8412	com
16	8244	em
17	7908	”
18	7465	para
19	7188	no
20	6765	é
21	6599	os
22	6541	?
23	6146	da
24	6135	por
25	5860	eu
26	5741	!

27	5488	era
28	5487	mas
29	5379	ele
30	4653	mais
31	4502	uma
32	4498	;
33	4271	as
34	4027	na
35	3966	me
36	3949	“
37	3508	como
38	3283	só
39	2853	tinha
40	2759	gente
41	2647	mesmo
42	2622	nem
43	2486	sem
44	2325	ao
45	2290	ou
46	2232	seu
47	2218	dos
48	2143	já
49	2021	ela
50	2002	tudo
51	1987	estava
52	1980	foi
53	1980	assim
54	1943	lá
55	1917	meu
56	1906	à
57	1809	muito
58	1787	bem
59	1651	até
60	1642	das
61	1604	quando

62	1580	senhor
63	1554	também
64	1520	todos
65	1503	sua
66	1462	agora
67	1457	ainda
68	1453	ser
69	1446	homem
70	1421	então
71	1382	dele
72	1352	tem
73	1325	minha
74	1290	lhe
75	1277	sempre
76	1271	tão
77	1261	nos
78	1208	dia
79	1206	quem
80	1194	ia
81	1189	Todo
82	1176	ali
83	1163	olhos
84	1151	tempo
85	1140	você
86	1137	outro
87	1132	’
88	1113	porque
89	1107	aí
90	1098	depois
91	1071	bom
92	1068	meio
93	1064	disse
94	1046	podia
95	1021	mim
96	1020	num

97	1012	ter
98	1007	outros
99	1007	isso
100	996	ver

Lista de frequência das 100 palavras mais usadas na obra de Guimarães Rosa. *Hyperbase* ©, versão 5.4.

Se considerarmos o tamanho total do *corpus* rosiano (58.647 vocábulos), uma lista de 100 vocábulos mais frequentes oferecida pelo *Hyperbase* seria insuficiente, pois indica menos de 1% da amostra do *corpus* (0,17%). Contudo, se contarmos a repetição de cada vocábulo da lista das 100 palavras, ou seja, se somarmos todas as frequências das 100 palavras, o quadro se apresenta de outra forma, são 638.527 ocorrências. Uma simples regra de três nos dá o resultado representativo de 58,34% do total de ocorrências do *corpus* (de 1.094.481 ocorrências). Desse modo, a perspectiva para observarmos a amostra de 100 vocábulos é completamente outra. Talvez essa amostra não nos possibilite extrair dados mais interessantes de cunho semântico, pois é uma constante que as primeiras ocorrências mais frequentes sejam as pontuações e as palavras gramaticais (*mots outils*). Por outro lado, Muller (1968) acredita que uma pequena amostra já representa grande parte de um discurso: [...] *un petit nombre d'unités lexicales (ou grammaticales) forment une grande partie de tout discours; on estime que les 50 unités les plus fréquentes, dans un idiome quelconque, couvrent 50% du texte* [...] (MULLER, 1968, p. 162).

Fundamentados nessas palavras de Muller, prosseguiremos então com a análise das formas mais frequentes. Podemos verificar os verbos que mais se destacam: "ser" e "ter", com o imperfeito como o tempo verbal mais empregado (tratamos anteriormente no item sobre evolução do léxico): "era" (5488), "tinha" (2853), "foi" (1980), "estava" (1987), "ser" (1453) (aqui há o problema da ambiguidade, pode ser verbo ou substantivo), "disse" (1064), "podia" (1046), "ter" (1012) e "ver" (996). Outra incidência verbal que surge é o presente do indicativo: "é" (6765), "tem" (1352). Dos substantivos temos: "gente" (2759), "senhor" (1580), "homem" (1446), "dia" (1208), "olhos" (1163), "tempo" (1151), "meio" (1068).

Interessante notar que, dos substantivos mais frequentes, existe uma relação de duas categorias: do "ser" como verbo ou substantivo (haja vista as incidências de "ser", "gente", "senhor", "homem", "olhos") e a do "tempo" ("dia" e "tempo").

As demais ocorrências dos 100 vocábulos mais usados são palavras gramaticais e dentre elas destacamos por ordem de classificação: "de" (36.379); "o" (29.856); "e" (23.767); "que" (23.693); "a" (22.046); "se" (14.590); "não" (12.835); "um" (9.281); "do" (9.009); "com" (8.412); "em" (8.244); "para" (7.465); "no" (7.188); "os" (6.599); "da" (6.146); "por" (6.135); "eu" (5.860); "mas" (5.487); "ele" (5.379); "as" (4.271); "na" (4.027); "dos" (2.218); "das" (1.642); "nos" (1.261).

O artigo de Carlos Maciel intitulado *"Repartição" e "perfil das palavras": a questão da presença/ausência nos estudos de vocabulário* - analisa as 162 formas mais frequentes de um *corpus* extraído da base de dados PORTEXT[72] com 81 textos de literatura brasileira (4.620.146 ocorrências, dentre elas 108.329 formas) abarcando cerca de quatro séculos de produção (de Gregório de Matos, Padre Antônio Vieira a nomes como Lima Barreto, compreendendo contos, romances e poesia). É interessante trazê-lo para o nosso estudo neste momento, pela importância do que diz o artigo acerca do tempo verbal e da relação temporal com os outros vocábulos. Maciel pode contribuir conosco com o seguinte resultado:

[72] PORTEXT é uma base de dados textuais em língua portuguesa criada em Nice-FR no final da década de 80, tem como pesquisadores: Ana Maria Vilhena, Tomás Ramires Pereira de Vilhena, Xuan Luong e outros. (MACIEL, 2000).

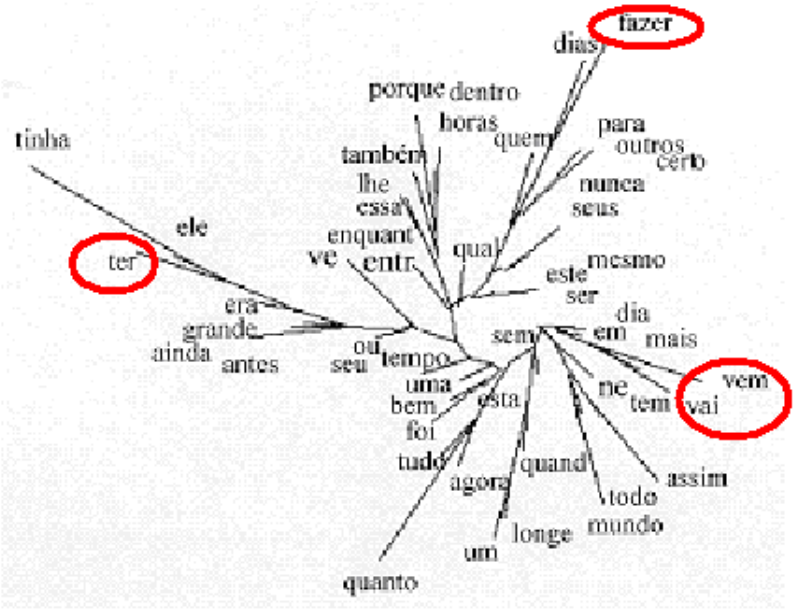


Gráfico 8: Análise em árvore. Adaptação de Maciel, (?), p. 435.

O gráfico apresentado é uma repartição de 93 formas (mais empregadas) extraídas de todas as obras componentes do PORTEXT. Esse tipo de apresentação ramificada (método Luong) apresenta as distâncias das ocorrências que compõem as obras do *corpus*. Para a análise, Maciel considerou 58 formas (artigos, sinais de pontuação e preposições foram excluídos, salvo a preposição "em"). Para nós, o que resultou de interessante no gráfico foi os três grandes "ramos" que delimitam os espaços entre os verbos "ter", "fazer", "ir" e "vir". Do gráfico, Maciel observa:

> - que o campo do verbo «ter» é o do passado (formas «tinha» e «era») e que este campo atrai na sua esteira as formas «antes» e «ainda». Este campo é também o da terceira pessoa «ele»;
> - que o campo do verbo «fazer» é o que carrega a marca do futuro («depois»); ele compreende igualmente as formas «horas», «dias», «quem» e «para», assim como todo um sub-grupo conduzido por

> «porque», no qual encontramos também a forma «enquanto»;
> - que o campo dos verbos «ir» e «vir» compreende as três formas verbais que estão no presente do indicativo: «vem», «vai» e «tem». A forma «agora» pertence (naturalmente) a este mesmo campo, que é também o das formas «quando» e «quanto», do indefinido «tudo», do demonstrativo «esta» e do substantivo «dia»;
> - que, por outro lado, a palavra «tempo» situa-se na intersecção dos três campos.

Faremos agora a mesma experiência com os textos rosianos, partindo dos dados que obtivemos do *ranking* das 100 formas mais empregadas (delas também excluiremos a pontuação, os artigos e as preposições).

Inicialmente, faz-se necessária uma explicação a respeito do procedimento estatístico que gera o gráfico a ser demonstrado. Ele resulta de cálculos de análise fatorial cujo método estatístico é descritivo e multidimensional, permitindo definir, para cada um dos quadrantes de um gráfico, as distâncias entre os elementos que o compõem.

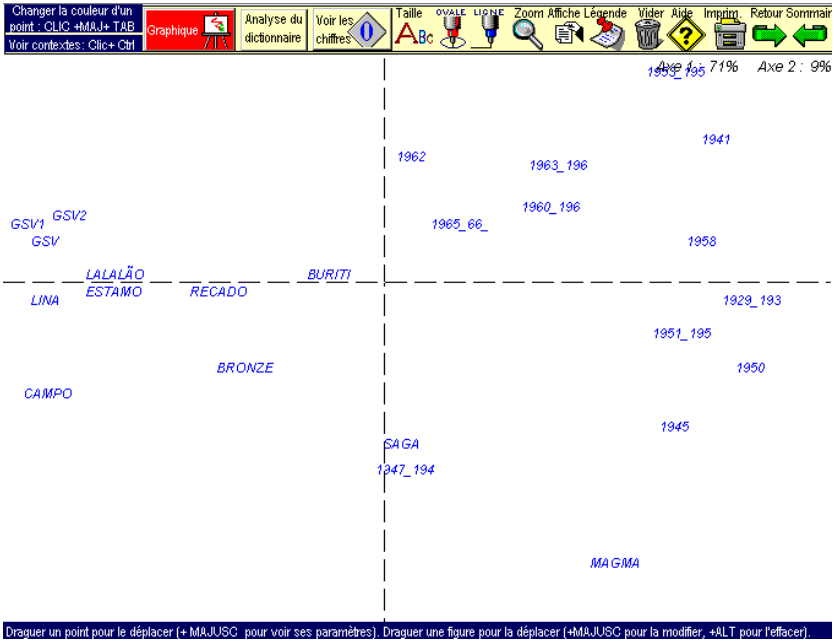
Segundo Lebart (1985), as técnicas de análise de dados ou métodos de estatística descritiva multidimensional são classificadas em duas grandes famílias complementares e podem ser aplicadas simultaneamente: os métodos fatoriais e os de classificação. Os primeiros utilizam cálculos de ajustes que recorrem à álgebra linear, produzindo representações gráficas (das quais os objetos descritos são ilustrados por meio de pontos sobre uma linha reta ou em um plano); já os métodos de classificação agrupam ou ordenam os objetos a serem descritos, como afirmam os autores:

> *Les méthodes factorielles*, largement fondées sur l'algèbre linéaire, produisent des représentations graphiques sur lesquelles les proximités géométriques usuelles entre points-lignes et entre points-colonnes traduisent les associations statistiques entre lignes et entre colonnes.
> [...]
> *Les méthodes de classification* automatique constituent à côté des méthodes factorielles [...] permettent de

> représenter les proximités entre les éléments d'un tableau lexical (lignes et colonnes) par des regroupements ou classes. (LEBART; SALEM, 1994, p. 80-111).

