Programa de Pós-Graduação em Lingüística

PPGL

TIPOLOGIA DE TRAÇOS LINGUÍSTICOS DE TEXTOS DO PORTUGUÊS DO BRASIL DOS SÉCULOS XVI, XVII, XVIII E XIX: UMA PROPOSTA PARA A CLASSIFICAÇÃO AUTOMÁTICA DE GÊNEROS TEXTUAIS

Jacqueline A. Souza

SÃO CARLOS
2010

Universidade Federal de São Carlos

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

Tipologia de traços linguísticos de textos do português do Brasil dos séculos XVI, XVII, XVIII e XIX: uma proposta para a classificação automática de gêneros textuais

Jacqueline A. Souza
Bolsista: FAPESP

Dissertação apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do Título de Mestre em Linguística.

Orientadora: Profa. Dra. Gladis M. B. Almeida (UFSCar)
Coorientadora: Profa. Dra. Sandra Maria Aluísio (USP)

São Carlos - São Paulo - Brasil

2010

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

S729tt

Souza, Jacqueline Aparecida de.
Tipologia de traços linguísticos de textos do português do Brasil dos séculos XVI, XVII, XVIII e XIX : uma proposta para a classificação automática de gêneros textuais / Jacqueline Aparecida de Souza. -- São Carlos : UFSCar, 2011.
164 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2010.

1. Linguística. 2. Linguística de corpus. 3. Aprendizado de computador. 4. Corpus histórico. I. Título.

CDD: 410 (20ª)

BANCA EXAMINADORA

Profa. Dra. Gladis Maria de Barcellos Almeida _____

Prof. Dr. Thiago Alexandre Salgueiro Pardo _____

Prof. Dr. Oto Araújo Vale _____

Agradecimentos

A minha mãe, Elenice Braga de Souza (*in memorian*), pelo amor irrestrito. A tudo que transmitiu durante os vinte cinco anos que convivemos, ao seu carinho, bondade, gentileza, educação, incentivo, apoio e cuidado. Aos meus queridos irmãos, Anderson e Daiane, pela preocupação, compreensão e carinho.

A tia Angela Maria Braga, pelo carinho, cuidado e apoio.

A professora Gladis M. B. Almeida, pela orientação, apoio, incentivo, pelas conversas e dicas. Além de professora exemplar é um modelo de comprometimento com a pesquisa. Obrigada pela oportunidade que me deu e pela paciência.

A Arnaldo Cândido Junior pelas orientações, pela disposição em mostrar o caminho a ser seguido nesta pesquisa, a todos os vários e-mails que respondeu, as dúvidas que esclareceu. Sem ele não seria possível realizar esta pesquisa.

A todos os meus amigos pelo companheirismo, pelas conversas sérias e não tão sérias: Madalice, Aline, Kelly, Débora, Eliane, Nadia, Cristina, Sirlei e Hélio Pajeú.

Ao Geterm (Grupo de Estudos e Pesquisas em Terminologia).

Ao professor Oto Vale.

À Mariazinha.

A todos que contribuíram para a realização desta pesquisa.

A agência financiadora da pesquisa, FAPESP, muito obrigada.

Sumário

Índice de figuras

Índice de tabelas

Índice de gráficos

Índice de textos

Resumo

Com base nos postulados metodológicos da Linguística de Corpus e nos conceitos de gênero, propostos por Swales (1990) e Biber (1995), esta pesquisa pretende descrever traços linguísticos característicos de textos históricos, correlacionando-os a seus respectivos gêneros, e propor uma tipologia de traços de forma que seja possível identificar o gênero de cada texto automaticamente. Para execução da pesquisa foi utilizado o corpus do português dos séculos XVI, XVII e XVIII do projeto *Dicionário Histórico do Português do Brasil* (programa Institutos do Milênio/CNPq – UNESP/Araraquara), constituído por 2.459 textos e 7.5 milhões de palavras. Para realizar uma descrição histórica, partiu-se de características sincrônicas obtidas a partir da tabela de traços contemporâneos elaborada por Aires (2005). No que tange à manipulação do corpus, utilizou-se o *Philologic*, o *Unitex* e desenvolveu-se uma ferramenta para extração e quantificação dos traços. Para fins de classificação, foram utilizados os algoritmos disponibilizados no Weka (*Waikato Environment for Knowledge Analysis*), tais como: *Naive Bayes, Bayes Net, SMO, Multilayer Perceptron e RBFNetwork, J48, NBTree*. A descrição foi realizada com base em 62 traços, os quais abarcam estatísticas baseadas no texto como um todo e em palavras, incluindo as classes de verbos, pronomes, advérbios, como também marcadores discursivos, expressões e unidades lexicais. Concluiu-se que os gêneros compartilham características linguísticas específicas, porém, também apresentam seus padrões próprios, como o uso de determinadas expressões e a frequência de unidades lexicais. Apesar das limitações e complicações em utilizar um corpus histórico, o desempenho dos classificadores com base nos traços levantados foi satisfatório, com a taxa de acerto 84% e 92% de classificação correta.

Palavras-chave: Lingüística de Corpus. Traços lingüísticos. Gêneros textuais. Classificação automática.

Abstract

Based on methodological postulates of the Linguistic of corpus and on the genre concepts, proposed by Swales (1990) and Biber (1995), this research intends to describe linguistic traces which are characteristic of historic texts and correlate them to their respective genres, as well as propose a typology of traces so that it is possible to automatically identify the genre. In order to execute the research, the corpus of the Portuguese of the centuries XVI, XVII and XVII of the project Historical Dictionary of the Portuguese in Brazil (program Institutes of the Millennium/CNPq – UNESP/Araraquara), which is constituted by 2,459 texts and 7,5 million words has been used. In order to realize a historical description, the study has started from synchronic characteristics obtained from the table of contemporary traces elaborated by Aires (2005). As for the manipulation of the corpus, it has been used the Philologic, the Unitex as well as another tool for the extraction and quantification of traces that has been developed. For the purposes of classification, algorithms available at Weka (Waikato Environment for knowledge Analysis) such as: Naive Bayes, Bayes Net, SMO, Multilayer Perceptron e RBFNetwork, J48, NBTree have been used. The description has been made based on the 62 traces, which include statistics based on a text as a whole and on words, including classes of verbs, pronouns, adverbs as well as discourse markers, expressions and lexical units. It has been concluded that the genres share specific linguistic characteristics. However, they also present their own standards with the use of specific expressions and the frequency of lexical units. Despite the limitations and complications in using a historical corpus, the performance of the classifiers based on the raised traces was satisfactory and the rate of correct classification was 84% and 92%.

Key words: Corpus linguistics. Features. Textual genre. Automatic classification

1. Introdução

Estudos sobre gênero discursivo – ou gênero textual, de acordo com a perspectiva de Marcuschi (2008) –, têm sido objeto de estudo em distintos ramos da Linguística, como Análise do Discurso, Linguística Aplicada, Linguística Textual, Linguística de Corpus e, no nosso caso aqui, em Processamento Automático de Língua Natural (PLN).

Para cada ramo que se quer empregar o termo gênero, é necessário ter em conta qual perspectiva está por trás da noção de gênero. O pesquisador pode optar, por exemplo, pela perspectiva interacionista e dialógica de Bakhtin (2000), ou pela histórico-cultural e sistêmica de Biber (1988), ou a sistêmico-funcional de Halliday e Hasan (1989), ou a socio-retórica de Swales (1990), ou ainda a sociodiscursiva de Bronckart (1999).

Na esteira dos trabalhos desenvolvidos com o intuito de descrever e classificar gêneros no cenário brasileiro, cumpre citar duas pesquisas relevantes. A primeira é a dissertação de Kauffmann (2005), intitulada *O corpus do jornal: variação linguística, gênero e dimensões da imprensa diária escrita*. Essa pesquisa teve como objetivo identificar empiricamente semelhanças e diferenças de natureza linguística nos textos e entre os gêneros de um jornal. Para sua consecução, o autor empregou recursos teórico-metodológicos proporcionados pela Linguística de Corpus. Além disso, para análise, utilizou a Análise Multidimensional (AMD), que leva em conta procedimentos estatísticos, buscando encontrar grupos coocorrentes de características e categorias linguísticas. Além disso, propôs uma tipologia dos gêneros presentes no jornal. Sua principal contribuição para esta pesquisa que ora se apresenta são as perspectivas teóricas abordadas e a própria metodologia de análise do corpus.

A segunda pesquisa importante é a tese de Aires (2005), intitulada *Uso de marcadores estilísticos para a busca na Web em português*, cujo objetivo foi classificar automaticamente gêneros contemporâneos em português, de maneira a tornar possível categorização de textos na Web.

Ressalte-se que esses dois trabalhos versaram sobre gêneros contemporâneos baseados em corpora atuais do português, não havendo ainda nenhum estudo sobre a classificação automática de gêneros baseada num corpus histórico do português do

Brasil. Assim, o propósito desta pesquisa é levantar os traços[1] (ou características) linguísticos de textos históricos, correlacionando-os aos respectivos gêneros, e propor uma tipologia de traços de forma que seja possível identificar o gênero de cada texto de forma automática, ou seja, por meio dos tipos de traços (lexicais, morfossintáticos, expressões) realizar a classificação automática dos gêneros. Ao caminhar na direção desses objetivos, será possível responder as seguintes questões de pesquisa:

1. as características dos gêneros comutam de um contexto para outro?
2. é possível a convergência de características linguísticas determinantes dos gêneros de forma a se obter uma classificação automática coerente de gêneros textuais?
3. como lidar com textos históricos, uma vez que a referência de mundo do classificador humano é a do contexto atual?
4. quais características do português sobressaem em textos de um corpus histórico? Elas são as mesmas características dos textos de um corpus contemporâneo?

Diferentemente de pesquisas anteriores, o corpus utilizado aqui foi desenvolvido no âmbito do projeto *Dicionário Histórico do Português do Brasil* (programa Institutos do Milênio/CNPq – UNESP/Araraquara), composto por textos escritos entre os séculos XVI e XIX (especificamente até 1808 com a chegada da família real ao Brasil), totalizando atualmente 2.458 textos e 7.5 milhões de formas simples, sem anotações morfossintáticas. Outros corpora considerados históricos no Brasil são:

- Tycho-Brahe: composto por 52 textos (2.356,811 palavras) criados entre os séculos XVI e XX, com dois milhões de palavras e com anotação morfológica e sintática.
- Corpus do Português: composto por textos das variantes brasileira e portuguesa, escritos entre os séculos XIV e XX, formado por 45 milhões de palavras e inclui textos do corpus Tycho-Brahe, do Lácio-Web, entre outros projetos.
- BIT-PROHPOR[2] (Projeto Banco Informatizado de Textos do Programa para a História da Língua Portuguesa): com o objetivo de construir corpora informatizado relativos à história do Português para trabalhos de

[1] Em inglês o termo é *feature*.

[2] http://www.prohpor.ufba.br/projetos.html

> natureza lingüística, os textos integrantes, literários e não-literários, foram produzidos entre os séculos XIII e meados do XVI, devidamente digitalizado. È coordenado pela Profa. Dra. Rosa Virgínia Mattos e Silva, Instituto de Letras da Universidade Federal da Bahia.

Uma vez que a pesquisa trata de uma descrição diacrônica, parte-se de características sincrônicas apresentadas na tabela de traços linguísticos contemporâneos[3] elaborada por Aires (2005). A autora sugere o levantamento de estatísticas baseadas em palavras (itens lexicais diferentes, iniciados por letra maiúscula, tamanho das palavras, etc.), estatísticas baseadas no texto como um todo (número de caracteres, frases, tamanho do texto) e outras estatísticas como frequência de pronomes, advérbios, verbos, marcadores discursivos, operadores argumentativos e expressões específicas, totalizando 46 sugestões de traços.

Os processadores de corpus utilizados nesta pesquisa foram o *Philologic*[4] e *Unitex*[5], ambos adaptados por pesquisadores do Núcleo Interinstitucional de Linguística Computacional[6] (NILC). Os dois programas foram escolhidos por atenderem às necessidades e peculiaridades deste tipo de pesquisa, de modo a auxiliar na busca, extração e recuperação de textos e fragmentos de textos, como também facilitar no levantamento dos traços.

No que tange à classificação ou categorização de gêneros, há inúmeras opiniões abalizadas que defendem a impossibilidade de se realizar essa tarefa, já que estão implicadas aí a visão de mundo de quem classifica e de quem elabora a classificação, aspectos ideológicos ou mesmo no simples fato de que ao classificar, desconstrói-se a noção de gênero (de acordo com a concepção bakhtiniana). Assim, há de se concordar com Marcuschi (2004) quando afirma que "qualquer proposta de classificação de fenômenos, que pretenda uma validade mínima, deve explicitar os princípios teóricos e metodológicos que a sustentam." Nesse sentido, esta pesquisa procura utilizar uma concepção de gênero que torna possível uma classificação, daí referir *conceito operacional de gênero*, ou seja, operacional para esta pesquisa.

Do ponto de vista computacional, especificamente no que se refere à

---

[3] No apêndice A encontra-se uma versão resumida da tabela de traços contemporâneos de Aires (2005)
[4] http://www.lib.uchicago.edu/efts/ARTFL/philologic/
[5] http://www-igm.univ-mlv.fr/~unitex/
[6] O Núcleo Interinstitucional de Lingüística Computacional foi criado em 1993 para promover a pesquisa e projetos de desenvolvimento na área de lingüística computacional e processamento de língua natural (PLN), tais como o desenvolvimento de corpus, classificação automática, tradução automática, ferramenta de apoio a escrita.

categorização automática, esta pesquisa se insere no âmbito do Aprendizado de Máquina Supervisionado, subárea da Inteligência Artificial dedicada a pesquisar métodos computacionais relacionados à aquisição de novos conhecimentos, habilidades e formas de organizar o conhecimento existente. Os algoritmos classificadores utilizados foram os disponibilizados no ambiente WEKA (*Waikato Environment for Knowledge Analysis)*[7] e que tiveram melhor desempenho durante a fase de treinamento, a saber: *Naive Bayes, Bayes Net, SMO, Multilayer Perceptron e RBFNetwork, J48, NBTree.* Tais classificadores apresentaram ótimo desempenho durante a fase de classificação.

Para utilização dos classificadores, é pré-requisito que os traços de cada texto estivessem no formato ARFF (*Attribute-Relation File Format*), para isso, desenvolveu-se um extrator de traços, que além de quantificar cada traço que constituía um texto, gerava o arquivo ARFF. Tal arquivo foi fundamental para o refinamento dos traços.

Ainda que houvesse a dificuldade em manipular um corpus histórico, devido a suas peculiaridades, a variação de grafia e junção, os propósitos foram alcançados. Os traços utilizados para descrever gêneros foram 62, incluindo unidades lexicais e expressões. Concluiu-se que os traços que descrevem textos do português contemporâneo também podem descrever textos históricos, com ressalva apenas para expressões e unidades lexicais que talvez possam mudar de período para período, mas isso pode ser a proposta de outro trabalho.

O valor desta pesquisa está no trabalho descritivo em si e na própria metodologia para pesquisar traços linguísticos em corpus histórico. A descrição do português histórico expõe hábitos culturais linguísticos, além de elucidar como a sociedade estava sendo formada no Brasil. Além disso, é relevante mencionar a importância dos resultados obtidos com a classificação automática, pois constatou-se que é possível classificar gêneros histórico com bons resultados.

Esta pesquisa está estruturada da seguinte forma: no Capítulo 2, apresentam-se as principais abordagens de gênero, os principais estudos referente a análises e os termos empregados nesta pesquisa. No Capítulo 3, é oferecida uma breve apresentação sobre categorização de textos, bem como algumas pesquisas existentes, conceitos e métodos. No Capítulo 4, descreve-se detalhadamente toda a metodologia empregada nesta pesquisa, as etapas percorridas, a importância e a utilização dos programas *Philologic*,

[7] http://www.cs.waikato.ac.nz/ml/weka/

Unitex e o extrator de traços. Por fim, apresentam-se, no Capítulo 5, os resultados da descrição e, no Capítulo 6, os resultados da classificação. A pesquisa se encerra no Capítulo 7, com as conclusões.