Tony Berber Sardinha (2004), pesquisador da área de linguística de *corpus* no Brasil, define a análise multidimensional como uma abordagem para estudos de *corpus* que usam procedimentos estatísticos, principalmente análises fatoriais, visando mapear associações que possam existir entre conjuntos variados de traços linguísticos de um *corpus* a ser analisado.

Para melhor entendimento, exemplificaremos com um gráfico resultante da análise fatorial sobre os dados da obra completa de Rosa e demonstraremos a seguir a distribuição dos textos rosianos que retrata a distância lexical (ou seja, a distância entre os textos de acordo com o conteúdo lexical que cada um carrega) em dois eixos 1 e 2[73]:



73 Demonstramos os resultados dos eixos 1 e 2, pois a soma desses eixos resulta em 80% do total do *corpus*, são os eixos que mais representam o vocabulário do *corpus*.

Gráfico 9: Análise fatorial sobre os eixos 1 e 2 da obra de Guimarães Rosa. *Hyperbase* ©, versão 5.4.

A seguir o mesmo método, porém com a reunião dos textos, segundo as edições publicadas e conhecidas pela crítica de Rosa. Dessa forma, temos outro viés, agora mais facilitado pela visualização imediata de cada obra no gráfico:

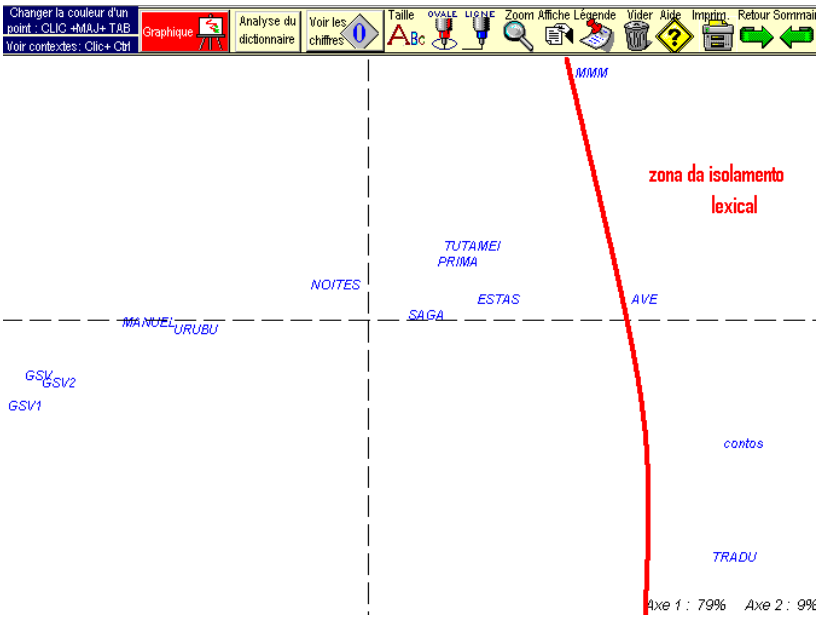


Gráfico 10: Análise fatorial sobre os eixos 1 e 2 da obra de Guimarães Rosa. *Hyperbase* ©, versão 5.4.

Nos quadrantes da esquerda, vemos claramente as obras *Grande sertão: veredas* e *Corpo de baile*, e que ao centro "Noites do sertão" se aproxima de *Sagarana*. Nos quadrantes da direita do gráfico percebemos que *Tutameia*, *Primeiras estórias* e *Estas estórias* compartilham da mesma região lexical. *Ave, palavra* distancia-se um pouco e se posiciona numa "zona de isolamento" - é perceptível a linha que caracteriza o vocabulário rosiano, pois as outras obras que ficaram à direita da linha vermelha são obras que sofreram outras influências em sua composição: os contos de um escritor principiante, o capítulo de um romance policial (*MMM*) desenvolvido em reunião com outros

escritores e a tradução condensada de um romance canadense. No gráfico anterior, podemos detectar o mesmo resultado nas datas de 1929-1930, 1958 e 1960-1961.

Voltemo-nos agora às análises mais detalhadas do léxico, ou seja, apresentamos um gráfico de fatorial que além de trazer as obras, inclui a distribuição das unidades lexicais que retiramos da lista das 100 palavras mais utilizadas por Rosa. Iniciaremos pela apresentação dos resultados em fatorial dos eixos[74] 1 e 2, que comportam cerca de 58% do vocabulário do corpus:

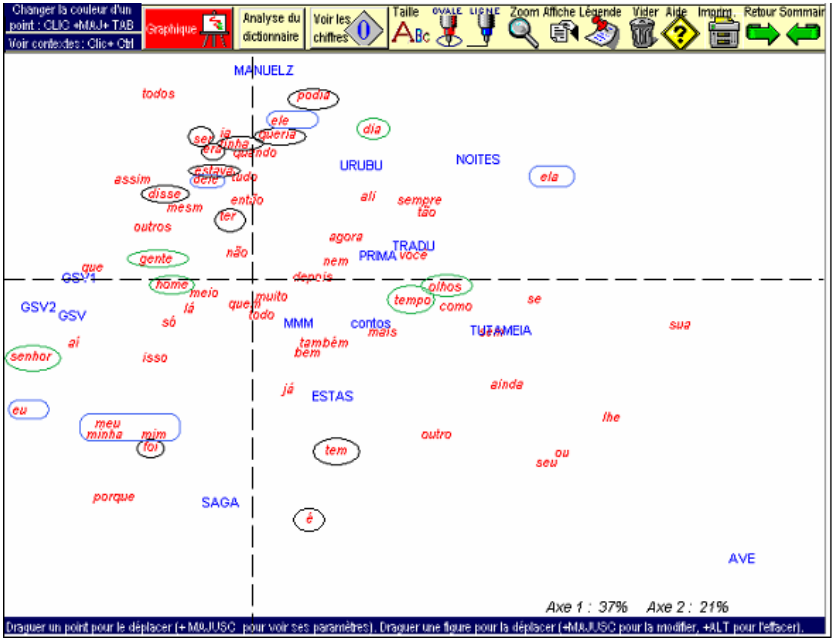


Gráfico 11: Análise fatorial sobre as formas nos eixos 1 e 2 da obra de Guimarães Rosa. *Hyperbase* ©, versão 5.4.

No quadrante esquerdo inferior encontram-se as partes que caracterizam a narrativa em primeira pessoa (“meu”, “minha”, “mim”, “eu”) aspecto representado fortemente por *Grande sertão: veredas* e *Sagarana*, interessante perceber que o único verbo do quadrante é o perfeito do indicativo “foi” cuja conjugação pertence à terceira pessoa

74 Não trabalhamos com os outros dois eixos disponibilizados pelo *Hyperbase* por apresentarem menor abrangência de vocabulário do *corpus*.

(mais presente em *Corpo de baile*); os advérbios de lugar "aí" e "lá" se encontram no mesmo quadrante. Os quadrantes superiores trazem, de um modo geral, em maior número os verbos ("disse", "estava", "tinha", "queria", "podia", "era", "ser", "ter"), as outras três ocorrências verbais estão nos quadrantes inferiores ("foi", "tem", "é"). Aqui, confirma-se a tese ilustrada por Maciel, citada anteriormente, e que vamos repetir: "campo do verbo «ter» é o do passado (formas «tinha» e «era») [...] Este campo é também o da terceira pessoa «ele»" (MACIEL, 2005, p. 435). Se observarmos no gráfico, essas formas estão igualmente reunidas nos quadrantes superiores e bem próximas ao eixo vertical.

É possível retirar conclusões a respeito dessa bipartição entre as categorias *ser* e *tempo* que detectamos pelas análises fatoriais, pois a presença marcante de advérbios temporais ("quando", "sempre", "depois", "agora", "já", "ainda") e a própria incidência dos substantivos ("dia", "tempo") nos levam em direção a tal tema. Para a temática do *ser*, temos os substantivos ("gente", "homem", "senhor") e os pronomes relacionados à primeira pessoa do singular ("eu", "meu", "minha", "mim"), e outros pronomes ("ela", "ele", "você", "seu", "sua", "quem").

Sobre essa questão do ser em Guimarães Rosa, Eduardo Coutinho (1993) já afirmou que o diferencial de Rosa, em relação aos demais regionalistas é que estes dão ênfase à paisagem, ao pitoresco, e a representação do homem é percebida em um plano secundário como mero pertence da região em foco. Rosa faz o contrário disso: o homem é o centro e é por meio dele que enxergamos a paisagem. "O homem não é mais retratado apenas em seus aspectos típicos ou específicos, mas antes apresentado como um ser múltiplo e contraditório e em tantas de suas facetas quanto possível" (COUTINHO in ROSA, 1993, p. 17). O mesmo ocorre com a paisagem que se mostra, para além de uma geografia, como uma "região humana, existencial, viva e presente na mente de seus personagens – uma região que só pode ser definida como uma espécie de microcosmo" (COUTINHO in ROSA, 1993, p. 17).

Essa ênfase no homem, sublinhada por Coutinho (1993), também detectamos no levantamento estatístico que recentemente comentamos, tanto nos resultados dos coeficientes de evolução do vocabulário, nas altas frequências, como nos cálculos de distância lexical, pois encontramos mais vocábulos que se direcionam para a temática do ser, do homem, do jagunço e suas preocupações existenciais, como o tempo (resultado nas fatoriais), o amor e os problemas da alma (resultado visto na evolução do vocabulário).