2. Conceitos de gêneros

A expressão "gênero" circula por diversas áreas do conhecimento, entre vários estudiosos, sejam eles linguistas, linguistas computacionais, sociólogos, especialistas em ensino e aprendizagem, o que corrobora uma abordagem cada vez mais multidisciplinar dos estudos de gêneros. Desse modo, analisar gêneros pode implicar em analisar textos, discursos, descrição da língua, visão da sociedade em uma perspectiva histórica ou não, realizar categorização, taxonomia e assim por diante. Dado o conceito variável de gênero, há diversas perspectivas teóricas a partir das quais se pode analisá-lo ou, em alguns casos, até mesmo classificá-los. Segundo Marcuschi (2008), essas perspectivas tentam abarcar as idéias de que gênero é:

- uma categoria cultural;
- um esquema cognitivo;
- uma forma de ação social;
- uma estrutura textual;
- uma forma de organização social;
- uma ação retórica.

Além do fato de que cada uma dessas características pode orientar a observação, o gênero também pode ser tudo isso ao mesmo tempo. Nesse sentido, Miller (1984) afirma que os gêneros são uma "forma de ação social". Eles são um "artefato cultural" importante como parte integrante da estrutura comunicativa de nossa sociedade. Os gêneros seriam "ações retóricas tipificadas" produzidas em resposta a situações sociais recorrentes e, enquanto "artefatos culturais", seriam instrumentos aptos para se desenvolver ações sociais em situações específicas, que se definem por objetivos comunicativos, audiência, regularidades formais e conteúdos. A autora vê no gênero um constituinte específico e importante da estrutura comunicativa da sociedade, de modo a constituir relações de poder bastante marcadas, sobretudo em instituições.

Bronckart (1999), sob uma perspectiva sociodiscursiva, vê os gêneros como "instrumentos de adaptação e participação na vida social e comunicativa" e sua apropriação "é um mecanismo fundamental de socialização, de inserção prática nas atividades comunicativas humanas". Teoriza o gênero como uma entidade textual-

discursiva de caráter psicológico e observa que os gêneros operam como formas de legitimação discursiva.

Diferente de Bronckart, Bhatia (1997) tem uma perspectiva socio-retórica voltada para o ensino de uma segunda língua. O autor trabalhou com análise de gêneros profissionais e acadêmicos, como Swales, e indagava por que determinado gênero é escrito de determinada maneira. Ele mesmo observou que na resposta a esta pergunta haveria não apenas questões socioculturais e cognitivas, mas também questões de ordem comunicativa e estratégias convencionais para atingir determinados objetivos. Constatou ainda que tanto os aspectos psicológicos quanto cognitivos contribuem para a dinamicidade dos gêneros, favorecendo, dessa forma, que sua descrição seja baseada no uso da língua.

No que se refere à análise de gênero, Bhatia (1993) sugere três tipos de orientação:

1. **Linguística:** apesar de se interessar pela natureza dos gêneros, na forma como os propósitos sociais são cumpridos e nos contextos em que são usados, o autor privilegia o estudo dos traços linguísticos (gramaticais, lexicais, estilos, registros, aspectos discursivos e retóricos).
2. **Sociológica:** possibilita pesquisar como determinado gênero define, organiza e comunica a realidade social.
3. **Psicológica e psicolinguística:** focaliza os aspectos táticos ou estratégicos da construção de gêneros. O aspecto psicolinguístico da analise de gênero tem a ver com a estruturação cognitiva, o aspecto tático refere-se às escolhas estratégicas individuais que o escrevente faz a fim de tornar a sua escrita eficaz e, desse modo, alcançar seus propósitos comunicativos.

O autor ainda sugere uma metodologia para uma análise abrangente de qualquer gênero, a qual é apresentada de forma sistematizada em sete passos, que poderão ser seguidos total ou parcialmente pelo analista, conforme seu interesse. Os sete passos são os seguintes:

1. colocar o texto-gênero num contexto situacional;
2. levantar a literatura existente sobre o gênero em questão;
3. refinar a análise contextual/situacional;
4. selecionar o corpus;
5. estudar o contexto institucional;
6. definir níveis de análise linguística;

7. reunir informações especializadas para a análise de gênero.

O autor finaliza, postulando que todos os gêneros têm uma forma e uma função, assim como estilo e conteúdo, mas sua determinação se dá basicamente pela função e não pela forma.

Além dos autores acima mencionados, há aqueles que merecem uma descrição mais pormenorizada, já que seus estudos e/ou reflexões influenciaram e serviram como referência teórica para muitas pesquisas linguísticas, são eles: Bakhtin, Halliday e Hasan, Swales e Biber, os quais serão tratados individualmente nas subseções seguintes.

2.1. Bakhtin

Segundo Bakhtin (2000), a língua tem a propriedade de ser dialógica, ou seja, todos os enunciados no processo de comunicação, independentemente de sua dimensão, são dialógicos. As relações de sentido que se estabelecem entre dois enunciados constituem o dialogismo, daí a assertiva de que os enunciados não existem fora das relações dialógicas. Isso significa afirmar que o enunciador, para constituir um discurso, leva em conta o discurso do outro, que está presente no seu.

Bakhtin desenvolveu seus estudos sobre os gêneros não em uma perspectiva de classificação da espécie, como na visão aristotélica, mas sob a ótica do dialogismo no processo comunicativo, ou seja, considerou as relações interativas como processos produtivos de linguagem, redimensionando a pragmática na perspectiva do dialogismo, conceito-chave para compreendê-lo. O autor não se interessa pelas propriedades formais, mas vincula a utilização da linguagem, na forma de enunciado, com atividades humanas, ou seja, os indivíduos agem em determinadas esferas de atividades – como em uma reunião de trabalho, em uma confraternização, nos lares, na igreja, na política – e essas atividades implicam a utilização da linguagem na forma de enunciado, ou seja, cada esfera de ação ocasiona o aparecimento de certos tipos de enunciados, que se estabilizam e mudam em função de alterações nessas esferas (FIORIN, 2006).

Dada a riqueza e a variedade dos gêneros, que abarcam a totalidade do uso da linguagem, Bakhtin (2000) os separa em dois grupos: **gêneros primários e secundários**. Os primários são os gêneros da vida cotidiana, predominantemente orais, isto é, pertencem à comunicação verbal espontânea, podem ser controlados diretamente

na situação discursiva, como um relato familiar, uma conversa telefônica, bilhetes, cartas, *e-mail*, *chat*, diálogos. Os gêneros secundários são textos geralmente mediados pela escrita, pertencentes à esfera da comunicação cultural mais elaborada, que fazem parte de um uso mais oficializado da linguagem como o discurso científico, jurídico, político, jornalístico, o teatro, o romance, etc. Fiorin (2006) observa que há entre os gêneros secundários aqueles que não são unicamente escritos, como o sermão, a poesia lírica, o discurso parlamentar, a comunicação científica, o artigo científico, as autobiografias e as memórias.

No entanto, existe uma interdependência dos gêneros, uma vez que os secundários valem-se dos primários ou, ainda, podem imitá-los em sua estrutura composicional, temática e de estilo, características estas que constroem o todo que constitui o enunciado.

A estrutura composicional refere-se ao modo como o texto é organizado, sua estrutura. O conteúdo temático não significa o assunto de um texto, mas o domínio de sentido de que se ocupa o gênero, como as cartas de amor, que apresentam o conteúdo temático de relações amorosas. E o estilo diz respeito a uma seleção de meios linguísticos, ou seja, uma seleção de meios lexicais, fraseológicos e gramaticais em função da imagem do interlocutor e de como se presume sua compreensão responsiva ativa do enunciado, como um estilo oficial, que usa formas respeitosas (requerimentos, discursos parlamentares, etc.) (FIORIN, 2006).

Diante do exposto sobre dialogismo, esfera de atividades, gêneros primários e secundários e a tríade estrutura composicional, conteúdo temático e estilo que compõem o todo de um enunciado, Bakhtin (2000) postula que:

> "A utilização da língua efetua-se em forma de enunciados (orais e escritos), concretos e únicos, que emanam dos integrantes duma ou doutra esfera da atividade humana. O enunciado reflete as condições específicas e as finalidades de cada uma dessas esferas, não só por seu conteúdo (temático) e por seu estilo verbal, ou seja, pela seleção operada nos recursos da língua, recursos lexicais, fraseológicos e gramaticais, mas também, e, sobretudo, por sua construção composicional. Estes três elementos (conteúdo temático, estilo e construção composicional) fundem-se indissoluvelmente no todo do enunciado, e todos eles são marcados pela especificidade de uma esfera de comunicação. Qualquer enunciado considerado isoladamente é, claro, individual, mas cada esfera de utilização da língua elabora seus tipos relativamente estáveis de enunciados, sendo isso que denominamos gêneros discursivos" (BAKHTIN, 2000, p. 279).

É importante ressaltar que tipos relativamente estáveis de enunciados implicam a necessidade de considerar a historicidade dos gêneros, sua incessante alteração, à

medida que as esferas de atividade se desenvolvem e ficam mais complexas, indicando uma imprecisão das características e das fronteiras dos gêneros.

Portanto, sua perspectiva rompe com o caráter normativo dos gêneros, vistos como um rol de propriedades formais, fixas e imutáveis. Ao pensar os gêneros na perspectiva do seu processo de produção, o autor renovou o conceito de gênero sob uma visão socio-retórica, em que os gêneros são passíveis de flexibilização, dependendo do contexto enunciativo. Ou seja, os gêneros estão sempre vinculados ao domínio da atividade humana, refletindo suas condições específicas e finalidades a partir de contextos sociais e históricos. Importa deixar claro que essa perspectiva torna incoerente uma proposta de classificação dos gêneros discursivos, razão pela qual esse autor não foi escolhido para embasar esta pesquisa, mas ainda assim, julgou-se conveniente retratá-lo, dado o pioneirismo de sua abordagem.

2.2. Halliday e Hasan: perspectiva sistêmico-funcional

> "As atividades humanas são realizadas comumente através de práticas discursivas escritas, ou seja, gêneros textuais. Os gêneros nos ajudam a navegar dentro dos complexos mundos da comunicação escrita e da atividade simbólica, porque ao reconhecer uma espécie de texto, reconhecemos muitas coisas sobre a situação social e institucional, as atividades propostas, os papéis disponíveis ao escritor e ao leitor, os motivos, as idéias, a ideologia e o conteúdo do documento e o lugar onde isso tudo pode caber em nossa vida". (BAZERMAN, 2005, p.84).

De acordo com a epígrafe, compreender interações práticas, funcionais que ocorrem em um gênero faz-se necessário uma vez que se busca compreender a linguagem como sistema mediador de ações, como parte de um tipo de atividade social recorrente em determinado ambiente. Dessa forma, entender a funcionalidade de um gênero é importante para compreender determinada prática discursiva.

Essa concepção direciona os estudos dos gêneros sob a perspectiva da linguística sistêmico-funcional (HALLYDAY e HANSAN, 1989). A abordagem sistêmico-funcional é capaz de mostrar como os textos se estruturam para construírem significados, já que a língua é um sistema semiótico, isto é, os falantes têm possibilidades de escolhas em diferentes níveis no sistema linguístico, como o léxico-gramatical, o fonológico, o fonético e o semântico (CONTO, 2003).

Halliday e Hasan (1989) e Eggins e Martin (1997) são os principais autores da linha sistêmico-funcional. Os primeiros explicitam como são conferidos ao texto funções de cunho sociocultural, enquanto Eggins e Martin (1997) aprofundam as diferenças conceituais entre gênero e registro e investigam a capacidade de ambos captarem a variação funcional do texto. Para autores que se basearam nas idéias de Halliday, como Eggins e Martin, e estudaram a questão do gênero, há uma instância que mediaria o gênero no âmbito da linguagem: o registro que seria "uma configuração de significados que são tipicamente associados com uma configuração particular de campo, modo e relações" (HALLIDAY; HASAN, 1989, p. 38-39). No entanto, a Configuração Contextual (CC), abordada na perspectiva de Halliday e Hasan (1989, p. 56), é a representação de uma atividade social, cujas características podem ser usadas para fazer previsões acerca de um texto. A CC é formada por três variáveis segundo Motta-Roth e Heberle (2005):

1. **Campo**: é possível identificar a atividade social envolvida e seu objetivo.
2. **Relação:** mostra quem são os participantes envolvidos, a relação e a distância social existente entre eles.
3. **Modo:** demonstra o papel da linguagem, o canal e o meio utilizado para efetivar a mensagem.

A Estrutura Genérica Potencial (EGP) é constituída pela função ideacional, ou seja, as escolhas léxico-gramaticais dos interlocutores; pela função interpessoal, que é a maneira como os interlocutores usam a linguagem para interagir; e a função textual, compreendida como a relação entre os aspectos semânticos e gramaticais do texto.

Para Hallyday e Hasan (1989):

> "(...) um gênero é reconhecido pelos significados a ele associados; em realidade, o termo gênero é uma forma abreviada de uma expressão mais elaborada, potencial semântico específico de gênero. (...) Gêneros variam em delicadeza do mesmo modo que contextos os fazem. Mas, para que alguns textos pertençam a um gênero específico, é necessário que sua estrutura seja uma realização de uma determinada EGP. (...) Daí decorre o fato de que os textos que pertencem a um mesmo gênero possam variar em relação a suas estruturas, o único aspecto em relação ao qual eles não podem variar sem consequência à sua alocação genérica são os elementos obrigatórios e a disposição da EGP" (HALLIDAY; HASAN, 1989).

Portanto, ao observar que a EGP é o conjunto de características funcionais em comum entre os textos, identificáveis por um conjunto de indivíduos em um contexto determinado, compreende-se que o texto, por meio dos gêneros, não é apenas um produto, mas um processo que reflete as escolhas efetuadas pelo falante.

Dessa forma, o foco das preocupações dos funcionalistas não é somente a estrutura da língua, mas também a função das categorias léxico-gramaticais, uma vez que as escolhas realizadas pelos falantes estão condicionadas a um determinado contexto de situação e cultura. Assim, pode-se considerar que essa perspectiva sistêmico-funcional possui e fornece ferramentas que, a partir dos resultados, demonstram a relevância da análise dos elementos textuais, contextuais e de suas funções, conjunto de fatores constituintes dos gêneros.

Uma vez que as escolhas ou opções linguísticas realizadas pelo indivíduo no ato comunicativo influenciam e sofrem influência do contexto social e cultural, Pereira e Almeida (2002) afirmam que:

> "Uma realidade se constrói à medida que se fazem opções dentro da língua e à medida que essas opções influenciam, determinam ou legitimam comportamentos na sociedade. (...) a realidade é um reflexo das escolhas que fazemos quando produzimos linguagem, ou ainda, que aquilo a que chamamos de realidade se constrói pela linguagem" (PEREIRA, ALMEIDA, 2002, p. 245)

Por meio dos estudos de gêneros, é possível perceber certas características textuais e contextuais e, a título de complementação, Bazerman *et al.* (2005) afirmam:

> "Na Linguística, as preocupações com a linguagem em uso e a análise do discurso têm renovado o interesse no gênero como meio de organizar os aspectos linguísticos em relação à ação situada. (...) temos estudos da maneira como elementos semânticos e sintáticos se agregam em diferentes gêneros e das maneiras como a organização interna dos gêneros revela o processo linguístico dos eventos numa série de movimentos tipificados, descritíveis em termos formais e funcionais" (BAZERMAN *et al*, 2005, p.58).

Portanto, essa perspectiva sistêmico-funcional valida a relevância da pesquisa, a questão da variação entre os gêneros, as escolhas linguísticas, bem como a padronização dessas características condicionadas por aspectos referentes ao domínio e ao contexto de uso.

2.3 Swales

Devido à sua proposta de análise de gênero, Swales apresenta uma concepção basilar para esta pesquisa. Em sua obra *Genre Analysis: English academic and research Settings* (1990), o autor oferece uma abordagem para o ensino de inglês acadêmico e

para a pesquisa. Elabora conceitos como comunidade discursiva, gênero e aprendizado de línguas, cujo objetivo é desenvolver uma competência comunicativa de nativos e não nativos no contexto acadêmico. Considera o papel que os textos desempenham no contexto e o propósito comunicativo que molda o gênero, determinando sua estrutura interna e impondo limites quanto às possibilidades de ocorrências linguísticas e retóricas.

Antes de apresentar sua concepção de gênero, é relevante discorrer sobre a noção de *comunidade discursiva*, uma vez que está ligada a de gênero.

Pode-se compreender que *comunidade discursiva* se refere a um conjunto de indivíduos que compartilham as mesmas atividades profissionais ou recreacionais e têm objetivos comuns, portanto, compartilham um conjunto comum de significados. Esses significados podem ser inacessíveis a não-membros dessas comunidades que, dependendo da área de atuação, são quase herméticas, não permitindo assim o acesso de outros.

Essa noção é fundamental para fazer com que o propósito comunicativo de determinado texto seja atingido, uma vez que ela restringe as possibilidades de ocorrências linguístico-discursivas e pressupõe padrões historicamente previstos e realizados por seus outros membros, ou seja, para que o gênero seja reconhecido entre os membros dessa comunidade. Swales (1990) propõe a seguinte definição para *comunidades discursivas*:

> "(...) redes sócio-retóricas que se formam a fim de atuarem em prol de conjuntos de objetivos comuns. Uma das características que os membros estabelecidos dessas comunidades discursivas possuem é a familiaridade com determinados gêneros que são usados no favorecimento desses conjuntos de objetivos. Em consequência disso, os gêneros são propriedades das comunidades de discurso, ou seja, pertencem às comunidades discursivas e não aos indivíduos, a outros tipos de agrupamentos ou a comunidades de fala mais abrangente" (SWALES, 1990 *apud* SILVEIRA, 2005, p 86).

A fim de esclarecer o conceito de comunidade discursiva, Swales (1992) apresenta suas características:

- a comunidade discursiva possui um conjunto perceptível de objetivos que podem ser consensuais ou distintos, mas sempre relacionados;
- possui mecanismo de intercomunicação entre seus membros;
- usa mecanismos de participação para uma série de propósitos, tais como promover o incremento da informação e do *feedback* para canalizar a inovação, para manter o sistema de crenças e de valores da comunidade e para

aumentar o espaço do profissional;

- utiliza uma seleção crescente de gêneros no alcance de seu conjunto de objetivos e na prática de seus mecanismos participativos;
- já adquiriu, mas continua buscando uma terminologia específica;
- possui uma estrutura hierárquica explícita ou implícita que orienta os processos de admissão e de progresso dentro dela.

Convém lembrar que essas características referem-se ao tipo de comunidade discursiva com que o autor trabalhou, no caso, a comunidade científica.

Portanto, a partir do que foi exposto, pode-se observar que o gênero se estabelece dentro de uma comunidade discursiva e ela se torna responsável por ele, ou seja, a comunidade discursiva desenvolve determinados gêneros e a existência de gêneros específicos configura grupos sociais como comunidade discursiva, por compartilhar propósitos comunicativos efetivados por meio dos gêneros pertinentes a ela. Essa noção diz respeito àqueles que trabalham usualmente ou profissionalmente com um determinado gênero e que, deste modo, têm um maior conhecimento de suas convenções. Portanto, dominar razoavelmente os gêneros de uma comunidade discursiva é essencial para que um indivíduo faça parte dela. Em outras palavras, é necessário manipular as convenções comunicativas e pragmáticas de determinada comunidade. Como observa Bonini (2001), conhecer o padrão linguístico particular de certo grupo de indivíduos que atuam comunicativamente mediante propósitos compartilhados é requisito não só para a adesão à comunidade discursiva quanto para a ascensão em sua estrutura hierárquica de participação.

O gênero então possui uma estrutura interna resultante de longa experiência por parte dos membros da comunidade. De acordo com Swales (1990), os gêneros são sócio-retoricamente construídos e não somente objetos textuais mais ou menos semelhantes. São eventos codificados, inseridos em processos sociais comunicativos compartilhados pelas comunidades em que ocorrem e reconhecidos por seus membros como legítimos.

Segundo o autor, há um pressuposto sobre o que seria a estrutura esquemática do discurso, e isso mapearia as escolhas dos enunciadores a respeito de conteúdo e estilo. As opções de escolha são limitadas pela própria comunidade discursiva e pelos outros textos que formam um inventário de exemplos e de consulta, principalmente por indivíduos novos nas comunidades. Diante disso, para Swales (1990):

> "Um gênero compreende uma classe de eventos comunicativos, cujos membros de uma comunidade discursiva compartilham os mesmos propósitos comunicativos. Tais propósitos são reconhecidos pelos membros especialistas da comunidade discursiva de origem, constituindo a racionalidade do gênero. Essa racionalidade modela a estrutura esquemática do discurso influenciando e restringindo as escolhas de conteúdo e de estilo. O propósito comunicativo, além de ser um critério privilegiado, também opera para manter o escopo de um gênero, quando concebido como uma ação retórica comparável. Afora o propósito, os exemplares de um gênero exibem vários padrões de semelhanças em termos de estrutura, estilo, conteúdo e audiência. Se num exemplar forem realizadas todas as expectativas sobre a caracterização de um determinado gênero, esse exemplar será considerado como prototípico pela comunidade discursiva em que ele circula. Os nomes dos gêneros herdados e produzidos pelas comunidades discursivas e importados por outras constituem uma valiosa comunicação etnográfica, mas necessita de posterior validação" (SWALES, 1990 *apud* SILVEIRA, 2005, p.91).

A partir desse conceito, subentende-se que:

- como já mencionado, um gênero é uma classe de eventos comunicativos – em que se devem considerar não apenas o discurso em si mesmo, e seus participantes, mas também o papel desse discurso e os entornos de sua produção e recepção, inclusive os aspectos culturais e históricos;
- o principal traço caracterizador que transforma uma coleção de eventos comunicativos num gênero é algum conjunto compartilhado de propósitos comunicativos e não as semelhanças de formas ou qualquer outro critério. Exceto em casos excepcionais, os gêneros são veículos comunicativos que visam ao cumprimento de objetivos. O autor reconhece que a tarefa de identificar propósitos comunicativos pode ser fácil em determinados gêneros, mas pode ser complicada em outros. Além disso, um gênero pode ter vários propósitos comunicativos, assim como certos tipos de gêneros não aceitam o critério do propósito comunicativo;
- há variação dos gêneros em sua prototipicidade, ou seja, expressões diversas circundadas pela estrutura retórica prevista;
- é o sistema (inerente ao gênero) que estabelece limitações às contribuições linguístico-discursivas disponíveis em função do conteúdo do texto, do posicionamento do autor e da forma textual compartilhada pelos seus pares. Desse modo, uma vez estabelecidos o propósito do gênero por parte dos membros da comunidade discursiva, existe um conjunto de convenções características e, em alguns casos, limitadoras. Essas convenções são

dinâmicas, isto é, elas se modificam ao longo do tempo, mas mesmo assim continuam a exercer influências dependendo do momento histórico;

- há uma nomenclatura usada pela comunidade discursiva que é responsável por dar nomes a classes de eventos comunicativos que os membros da comunidade discursiva reconhecem como exercendo ação retórica recorrente.

Diante do que foi exposto, o conceito de gênero de Swales (1990) apresenta uma perspectiva linguística e social, de modo que é preciso identificar as diferentes partes (características) que formam a estrutura genérica, e isto só é possível se o ponto de partida for a produção textual. É a partir do texto que se podem observar determinadas características recorrentes as quais consolidam um dado gênero.

2.4 Biber: registro/gênero

Sob a perspectiva histórico-cultural e sistêmica, Biber (1988) investigou a língua utilizando como suporte teórico inicialmente o conceito de gênero que, para ele, é geralmente determinado com base nos objetivos dos falantes e na natureza do tópico tratado, sendo, assim, uma questão de uso e não de forma. Posteriormente, optou por utilizar o termo registro, como se preponderasse o viés sociolinguístico que a palavra registro carrega. Além disso, Biber (1995) também considera registro/gênero como uma categoria mais ampla e abstrata, que congrega vários sub-registros, ou seja, são categorias de textos definidas situacionamente (BIBER, 1995).

Ainda referente a termos empregados por Biber, ele difere gênero/registro de tipos de texto, uma vez que estes seriam categorias linguisticamente definidas. No contexto da Análise Multidimensional (AMD), os tipos de textos corresponderiam aos textos cuja distribuição se concentra ao longo das dimensões e que formam grupos linguisticamente semelhantes (KAUFFMANN, 2005).

O conceito de Biber, embora seja semelhante ao de Swales (1990), enfatiza o fato do "gênero/registro ser uma variedade definida por variáveis situacionais e não apenas linguísticas" (BERBER SARDINHA, 2004). Ou seja, esse conceito é amplo, abrange situações variadas, desde um sermão até uma nota de aula, um *e-mail*, certidão, conversa ao telefone, etc. Todas essas ações são consideradas um registro/gênero.

2.4.1 Estudos da variação linguística e a Linguística de Corpus

Biber, Conrad e Reppen (1998), no capítulo 6 da obra *Corpus Linguistics: investigating language structure and use*, abordam questões sobre a descrição de características da língua (lexicais, gramaticais ou traços discursivos [*discourse features*]). No entanto, afirmam que pesquisadores têm estudado questões relacionadas à descrição de registros[8] ao invés de características linguísticas individuais. O termo registro é usado como um *cover term*, ou seja, um conceito impreciso para variedades definidas por suas características situacionais. Alguns registros podem ser mais específicos, como uma novela, ou uma metodologia de um artigo científico, assim como podem ser mais gerais, como uma conversa ao telefone. Segundo os autores, os registros se diferem de dialetos, pois são definidos de acordo com sua situação de uso, considerando seu propósito comunicativo, tópico, interatividade, modalidade. Já dialetos são definidos por suas associações com diferentes grupos de falantes.

Os autores também afirmam que dominar uma escala de registros é essencialmente importante para um falante competente e fluente de uma língua. Evidentemente que ninguém domina um único registro, pois no decorrer de um dia, fala-se e escreve-se sob a orientação de uma escala de registros. Por exemplo, a linguagem que se usa para escrever um artigo é diferente da usada para conversar com um amigo ou parceiro, e esses registros são diferentes daqueles utilizados para se discutir com um professor ou escrever uma carta para a mãe.

Dada a necessidade de permuta entre os registros, a aquisição das características do registro são fundamentalmente importantes para estágios de desenvolvimento no que se refere a aprendizado, por exemplo: uma criança usa uma linguagem para se comunicar com amigos, diferente da que usa com adultos; ou para um estudante de ensino fundamental que descobre as narrativas escritas, cujo registro é diferente de uma conversação, ou ainda, para um estudante universitário que descobre que os modos de dizer de um artigo de Biologia são diferentes das narrativas pessoais escritas.

Para entender o processo de aquisição do registro e as maneiras que o professor utiliza para facilitar este processo, é preciso primeiro descrever as características linguísticas de diferentes registros, identificando suas similaridades e diferenças.

[8] Os autores denominam registro, o que nesta pesquisa denomina-se gênero. Desse modo, o emprego do termo "registro" aqui é para manter a terminologia dos autores.

Embora pesquisadores tenham reconhecido a necessidade de descrições, encontrou-se a dificuldade de conseguir descrever sem uma abordagem baseada em corpus. Para tanto, para estudar registro há três requisitos importantes, segundo os autores:

1. inclusão de um grande número de textos;
2. consideração de uma ampla escala de características linguísticas;
3. comparação das características que identificam registros.

A inclusão de um grande número de textos é importante porque estudos de registros baseados em poucos textos são provavelmente imprecisos.

Segundo, em estudos baseados em apenas alguns registros, as características selecionadas não fornecem descrições detalhadas do mesmo, e as descrições baseadas em tais estudos são provavelmente insuficientes. É raro que um registro seja identificado por ocorrências de uma característica linguística diferente encontrada somente neste registro, ou seja, é raro que um registro seja identificado por uma característica encontrada nele próprio.

Preferencialmente, registros compartilham muitas características linguísticas, como a presença de classes de palavras: substantivos, pronomes, verbos, adjetivos, etc., e eles são diferenciados pelo uso relativo dessas características, isto é, as diferenças sistemáticas no uso relativo de traços linguísticos fornecem características preliminares de diferenças que podem identificar registros.

Finalmente, análises de registros exigem uma abordagem comparativa: é necessária uma base para comparação, cuja finalidade é saber se o uso de um traço em um registro é raro ou comum.

Uma frequência particular, específica, não é nem comum e nem rara. Por exemplo: pode-se somente interpretar a frequência de um determinado traço em determinado registro em relação a outro registro. Uma contagem de frequência para orações relativas em um registro particular é somente significativa em comparação com outro registro.

De acordo com os autores, essas análises são difíceis de serem realizadas manualmente, pois é extremamente demorado analisar à mão mesmo uma ou duas características linguísticas. É quase impossível identificar as ocorrências de características linguísticas múltiplas em uma única leitura do texto. Para experimentar essa difícil tarefa, os autores sugerem o seguinte exercício: escolher uma página de um livro e ler apenas uma vez, tentando identificar todas as ocorrências de cinco características linguísticas:

1. nominalizações;
2. construções passivas;
3. verbos modais;
4. orações relativas;
5. frases preposicionais.

Os autores sugerem cronometrar quanto tempo esse tipo de tarefa exige. Ao terminar, deve-se comparar com algum colega que tenha feito a mesma atividade para determinar o tempo gasto em suas análises. Finalmente, é preciso imaginar o quão difícil seria manter o mesmo tempo de análise, para escrutinar uma dúzia de características linguísticas diferentes, analisando textos provenientes de registros diferentes.

Essa abordagem apresentada pelos autores é de fundamental importância para este trabalho, uma vez que orienta uma metodologia de análise dos gêneros, sobretudo no que se refere aos três requisitos apresentados acima, mais especificamente a comparação entre os gêneros para definir um traço[9] recorrente.

2.4.2 Aplicação de técnicas baseadas em corpus para análises de registros

Para a realização dos estudos mencionados acima, é preciso que grandes quantidades de dados linguísticos (provenientes de um corpus) sejam compilados e organizados, de modo a serem posteriormente analisados com o auxílio de ferramentas computacionais.

Nesse sentido, Conrad, Biber e Reppen (1998) enfatizam que as técnicas baseadas em corpus facilitam muito a realização de estudos detalhados sobre registros. Os computadores tornaram possível armazenar um grande número textos, analisar um grande número de características linguísticas presentes nos textos e fazer comparações dos resultados entre os registros, permitindo um estudo mais sistemático das particularidades de cada registro.

A título de ilustração, os autores apresentam algumas questões de pesquisa que confirmam a relevância de se manter os três requisitos apresentados anteriormente:

[9] Nesta pesquisa, denomina-se *traço* o que os autores chamam de *característica*.

1. Como registros falados e escritos diferem em relação ao emprego das orações subordinadas?

Esta questão de pesquisa ilustra as técnicas baseadas em corpus usadas para estudar diferenças de registro no que diz respeito a um único tipo de construção: as orações subordinadas. No passado, muitos estudos supunham que as partes constitutivas dessas orações tinham finalidades e distribuições similares. No entanto, hoje pesquisas mostram que o uso dessas orações varia dependendo do registro. Além disso, os autores demonstram como a inclusão de poucos textos ou poucas características linguísticas pode conduzir a caracterizações imprecisas do registro.