No próximo item veremos outra abordagem de gráfico sobre as ocorrências, o método de Xuan Luong, que resulta em medição das

distâncias lexicais de um texto para outro e que, diferente da exposição entre quadrantes, o método ilustra os resultados das distâncias em ramificações.

6.6 DISTÂNCIA LEXICAL

As distâncias textuais permitem fazer um julgamento em termos de proximidade e distância – as chamadas conexões por Muller (1977)[75] - como "a intersecção do vocabulário de dois textos" do ponto de vista de seu valor lexical. Essa ideia de correlação lexical já havia sido comentada, no final da década de 1950, por Pierre Guiraud (1959):

> […] *on pourrait établir un tableau de corrélations lexicales entre les différentes oeuvres en les prenant deux à deux pour voir les mots qu'elles ont en commun et ceux qu'elles ont en propre; mais c'est un travail énorme.* (GUIRAUD, 1959, p. 129).

No momento em que Guiraud escreveu essas palavras, o trabalho realmente seria enorme, pois o cálculo da distância deveria ser feito manualmente. Porém, hoje em dia temos aplicativos que, oriundos de tecnologias de programação, realizam cálculos sofisticadíssimos que são suficientes para essa tarefa. No *Hyperbase*, por exemplo, temos dois métodos que implementam o cálculo de distância lexical: o de Jaccard que considera as presenças/ausências dos textos a serem medidos e o de Labbé que leva em consideração as frequências reais e teóricas. Essas distâncias podem ser traduzidas graficamente por meio de análises fatoriais de correspondências ou em árvores[76] (*arborées*). *Grosso modo*, os cálculos operam geralmente sobre a distinção entre conexão de vocabulários (distância sobre V) e conexões de textos (distância sobre N).

Tomando os textos, dois a dois, os cálculos consideram a presença ou a ausência de vocábulos em cada uma das obras, sem levar em conta a sua frequência. Desse modo, uma palavra contribui para a aproximação de duas obras, se ela for comum às duas, ou irá afastá-las caso a palavra seja específica de apenas uma delas. Tal cálculo não foi

---

[75] MÜLLER, 1977 *apud* BRUNET, 2003. *En savoir plus sur Hyperbase*. Disponível em <http:// textopol.free.fr/HYPERBASE2.HTM> acesso em 12 mai. 2011.

[76] Vimos um exemplo desse tipo de gráfico anteriormente sobre o estudo do professor Carlos Maciel (?, p. 435).

apenas estabelecido para mostrar as palavras ausentes das obras A e B, mas para apontar também as que estão presentes nas outras obras do conjunto. Em suma, tal método considera a parte comum do vocabulário particular das obras cuja distância é buscada; a frequência das formas ausentes nas obras A e B, porém presentes em outras partes do *corpus*; e a extensão do vocabulário de cada obra.

Sobre esse método Jaccard (distância lexical exercida sobre V), Brunet[77] explica nos seguintes termos:

> *Il se borne à établir, pour deux textes à comparer, le rapport entre les mots qui sont communs aux deux textes et ceux qui n'appartiennent qu'à l'un des deux. Chacun des deux quotients (dont la somme constitue la mesure de la distance) est le rapport, pour un texte donné, du vocabulaire exclusif au vocabulaire total. Il évolue nécessairement entre 0 et 1. La somme a donc pour limites 0 et 2 (et la moyenne 0 et 1). Pour chaque paire considérée, la distance obtenue tient compte de l'étendue de l'un et l'autre vocabulaires, selon la formule:*
>
> $$d = ((a-ab)/a) + ((b-ab)/b),$$
>
> *où ab désigne la partie commune aux vocabulaires a et b (a-ab et b-ab recouvrant les parties privatives). Dans cette formulation améliorée la somme se situe autour de 1 et reste insensible aux différences d'étendue des deux textes mis en parallèle. Observons en effet que les deux quotients évoluent en sens inverse et d'un même pas, quand s'accroît l'inégalité d'étendue des textes. En une telle situation le plus petit texte aura du mal à affirmer son indépendance face au plus gros, et son quotient d'exclusivité se rapprochera de zéro. Mais pour la même raison, le texte le plus long aura un gros contingent de termes exclusifs qui échapperont par la force des choses au plus petit, et son quotient d'exclusivité*

77 BRUNET, E. Peut-on mesurer la distance entre deux textes? *Corpus* Nº 2: La distance intertextuelle - décembre 2004. <http://corpus.revues.org/index30.html>. Acesso em: 27 abr. 2009.

> *tendra vers 1. Au total on observera une neutralisation mutuelle de ces deux mouvements opposés. (BRUNET, 2004).*

Resumindo, trata-se de encontrar a relação que existe entre as obras tendo como recurso cálculos sobre as palavras que sejam comuns ou próprias, de uma obra e de outra. Ou seja, duas obras são consideradas próximas segundo o vocabulário que compartilham e que se diferenciam em relação às outras obras do *corpus*.

O outro método estatístico utilizado pelo *Hyperbase* é o de Dominique Labbé, que trabalha com a conexão lexical (distância sobre N). Ela visa comparar a superfície dos textos levando em conta as frequências de emprego das palavras. Busca-se avaliar quantas palavras são comuns às obras submetidas à análise. Para cada palavra, calcula-se a diferença entre frequência teórica e frequência observada (real). Esse índice não distingue as diferenças de tamanho de textos, porém, ele não pode ser aplicado sobre textos muito pequenos (menos de 1000 palavras), pois o algoritmo se voltará mais às baixas frequências, como os *hapax* e palavras particularmente raras dos textos.

Na conexão lexical se calcula: a extensão do vocabulário de obra a obra; a extensão do vocabulário de duas obras reunidas num mesmo conjunto (vocábulos e ocorrências); a parte comum às duas obras e a parte exclusiva de cada uma. É sobre os dados brutos da distribuição lexical - vimos no item anterior-, que se estabelece a distância lexical entre os textos, e a partir deles também é que chegamos a outra leitura gráfica mais plausível para a interpretação.

Além dos métodos de Jaccard sobre a distância lexical que se exerce em V, e o método de Labbé, no que compete à conexão lexical (distância sobre N), Brunet se refere ainda a outro método, a análise *arborée* de Xuan Luong[78] (conexão dos textos e de seus vocabulários). Trata-se de uma técnica de classificação que se chama "análise em árvore" desenvolvida na sua tese[79] e aplicada no *Hyperbase*. O algoritmo de Luong, *grosso modo*, produz gráficos que refletem a proximidade dos objetos estudados (textos) a partir de uma distância (cuja elaboração é de Labbé). Para a leitura dos dados, todos os cálculos se resumem em uma representação gráfica, que pode ser visualizada em dois formatos: retangular ou radial.

---

78 Xuan Luong é professor pesquisador da *Université Nice-Sophia Antipolis* e também desenvolve pesquisas no laboratório BCL (*Bases, Corpus, Langage*) de Nice.

79 *Méthodes d'analyse arborée. Algorithmes. Applications*. Université de Paris 5, 1988.

O método de representação em árvore consiste em materializar, sobre um plano, um gráfico de ramificações que demonstram a distância de uma obra a todas as outras, e de obras em pares, traduzindo diretamente essa distância pelo comprimento dos segmentos que leva de uma a outra obra, ou seja, de uma folha do final de uma ramificação a uma outra folha. Essas distâncias apresentam uma visualização mais simples e fácil de interpretar, pois são diretamente transpostas pela representação em árvore, sendo proporcionais ao comprimento dos segmentos.

Verificaremos a seguinte representação gráfica em árvore sobre as formas (as conjunções, substantivos e verbos) que destacamos anteriormente (das temáticas do "ser" e do "tempo"), para que toda a explanação a respeito dos métodos supracitados sejam melhor compreendidos:

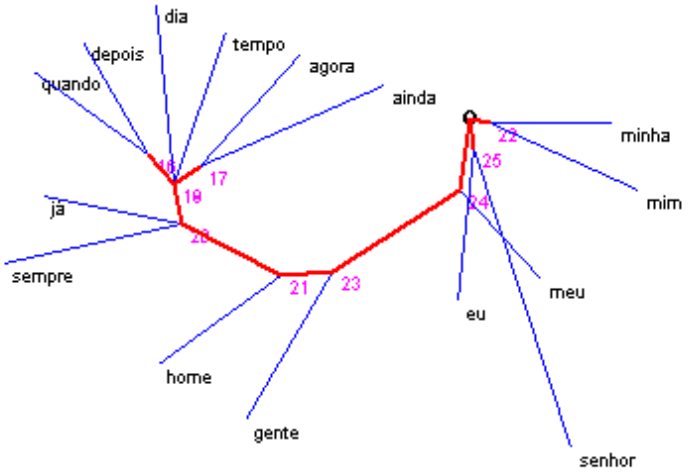


Gráfico 12: Análise em árvore sobre formas da obra de Guimarães Rosa.
Hyperbase ©, versão 5.4

Este tipo de gráfico, que apresentamos anteriormente sobre o estudo de Maciel (2005, p. 435) no item 4.5 deste trabalho, ilustra de forma bastante distintiva essa divisão que vimos tratando entre o *ser* e o

tempo. Percebemos que os nós 21, 22, 23 e 25 compartilham toda a temática do *ser* e para o lado esquerdo da árvore temos em leque a distribuição da temática do *tempo*.

6.7 DISTRIBUIÇÃO DE FREQUÊNCIAS

Outra ferramenta de trabalho para a análise estatística textual é a distribuição de frequência permite estudar as proporções relativas baixas, médias e as altas. Muller (1968) explica que, independente da natureza do texto, ou do idioma, a tabela que irá representar a distribuição de frequências obedece algumas constantes:

> *Les plus frappant, c'est que les effectifs décroissent quand la fréquence croît. Evidemment, quand les effectifs deviennent faibles, pratiquement à partir de 20 environ, des irrégularités se produisent; on trouve une fréquence qui a un effectif plus élevé que la fréquence inférieure; mais il est visible que ces irregularités relèvent des variations aléatoires, et ne mettent pas en cause la loi générale. On peu donc affirmer avec assurance que si dans un texte ou un corpus il y a 100 vocables de fréquence f, il y en plus de 100 de fréquence f – 1, et moins de 100 de fréquences f + 1. […]*
> *Quand la fréquence augmente, les effectifs correspondants tendent vers l'unité; à partir d'un certain point, on voit des effetifs très faibles (1, ou un petit nombre d'unités) alterner avec 0, c'est-à-dire que certaines fréquences ne sont pas représentées, ce qui n'arrive jamais au début du tableau; à mesure que ces intervales entre fréquences représentées augmentent en nombre et en amplitude, les effectifs supérieurs à 1 deviennent de plus en plus rares.* (MULLER, 1968, p. 160-62).