2. Como se realizam os discursos falado e escrito em inglês? Especificamente, quais os padrões no uso de determinadas características linguísticas que são importantes para diferenciar os principais registros falados e escritos? No que se refere a essas características, os registros falados e escritos são diferentes ou similares?

Por muitos anos, as diferenças entre o discurso falado e escrito foram um tópico de interesse dos linguistas. Responder a esta questão de pesquisa exige uma análise muito mais detalhada de características linguísticas do que a primeira pergunta. Em razão disso, os autores introduziram a Análise Multidimensional (AMD) para investigar a variação de registro. Trabalhando no âmbito do estudo do corpus, a AMD permite a investigação dos padrões principais da variação em características linguísticas dos registros falado e escrito.

3. Como textos de áreas acadêmicas diferentes se distinguem no que diz respeito aos padrões da variação linguística?

Essa questão de pesquisa interessa ao uso do inglês para fins específicos. Atualmente, pesquisas mostram como artigos de duas áreas acadêmicas diferentes variam em seus padrões de uso da linguagem, e como esses padrões linguísticos estão relacionados às finalidades e métodos de cada área. Essa questão mostra como a AMD pode ser usada para investigar registros específicos.

4. Quanto as seções dos textos de um único registro acadêmico variam linguisticamente?

Essa pergunta ilustra a análise de registros muito específicos, por exemplo, as seções dentro dos artigos de Biologia. Nesse caso, aplica-se a estrutura multidimensional a cada uma das seções dentro dos artigos de Biologia, centrando-se mais detalhadamente sobre a variação nas seções: a introdução, os métodos, os resultados e a discussão.

Finalizando, Biber, Conrad e Reppen (1998) afirmam que, para todas as perguntas acima, certamente, em todas as análises da variação do registro, deve-se enfatizar que as técnicas quantitativas não são suficientes. Particularmente, as interpretações qualitativas são necessárias para examinar as bases funcionais que determinam os padrões de características linguísticas.

2.4.2.1 Análise Multidimensional de Variação de Registro (AMD)

A Análise Multitraço e Multidimensional de Variação de Registro (*Multifeature Analysis of Register Variation*) foi criada por Douglas Biber com o objetivo de permitir uma descrição rica e complexa de corpora inteiros de textos por meios estatísticos, bem como a extração precisa de características textuais em comum entre corpora. O nome dessa abordagem deriva do conceito de dimensão de variação. Dimensão, nesse caso, constitui um conjunto de traços que subjazem a um corpus. Esse método de análise possibilita utilizar concomitantemente uma variedade de traços linguísticos empregados na análise textual e aplicar a codificação desses traços a um número de textos maior do que se poderia fazer manualmente, utilizando computadores e técnicas estatísticas. Como para Biber o ideal é combinar a descrição firmada em características situacionais com a descrição baseada em traços linguísticos, a AMD fornece o instrumental para a identificação de padrões de coocorrência dos dois tipos de características, visando a caracterização de uma língua, ou de um conjunto de tipos de textos, de modo abrangente. Por meio dela, a variação entre textos e registros pode ser mais adequadamente descrita por meio de múltiplos parâmetros, possibilitando a utilização de um aparato quantitativo de descrição, o qual permite a especificação da coocorrência dos traços linguísticos de modo preciso (BERBER SARDINHA, 2004).

O processo estatístico da AMD é chamado de Análise Fatorial (AF), que considera a frequência das variáveis utilizadas em todos os textos do corpus para buscar

correlações entre elas e extrair um determinado número de fatores. Ressalte-se que um fator enfeixa variáveis coocorrentes, e cada fator é responsável por uma parcela da variação linguística observada no corpus de estudo. Portanto, para Biber, a partir da interpretação dos fatores, é possível identificar as dimensões[10] (KAUFFMANN, 2005). Sendo assim, é necessário recorrer à utilização de técnicas qualitativas de interpretação, uma vez que as dimensões são rotuladas, e constatar que se combina a análise de nível macro com análise de nível micro, já que a microdescrição dos traços de cada texto permite a indução dos macroagrupamentos textuais ou genéricos (BERBER SARDINHA, 2004).

Berber Sardinha (2004) afirma que:

> "A variação entre registro era investigada por meio de poucos parâmetros (por exemplo, formalidade ou planejamento) e, por conseguinte, a distinção que se fazia entre textos era incompleta, pois privilegiava apenas uma das muitas diferenças que podem existir entre os textos. O emprego de poucos parâmetros também tinha o efeito de polarizar a descrição, assim havia uma tendência para descrever textos através de dois opostos, por exemplo, formal X informal, ou planejado X espontâneo. Por último, a descrição da coocorrência feita através de meios intuitivos podia ser falha, já que o analista não oferecia uma descrição objetiva dos traços que supostamente ocorriam" (BERBER SARDINHA, 2004, p.301).

Portanto, com a AMD deve-se utilizar uma quantidade maior de parâmetros para permitir uma comparação mais abrangente. O analista deve dispor de um conjunto que inclua o maior número possível de características linguísticas, já que o aumento da quantidade de parâmetros implica um número maior de traços linguísticos necessários para cobrir a maior gama de características.

2.4.2.1.1 Pressupostos da AMD

De caráter essencialmente quantitativo e computacional, os pressupostos da AMD são os seguintes (BERBER SARDINHA, 2004):

- permitir descrever a língua por meio de um conjunto variado e extenso de características linguísticas;

[10] Para enfatizar, dimensão é um conjunto de traços que subjazem a um corpus.

- basear-se na análise fatorial;
- utilizar uma maior quantidade de características linguísticas, de forma a aumentar a quantidade de parâmetros de comparação entre corpora;
- não descartar a utilidade de técnicas qualitativas de interpretação, uma vez que as dimensões são rotuladas;
- combinar análise macro com análise micro, já que a microdescrição dos traços de cada texto permite a indução dos macroagrupamentos textuais ou genéricos;
- ser de caráter cumulativo, pois permite a descrição de banco de dados em crescimento;
- ser flexível, pois acomoda diversos traços linguísticos.

2.4.2.1.2 Etapas na realização de uma Análise Multidimensional

De modo geral, as três etapas básicas para execução de uma análise multidimensional são (BIBER, CONRAD, REPPEN, 1998; BERBER SARDINHA, 2004):

1) revisão da literatura em busca de traços linguísticos relevantes, coleta do corpus e codificação dos textos de acordo com o elenco de características linguísticas selecionadas;
2) análise fatorial, fase em que é feito o agrupamento das características linguísticas em fatores e a interpretação funcional desses fatores a fim de descobrir um traço comunicativo dominante subjacente ao fator, dando origem às dimensões;
3) cálculo de escores de cada texto em relação a cada fator e a interpretação das dimensões à luz dos textos que as compõem.

Sendo assim, seguem as principais etapas da AMD:

- levantamento das características linguísticas relevantes através de uma ampla consulta à literatura;
- coleta ou adoção de um corpus de dados linguístico representativo e compatível com as metas da análise;

- transformação das características linguísticas em variáveis quantificáveis;
- codificação dos dados baseada nas variáveis selecionadas, com a utilização de ferramentas computacionais para análise automática, semiautomática ou mesmo manual;
- conferência manual da codificação feita por computador para checar sua exatidão;
- padronização das frequências para permitir a comparação entre variedades (textos, registros ou corpora) de extensões diferentes;
- computação de frequências médias de cada variável;
- análise fatorial inicial para obter os pesos (*loadings*) de cada variável em cada variedade;
- determinação do número de fatores por meio da aplicação de técnicas como observação dos valores *eigen (eigenvalues)* em um gráfico *scree (scree plot)*;
- análise fatorial posterior, com a rotação dos fatores;
- cálculo de escores de cada texto por fator pela padronização dos escores com base na média e no desvio padrão;
- cálculo de escores médios de cada variedade por fator;
- interpretação de cada fator e rotulação das dimensões.

No Brasil, além do projeto de Kauffmann (2005), que aplicou a AMD, há o projeto CORPOBRAS. Esse projeto, coordenado por Lucia Pacheco de Oliveira e desenvolvido na PUC-RJ, tem a finalidade de fornecer subsídios para o estudo de diversos gêneros do discurso (oral e escrito). Atualmente, o corpus é constituído de cerca de 660.000 palavras, mas pretente atingir 1 milhão de palavras, e conta com 21 gêneros discursivos. Além disso, como pretende fornecer resultados de estudos sincrônicos, o corpus enfoca textos contemporâneos, considerando os textos de domínio acadêmico, comercial e jornalístico (artigos científicos, circulares, notícias, editoriais, etc) da última década do século passado e os primeiros anos deste século (1990-2006). Já no caso do domínio literário e pessoal, ou seja, romances, contos, crônicas, cartas pessoais, o corpus considera um escopo maior, mas ainda dentro da contemporaneidade – de 1901 a 2001 (OLIVEIRA, 2003).

Diante do que foi exposto, a importância da AMD para este projeto deve-se às suas etapas básicas, que se enquadram no tipo de análise de gênero empreendido por

esta pesquisa, e aos seus pressupostos, os quais orientam uma análise bastante detalhada de gêneros para fins de classificação. Ressalte-se ainda a adoção de parâmetros para comparação de corpora, bem como a utilização de técnicas qualitativas de interpretação.

2.5 Marcuschi

Semelhante aos conceitos de Swales e Biber, Marcuschi (2008) afirma que o gênero textual diz respeito aos textos materializados em situações comunicativas recorrentes. São entidades empíricas em situações comunicativas e se expressam em designações diversas, não havendo uma categoria fixa de gêneros. O autor afirma que os gêneros textuais "são os textos que encontramos em nossa vida diária e que apresentam padrões sociocomunicativos característicos definidos por composições funcionais, objetivos enunciativos e estilos concretamente realizados na integração de forças históricas sociais, institucionais e técnicas" Marcuschi (2008, p.4). Desse modo, alguns exemplos de gêneros textuais seriam: sermão, diário, telefonema, carta, relatório, receita, reportagem, ata, depoimento, etc. e, no contexto da tecnologia digital, seriam os *e-mails*, salas de bate-papo, *blogs*, etc. Portanto, deve-se atentar para o contexto de uso do gênero em uma perspectiva sócio-retórica.

2.6 Termos e conceitos empregados

Antes de prosseguir com as perspectivas teóricas acerca desta pesquisa, é importante discorrer sobre a utilização e uso de alguns termos e conceitos empregados em pesquisas desta natureza.

Um dos termos empregados é *traço* ou *feature* que, de acordo com Berber Sardinha (2004), são elementos linguísticos pertinentes à análise e que podem ser quantificados, tais como a frequência de determinadas classes de palavras, de determinadas unidades léxicas ou expressões fixas, etc. Tais traços são escolhidos mediante pesquisa na literatura disponível e devem representar um aspecto funcional no

nível do texto. No âmbito da AMD, os traços são chamados de variáveis quando ocorre o processo estatístico que considera a frequência dessas variáveis utilizadas em todos os textos do corpus no qual se busca as correlações entre elas. Embora não seja aplicada a AMD nesta pesquisa, é importante citá-la devido às suas etapas, pressupostos e metodologia.

Outro termo que é importante descrever é *característica.* Berber Sardinha (2005) sugere dois tipos de características: a linguística e a não linguística (ou situacionais como denominam Biber, Conrad e Reppen, 1998). As características linguísticas são traços que se escolhe quantificar e as não linguísticas/situacionais são aquelas preexistentes e que descrevem as características de uso de uma variável, por exemplo, formalidade ou propósito comunicativo.

Tipo textual ou tipo de texto também são termos bastante empregados em pesquisas de descrição linguística. De acordo com Berber Sardinha (2005), tipo de texto "designa um conjunto de textos formados exclusivamente com base em critérios linguísticos". Para Marcuschi (2008), tipo textual:

> "... designa uma espécie de construção teórica (em geral uma sequência subjacente aos textos) definida pela natureza linguística de sua composição (aspectos lexicais, sintáticos, tempos verbais, relações lógicas, estilo). O tipo caracteriza-se muito mais como sequência linguística (sequências retóricas) do que como textos materializados: a rigor são modos textuais. Em geral, os tipos textuais abrangem cerca de meia dúzia de categorias conhecidas como: narração, argumentação, exposição, descrição, injunção. O conjunto de categorias para designar tipos textuais é limitado e sem tendência a aumentar" (MARCUSCHI, 2008, p.154).

Similar a este conceito, Silva (1997) afirma que tipos de textos são estruturas disponíveis na língua, formas linguísticas convencionais de que o falante dispõe na língua quando quer organizar o discurso. Tais formas caracterizam-se por determinados traços linguísticos, como tempo/aspecto/modo verbal, pessoa do discurso predominantemente referida, tipo de predicado, unidade semântica básica, unidade sintática básica. A autora distingue tipo textual de gênero textual afirmando que o gênero faz a utilização dessas estruturas em situações reais de comunicação, ou seja, instâncias de uso em que elas aparecem sob uma organização típica, associadas a diferentes situações comunicativas; por exemplo, estruturas narrativas podem aparecer em diferentes gêneros como história, reportagem policial, etc.

Outro conceito importante encontrado na literatura que serviu como base para elaborar uma tipologia dos gêneros textuais históricos é *domínio discursivo*. Ele

constitui muito mais uma esfera da atividade humana no sentido bakhtiniano do termo do que um princípio de classificação de textos e indica instâncias discursivas (por exemplo: discurso jurídico, jornalístico, religioso, etc.). Não abrange um gênero em particular, mas dá origem a vários deles, já que os gêneros são institucionalmente marcados, do ponto de vista de alguns teóricos como Miller (1984).

Para finalizar, vale ressaltar que o termo *traço linguístico* empregado nesta pesquisa refere-se a elementos linguísticos recorrentes nos gêneros e quantificáveis, que devem abarcar elementos morfológicos, sintáticos, e lexicais.

3. Categorização de textos

No que se refere ao aspecto computacional desta pesquisa, estamos no âmbito do Processamento Automático de Língua Natural (PLN), definida pelo Comitê Especial de Processamento de Língua Natural (CE-PLN)[11] como a área que:

> "lida com problemas relacionados à automação da interpretação e da geração da língua humana em aplicações como Tradução Automática, Sumarização Automática de Textos, Ferramentas de Auxílio à Escrita, Perguntas e Respostas, Categorização Textual, Recuperação e Extração de Informação, entre muitas outras, além das tarefas relacionadas de criação e disponibilização de dicionários/léxicos e *corpus* eletrônicos, desenvolvimento de taxonomias e ontologias, investigações em linguística de *corpus*, desenvolvimento de esquemas de marcação e anotação de conhecimento linguístico-computacional, resolução anafórica, análise morfossintática automática, análise semântico-discursiva automática". (Comitê Especial de Processamento de Língua Natural).

Nesse sentido, a pesquisa trata da tarefa de categorização de textos, baseada em gênero via métodos de Aprendizado de Máquina Supervisionado. A seguir, são apresentados alguns conceitos basilares para a pesquisa.

3.1 Trabalhos na área de Categorização de Textos

A categorização de textos teve suas pesquisas datadas a partir da década de 60. Contudo, somente nos anos 90 observa-se que recebeu maior atenção dos pesquisadores, em decorrência do aumento e desenvolvimento de ferramentas mais eficientes.

Na literatura foram encontradas diversas definições de categorização, dentre elas a definição de Peixoto, Batista e Capelo (2004), que afirmam que "a categorização de textos é uma ferramenta utilizada para classificar automaticamente um conjunto de documentos numa ou mais categoria preexistentes, não tendo outra finalidade senão recuperar a informação". Tal definição é bastante abrangente ao retomar a questão da categoria, classe e finalidade, fatores primordiais para o desenvolvimento de qualquer ferramenta, em conjunto com a sua aplicação.

[11] http://www.nilc.icmc.usp.br/cepln/

Para Rigo, Oliveira e Barbieri (2007), a categorização "trata-se de uma técnica que possibilita auxílio na localização de resultados desejados em pesquisas e em sistemas de recomendação, minimizando a sobrecarga de informação. Consiste no processo de classificar automaticamente um conjunto de documentos em uma ou mais categorias pré-existentes facilitando a busca seletiva de informações".

Moens (2000) afirma que o homem executa a categorização de texto lendo o texto e deduzindo as classes de expressões específicas e seus padrões de contexto. A categorização automática simula este processo e reconhece os padrões de classificação como uma combinação de características de texto. Esses padrões devem ser gerais o bastante para ter grande aplicabilidade, mas específicos o suficiente para serem seguros quanto à categorização de grande quantidade de textos.

Outra definição é a de Rizzi *et al*. (2000), os autores afirmam que a categorização de textos é uma técnica utilizada para classificar um conjunto de documentos em uma ou mais categorias existentes e geralmente é utilizada para classificar notícias, resumos e publicações.

Em síntese, ora categorização aparece como técnica ora como ferramenta. Pode-se constatar que categorização de textos é uma técnica, disponibilizada, suportada, implementada por uma ferramenta, usada em Processamento Automático de Língua Natural (PLN) para classificar automaticamente um conjunto de documentos ou textos numa ou mais categorias preexistentes, orientadas tanto a partir de critérios pré-determinados, como por características linguísticas, assunto, conceitos, por fim, dependente de sua própria aplicabilidade.

3.2 Aprendizado de máquina (AM)

Pelo exposto e pelo objetivo da pesquisa no que tange à classificação automática, este trabalho se apóia nos pressupostos teóricos metodológicos advindos do PLN, mais especificamente na subárea de Aprendizado de Máquina (AM). Nas palavras de Rezende (2003), AM é uma subárea da Inteligência Artificial que pesquisa métodos computacionais relacionados à aquisição de novos conhecimentos, novas habilidades e novas formas de organizar o conhecimento já existente. Monard e Barauskas (2003)

compartilham com essa definição ao afirmarem que o AM vale-se da construção de sistemas capazes de adquirir conhecimento de forma automática.

No Brasil, muitas pesquisas foram realizadas sobre Aprendizado de Máquina e classificação automática de gênero, assuntos, bem como seleção de atributos, a saber:

- Galho e Moraes (2003), que apresentaram um protótipo para a categorização automática de notícias, em português, utilizando a técnica de similaridade difusa.
- Moraes e Strube de Lima (2007), que estudaram categorização hierárquica de documentos, utilizando determinado algoritmo para classificar os assuntos de uma grande coleção de textos escritos em língua portuguesa, o corpus PLN-BR CATEG, criado no âmbito do projeto PLN-BR[12] (BRUCKSCHEN, et al., 2008).
- Silva e Vieira (2005, 2007), que estudaram os grupos gramaticais e sintáticos em categorização automática, e também a categorização de textos: Categorização de Textos da Língua Portuguesa com Árvores de Decisão e Informações Linguísticas.
- Matsubara (2004), que realizou uma pesquisa de mestrado sobre *Algoritmo de Aprendizado supervisionado Co-Training e sua aplicação na rotulação de documentos*.
- Sanches (2003), que desenvolveu, também no mestrado, o trabalho *Aprendizado de máquina semi-supervisionado: proposta de um algoritmo para rotular exemplos a partir de poucos exemplos rotulados.*
- Nogueira (2009), que pesquisou, também no mestrado, a avaliação de métodos não-supervisionados de seleção de atributos para mineração de textos.
- Martins (2003), que desenvolveu a tese intitulada *Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado.*

Como se observa, são várias as pesquisas sobre aprendizado de maquina e categorização. Ressalte-se que há diversos sistemas de aprendizado de máquina, os quais possuem características particulares e comuns que possibilitam sua classificação quanto à linguagem de descrição, modo, paradigma e formas de aprendizado. Sendo assim, as estratégias de aprendizado são as seguintes: Aprendizado por Hábito, Instrução, Dedução, Analogia e, por fim, Indução, que será realizada nesta pesquisa.

[12] http://www.nilc.icmc.usp.br/plnbr/

Segundo Monard (1997), tais estratégias apresentam crescente complexidade de inferência, na seguinte ordem:

1. **Aprendizado por hábito:** o aprendiz não precisa desempenhar nenhuma inferência sobre a informação fornecida. O conhecimento é diretamente assimilado pelo aprendiz.
2. **Aprendizado por instrução:** o aprendiz adquire conceitos de uma fonte, mas não copia diretamente a informação fornecida para a memória, ele engloba a seleção dos fatos mais relevantes e/ou uma transformação da informação fonte em formas mais apropriadas.
3. **Aprendizado por dedução:** o aprendiz adquire um conceito através de dedução sobre o conceito já adquirido.
4. **Aprendizado por analogia:** o aprendiz adquire um novo conceito modificando a definição de um conceito semelhante já conhecido
5. **Aprendizado por indução:** é a forma de inferência lógica que permite que conclusões gerais sejam obtidas de exemplos particulares. O aprendiz adquire um conceito fazendo inferências indutivas sobre os fatos apresentados.

O aprendizado indutivo pode ser dividido em supervisionado, não-supervisionado e semissupervisionado. Contudo, independente da estratégia, existem modelos comuns a todos. De acordo com Monard e Baranauskas (2003) são os seguintes:

- **Simbólico:** sistemas de aprendizado simbólico que buscam aprender construindo representações simbólicas de um conceito através de análise de exemplos e contra-exemplos desse conceito. As representações simbólicas estão tipicamente na forma de alguma expressão lógica, árvore de decisão, regras ou rede semântica.
- **Estatísticos:** sistemas que utilizam modelos estatísticos para encontrar uma boa aproximação do conceito induzido. Dentre os métodos estatísticos, destacam-se os de aprendizado Bayesiano, que utilizam um modelo probabilístico baseado no conhecimento prévio do problema, o qual é combinado com exemplos de treinamento para determinar a probabilidade final de uma hipótese.
- **Baseado em exemplos:** uma forma de classificar um exemplo é lembrar-se de outro similar cuja classe é conhecida e assumir que o novo exemplo terá a mesma classe.

- **Conexionista:** são as famosas Redes Neurais, as quais são construções matemáticas simplificados inspiradas no modelo biológico do sistema nervoso. Sua representação envolve unidades altamente interconectadas e, por esse motivo, o nome conexionismo é utilizado para descrever a área de estudo.
- **Evolutivo:** consiste de uma população de elementos de classificação que competem para fazer a predição.

3.2.1 Aprendizado de Máquina Supervisionado

Como já foi mencionado, o aprendizado de máquina supervisionado é por indução. Tem como objetivo induzir conceitos a partir de exemplos que estão pré-classificados, ou seja, os exemplos estão rotulados com uma classe conhecida, ou conceito. Já no aprendizado não supervisionado os exemplos não possuem uma classe correspondente. Nesse caso, as tarefas podem ser relacionadas com o agrupamento dos exemplos (ou *clustering*) com uma descrição compacta de um subconjunto de dados, denominado sumarização, ou com a caracterização, por meio de regras de associação, do quanto à presença de um conjunto de atributos implica na presença de algum outro conjunto distinto de atributos nos mesmos exemplos. O processo de aprendizado supervisionado é caracterizado pela apresentação de dados de treinamento a um algoritmo de aprendizado, o indutor. Cada exemplo possui uma classe associada. Há também o conceito de atributo, que é uma característica ou uma informação que visa descrever o exemplo, o qual pode ter ou não um rótulo associado. Esse rótulo é a classe do exemplo e representa um atributo especial que descreve uma instância do fenômeno de interesse, que é o conceito que se deseja induzir em tarefas de classificação (Martins, 2003).

Candido Jr. (2008), no contexto do corpus histórico, exemplifica que para um metadado, como data de edição, é possível criar as classes século XVI, século XVII e século XVIII. A partir de um conjunto de textos datados desses três séculos, um classificador pode ser treinado e utilizado para identificar novos textos sem datação conhecida, criados durante esses séculos. Isso é permitido por meio da descrição de traços linguísticos presentes nos textos.

Os métodos de classificação automática aplicados nesta pesquisa foram alguns dos algoritmos classificadores disponibilizados no ambiente Weka (*Waikato Environment for Knowledge Analysis*)[13], que é formado por um conjunto de implementações de algoritmos de diversas técnicas de Mineração de Dados. Entre os classificadores disponibilizados, são utilizados os seguintes:

- os que se utilizam de modelos estatísticos: Naive Bayes e Bayes Net; SMO[14]
- os baseados em redes neurais: Multilayer Perceptron e RBFNetwork;
- os que utilizam modelos simbólicos, particularmente, árvore de decisão: J48 e NBTree.

[13] http://www.cs.waikato.ac.nz/ml/weka/
[14] Sequential Minimal Optimisation

4. Metodologia

Neste capítulo, apresentar-se-ão todas as etapas metodológicas constitutivas da pesquisa, desde os primeiros exercícios com o corpus, até a tabela de traços adaptados ao contexto histórico[15]; os programas computacionais utilizados, suas vantagens e as dificuldades da pesquisa; bem como toda a interação necessária entre a linguista e o informata.

Uma vez que o objetivo do projeto é identificar os traços linguísticos recorrentes de textos em língua portuguesa dos séculos XVI, XVII, XVIII e XIX que permitirão a classificação em gêneros textuais, para sua consecução, empregaram-se alguns recursos metodológicos proporcionados pela Linguística de Corpus, tais como a obtenção de frequência das características. Acerca disso, formularam-se duas questões de pesquisa que expressam as linhas de investigação que orientam este trabalho:

1. quais são os traços linguísticos recorrentes no português dos séculos XVI, XVII, XVIII e XIX?
2. quais os traços específicos de cada gênero?

De acordo com Hoey (1991), ao estudar a relação entre o léxico e o texto, deve-se ressaltar a intensa ligação entre esses dois aspectos na produção comunicativa: “cada seleção textual coage nas escolhas lexicais possíveis e é nessa combinação entre escolhas lexicais e textuais efetuadas por escritores ou falantes que sua atividade é expressa”.

Semelhante a essa perspectiva, encontra-se a abordagem teórica de Halliday (1991), que defende que a linguagem é um sistema probabilístico em que certos padrões são mais frequentes que outros, o que permite descrever a probabilidade dos sistemas linguísticos, dados os contextos em que os falantes os empregam. Sua visão da linguagem enquanto sistema probabilístico pressupõe que, embora muitos padrões linguísticos sejam possíveis teoricamente, eles não ocorrem com a mesma frequência. Exemplo disso, no nível morfossintático, é a frequência de substantivos (no inglês e, certamente, no português) que é maior do que qualquer outra categoria, já que 25% das palavras são substantivos (KENNEDY, 1998). Desse modo, a probabilidade de um

[15] Apresentada também no Apêndice B.

padrão ser um substantivo é maior do que outra classe gramatical, e, portanto, os usos de elementos morfossintáticos ou lexicais não se realizam com a mesma frequência.

Dessa maneira, o mais importante da diferença de frequências entre os traços é o fato de essas diferenças não serem aleatórias. Se o fossem, então o uso de elementos morfossintáticos ou lexicais ao se realizarem com frequências diferentes não seria significativo, isto é, não acrescentaria informação a respeito da própria estrutura e constituição do enunciado, do texto. Entretanto, pelo contrário, há um mapeamento regular entre a frequência maior ou menor de um padrão e um contexto de ocorrência. Ou, nas palavras de Biber (1988, 1995), há uma correlação entre características linguísticas e situacionais (os contextos de uso). O conjunto da pesquisa desenvolvida por Biber apresenta evidências inequívocas de que conjuntos de padrões linguísticos variam sistematicamente com relação a textos típicos de contextos comunicativos específicos. Em outras palavras, a variação não é aleatória.

Diante disso, quando se diz que a variação não é aleatória, na verdade, está se afirmando que a linguagem é padronizada (*patterned*). A padronização se evidencia pela recorrência, isto é, uma colocação, coligação ou estrutura, que se repete significativamente, mostra sinais de ser na verdade um padrão lexical ou léxico-gramatical. A linguagem forma padrões que apresentam regularidade (se mostram estáveis em momentos distintos, isto é, têm frequência comparável em corpora distintos) e variação sistemática (correlacionam-se com variedades textuais, genéricas, dialetais, etc).

No caso do léxico, podem-se diferenciar as palavras entre aquelas de maior frequência e as de menor frequência, sendo que a diferença entre elas é relativa. Assim, algumas palavras têm frequência de ocorrência muito rara e, para que haja probabilidade de ocorrerem no corpus, é necessário incorporar-se uma quantidade grande de palavras ao corpus. Em outras palavras, quanto maior a quantidade de palavras (ou quanto maior for o corpus), mais probabilidade há de palavras de baixa frequência aparecerem.

Biber, Conrad e Reppen (1998) sugerem que há três requisitos importantes para estudar registros/gêneros: inclusão de um grande número de textos, consideração de uma ampla escala de características linguísticas e comparação entre os registros/gêneros, como já mencionado anteriormente.

Com base nisso, foi necessária uma observação preliminar do corpus em um nível macrolinguístico, combinada a uma análise microlinguística de identificação de padrões

recorrentes, com o auxílio de uma metodologia de pesquisa de natureza essencialmente quantitativa, para avaliação da frequência das características e categorias linguísticas coocorrentes, por isso, os procedimentos empregados são os sugeridos pela Linguística de Corpus.

Referente às etapas metodológicas, a primeira é iniciada a partir da tabela de traços contemporâneos elaborada por Aires (2005) e que, a partir dela, foi necessário refinar os traços a partir de um processo incremental e iterativo de buscas no corpus (*bootstrapping method*) (AUGER; BARRIÈRE, 2008, p. 5), finalizando com a geração do arquivo ARFF e treinamento com os algoritmos classificadores. Nas próximas seções, serão descritos o corpus, as etapas metodológicas em detalhes e os métodos de aprendizado de máquina supervisionados escolhidos nesta pesquisa. Finalizamos com a análise de contribuição dos traços escolhidos com o algoritmo de seleção de atributos InfoGainAttributeEval do ambiente WEKA (Witten e Frank, 2005), mostrando os mais distintivos no corpus de treinamento. Abaixo, encontra-se a Figura 1 que representa as etapas metodológicas realizadas, iniciada a partir da tabela de traços contemporâneos de Aires (2005)[16].