A distribuição de frequências é um recurso que serve tanto para delimitar um campo lexical específico de um autor, como também detectar algumas concentrações temáticas das obras em análise.

Como vimos no item 5.5 deste trabalho, grande parte dos vocábulos mais frequentes são as palavras gramaticais, pois são elas que organizam a lógica do discurso, permitindo-o veicular e transitar dentro

de uma coerência comunicativa, e por isso, carregam em si o seu grau de importância, é o que explica Roberto Busa:

> *Selon l'optique de la fonction discursive, dans chaque lexique on distingue en outre deux zones ou groupes de mots: l'un comprend les mots-véhicules, ceux "avec lesquels" on s'exprime ; l'autre les mots-messages, ceux qui précisent "ce que" l'on veut communiquer*. (BUSA, 1998).

O autor afirma que as palavras gramaticais não mudam segundo o argumento, pois elas não exprimem o conteúdo do discurso, mas a lógica do mesmo. Contudo, na estatística textual, principalmente na estilometria, as palavras gramaticais atingem uma importância significativa no conteúdo, é o que tentaremos explicar.

Maria Lobo (2001) explica brevemente a função das palavras gramaticais no enunciado, dizendo que elas podem ser significativas quando acompanhadas de outras (as lexicais), pois colaboram com a estruturação da frase, servindo para:

> - relacionar o enunciado com a situação de enunciação, indicando os participantes da comunicação, o espaço e o tempo em que ela se dá. São os dêiticos (eu, tu e suas variantes, aqui, aí, agora, possessivos e demonstrativos referentes à 1a e à 2a pessoa etc.);
> - substituir ou referir algum elemento presente no enunciado. São os anafóricos ou representantes (ele, demonstrativos não relacionados à 1a e 2a pessoas etc.)
> - atualizar os nomes, transformando-os de elementos do paradigma ou palavras de dicionário em termos de frase. São os determinantes, como, por exemplo, o artigo, pronomes adjetivos, numerais;
> - indicar quantidade e intensificação (numerais, pronomes indefinidos quantitativos, advérbios quantitativos);
> - relacionar palavras no sintagma (preposições) e operações na frase (conjunções e pronomes relativos);
> - estabelecer coesão textual, seja dentro de uma frase, seja entre frases diversas (anafóricos, conjunções). (LOBO; APRÍGIO, 2001).

Portanto, o emprego das palavras gramaticais se refere à sintaxe e à organização textual baseadas em regras gramaticais, contudo, essas regras sempre podem ser violadas na medida do efeito expressivo que se quer atingir, daí a questão do estilo:

> Palavras gramaticais podem perder, em certos empregos, esse valor gramatical e tornar-se meros elementos de realce ou ainda receber um valor nocional, aproximando-se das palavras lexicais. Também palavras lexicais podem perder seu valor nocional, gramaticalizando-se. (LOBO; APRÍGIO, 2001).

No caso de Guimarães Rosa, as palavras gramaticais exercem outras funções, vejamos a seguir alguns exemplos retirados do *corpus* em que podemos perceber essa inversão, cujo valor agregado ao advérbio (ou pronome, no caso dos exemplos a seguir) transforma a palavra gramatical em lexical, reforçando a questão do estilo:

```
ES1149b|   te , dançando . As iaiás também . O quando o dia já estava pronto para a
GS1942c|    mesmo sem juízo nenhum falável ; o quando no meio deles se trança um aj
G62018e|   senhor sabe - a briga de ventos . O quando um esbarra com outro , e se e
G62060e|   tribuído endireito em Zé Bebelo . O quando falou :
```

Lista de concordância da expressão “o quando” na obra de Guimarães Rosa. *Hyperbase* ©, versão 5.4.

```
RE1199d|    os caponetes nas dobras , sempre o sempre . Mesmo seo Jujuca se queixav
LI1387a|   nte . À hora , lá estavam fazendo o sempre o Pernambo e Placidino , e o
BU1684c|   mples satisfação , era qual , era o sempre . Aquela noite , como no mome
GS1966d|   de ferro de Joãozinho Bem - Bem - o sempre sem mulher , mas valente em q
G72204f|    de mim ! Quem é que era o Demo , o Sempre - Sério , o Pai da Mentira ?
112556d|    , abrindo as ventas , para tomar o sempre desacostumado forte cheiro de
112753d|   íamos , VÍNHAMOS - do jamais para o sempre . □
112756c|   ridamente - através de quem nós : o sempre : o CIMO ! □  … ” no meio do
143278b|   a coleção de milagres . O nunca é o sempre , escondido às nossas costas
```

Lista de concordância da expressão “o sempre” na obra de Guimarães Rosa. *Hyperbase* ©, versão 5.4.

```
11 156b|    repatriado , para a epilogação . O nada acontece muitas vezes . Assim -
G62015a|    um . Acho que nem dormia , comia o nada , nada , às pressas , pitava o
G62060d|   m frente da barriga - só esperava o nada virar coisas . □ Acontecesse o
G62094d|   lma . Dias inteiros , nada , tudo o nada - nem caça , nem pássaro , nem
G62132d|   nha nenhuma , o senhor sabe ? Sou o nada coisinha mesma nenhuma de nada
G72268f|   ever de crime . E o homem da égua o nada de tudo espiava , por mais inte
G72277a|   edo dele era medonho … Só achamos o nada dele … " - assim rendiam explic
112589a|    Nhinhinha , só , sentada olhando o nada diante das pessoas : - "Eu quer
112711b|    repatriado , para a epilogação . O nada acontece muitas vezes . Assim -
112714a|   inverso : se regresso . □ Muito é o nada nesta vida . □ E , dos três , q
112740b|   S - recapturada - a Menina . Ou - o Nada ? Nada , nada , a lua : □ Lua p
112759c|   / como se o céu não fosse curvo . O nada é muito vivente : os animais ,
112760a|   □ por não ser um sorriso … ) □  - o NADA . Liso , calmo , quieto , fresc
112760a|   , frio , morto , imperturbado - é O NADA . □  - Minha mãe ! Minha mãe !
122975b|    , sem carecer de que acontecesse o nada . Do jeito feito agora , no cor
143279a|   inho , ando , apoio - me : contra o nada , só minha memória trabalha , q
143285a|   uma definição " por extração " - "O nada é uma faca sem lâmina , da qual
143400d|    a ela lembrar os mortos . □ Ele - o nada a se fazer - pegado pelos entre
143404d|   o na base do passo . Passou - lhe o nada pela cabeça . Na rua , à vista
```

Lista de concordância da expressão "o nada" na obra de Guimarães Rosa.
Hyperbase ©, versão 5.4.

```
SA 462c|    í eu esperar notáveis coisas para o depois . □ Santana costuma dizer : -
ES1070d|   ifamá - la a mal . Morreu , sobre o depois , sua alma veio assombrar . M
GS1943e|   senhor sabe : nome - da - mãe , e o depois , quer dizer - meu pinguelo .
G62118d|   s só até uma parte - não entendia o depois - do - fim , o confrontante .
G62150a|    esquerda e rumo do norte . Desde o depois , o do poente mesmo . Com for
G62153b|   s embora da jagunçagem , que já é o depois - de - véspera , que os vivos
132993c|   e seguem , para lá , lá , em todo o depois - as das sombras matosas , e
133161i|   norme . Sozinho , entre o antes e o depois , como o sol se punha : amare
133170f|   sco . Precipitavam - se o antes e o depois . Fechou - se o círculo . Nad
```

Lista de concordância da expressão "o depois" na obra de Guimarães Rosa.
Hyperbase ©, versão 5.4.

```
BR1291a|    ueria era que se achasse para ele o quem das coisas ! □  A VOZ DO VIOLEI
LI1333g|   sido um sabiá ou um sofrê ; mesmo o quem - quem - que em toda baixada de
G62184e|    Eu pensava , como pensava , como o quem - quem remexe no esterco das va
112734b|   buraco de escaravelho . Não havia o quem que fosse , mas havia o por se
```

Lista de concordância da expressão "o quem" na obra de Guimarães Rosa.
Hyperbase ©, versão 5.4.

Essas listas de concordâncias servem como exemplos, ou mera constatação, sobre o desdém que se cria por vezes nos estudos de estatística de textos ao ignorarem a presença e a importância das palavras gramaticais, quando se voltam apenas ao estudo dos substantivos pelo seu valor semântico. Ainda assim, como pudemos verificar no *corpus* rosiano, é possível fazer leitura impregnada pelo teor semântico a partir de palavras gramaticais.

O processo de substantivação dessas palavras, bem como a criação de neologismos em Rosa já foi muito comentado pela crítica, e nosso objetivo não é discorrer sobre tal fenômeno. Porém, como também encontramos alguns advérbios e pronomes que nos levaram a identificar duas categorias semânticas, definindo aspectos para as temáticas do "ser" e do "tempo"- vimos ao longo deste capítulo - decidimos trazer à discussão as frequências das palavras gramaticais porque tais palavras exercem importância na ficção rosiana, e não devem ser consideradas apenas instrumentos de um discurso.

7 “ ... CADA UM O QUE QUER APROVA, O SENHOR SABE: PÃO OU PÃES, É QUESTÃO DE OPINIÃES”: O QUALITATIVO E O QUANTITATIVO NOS TEXTOS LITERÁRIOS

O todo sem a parte não é todo,
A parte sem o todo não é parte,
Mas se a parte o faz todo, sendo parte,
Não se diga, que é parte, sendo todo.
*(*Gregório de Matos)

A abordagem estatística que apresentamos aqui foi um ensaio, como tentativas sobre as possibilidades de leitura que a estatística de texto pode propiciar dentro da literatura. Quando realizamos contagem de palavras que um determinado autor utiliza, quando buscamos uma estrutura de frase mais comum, ou temas mais explorados, de igual forma, apelamos para o estudo estatístico, ainda que em um grau menor, básico e leigo, a contagem existe.

Contudo, com o auxílio de programas voltados para o estudo estatístico de vocabulário, podemos aprofundar mais essas análises, aumentar o grau estatístico e ainda, obter dados que a leitura humana não seria capaz de realizar. Vide o caso do cálculo de evolução do vocabulário, como poderia um leitor perceber, ao ler a obra completa de um dado autor, qual o vocabulário mais utilizado ao fim de sua carreira literária e quais palavras foram sendo deixadas de uso? Não seria impossível manualmente, mas daria muitíssimo trabalho. A leitura tradicional traz uma ideia do campo semântico utilizado, mas se fôssemos perguntar a um leitor balzaquiano, por exemplo, qual o vocabulário específico de Balzac, provavelmente a resposta não traria o mesmo rigor de um aplicativo voltado para esse tipo de extração de dados. É claro que questões mais evidentes qualquer leitura traria, como os temas mais expostos, mas a garimpagem de programas estatísticos traz informações que está para além do olhar humano.