[16] Ver seção 4.3

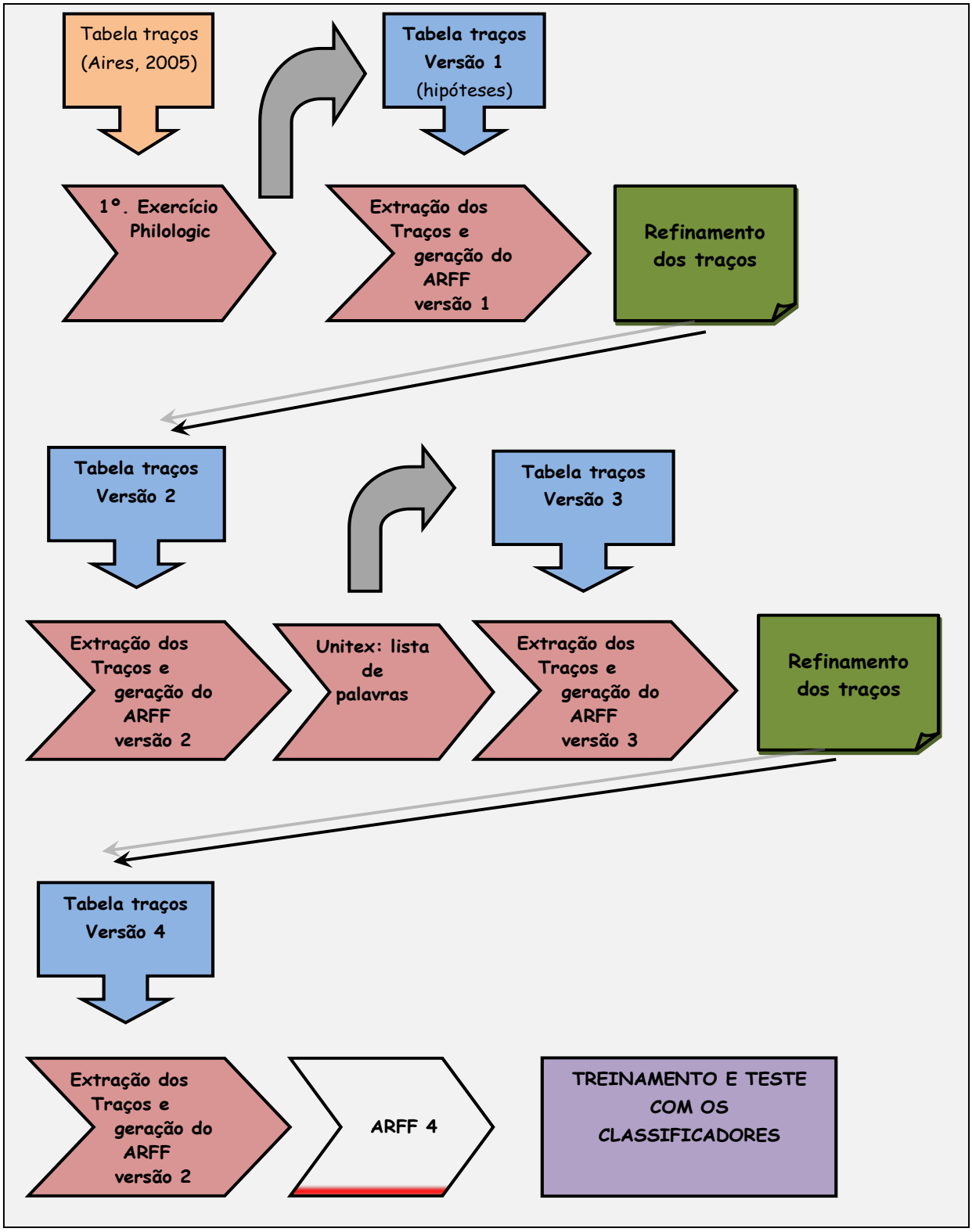

Figura 1 - Etapas de análise de traços no corpus DHPB para a escolha dos traços adequados para adequados para a classificação dos gêneros escolhidos

4.1 Descrição do corpus do projeto DHPB

Considerando a linguagem como um sistema global e probabilístico, pode-se afirmar que todo corpus é um fragmento de língua, entretanto, tem de ser capaz de representar o sistema global, em sua totalidade ou não, refletindo possibilidades de ocorrências de usos linguísticos significativos. Nessa perspectiva, descreve-se esquematicamente o corpus de estudo, de acordo com o número de textos, palavras e formas. Inicialmente, porém, apresenta-se o projeto responsável pela elaboração desse corpus, as instituições envolvidas e algumas peculiaridades de um corpus histórico.

O corpus utilizado foi *desenvolvido no âmbito do projeto intitulado Dicionário Histórico do Português do Brasil – séculos XVI, XVII, XVIII e XIX*, que integra o Programa Institutos do Milênio do CNPq. O projeto, coordenado inicialmente por Maria Tereza Camargo Biderman[17] (FCL, UNESP, Campus de Araraquara), tem como objetivo elaborar um dicionário do português do Brasil dos séculos XVI, XVII, XVIII e XIX a partir de *corpora*. Ressalte-se que o Brasil não conta com nenhuma obra lexicográfica sobre seu vocabulário nos primeiros tempos da formação do Português Brasileiro. Assim, essa seria uma obra pioneira e necessária. O projeto tem/teve como instituições parceiras a Universidade de Évora (Portugal), a Universidade de São Paulo (Campus de São Paulo e Campus de São Carlos), a Universidade Federal de São Carlos, a Universidade Federal do Rio Grande do Sul, Universidade Federal de Minas Gerais, a Universidade Federal do Mato Grosso do Sul, Universidade Federal da Bahia.

Esse projeto, ao propor a construção de uma obra lexicográfica sobre o vocabulário nos primeiros tempos de formação do Português do Brasil, ou seja, nos séculos XVI, XVII, XVIII e XIX, revela-se pioneiro, pois sua realização preencherá uma lacuna na cultura brasileira. Para tão vultosa tarefa, foi necessária a organização de equipes de pesquisadores provenientes de várias regiões do Brasil como também de Portugal, sobretudo no que se refere à busca por documentos e obras que constituiriam o corpus.

Para que o material recolhido adquirisse um formato digital e condições de uso, várias atividades e etapas foram realizadas, desde a digitalização de documentos de edições impressas dos séculos XVI, XVII, XVIII e XIX, convertidos primeiramente em

[17] Devido ao falecimento da pesquisadora em 29/05/2008, a coordenação do projeto está agora a cargo da Profa. Dra. Clotilde de Almeida A. Murakawa, da UNESP Araraquara.

arquivos de imagem para então serem transformados em arquivos de texto, até a digitação de documentos impróprios para a digitalização. Ressalte-se a enorme tarefa manual de limpeza, revisão e formatação dos textos digitalizados, já que as ferramentas de processamento de corpus não são capazes de tratar determinados formatos próprios dos textos escritos daquelas épocas. De acordo com Candido Jr. (2008), os problemas mais recorrentes diziam respeito a:

- a utilização de caracteres que caíram em desuso: ligaduras, consoantes acentuadas e símbolos gerais em Latim. Para solucionar isso, optou-se por utilizar o Unicode[18], uma vez que esses caracteres não são emitidos pelos padrões de codificação usuais;
- a grande quantidade de abreviaturas: o processamento de abreviaturas é particularmente difícil, pois elas são ambíguas e podem ter um grande número de significados. Além disso, sua presença no corpus limita o poder das ferramentas de extração e recuperação de informação;
- as diversas variações de grafia para uma dada palavra: os textos que constituem o corpus, por terem sidos escritos nos séculos XVI, XVII, XVIII e XIX, época em que não havia um sistema ortográfico unificado para o português, apresentam para uma única palavra diferentes grafias, dificultando as buscas automáticas, o que limita os resultados de busca por um padrão quando não se conhecem todas as variações;
- o problema das junções ou contrações: os principais problemas ocasionados pelas junções são as distorções na contagem de frequência e a necessidade de se criarem expressões de busca mais sofisticadas para uso no concordanciador, as quais fossem capazes de recuperar uma palavra dentro de uma junção. A solução para este problema foi separar as junções manualmente.

Ao final, o corpus compilado é constituído por 2.458 textos e 7,5 milhões de formas simples. Dentre os textos selecionados, encontram-se cartas de doação, depoimentos, diários, cartas de missionários jesuítas, inventários, autos, testamentos, relatos, documentos da inquisição católica, entre outros. A Tabela 1 apresenta outras informações.

[18] "Unicode é um padrão que permite aos computadores representar e manipular, de forma consistente, texto de qualquer sistema de escrita existente, ou seja, é um esforço para a criação de um padrão que suporte todos os idiomas contemporâneos". (Cândido Jr., 2007, p.23).

Dados	Valores
Tokens – número de itens ou palavras, incluindo suas repetições	16.505.808
Types – cada item ou palavra sem repetições	368.850
Formas simples	7.492.473
Formas simples únicas	368.529
Sentenças	287.570
Textos	2.458
Tamanho em MegaBytes (UTF-16)	82,2

Tabela 1 - Desenho do corpus

Uma vez que foi preciso comparar os gêneros para identificar os traços, foram elaborados 9 subcorpora, constituídos pelos seguintes gêneros: assento, auto de provimento, diário, escritura, parecer, registro, sermão, *termo* e notícia. Para essa seleção, o critério utilizado foi selecionar os gêneros que possuíssem um número equivalente de textos no corpus, aproximadamente vinte. Dessa forma, dez textos de cada gênero foram utilizados para treinar os classificadores e o restante para realizar os testes de classificação.

4.2 Tipologia de gêneros do português do Brasil dos séculos XVI, XVII, XVIII e XIX.

No âmbito do projeto do *Dicionário Histórico do Português do Brasil (DHPB)*, para melhor consulta do corpus, Souza, Aluísio e Almeida (2006) elaboraram uma tipologia textual manualmente. É essa tipologia que, neste trabalho, orienta a classificação dos gêneros e o domínio discursivo a que pertencem.

Para propor essa tipologia de modo que ela contivesse os tipos de textos existentes nas épocas abarcadas pela pesquisa, partiu-se do *Catálogo Alberto Lamego*. Nesse catálogo, estavam listados os manuscritos presentes no acervo do Instituto de Estudos Brasileiros da USP. Foi dessa lista que se retirou grande parte dos documentos dos séculos XVI, XVII e XVIII. Para não deixar nada de fora, considerou-se também a tipologia organizada para o Projeto Lácio-Web[19]. Juntando os tipos de textos

[19] Produto de uma parceria entre o Núcleo Interinstitucional de Linguística Computacional (NILC/ICMC-USP, São Carlos), o Instituto de Matemática e Estatística (IME-USP, São Paulo) e a Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH-USP, São Paulo), o projeto, financiado pelo CNPq e coordenado por Sandra Maria Aluísio, tem como objetivo divulgar e disponibilizar livremente na Web vários corpora do português brasileiro e ferramentas linguístico-computacionais.

encontrados no *Catálogo Alberto Lamego* e os tipos considerados no Lácio-Web, foi gerada uma lista inicialmente com 194 tipos.

Uma vez que se trabalha com documentos bastante diferentes do que hoje é conhecido, foi necessário pesquisar a definição de cada um. Feito isso, procurou-se organizar todos os tipos encontrados em grupos afins ou categorias. Partiu-se do esquema utilizado no Projeto Lácio-Web, o qual está dividido nos seguintes gêneros: Científico, Referência, Informativo, Jurídico, Prosa, Poesia, Drama, Instrucional, Técnico Administrativo. Entretanto, percebeu-se que esse esquema não incluía todos os gêneros, por isso, os tipos e as categorias foram adaptados.

Optou-se por utilizar a divisão e a terminologia propostas por Marcuschi (2002), em que propõe a denominação "domínio discursivo" para aquilo que no Lácio-Web era chamado de gênero, já que designa uma esfera ou instância da produção discursiva ou atividade humana. Por isso, fala-se em discurso jurídico, jornalístico, religioso etc. Dessa forma, organizou-se a tipologia em domínio, subdomínio e gêneros textuais.

Uma das dificuldades no desenvolvimento da tipologia foi estabelecer relações hierárquicas entre domínio, subdomínio e gêneros, principalmente no Domínio Jurídico, além da necessidade de criar outros domínios como o Domínio Pessoal.

Vale ressaltar que a atividade de agrupar, segregar, classificar é compreender, sob a ótica da tipologia textual, qual o conteúdo que as designações representam e, principalmente, o que elas possuem em comum para que as relações sejam estabelecidas. Dessa forma, foram organizados os Domínios Jurídico, Científico, Informativo, Referencial, Instrucional, Técnico-administrativo/Oficial, Religioso, Literário, além do Domínio Pessoal. A tipologia está constituída por aproximadamente 220 gêneros e 9 domínios (conferir a tipologia completa no Apêndice C).

Nos Quadros 1 a 9, apresentam-se trechos dos textos de cada gênero utilizado na pesquisa, a título de exemplificar a complexidade do tipo de corpus, a questão da variação de grafia e outros fenômenos. Os nomes dos gêneros foram preservados conforme titulado em cada documento.

1. Assento

Assento que fazem Aoprimeiro dia domez de Junho demil seis eentoz equarenta annoz nesta Cidade do Salvador Bahia detodoz os Santoz, e Cazas da Camera estando ahiprezentes os Juizes Vereadores, emais Officiaes della sendo chamado opovo que sam de Companhia escolhidoz segundo o costume lheforam lidas as propostas do Senhor Marques Visse Rey deste Estado emque pedia a Camera desta Cidade ordenasse que sem haver dillação alguma fizesse cobrar odinheiro que estivesse colhido doque esta Cidade tinha concedido para as crennas eque isto sefizesse por Repartição igual sem afeição de Pessoa alguma para com elle setratar comtoda abrevidade defazer quatro galiatoz Reforçadoz para guarda desta Cidade edesta Bahia eseu Reconcavo, por quanto entendem ospraticoz que esta he averdadeira fortificação eque se ouvira quando digo equese os ouvera quando aguerraveyo aqui o Innimigo nam confiscara, efizera osdannos quefez, eque lhesfosse defabrica , e apresto dos Galiatoz se empregaria tambem noque fosse mais conveniente para defensa do mar, ea Cidade onde nosmarcão as qualidades dosteres , eserá o Thezoureiro da Cidade o Cidadam mais antigo digo do mar, ea Cidade: Ordena o Marquez que oguardem tres chavez, eterá huma o Thezoureiro da Cidade o Cidadam maiz antigo, eoutra para que oditto Senhor Marquez Visse Rey ordenou para que constasse sempre a Sua Magestade como senam divertia este dinheiro do seu Serviço eanecessidade delle por seus Ministroz esemque sedespendera, esendo assim lida aproposta aoditto povo porque foi respondido que o assento que sefizera emque convieram asedarem os Secenta mil cruzados ao Conde da Torre de cujo Resto sepedem os quarenta, ecinco mil cruzados fora com condiçam (...)

Quadro 1: Exemplo de texto do gênero assento

2. Auto de provimento

Anno de Nascimento de Nosso Senhor Jesus Christo de mil e setecentos setenta e seis aos onze dias do mez de Março do dito anno nesta Villa de Curitiba em correição nas casas de aposentadoria do Doutor Antonio Barboza de Matos Coitinho, Ouvidor e corregedor desta Comarca aonde eu escrivão de seu cargo adiante nomeado fui vindo, e sendo ahy tambem presentes os Juizes ordinarios e mais officiaes da Camara desta Villa, e sendo ahy todos presentes para effeito de se proceder a Provimentos de Correição no que parecer a elle Doutor Ouvidor Geral e Corregedor para o bom regimento desta Republica, e utilidade do Bem Comum, os quaes provimentos são os seguintes, para o que elle dito Ministro mandou fazer este auto por mim escrivão que asignou no fim delles com os ditos Juizes ordinarios e officiaes da Camara, e en Pedro Martins escrivão da Ouvidoria geral e correição o escrevy.

Porque tem emtroduzido o abuzo dos recomendaveis provimentos do Dez.^or Rafael Pires Pardinho, não só hua total inabed.^o, mas ainda o concideravel prejuizo, que sentem alguns dos moradores desta V.^a e seu termo principalmente no esquecimento do Cap.^o 64 tão util â L.^ca quão recomendavel ao cuid.^o pellos rezaltantes enteresses dos criadores deste continente; se faz indispensavel recordar advertindo o mesmo de que deverão lembrarse enteressados os que sucedendo no governo da Republica deverão concervar illeza aquela lembrança de que o bom regimem e concervação das suas criaçoins mesmo que se acha estabelecida se faz precizo suscitar do esquecimento aquella lembrança tão justa como por este modo os continuados extravios, e roubos de que se lamentão os seus habitadores:

Proveo em pr.^o lugar que todos os criadores uzem de sua distinctiva marca e propria em todos os animais da sua criação : Em seg.^do que todos os compradores e negociantes que comprão gados, e outras qualidades de animais recebão hum escrito do vendedor declarando os que vende, sua qualidade, cores, e marcas, que leva, e o comprador os não poderá levar sem que me . . . . . licença da Cam.^a lhe faça certa com o escrito do vendedor a sua compra cujos escriptos serão guardados na arca da mesma Camr.^a para asim servir no conhecimento dos furtos, que nos d.^os animais se costuma fazer com notavel prejuizo de seus donos maquinado as mais das vezes pellos mesmos vendedores(...)