A necessidade de contar palavras segundo diversos procedimentos e de chegar a resultados de cálculos estatísticos ainda é pouco evidente no campo da teoria literária, haja vista a oferta quase nula de disciplinas ofertadas na graduação em Letras que se relaciona aos estudos estatísticos. Historiadores, analistas do discurso, tradutores, sociolinguistas já utilizam métodos quantitativos que renovam as abordagens tradicionais, por que privar os estudos literários de tal

vertente? Por que deixar os estudos literários à margem das transformações que as novas tecnologias contribuem para nosso meio? Não seria o momento de romper com o preconceito e aceitar a contribuição ou até mesmo o auxílio do quantitativo ao qualitativo na literatura?

No final da década de sessenta, focado na estatística lexical, Charles Muller diria:

> *On verra [...] qu'il existe des applications de cette méthode qui ne postulent pas un dépouillement intégral du texte étudié; mais on commencera par examiner ce dénombrement complet du vocabulaire, qui a l'avantage d'être "neutre", de fournir des matériaux bruts qui seront accessibles à des nombreux chercheurs, sans préjuger des exploitations possibles.* (MULLER, 1969, p. 30).

Na tentativa de se isentar de qualquer tendência ou de impressões que um leitor pode ter de uma obra, no que diz respeito ao estilo de escrita de um autor ou de um tema abordado preferivelmente por ele, é que optamos pela metodologia da estatística textual aplicada aos estudos literários. Se na linguística de *corpus* já foi afirmado que muitas vezes a intuição humana sobre o entendimento da linguagem foi inexata (SAMPSON 2001 *apud* SARDINHA 2002, p. 16), por que isentarmos os estudos literários? Não queremos aqui, descartar a importância da intuição, pois muitas vezes, ela é o ponto de partida de muitas pesquisas, contudo, já que temos as ferramentas para ir além da intuição, por que não utilizá-las? Por meio delas, a obra literária se torna um grande corpo que poderá ser dissecado em todas as suas partes, explorado em cada pormenor, em cada detalhe, desde o lugar mais frequente da vírgula em uma sentença à maior incidência de uma palavra em todo o conjunto. Contudo, o crítico deve ter em mente que a vírgula faz parte de um todo e que esse todo é um texto literário. Agora, se a estatística busca exatamente a imparcialidade, não cabe então ao pesquisador de Letras, muito acostumado a exercer leituras tendenciosas e impressionistas, a tentativa de ser parcial ou tendencioso com os dados que ele obtém por meio da estatística textual. Para Freitas (2007), nos estudos literários, o pesquisador já traça seus objetivos e intuições por uma série de fatores: a obra, a bibliografia crítica, informações extratextuais e contextuais de que já dispõem de uma tradição secular de estudos; e que cada caso de investigação demandará as ferramentas necessárias para o desenvolvimento da pesquisa.

Por vezes, podemos perder o foco da pesquisa pelos muitos resultados e dados que surgem e também, por isso, vários fatores devem ser considerados, para que o todo não seja interpretado apenas pela parte. Todo dado é resultante de interligações de cálculos, por exemplo, se, no resultado de um quadrante (oriundo de um gráfico de análises fatoriais), aparecerem as obras A, D e G e, no outro quadrante, as obras B, C, E e F, isso expressa os cálculos de aproximações de vocabulário em comum e de vocabulário distinto. Para determinar o vocabulário que está em cada quadrante, é necessário recuar, e buscar a matéria-prima desse resultado, ou seja, a lista de palavras de cada obra. Dessa lista fornecida, podemos reconhecer as palavras que foram mais empregadas e com isso determinar qual vocabulário responsável pela aproximação ou distância de cada obra. Citamos esse exemplo para ilustrar a sinergia que há entre os mecanismos estatísticos de um aplicativo, que, para o entendimento dos resultados ser mais coeso, será necessário ir às partes, tomando o cuidado de não vê-las separadamente.

A imparcialidade dos dados fornece outra forma de ler o texto literário, a exemplo disso, mostraremos o seguinte gráfico:

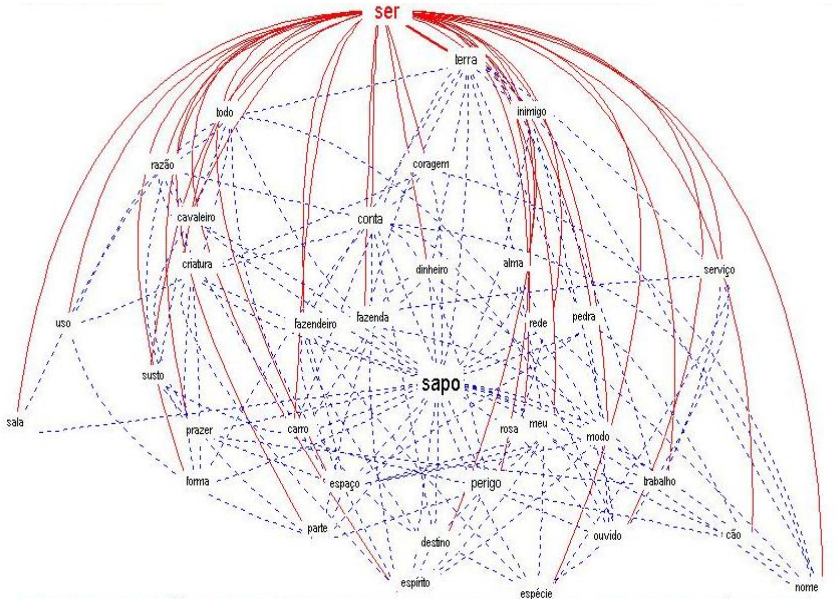


Gráfico 13: Correlações da forma-polo “ser” na obra de Guimarães Rosa.
Hyperbase ©, versão 2011.

Trata-se de um gráfico sobre as associações entre palavras na obra completa de Guimarães Rosa, nesse caso, a palavra-polo escolhida para estabelecer as associações foi "ser". O gráfico representa as palavras conforme seus empregos nas obras e suas associações ao vocábulo "ser" em todo o conjunto das obras. A quem ou a quê a palavra "ser" na obra de Rosa se relaciona? As linhas vermelhas representam as fortes ligações e os traços pontilhados em azul as fracas. Da mesma forma que as conexões vermelhas se distribuem com a forma-polo "ser", as linhas tracejadas em azul se conectam com a forma "sapo", e todas elas de alguma maneira se relacionam com as palavras "ser" e "sapo". Por exemplo, "ser" está ligado à "terra" que está diretamente ligado à "sapo". A partir daí, podemos observar como a estatística textual trabalha na micro e na macroestrutura do texto.

Na perspectiva micro, focalizando o léxico, podemos investir também em confirmações mais rigorosas de impressões que surgem sobre nós leitores. É possível, por exemplo, verificar se um escritor utiliza de modo peculiar uma dada conjunção no início de frases, fato que pode ser intuído ou percebido em formas de leituras tradicionais, mas que temos como, de imediato, verificar de que modo tal fenômeno acontece concretamente. Um aplicativo voltado para estatística e estilometria poderá organizar as informações de maneira que se confirmem essas impressões iniciais e, a partir daí, se viabilizem novas interpretações. E essa organização implica sequência de textos, extrações de dados, análises e interpretações desses dados e estratégias de leitura que se diferenciam da forma tradicional de se ler. Contudo, vale lembrar que, apesar de estratégias diferenciadas, nada impede de chegarmos aos mesmos resultados, pois muitas vezes a intuição pode ser confirmada. É o que vimos no capítulo 6 desta tese quando confrontamos algumas análises de Sperber (1982), Covizzi (1978) e Daniel (1968), chegando, por vezes, às mesmas conclusões, mas trilhando percursos muito distintos.

Por outro lado, se pretendermos analisar a macroestrutura do *corpus*, dispomos de ferramentas que viabilizam caminhos panorâmicos. É o caso das análises sobre as fatoriais ou as análises em árvore. O texto literário se transforma não apenas em números ou cálculos fatoriais, mas em um novo universo que retrata desde uma ínfima incidência (aquela palavra que foi escrita uma única vez por seu autor) dentro de uma esfera de mais de cinquenta e oito mil palavras (no caso de Rosa) ao vocabulário de preferência ou habitual. É por meio desse novo universo apresentado que são oferecidas à obra outras perspectivas de leitura.

Contudo, vale lembrar que o texto não é um conjunto de estatísticas; texto é o que construímos a cada leitura da obra; tampouco a obra é um conjunto de estatísticas; a obra é o que nos entrega seu autor. O que queremos demonstrar aqui é que, para ir da obra ao texto, podemos passar por diversos caminhos (que não são necessariamente excludentes, muito pelo contrário), e um desses caminhos é a estatística textual.

Geoffrey Rockwell (2003), em artigo intitulado *What is the text analysis, really?*, comenta que as ferramentas tecnológicas disponíveis para a análise de textos exige do pesquisador maior complexidade nas formulações de suas pesquisas, ou seja, essas ferramentas de tal forma, atingem o modo de pensar do pesquisador. Para o autor, nós temos que pensar não apenas sobre como representar o texto mas também sobre o ato de analisar e quais ferramentas empregar essas análises com o computador. A lógica das ferramentas, apesar de (ou por causa de) buscar tranparência no uso, pode aumentar ou restringir diferentes tipos de leitura, que por sua vez os torna uma melhor ou pior para práticas de leitura crítica incluindo a realização da crítica. Contudo, vale lembrar que tampouco há utilidade alguma nessas tecnologias, se o pesquisador não estiver munido de questionamentos e reflexões, elas não são um "passe de mágica", e não vieram para substituir nenhum esforço de reflexão.

7.1 DEVE A CRÍTICA JUSTIFICAR O USO DESSA METODOLOGIA?

Não bastam apenas os dados, é preciso interpretá-los e ainda dar sentido aos mesmos, pois, como verificamos neste trabalho, podemos fazer diferentes graus de leitura dos dados: a descritiva e a analítica, e ambas voltadas para a área em que se quer trabalhar, no nosso caso, a teoria literária. Se ficarmos no primeiro nível, na descrição sintética dos dados, talvez não nos servisse muito, pois qual seria a importância em saber, por exemplo, o vocabulário total de Guimarães Rosa? Para definir um dicionário rosiano e com isso identificar a variedade lexical do autor? Mas, além disso, que importância traria essa informação para os estudos literários?

A outra leitura, a analítica, se torna mais interessante, pois além da descrição ela implica análise dos dados. Desse modo, o pesquisador dá um passo a mais e com isso a investigação se torna mais significativa para contribuir com a área em que se destina. As análises dos dados para a literatura exigem do investigador certa bagagem de leitura não apenas da obra a ser pesquisada, como também de estatística, pois de que

adiantariam descrever os dados sem poder aplicá-los? Análise aqui significa, por exemplo, observar um quadro resultante de uma fatorial que mostra uma obra A se distanciando de outra obra B e identificar o porquê dessa distância, como ela aconteceu, quais os elementos que a tornaram visível. Não queremos aqui afirmar que o usuário do aplicativo deva dominar os cálculos hipergeométricos – e esse nem é o objetivo -, mas no mínimo conhecer alguns princípios de estatística e, principalmente, saber qual a função de cada ferramenta oferecida pelo programa, pois nas ferramentas mais avançadas (os desvios, a riqueza lexical que determina qual obra é mais rica ou pobre em relação a outras; o crescimento do vocabulário etc.) estão as variáveis estilísticas e a partir delas resultará uma leitura mais significativa que legitimará o estudo na área da literatura.

Agregar aos resultados e a toda essa matemática envolvida razões que justifiquem a manobra do pesquisador de literatura, exige do mesmo criatividade, potencialidade de leitura e interpretação, a partir disso é que o resultado do gráfico em árvore ganhará sentido para a teoria literária. O que um crítico literário percebe além de resultados numéricos? Outra questão que deve ser compreendida é a tendência que temos, de ainda assim, com os resultados em mãos, com os dados analisados, retroceder (se é que assim podemos considerar) ao processo e intuir sobre os resultados, comportamento muito comum, haja vista a prática secular de refletir dessa maneira. Pois podemos partir de uma intuição para investigar e não seria interessante obter os dados e intuir? Mas aqui sobrevoa um paradoxo, porque o pesquisador de literatura deve utilizar o seu repertório de leitura, para assim, exercer a sua criatividade sobre os dados estatísticos.

Esta forma indireta de se chegar ao texto literário faz uma radiografia panorâmica do vocabulário e possibilita encontrar muitos pontos a serem trabalhados. Um dos grandes obstáculos é selecionar quais dados desenvolver. Uma vez encontrada a proposta de análise, devemos buscar as ferramentas estatísticas necessárias para o desenvolvimento da proposta. Alguns pesquisadores preferem lançar o *corpus* no aplicativo e a partir dos dados que o mesmo gerar, fazer as leituras das características extraídas dos resultados, é um método que não apresenta, *a priori*, uma hipótese, uma dúvida, um motivo. Outros já preferem o contrário, buscar nos dados a verificação de uma ideia, de uma proposta estabelecida anteriormente por meio de leituras sobre a obra a ser examinada. É o que afirma Freitas (2007):

> [...] não pode haver uma crítica que prescinda do conhecimento profundo da obra sem que haja uma teoria que sustente a organização e a interpretação desses dados. Antes mesmo de extraí-los, o pesquisador tem que ter em mente que tipo de informação deseja extrair do *corpus*, de acordo com os objetivos de seu estudo. Dessa posição aparentemente simples, surgem as principais questões metodológicas do campo. O que contar? Que recursos estatísticos escolher para organizar o que foi contado e dar uma resposta satisfatória às dúvidas ou questionamento que pode ajudar a solucionar? Os dados coletados são suficientes? Deve-se usar amostra? De que tamanho? (FREITAS, 2007, p. 59).

Nosso trabalho vai ao encontro dos dois tipos supracitados de investigação estatística. Pois, nossa hipótese, além de desvendar as características gerais do vocabulário rosiano, foi testar hipóteses de três críticas literárias a respeito do léxico preferencial do autor. Pois a crítica chega intuitivamente a conclusões passíveis de serem contadas automaticamente para objetivar melhor as apreciações estilísticas. Vale dizer que nossos questionamentos permearam o real vocabulário (no sentido estatístico) de preferência do autor. Constatamos existir uma evolução lexical que se distingue em três períodos oscilatórios: não-rural, rural e não-rural novamente, e que grande parte das obras cuja temática é o sertão apresenta déficit de vocabulário. Outra característica que encontramos foi *Ave, palavra* que trazer um repertório lexical mais diversificado e com mais entradas de *hapax* e que o repertório geral de vocabulário rosiano se concentra em dois grandes grupos temáticos: *ser* e *tempo*.

8 "... AO FIM RETOMO, EMENDO O QUE VINHA CONTANDO": DESDOBRAMENTOS

Neste capítulo final resumiremos em três partes o que esta tese traçou, ou seja, as características gerais e específicas do vocabulário de Guimarães Rosa, as comparações entre os resultados e a crítica literária e por último, uma breve contribuição à metodologia.

8.1 DAS CARACTERÍSTICAS GERAIS E ESPECÍFICAS DO VOCABULÁRIO ROSIANO

Para cumprir a tarefa de reunir abrangentemente o léxico (1.094.481 ocorrências, 58.647 vocábulos) de Guimarães Rosa dentro de um percurso cronológico de sua produção literária, utilizamos uma ferramenta informatizada chamada *Hyperbase*, desenvolvida pelo laboratório BCL da Universidade de Nice, e por meio dela, fizemos nossas leituras de análises da estatística de textos.

Verificamos que mais da metade do vocabulário (52,81%) não se repete, demonstrando ser bastante rico e diversificado e representando maior riqueza lexical nos textos curtos, os textos com mais diálogos apresentaram menor diversificação. Concluímos que a maioria das obras apresenta mais déficit de vocabulário que excedente, principalmente as que retratam a temática do sertão (*Sagarana*, *Grande sertão: veredas* e *Corpo de baile*). Os resultados dos *hapax* também nos levaram à mesma conclusão, apresentando *Ave, palavra* como a obra mais diversificada. A partir dos resultados analisados, vimos que a riqueza lexical é avaliada em relação a critérios como o gênero literário, o estilo e a temática, porém, no caso de Rosa o tema não parece contribuir com a diversificação do léxico.

Gostaríamos de salientar que o tema sertanejo é bastante presente nas obras *Tutaméia*, *Estas estórias* e *Primeiras estórias*. Porém, a ênfase no vocabulário é dada nas três obras destacadas (*Sagarana*, *Grande sertão: veredas* e *Corpo de baile*).

Do mesmo modo, estudamos a pontuação rosiana por meio da ferramenta que analisa a evolução do vocabulário, e, percebemos que, ao longo de sua produção, Rosa usou mais a vírgula e o ponto-final, enquanto abandona o emprego da exclamação. Quanto à interrogação, o sinal apresenta tendências deficitárias no início e no fim da linha cronológica, porém tem seu auge positivo nas grandes obras *Grande sertão: veredas* e *Corpo de baile*. Sobre os verbos, podemos afirmar que

os tempos verbais mais utilizados são: o subjuntivo do presente e o futuro do pretérito, por outro lado, o pretérito imperfeito foi menos empregado no final da produção.

Pelo levantamento que fizemos a respeito do crescimento de vocabulário concluímos que, em Rosa, não é o tamanho da obra que vai definir a sua riqueza, mas sim a temática.

Nas análises fatoriais constatamos que as obras mais distantes da linha que permeia a característica rosiana foram as de naturezas distintas: os contos de um escritor principiante (1929-1930), o capítulo de um romance policial (MMM - 1958) desenvolvido em parceria com outros escritores e a tradução condensada de um romance canadense (1960-1961).

Detectamos, tanto nos resultados dos coeficientes de evolução do vocabulário, nas altas frequências, como nos cálculos de distância lexical, mais vocábulos que se direcionam para a temática do *ser* e suas preocupações existenciais, tais como o *tempo*, o *amor* e a *alma*. Do mesmo modo, redefinimos a importância das palavras gramaticais, pois além das mesmas cumprirem com o funcionamento da lógica do discurso, elas também possuem carga semântica. Encontramos alguns advérbios e pronomes que identificaram dois campos semânticos: "ser" e "tempo".

8.2 DAS SOBRE-HIPÓTESES DE SPERBER, COVIZZI E DANIEL

Ademais da descrição sobre as características gerais e específicas do vocabulário rosiano, nossa tese foi também confirmar ou complementar intuições oriundas de leituras convencionais, por isso, nos apoiamos em Sperber (1982), Covizzi (1978) e Daniel (1968). Das confirmações de Sperber, buscamos trabalhar com as análises sobre o crescimento e evolução de vocabulário. Vimos com Sperber (1976), que no processo de estruturação da obra rosiana, o estilo de Rosa se sobressai ao tema. Analisamos o resultado sobre a variação de vocabulário, o gráfico mostrou déficit nas obras que retratam o sertão. Nas outras obras (as não-rurais) houve uma forte diversificação do léxico. Concluímos que, a temática em Rosa é um elemento de estilo, contudo ela não contribui para a diversificação do léxico.

Sob a perspectiva de Covizzi (1978), vimos pelo levantamento de vocabulário por ocorrências de *hapax*, concluímos que a fase

expressiva[80] de Rosa, a qual teria o seu cume em *Grande sertão: veredas*, não ocorreu em termos lexicais, e que a fase explicativa se utilizou de uma diversidade lexical muito maior que a fase expressiva. Por meio dos indícios verbais extraídos dos coeficientes positivos e negativos sobre a evolução do vocabulário, acordamos com Covizzi (1978), que, ao final da carreira literária, Rosa compôs mais locuções discursivas e comentários. Ainda sobre as afirmações de Covizzi (1978) a respeito do percurso narrativo de Rosa, a autora conclui que a partir dos textos de 1956 há um processo de diluição do teor regional como em *Primeiras Estórias*. Para analisar essas afirmações, fizemos um levantamento de vocabulário pelas ocorrências de *hapax*, desse modo conseguimos verificar o comportamento da entrada de novos *hapaxes*, caracterizando assim o processo de criação e renovação de vocabulário.

Para as análises que fizemos sobre as afirmações de Daniel, buscamos primeiramente, verificar se houve uma ruptura do léxico em duas fases, rural e não-rural. Utilizamos a ferramenta que mede a evolução e outra que aborda a distância lexical. Dos resultados obtidos, verificamos que há dois momentos de ruptura no material linguístico de Rosa, apresentando uma grande diferença de aparecimento de *hapaxes* e riqueza lexical, ou seja, de diversidade vocabular. As obras de maior fôlego (*Sagarana*, *Grande sertão: veredas* e *Corpo de baile*) apresentaram um vocabulário mais restrito, enquanto que as outras obras mostraram um vocabulário mais variado.