Quadro 2: Exemplo de texto do gênero auto de provimento

3. Diário

DIÁRIO E MARCHA DA COMPANHIA DE QUE É CAPITÃO ESTÊVÃO RIBEIRO BAIÃO.
Dias de marcha Julho de 1769 Amanheceu o dia 20 de julho de 1769, quinta-feira, com sinais de chuva; porém desfazendo-se as nuvens, se mostrou sereno e saiu o sol indicando bom tempo; e por isto, como estava tudo pronto, se correu a caixa, e mandando-se perfilar a esquadra de Inácio da Mota, que se compunha de gente dos Campos Gerais, entre os quais iam os picadores, e guias, lhes encomendou o Sr. D. Afonso, a felicidade e deligência, e tomando as munições de um, e outro gênero, que pôde levar, desfilou, e se pôs em marcha subindo a lombada, que fica defronte dos quartéis da parte do sul, e corre de leste para oeste pouco mais, ou menos, e indo do mesmo modo com a cara ao sudueste, foi a dita esquadra pouco a pouco dobrando para noroeste, e seguindo êste rumo até a entrada do mato. Aí pousaram, e andariam légua e quarto.
Saiu esta esquadra às 11 horas do dia, e foi dando salvas, e vivas El-Rei Nosso Senhor. Sinal de alegria, com que começavam o serviço. Esta esquadra era dos picadores.
No mesmo dia às 2 horas da tarde perfilou-se a esquadra de Francisco de Oliveira Franco, e se pôs em marcha dando salvas, e vivas com muito prazer. Nesta saiu o sargento do número. Esta esquadra tôda era composta de homens eminentes caçadores. Seguiu a mesma marcha, e o mesmo pouso, que a primeira. Pelo que respeita ao sustento da tropa tem servido bem, e principalmente o soldado Pedro da Cruz.
2 Na segunda-feira 21 do dito mês, e ano, feita a mesma evolução, saiu a esquadra de Miguel Fernandes, fazendo e dando os mesmos sinais de alegria ; seguiu a mesma derrota, e fêz o mesmo pouso : saiu às 9 horas. (...)

Quadro 3: Exemplo de texto do gênero diário

4. Escritura

Saibão quantos este publico instrom^to de escriptura de distrato de outra virem que no anno do nascimento de nosso Senhor Jezus christo de mil e seis sentos e quarenta e nove annos aos trinta e hum dia do mes de Julho do dito anno nesta Cidade do Salvador Bahya de todoz os Santos e pouzadas do Lecenseado Hieronimo de Burgoz Juis dos orfaós desta Cidade a Donde eu taballião aodiante nomeado fui ahy apareserão a esto prezentes e outorgantes partes, a saber Pedro Borgez Pacheco tutor da orfam filha que ficou de Francisco Rodrigues morador no termo desta Cidade, e o reverendo Padre Frey Pedro de Jezus procurador Geral do Mosteyro de São Bento desta Cidade pesoas de mim taballiãm reconhecidas, e logo por o dito Padre frey Pedro de Jezus foi dito em minha prezensa, e das testemunhas aodiante nomeadas, que sendo vivos Antonio Fernandes, a sua molher Maria Rodrigues de oLiveyra fizeráo Venda de húas Cazas com suas Logias a retro aberto citas na Ladeyra que vay do Guindaste para a Praya desta Cidade, a Francisco Rodrigues Roza de que fizeráo escritura nas notas do taballião João de freytas aos oitos dias do mes de Dezembro de mil e seis sentos e trinta e quatro annos, e porquanto o dito Antonio fernandes, e sua mulher Maria Rodriguez de oLiveyra erão falecidos da vida prezente, e a seu Nosteyro ficara por seu herdeiro, e o dito Francisco Rodrigues Roza tambem era falecido da vida prezente a quem ficara huá filha orfam de quem era tutor o dito Pedro Borgez Pachequo elle dito frey Pedro de Jezus como procurador Geral do dito mosteyro trazia ante o dito Juis dos orfaós o Lecenseado Hieronimo de Burgos Sento e sincoenta e sinco mil reis, preço em que forão vendidas as ditas Cazas a retro aberto, e requeria ao dito Juis lhe ouvesse por desaretradas as dittas Cazas; e lhe manda se fazer escriptura de distrato o que visto pello dito Juis, e se comtar o dinheyro em sua prezensa, e de mim taballião, e estar o dito tutor prezente mandou se fizesse esta escriptura de distrato pella qual e como o dito tutor havião por distratado a escriptura de venda, (...)

Quadro 4: Exemplo de texto do gênero escritura

5. Parecer

PARECER DE JOÃO DE ABREU DE CASTELO BRANCO, GOVERNADOR DO ESTADO DO MARANHÃO E GRÃO-PARÁ

Senhor.
Haja Vista o Procura.^dor da Coroa Lisbo.^a ocidental 11 de Dezem.^bro de 1738.
Este negoci.^o se deve consultar a S. Mg.^de sendo, como He, justificado fundamen.^to par.^a se mover, e fazer esta guerra, Recomendam= dose ao governador a disponHa em modo que fiquem vencedoras as armas de S. Mg.^de e castigados, e temerosozos estes gentioz.
Como na devassa, que com esta Ha de ser prezente a V. Mag.^de Se inquirio sobre Cazos e Nascoens' differentes; e succedidos em partes muy distantes Huma da outra, se me Offerece em primeiro lugar Reprezentar a V. Mag.^de que nao' sõ me parese justo, mas muy precizo, que se fassa guerra as Nascoen's que infestao' o Rio dos Tocantins' na forma que se vẽ da mesma devassa, e que mais notoriamente me consta por outras informaçoens'; sendo muito par.^a temer que allem dos damnos que fas este gentio pella sua VezinHanca, a estas povoacoens', esteja fazendo continuo estrago em todas as pessoas, que inadvirtida E ignorantemente descem das Minas novas de Sao' Feliz por aquelle Rio, de que algumas tem escapado quazi milagrosamente Havendo grande Conjectura de que muitas terao' perecido sem ficar quem dê a noticia; Ao que parese se deve acudir com a mayor brevidade e promptidao' possivel, E particularmente sendo Certo que os Homens' que descem por aquelle Rio abaixo, nao' fazem agravo ou offenssa alguma aquellas Nascoens' de gentio.
Pello que toca as Hostilidades que o Gentio da Nascao' Mura tem Cometido no Rio da Madeira, Como da devassa Conste que o dito gentio tem (...)tado Mortes, E que para esta barbaridade se lHe nao' deu Cauza; E que
Respondase ao Governad.^or que não esta em termos de se Reputar justas e necessari.^as estas guerras e quanto aos Indios do Caminho par.^a os Tocantis se deve ter Cuidad.^o em não adiantar os pouoacoins por aquella par.^te par.^a melhor se observar a prohibicão daquelle Caminho em que se Reputão os maiores inconvenientes / na forma das Repetidos ordens / Lixbo.^a oCidental 28 de Janeir.^o de 1739 por este Respeito estã impedido o UZo, e Comercio do mesmo Rio; E allem disto me conste que estao' atemorizados os Indios das Missoens', E os Missionarios que estao' situados no mesmo Rio da Madeira, me parece que serâ justo, e conveniente ao servico de V. Mag.^de que depois de Execcutada a guerra com os Tocantins' se proceda a fazella no Rio da Madeira. V. Mag.^de mandarâ o que. for servido.
Bellem do Parâ treze de Outubro de mil SetteCentos trinta e oito.
João de Abreu de Castelo Branco

Quadro 5: Exemplo de texto do gênero parecer

6. Registro

N.^o 504 Data e sesmaria de D. Ponciana de Souza Barbalho e Manoel Lopes de Aguiar, de tres leguas de terra no rio Choró, concedida pelo Capitão-mor João de Teve Barretto e Menezes, em 22 de abril de 1746, ás folhas 13 a 13v. do Livro 13 das sesmarias Registo da carta de Datta e Sismaria de Donna Ponciana de Souza Barbalho e Manoel Lopes de Aguiar em o Rio do chorô nas Testadas de Domingos Lopes delgado pello Rio aSima ditto em 22 de Abril de 1746 João de Teyve Barretto e Menezes Fidalgo cavallejro da caza de S. Magestade capitam Major eGovernador da Capitania do Ceará grande pello ditto Snor q Deos goarde etc Faço saber aos q esta minha carta de Datta esismaria Virem q amim me emviou adizer por sua petição Dona Ponciana de Souza Barbalho e Manoel Lopes de Aguiar cujo theor he oSeguinte Senhor capitão major eGovernador Dizem Dona Ponciana de Souza Barbalho e Manoel Lopes de Aguiar moradores nesta Capitania q elles Suplicantes tem seos gados Vacuns ecavallares enão tem terras bastantes para Os puder aComodar eporq Se achão terras deVoLutas enão pedidas pello Rio do chorô nas testadas de Domingos Lopes delgado pello ditto Rio aSima com tres Legoas decomprido ehua deLargo fazendo piam no posso do Serrotte em cujo citio Sepodem os Suplicante aCommodar ecriar seus gados Vacuns eCavallares emq augmenta os Dizimo Reaes portanto. Pedem aVS. lhe conceda em nome de S. Magestade por Datta esismana tres Legoas deterra decumprido ehua de Largo no citio q apontão edecLarão na sua petissão para Sy eseus Erdeyros Accendentes edessendentes noq recebera mercê" Informe os officiaes da camera da Villa do Aquiraz Villa da Fortalleza Vinte seis de Março, de mil Sette centos equarenia eseis" Menezes Snor capitam mayor eGovernador. Não nos consta enem das Sobre dittas terras emq aSuplicante em sua petição fas menção Sabemos se tem ou não empedimento isto he oq aVS. pudemos informar q mandara oq for servido escripta era camera dopr.^o de Abril demil Sette centos quarenta eseis annos nesta Villa de S. Jozê da Ribamar do Aquiras" Manoel Ribeiro Bessa" Francisco Pereira Facanha" Antonio de Barros Martins" Manoel da Costa doValle" Vista aimformação passe cartta de Datta aos Suplicantes na forma do Estillo eOrdem de S. Magestade" Menezes" Hej p bem de conceder" como pella Prezente ofaço" em nome de S. Magestade as terras q os Suplicantes pedem (...)

Quadro 6: Exemplo de texto do gênero registro

7. Sermão

SERMÃO DA Dominga Vigesima-Segunda P0ST PENTEC0STEN Cujus est imago hæc, et superscriptio ? Dicunt ei : Cæsaris.
I Não ha terra mais difficultosa de governar que a patria, nem ha mando mais mal soffrido, nem mais mal obedecido, que o dos iguaes. Vivendo os Hebreus governados por Deus, o qual no Propiciatorio respondia a todas suas consultas, e ordenava em voz clara o que se havia de fazer, foram elles tão mal aconselhados, que quizeram ser governados por homens, como as outras nações ; e sendo tão soberbos, que desprezavam a todas em tudo o mais, n'este ponto, que era a ara maior prerogativa, pediram ser similhantes a ellas : Constitue nobis regem, sicut et universæ habent Nationes. Os primeiros governadores, pois, que Deus lhes concedeu com poder e soberania real, foram Saúl e David: Saúl que andava buscando as jumentas que se perderam a seu pai, e David que andava guardando as ovelhas do seu. Não fez differença das qualidades, porque todos eram filhos de Abrahão; nem a fez tambem dos officios, porque todos n'aquelle tempo viviam de suas lavoiras, e dos seus pastos.
Só teve attenção ás pessoas e aos talentos ; porque assim Saúl como David debaixo do seu saial eram homens de tão grandes espiritos, como logo mostraram as suas obras. Mas quaes foram os applausos com que foi recebida n'aquella republica, depois de tão apertadas instancias, a eleição d'estes dois governos ? A terra era a patria, e os eleitos eram iguaes (como dizia) e não bastou que fôsse um Saúl e outro David, para serem bem aceitos. Alegraram-se os parentes, murmuraram os estranhos, o os demais (que eram quasi todos) ficaram descontentes. Não digo o que disseram, porque as coisas não eram para dizer, nem são para ouvir; só digo que estamos no mesmo caso. Temos repartido este nosso Estado em dois governos iguaes, e debaixo de suas cabeças, ambas naturaes da mesma terra, sem ser a da Promissão; e assim da parte das cabeças, como dos membros, assim da parte dos novos governadores, como dos subditos, se podem recear, como já se temem, não pequenos inconvenientes. O recurso está longe, o remedio não póde chegar senão tarde ; entretanto só vos peço que tomeis o melhor conselho. A obrigação dos prègadores, a quem a Escriptura chama Anjos da paz, é serem ministros da união e concordia ; e porque esta devemos desejar todos, como bons christãos, como bons republicos, e como bons vassallos ; para eu satisfazer á minha obrigação, não me occorre outro meio mais efficaz, que declarar a uns e a outros as suas. O meu intento será este, o Evangelho a guia, a Intercessora para a graça a Virgem Senhora nossa. Peçamol-a com aquella attenção que requer tão importante materia. Ave Maria.

II Perguntando Christo Senhor nosso, como Mestre da Lei, se era licito aos Hebreus pagar tributo ao Cesar imperador dos Romanos, respondeu que lhe mostrassem primeiro a moeda do tributo : Ostendite mihi numisma

Quadro 7: exemplo de texto do gênero sermão

8. Termo

ATAS DA CÂMARA Termo de Uistoria Aos vinte e quatro do mes de Majo de mil e Seis centos | e seSsenta e sette annos foraõ os officiais da Camera abai | xo aSsinados Iuis do Pouo, E Misteres por vertude da petiçaõ E despacho atras aAgoa dos meninos; e sendo prezen | tes os Moradores do ditto distrito contheúdos na ditta pe | ticaõ, E o supplicado Ioaõ Martins Frances que em sua | peSsoa o ouue per citado pera a ditta vistoria, a qual se fez | n[a] forma seguinte "viraõ o ditto Iuis e Mais officiais | da Camera a fonte E agoa de que se trata na petiçaõ, E acha | raõ auela diuertido o ditto Ioaõ Martins Frances do caminho E baixa per onde a ditta agoa corria antiguamente | E cahia qua fora na Rua junto do mar que era entre | as ca[s]as delle supplicado, E as que forão de francisco Pereira Carnejro, E aleuou e emcaminhou per outro Caminho | donde uem a cahir junto da porta de seu quintal que | fora na Rua, E no ditto quintal bem junto da Correnteza | da agoa tem o ditto Ioaõ Martins hu Curral donde de Reco | lhe gado seu que pasta e tras no ditto quintal, o qual faz | damno ao rego per onde corre a ditta agoa em rezão a in | tupir e sujala, perder esta agoa, em rezão de a intupir e sujala, perder esta agoa donde bebem os Moradores | daquele districto, e seruir como serue pera as auguadas d[o]s Nauios, e ser muito importante pera hua cousa e outra | e sobretudo bem comum pera o pouo; O que uisto pelos dittos | officiais da Camera mandaraõ abrir quinse palmos de | Estrada de huã parte e da outra; a Saber sinco palmos pera se | tomarem E recolherem todos os olhos de agoa que estam | E nascem na ditta terra, E sinco palmos de cada parte em | estrada [p]era se poder andar E alimpar athe onde estiue | rem os [d]ittos olhos dagoa qu[e] he junto a agoa bruca, e que | a ditta Estrada correSse direi[t]a donde caje, pera o pouo se aprouei | tar della E os Homens do Mar em fora pera suas augua | das, E pera todos os Mais que quizerem uzar della por | ser liure e comua a todos, E porquanto s[e] achou que o ditto | Ioaõ Martins, tras g[a]do na ditta terra e tem Curral ne | lla, E cõ o gado sujaua E intupia a Seruentia da agoa | Mandar[õl o nottificaSse Eu escriuaõ que tiraSse o ga | [do] E Curral da ditta terra com penna de seis mil reiz, a que | E[u] Escriuaõ satisfis e notifiquei ao ditto Ioaõ Martinz | [em] sua peSsoa tiraSse logo o gad[o] da ditta terra E o ditto Cu | rral com p[enna] de seis mil reis de que tudo min[ha] | fe E outro sy [Manda]raõ que os Moradores daquelle distrito da | goa dos Meninos [t]iueSsem sempre a estrada limpa de huã | parte E da outra, E o rrego por onde corre agoa da Mesma | manejra, e naõ o fazendo aSim procederiaõ contra elleS | E que o ditto Ioaõ Martins, agora, nem em nenhu tempo | [lh]e impida nem poSsa impedir a ditta Seruentia (...)