8.3 DO VELHO REFORMADO PELO NOVO

A estilometria como uma nova textualidade tem configuração híbrida e multidisciplinar e exige do pesquisador das Letras uma postura humilde em saber reconhecer suas limitações e assumi-las para que o quarteto mínimo necessário ao desenvolvimento dos estudos aconteça: o conhecimento do estatístico, do informático, do linguista, e no nosso caso, da crítica literária. Como indica Rockwell (2003), já viemos de uma longa tradição de leitura, editoração e prática artística, teórica e crítica com o texto impresso. Contudo, o problema maior está no posicionamento crítico, não se trata apenas de direcionar a crítica do impresso para o digital, pois o momento de transposição não é imediato,

[80] Vale lembrar que Covizzi (1978) considerou para a sua análise duas categorias "expressão" e "explicação". Em nossa tese, retomamos essas categorias para prosseguir com o trabalho da pesquisadora, porém, sob outro viés e alcançando a obra completa de Guimarães Rosa.

se na Europa ou nos Estados Unidos isso já é uma realidade em debates universitários, ainda no Brasil, na área de literatura, temos muito a nos envolver e desenvolver.

O velho texto literário reencapado pelo digital demanda uma crítica também reformada, é o material digital ou digitalizado para uma hermenêutica também digital. Trata-se de reapreender a lidar com a literatura, empregando um outro olhar, por meio de ferramentas digitais. Adaptar-se a novas ferramentas é estar aberto a novos questionamentos, pois novos tipos de textos também surgem: temos os textos híbridos gerados por computadores, escrita automática, vide o caso do SINTEXT[81], projeto exercido pelo professor Pedro Barbosa em Portugal. Temos as criações digitais, a poesia digital, blogs etc. A literatura digital e a literatura digitalizada (a que nasceu no meio impresso mas foi trazida ao digital) se deparam com metodologias da velha crítica, e é preciso recauchutá-la. Isto não significa abrir mão de todo o conhecimento postulado nos estudos literários, isso não seria construtivo, mas se aliar ao conhecimento que há séculos se vem acumulando e se repetindo.

Nem tanto apocalíptico nem tanto integrado. Não se trata aqui de uma tecnofilia exacerbada, mas uma chamada aos tecnófobos. É possível, cada um ao seu modo, instrumentalizar-se para dar conta da literatura que surge na tela, seja ela digital, ou como no nosso caso, digitalizada e numérica.

A estilometria alcança com longos braços a massificação de textos, pois a soma da ampliação de capacidade de memória dos computadores, mais a potencialidade das ferramentas tecnológicas para análise de textos repercutem na viabilidade do pesquisador em mapear tantos textos ao mesmo tempo. Contudo, *more is more but not better*, e as ferramentas disponíveis criam apenas possibilidades de interpretação sobre a obra literária, é do estudioso munido também de correlações temporais, contextuais ou de outra natureza que irão surgir as leituras.

No fim, o senhor me completa.
(Riobaldo)

81 Para maiores informações: <http://www.pedrobarbosa.net/SINTEXT-pagpessoal/SINTEXT.HTM>. Acesso em 24 jul. 2012.

REFERÊNCIAS

ALONSO, Dámaso. *Poesia española. Ensayo de métodos y límites estilísticos*. Madrid: Gredos, 1966.

BAGAVANDAS, Mappillairaju; MANIMANNAN, Ganesan. *Quantification of stylistic traits: a statistical approach*. Louvain, França: 7es Journées internationales d'Analyse statistique des Données Textuelles (JADT), 2004. p. 71-78.

BALLY, Charles. *Traité de Stylistique Française*. Paris: Klincksieck, 1951.

BARBETTA, Pedro Alberto. Estatística aplicada às ciências sociais. Florianópolis: UFSC, 1998.

BEAUDOUIN, Valérie. *Statistique textuelle: une approche empirique du sens à base d'analyse distributionnelle*. *Texto*! set. 2000. Disponível em: <http://www.revue-texto.net/Inedits/Beaudouin_Statistique.html>. Acesso em: mai. 2011.

BENZÉCRI, Jean-Paul. *Pratique de l'analyse des données : linguistique et lexicologie*. In : *Mots*, oct. 1982, n. 5. p. 223-224.
BERNARD, Michel. *Introduction aux études littéraires assistées par ordinateur*. 1. ed. Paris: Presses Universitaires de France, 1999.

BOLLE, Willi. Fórmula e fábula. Teste de uma gramática narrativa, aplicada aos contos de Guimarães Rosa. São Paulo: Perspectiva, 1973.
BOSI, Alfredo. História concisa da literatura brasileira. São Paulo: Cultrix, 1994.

BRANDÃO, Saulo Cunha de Serpa. Atribuição de autoria. Um problema antigo, novas ferramentas. Revista Texto Digital, Florianópolis, ano 2, n. 1, 2006. Disponível em: <http://www.textodigital.ufsc.br/num02/saulo.htm>. Acesso em: 4 abr. 2011.

BRUNEL, Pierre (Org.). Dicionário de mitos literários. Trad. Carlos Sussekind. 3. ed. Rio de Janeiro: José Olympio, 2000.

BRUNET, Étienne. *HYPERBASE ©. Logiciel hypertexte pour le traitement documentaire et statistique des corpus textuels. Manuel de référence. Version 8.0 et 9.0*. Nice, França: Université Nice Sophia Antipolis, 2011.

_____. *Peut-on mesurer la distance entre deux textes? Corpus. La distance interxtextuelle*, nº 2, dez. 2003. Disponível em: <http://corpus.revues.org/index30.html>. Acesso em: 27 abr. 2009.

_____. *En savoir plus sur Hyperbase*. Disponível em: <http://textopol.free.fr/HYPERBASE2.HTM>. Acesso em: 12 mai. 2011.

_____. *Le vocabulaire de Hugo.* Paris-Genève: Champion-Slatkine, 1988.

_____. *Le vocabulaire de Proust : l'étude quantitative*. Genève: Champion-Slatkine, 1983.

BUSA, Roberto. *Dérnieres rèflexions sur la statistique textuelle*. *Journées Internationales d'Analyse statistique des Données Textuelles* (JADT), Nice, 1998. Disponível em: <http://lexicometrica.univ-paris3.fr/jadt/jadt1998/busa.htm>. Acesso em: 11 jan. 2012.

CAMARGO, Maria Tereza de Almeida. Estatística linguística. Alfa: Revista de Linguística. Faculdade de Filosofia, Ciências e Letras de Marília, v. 11, p. 117-128, 1967. Disponível em: <http://seer.fclar.unesp.br/alfa/issue/view/270/showToc>. Acesso em: 10 mai. 2011.

CASCUDO, Luís da Câmara. Dicionário do folclore brasileiro. 9. ed. São Paulo: Global, 2000.

CASTRO, Nei Leandro de. Universo e vocabulário do Grande Sertão. Rio de Janeiro: José Olympio, 1970.

CHEVALIER, Jean; GHEERBRANT, Alain. Dicionários de símbolos: (mitos, sonhos, costumes, gestos, formas, figuras, cores, números). Trad. Vera da Costa e Silva et al. 16. ed. Rio de Janeiro: José Olympio, 2001.

CORRÊA, Nereu. A tapeçaria linguística d'os Sertões e outros estudos. São Paulo: Quíron, 1978.

COSTA, Ana Luiza Martins. Veredas de Viator. In: Cadernos de Literatura Brasileira: João Guimarães Rosa. N. 20-21. RJ, Instituto Moreira Salles, p. 10-58, 2006.

COUTINHO, Eduardo F. Guimarães Rosa: um alquimista da palavra. Prefácio a João Guimarães Rosa: ficção completa. In: ROSA, G. Obras completas. Rio de Janeiro: Nova Aguilar, 1994. 2vols. Vol. 1, p. 11-24.

COVIZZI, Lenira Marques. O insólito em Guimarães Rosa e Borges. São Paulo: Ática, 1978.

CÚRCIO, Verônica Ribas. Sintaxe da Frustração: análise estatística do estilo de Kafka. 99 f. Dissertação (Mestrado em Teoria Literária) - Universidade Federal de Santa Catarina, Centro de Comunicação e Expressão. Programa de Pós-Graduação em Literatura, Florianópolis, 2007.

DANIEL, Mary Lou. João Guimarães Rosa*:* Travessia literária. Rio de Janeiro: José Olympio, 1968.

EDUMÉTRIE. [*Site*]. Neuchâtel, Suisse: *Institut de recherche et de documentation pédagogique* (IRDP), 20-?. Disponível em: <http://www.irdp.ch/edumetrie/index.htm>. Acesso em: 2 jan. 2012.

FERRAND, Nathalie *et al*. *Banques de données et hypertextes pour l'étude du roman*. Paris: PUF, 1997.

FERREIRA, João Martins. Contribuições da estatística, matemática e informática em análises linguísticas e semióticas. In: SARDINHA, T. B. (Org.). A língua portuguesa no computador. São Paulo: Mercado das Letras, 2005. p. 249-267.

FREITAS, Deise Joelen Tarouco de. A composição do estilo do contista Machado de Assis. 204 f. Tese. (Doutorado em Teoria Literária) - Universidade Federal de Santa Catarina, Centro de Comunicação e Expressão, Programa de Pós-Graduação em Literatura, Florianópolis, 2007.

GALVÃO, Walnice Nogueira. Mitológica Rosiana. São Paulo: Ática, 1978.

_____. As formas do falso. São Paulo: Perspectiva, 1972.

GAMA, Mônica Fernanda Rodrigues. Sobre o que não deveu caber. Repetição e diferença na produção e recepção de Tutaméia. 195 p. Dissertação (Mestrado em Língua e Literatura Francesas) – Universidade de São Paulo, 2008.

GONÇALVES, Lourdes Bernardes. Linguística de corpus e análise literária: o que revelam as palavras-chave. In: TAGNIN, S. E. O.; VALE, O. A. Avanços da linguística de corpus no Brasil. São Paulo: Humanitas, 2008. p. 387 - 406.
GUIRAUD, Pierre. A estilística. São Paulo: Mestre Jou, 1970.

_____. *Essay de stylistique*. Paris: Klincksieck, 1985.

_____. *Problèmes et méthodes de la statistique linguistique*. Dordrecht: D. Reidel, 1959.

HACQUARD, George. Dicionário de mitologia grega e romana. Asa: Rio Tinto, 1996.

HOCKEY, Susan. *Electronic texts in the humanities*. London: New York: Oxford University Press, 2004.

HOOVER, David L. *Language and style in The Inheritors*. Lanham: UPA, 1999.

KENNY, Anthony. *The computation of style*. *Un introduction to statistics for students of literature and humanities*. Pergamon Press: Oxford and New York, 1982.

LEBART, L.; SALEM, A. *Statistique textuelle*. Dunod: Paris, 1994.

LEBART, L. *et al*. *Tratamiento Estadístico de Datos*. Barcelona, Espanha: Marcombo Boixareu Editores, 1985.

LEECH, Geoffrey N.; SHORT, Michael H. *Style in Fiction*. New York: Longman, 1990.

LEONEL, Maria Célia. Guimarães Rosa. Magma e gênese da obra. São Paulo: UNESP, 2000.

LOBO, Maria A. C.; APRÍGIO, Carina Rejane. Estilística da palavra. In: Anais do V Congresso Nacional de Linguística e Filologia. 2001. Disponível em: <http://www.filologia.org.br/vcnlf/anais%20v/civ2_04.htm>. Acesso em: 11 jan. 2012.

LORENZ, G. Diálogo com Guimarães Rosa. In: COUTINHO, E. Guimarães Rosa. Fortuna Crítica. Rio de Janeiro: Civilização Brasileira, 1983. p. 67-92.

MACIEL, Carlos Alberto Antunes. *Richesse et evolution du vocabulaire d'Érico Veríssimo (1905-1975 – Porto Alegre, Brésil)*. Paris: Champignon; Genéve: Slaktine, 1986.

_____. *La base PORTEXT*. 2000. Disponível em: <http://ancilla.unice.fr/~brunet/pub/index.html>. Acesso em: 23 dez. 2011.

_____. "Repartição" e "perfil das palavras": a questão da presença/ausência nos estudos de vocabulário. In: Homenagem a Maria Emília Ricardo Marques. Universidade Aberta. 2005. Disponível em: <https://repositorioaberto.uab.pt/bitstream/10400.2/377/1/Des(a)fiando%20Discursos429-440.pdf.pdf>. Acesso em: 23 dez. 2011.

MANNING, Christopher D.; SCHÜTZE, Hinrich. *Foundation of statistical natural language processing*. Massachusets: MIT, 1999.

MARQUES, Oswaldino. O Repertório Verbal. In: COUTINHO, E. Guimarães Rosa. Fortuna Crítica. Rio de Janeiro: Civilização Brasileira, 1983.

MARTINS, Nilce Sant'Anna. O Léxico de Guimarães Rosa. São Paulo: EDUSP, 2001.

MCENERY, Tony; WILSON, Andrew. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1997.

MULLER, Charles. *Initiation à la Statistique Linguistique*. Paris: Librairie Larousse, 1968.

NUNES, Benedito. Guimarães Rosa. In:_____. O dorso do tigre. 2. ed. São Paulo: Perspectiva, 1976. p. 142-210.
_____. O tempo na narrativa. São Paulo: Ática, 1988.

OLIVIER, Andrew. *Retour au Père Goriot: ou ce que nous apprend la statistique. Journées Internationales d'Analyse statistique des Données Textuelles* (JADT), Nice, 1998. p. 467-486. Disponível em: <http://lexicometrica.univ-paris3.fr/jadt/jadt1998/oliver.htm>. Acesso em: 2012.

PARO, Sandra Regina. Crítica Textual em Tutaméia – Terceiras Estórias. No Prosseguir, a travessia rítmica. 160 f. Dissertação (Mestrado em Letras – Literatura e Crítica Literária) - Universidade Católica de Goiás, Goiânia, 2008.

PROENÇA, M. Cavalcanti. Trilhas no Grande Sertão. In:_____. Augusto dos Anjos e outros ensaios. Rio de Janeiro: Grifo, 1976. p. 155-239.
RAMOS, M. L. Análise estrutural de Primeiras Estórias. In: COUTINHO, E. Guimarães Rosa. Fortuna Crítica. Rio de Janeiro: Civilização Brasileira, 1983. p. 514-519.

REINERT, Max. *ALCESTE: Une méthodologie d'analyse des données textuelles et une application: Aurélia de Gérard de Nerval*, 1990. In: BEAUDOUIN, Valérie. *Statistique textuelle: une approche empirique du sens à base d'analyse distributionnelle*. Texto! set. 2000. Disponível em: <http://www.revue-texto.net/Inedits/Beaudouin_Statistique.html>. Acesso em: jan. 2013.

RICOEUR, Paul. Tempo e narrativa. v. 2. Campinas: Papirus, 1995.

ROCKWELL, Geoffrey. *What is text analysis, really? Literary and Linguistic Computing*, v. 18, n. 2, p. 209-219, 2003.

RONÁI, Paulo (Org.). Os prefácios de Tutaméia. In: ROSA, J. G. Tutaméia. Rio de Janeiro: José Olympio, 1976. p. 193-201.

_____. Trajetória de uma obra. In: ROSA, G. Seleta. Rio de Janeiro: José Olympio, 1973.

ROSA, João Guimarães. Ave, palavra. Rio de Janeiro: José Olympio, 1970.

ROSA, João Guimarães. Carta a Harriet de Onís: 4 nov. 1964. [S. l.]: [s. n.], 19-?.

_____. Corpo de baile. Rio de Janeiro: José Olympio, 1976a.

_____. Estas estórias. Rio de Janeiro: José Olympio, 1976b.

_____. Ficção completa. v. 1. Rio de Janeiro: Nova Aguilar, 1995a.

_____. Ficção completa. v. 2. Rio de Janeiro: Nova Aguilar, 1995b.

_____. Grande Sertão: Veredas. Rio de Janeiro: José Olympio, 1974.

_____. Magma. Rio de Janeiro: Nova Fronteira, 1997.

_____. Primeiras estórias. 12. ed. Rio de Janeiro: José Olympio, 1981.

_____. Primeiras estórias. Rio de Janeiro: José Olympio, 1974.

_____. Sagarana. Rio de Janeiro: Nova Fronteira, 1993.

_____. Sagarana. Rio de Janeiro: José Olympio, 1980.

_____. Tutaméia. Rio de Janeiro: José Olympio, 1967.

_____. No Urubuquaquá no Pinhém. Rio de Janeiro: Nova Fronteira, 1984.

SARDINHA, Tony Berber. Linguística de corpus. São Paulo: Manole, 2004.

SPERBER, Suzi Frankl. Caos e cosmos. Leituras de Guimarães Rosa. São Paulo: Duas Cidades, 1976.

_____. Guimarães Rosa: signo e sentimento. São Paulo: Ática, 1982.

SPITZER, Leo. *Linguística e historia literaria*. Madrid: Gredos, 1968.

VERSIANI, Ivana. Para a sintaxe de Grande Sertão: veredas – valores do subjuntivo. In: COELHO, Nelly Novaes; VERSIANI, Ivana. Guimarães Rosa. São Paulo: Quíron, 1975. p. 79 – 142.

WEINRICH, Harald. *Estructura y función de los tiempos en el lenguaje*. Madrid: Gredos, 1968.

WELLBERY, David E. Neoretórica e desconstrução. Rio de Janeiro, Editora da UERJ, 1998.

WELLEK, René; WARREN, Austin. *Teoría literaria*. Madrid: Cátedra, 1966.

YVANCOS, José Maria Pozuelo. *Teoría del Lenguaje Literario*. Madrid: Cátedra, 1994.

REFERÊNCIAS CONSULTADAS

ARAÚJO, Heloísa Vilhena de. As Três Graças. Nova contribuição ao estudo de Guimarães Rosa. São Paulo: Mandarim, 2001.

ARROYO, Leonardo. A cultura popular em Grande Sertão: Veredas. Rio de Janeiro: José Olympio, 1984.

BOLLE, Willi. Fórmula e fábula. Teste de uma gramática narrativa, aplicada aos contos de Guimarães Rosa. São Paulo: Perspectiva, 1973.

BOSI, Alfredo. História Concisa da Literatura Brasileira. São Paulo: Cultrix, 1993.

CAMPOS, Vera Mascarenhas de. Borges & Guimarães. Na esquina rosada do Grande Sertão. São Paulo: Perspectiva, 1988.

CÂNDIDO, Antonio. O homem dos avessos. In: _____. Tese e antítese. São Paulo: Nacional, 1978. p. 119 – 139.
_____. Formação da Literatura Brasileira. Rio de Janeiro: Itatiaia, 1993.

COUTINHO, Afrânio. A Literatura no Brasil. Rio de Janeiro: José Olympio, 1986.

GARBUGLIO, José Carlos. O mundo movente de Guimarães Rosa. São Paulo: Ática, 1972.

MACEDO, Tânia. Guimarães Rosa. São Paulo: Ática, 1996.

MACHADO, Ana Maria. Recado do Nome. Leitura de Guimarães Rosa à luz do nome de seus personagens. Rio de Janeiro: Imago, 1976.

MIKETEN, Antonio Roberval. Travessia de Grande Sertão: Veredas. Brasília: Thesaurus, 1982.

PAZ-ANDRADE, Valentin. A galeguidade na obra de Guimarães Rosa. São Paulo: DIFEL, 1983.

PEREZ, Renard. Em Memória de João Guimarães Rosa. Rio de Janeiro: José Olympio, 1968.

ROCHA, Karina Bersan. Veredas do amor no Grande Sertão. Nova Friburgo: Imagem Virtual, 2001.

ROCHA, Luiz Carlos de Assis. Teoria sufixal do léxico português aplicada às formações nominais de Guimarães Rosa. Tese (Doutorado em Letras – Letras Vernáculas) - Faculdade de Letras UFRJ, Rio de Janeiro, 1992.

SANTOS, Julia Conceição Fonseca. Nomes dos personagens em Guimarães Rosa. Rio de Janeiro: INL, 1971.

SANTOS, Wendel. A construção do romance em Guimarães Rosa. São Paulo: Ática, 1978.

SIMÕES, Irene Gilberto. Guimarães Rosa: as paragens mágicas. São Paulo: Perspectiva, 2003.

TEYSSIER, Paul. A língua de Guimarães Rosa. In: [*ASSOCIATION Freudienne Internationale*]. Um inconsciente pós-colonial. Se é que ele existe. Porto Alegre: Artes e Ofícios, 2000. p. 77-86.

VALENTE, Luiz Fernando. *Fiction and the reader: the prefaces of Tutaméia*. *Hispanic Review*, v. 56, n. 3, p. 349-362, 1988. Disponível

em: <http://www.jstor.org/stable/474024?origin=JSTOR-pdf>. Acesso em: 6 jun. 2012.

XISTO, Pedro; CAMPOS, Augusto de; CAMPOS, Haroldo de. Guimarães Rosa em três dimensões. São Paulo, 1970.