Quadro 8: Exemplo de texto do gênero termo

9. Notícia

Folio 13r [III] NOTICIA DE VARIOS AROMAS, Q' SE CONHECEM NO BRAZIL.

Ambar - Hé húa masa, q' se acha p^las praias do Mar de cores diferentes; br^co chamado Ambargri, pardo chamado Mexoeiro, e preto q' hé o infimo, todos cheios de fragrancia. Os Indios o vão pescar no fundo do Mar, e dizem q' são húas arvores nascidas no mesmo fundo do Mar com os troncos curtos e grosos esgalhados, a q' os mesmos brasos servem de folhas, brotão estas de si húa resina, q' despegada, e sahida nas praias hé o Ambargri q' os peixes comem, e q^to mais corrupto, mais negro. Acha-se tãobem pelas praias das Terras de Paria e Panamá, Golfo Mexicano, Costas de Florida, e Virginia.
Balsamo - Hé um arvoredo m^to alto e frondozo, chamado Caboreúba; engrosa o tronco athé 3 palmos, de 2 Castas, hum vermelho e outro pardo, ambos bons de lavrar, e m^to duraveis: tem ambos as folhas do comprim^to de 1 dedo, e largura de 2, a casca do páo toda reigada com seos intercascos. Desta lansa o precioso licôr nos mezes de Junho, Julho, Agosto e Setembro, com tanta abund^ca, q' ensopa a terra; entrando as aguas não estila mais: o licôr hé groso, como mel, e se apanha com húa colher; e não querendo q'q^r conservar a arvore, a corta, e mete-se húa ponta em húa fogueira, e a outra em húa vazilha p^r apanhar o balsamo, q' escorre em bica.
Cupaúba - Hé um tronco q' engrosa mais do q' o balsamo; há de 3 especies; hum chamado oleo pardo, p^r ser a madeira desta côr; os dois tem a madeira vermelhosa m^to duravel; hum de casca liza, e outro sarabulhenta, e intercascada; as folhas como de Limoeiro: Tem o oleo

Folio 14r [Está encadernado errado; deveria ser o Folio 13v]
no centro do madeiro, e p^a se-colher hade ser nos mezes de Junho athé Setembro, p^a o q' se dá hum furo no pé da arvore, q' chegue ao meio com trado ou machado, p^lo q' brota o oleo, q' das arvores velhas hé melhor, e mais abund^te. (...)

Quadro 9: Exemplo de texto do gênero notícia

4.3 Tabela de traços contemporâneos

Outro fator importante na pesquisa era definir como iniciar a descrição linguística dos textos. A melhor maneira seria ler os textos de cada gênero e anotar, simplesmente observar as palavras mais frequentes, ou observar que as palavras e as frases de um sermão são diferentes de um assento? Não obstante a importância de ler alguns textos e observar a sua formação, semelhante à proposta desta pesquisa, Biber e Finegan (1993) descreveram a variação diacrônica de três gêneros da língua inglesa, do século XVI até o presente e, posteriormente, Biber (1998) publicou os resultados de uma descrição ampliando para sete gêneros. No entanto, a metodologia usada para a identificação das dimensões diacrônicas de variação não foi explicitada. Assim, Berber Sardinha (2004) sugere que estudos de descrição diacrônica iniciam-se por meio de características sincrônicas, em que os textos históricos se encaixam, ou seja, em vez de iniciar com características compartilhadas de cada texto e partir para o agrupamento dessas características, inicia-se com a comparação dos textos históricos, com as características preexistentes relativas à descrição do português contemporâneo. Com base nisso,

aplicou-se a tabela de traços linguísticos de Aires (2005)[20] nesta pesquisa e, posteriormente, foram realizadas as devidas adaptações.

Aires (2005) escreveu sua tese, intitulada *Uso de marcadores estilísticos para a busca na Web em Português*, com o objetivo inicial de investigar a utilização do PLN na Recuperação da Informação (RI) de textos em português provenientes da web. Por um lado, aplicar técnicas de PLN do português na RI, por outro, lançar alguma luz sobre as características, que se supõem diferentes, da web brasileira e dos usuários brasileiros e/ou em português. Foi a primeira tese que partiu dos problemas dos usuários em português em vez de aplicar técnicas já desenvolvidas para o inglês ou para a web em geral. Após alguns estudos preliminares, a autora optou por implementar um metabuscador e estudar a categorização das respostas em esquemas de classificação que fossem compreensíveis e úteis aos usuários. Assim, o principal objetivo de sua tese foi estudar para o português que categorizações dos textos e páginas web permitem uma forma mais fácil de organização dos resultados de uma busca e como obter automaticamente essa categorização. Para isso, investigou o uso de características estilísticas de um corpus de páginas web classificadas segundo as necessidades que satisfizessem os usuários.

Para a classificação automática dos textos em gêneros, utilizou os gêneros do corpus Lácio-Ref, um corpus aberto e de referência do português contemporâneo do Projeto Lácio-Web, composto por textos em português brasileiro escritos na norma culta. Utilizou o algoritmo da classificação J48[21], disponibilizado na coleção de algoritmos de aprendizado de máquina Weka, que é uma versão do algoritmo C4.5, usado para gerar uma árvore de decisão, que por sua vez pode ser usada para classificação.

Aires (2005), ao elaborar uma *tabela de traços linguísticos*[22], sugere levantar estatísticas baseadas em palavras (itens lexicais diferentes, iniciados por letra maiúscula, tamanho das palavras, etc.), estatísticas baseadas no texto como um todo (número de caracteres, frases, tamanho do texto) e outras estatísticas, como: pronomes, advérbios, verbos, marcadores discursivos, operadores argumentativos e expressões específicas, totalizando 46 sugestões de traços linguísticos.

---

[20] Essa tabela está explicitada no Apêndice A.

[21] Vale ressaltar que nesta pesquisa foi utilizado o mesmo algoritmo devido aos resultados satisfatórios que Aires (2005) obteve. Quem aplicou o algoritmo foi Arnaldo Cândido Jr., mestre em Ciências da Computação pelo ICMC – USP, São Carlos, e pesquisador do NILC.

[22] Ver apêndice A

Com base em sua tabela, é que se iniciou a investigação e levantamento de traços linguísticos recorrentes no português do Brasil dos séculos XVI, XVII e XVIII, passando por adaptações para o contexto histórico. A seguir, são mencionadas peculiaridades de algumas características sugeridas na tabela de Aires (2005).

4.3.1 Estatísticas baseadas em palavras

- estimativa de itens lexicais diferentes, dado pelo número de itens lexicais diferentes dividido pelo número de itens (*type/token ratio*);
- estimativa de itens lexicais diferentes, porém, considerando-se apenas os itens iniciados por letra maiúscula (*capital type token ratio*);
- número de dígitos;
- tamanho médio das palavras em caracteres;
- número de palavras longas (com mais de 6 caracteres);

Nesta pesquisa, essas estatísticas foram levantadas automaticamente com o uso de uma ferramenta computacional denominada extrator de traços, descrita na Seção 4.7.1.

4.3.2 Estatísticas baseadas no texto como um todo

- número de caracteres;
- tamanho médio das frases em caracteres;
- número de frases;

Há alguns casos de pontuação que poderiam gerar problemas, como os exemplificados abaixo:

1. vos inhonorastis me**. .** Mas

   Entre uma frase e outra há o uso inadequado do ponto final.

2. perto uma da**.** Outra

Apesar desses casos problemáticos, o extrator de traços desenvolvido está preparado para ser executado nesse cenário.

Ainda referente às estatísticas baseadas no texto como um todo, a tabela sugere levantar um tamanho médio das frases em palavras e um tamanho do texto em palavras. Tais características também são fornecidas pelo extrator de traços.

4.3.3 Outras estatísticas

Outras estatísticas referem-se às características lexicais e morfossintáticas, além de expressões que podem ser específicas de um gênero ou domínio.

- número de ocorrências das expressões "acho", "acredito que", "parece que" e "tenho impressão (de) que";
- verbo SER (nas formas "é" e "são") (devem ser considerada as grafias como "he", "sao", entre outras);
- pronomes na primeira pessoa (eu, nós);
- pronomes na segunda pessoa;
- pronomes na terceira pessoa (ele e ela, plural e singular) – a análise dos pronomes de primeira, segunda e terceira pessoa deve ser manual, após realizar a busca no *Philologic*[23], para verificar as variações de grafia e depois no *Unitex*[24], considerando as variações das grafias encontradas;
- frequência e tipo de pronomes demonstrativos;
- frequência e tipo de pronomes indefinidos;
- frequência e tipo de pronomes interrogativos;
- frequência e tipo de preposições;
- advérbios (lugar, tempo e terminados em -mente);
- frequência e tipo de interjeições;
- operadores argumentativos;
- marcadores discursivos "agora", "da mesma forma", "de qualquer forma", "de qualquer maneira" e "desse modo".

Além dessas sugestões, dentre outras estatísticas, a tabela original contempla os seguintes traços:

[23] O programa será detalhado na Seção 4.4.
[24] O programa será detalhado na Seção 4.5.

- amplificadores (*amplifiers*) – alguns exemplos são: “absolutamente”, “extremamente” e “completamente”;
- *conjuncts* – alguns exemplos são: “além disso”, “consequentemente”, “assim” e “entretanto”;
- *downtoners* – alguns exemplos são: “com exceção”, “levemente”, “parcialmente” e “praticamente”;
- enfáticos (*emphasizers*) – alguns exemplos são: “definitivamente”, “é óbvio que”, “francamente” e “literalmente”;
- verbos suasivos (persuasivo) como aderir, crer e dar;
- verbos privados como ter e guardar;
- verbos públicos como abolir, promulgar e mencionar;
- contrações;
- conjunções causais, finais, proporcionais, temporais, concessivas, condicionais, conformativas, comparativas e consecutivas.

Todas as características elencadas em “Outras estatísticas” podem ser consideradas as mais complexas para serem identificadas e quantificadas, pois há sempre que levar em consideração a variação de grafia, as abreviaturas e o fato de os processadores de corpus não possuírem desambiguadores lexicais para corpus histórico.

Outro fator importante a ser ressaltado é que alguns traços, como a ocorrência das expressões “acho”, “acredito que”, “parece que” e “tenho impressão (de) que” deverão ser alterados para se adequarem ao contexto de uma descrição de textos antigos, assim como os verbos, pois serão identificadas outras expressões, adequadas a cada gênero e ao contexto histórico.

4.4 Primeiro exercício com o corpus

Em posse da tabela de Aires, iniciou-se o primeiro exercício com o corpus, sem haver ainda determinado quais os gêneros seriam estudados, sem ainda sistematizar como se daria a pesquisa. Os processadores de corpus utilizados foram o *Philologic* e o *Unitex*, já adaptados por pesquisadores do NILC com o objetivo de manipular o corpus do DHPB. Esses programas foram escolhidos por atenderem às necessidades e

peculiaridades deste tipo de pesquisa, de modo a auxiliar na busca, extração e recuperação de textos e fragmentos de textos.

4.4.1 *Philologic*

O *Philologic* é um ambiente Web desenvolvido pelo projeto ARTFL *(American and French Research on the Treasury of the French Language)* na Universidade de Chicago, em colaboração com sua Biblioteca, e prevê um sofisticado recurso de busca e recuperação de informações, bem como gerenciamento (sistema de relatórios) de enciclopédias, dicionários e até mesmo sistemas multimídias (sons, vídeos, imagens). Foi inicialmente desenvolvido para gerenciar textos em francês, no entanto, devido à codificação Unicode, diversos idiomas são tratados. Além de ser acessível via Web, os recursos que o ambiente oferece são: um concordanciador, um gerador de colocações, um contador de frequências e um buscador de dados de cabeçalho, que é capaz de listar os documentos do corpus e formar subcorpora. A Figura 2 apresenta a interface inicial do *Philologic*.

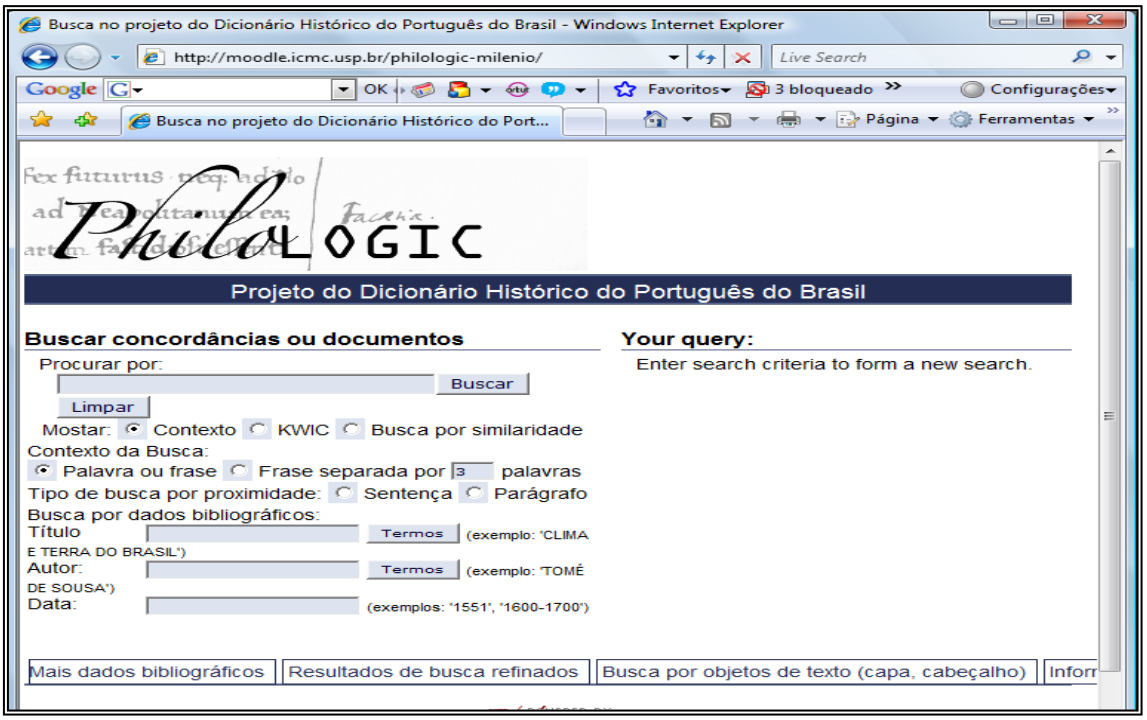

Figura 2: Tela Inicial do Philologic

Esta figura representa a primeira tela do *Philologic*, em que o usuário tem a opção de realizar buscas por palavras ou expressão, escolher qual melhor forma de apresentação dos resultados, ou seja, mostrar o contexto, mostrar na forma KWIC[25], ou ainda realizar uma busca por similaridade, uma vez que não havia ortografia nos séculos XVI, XVII e XVIII. Outra opção importante que o programa oferece é a busca por dados bibliográficos (título, autoria e data).

A pesquisa no *Philologic* ficou dividida em cinco etapas distintas:

1) delimitação de um subcorpus;
2) expansão de palavras (tratamento de expressões regulares);
3) busca por palavras indexadas;
4) extração de texto;
5) resolução de hyperlinks e formatação (em HTML).

É importante ressaltar que, após a definição de um subcorpus, é possível pesquisar por um único termo ou frase. Ao se analisar os índices das palavras em uma base de dados relacional, o *Philologic* extrai blocos de texto, contendo o termo de pesquisa com ligações a grandes blocos de texto. Esses excertos são formatados para serem exibidos em um navegador da web e, por vezes, incluem links para imagens, gravações de som, outros textos, ou mesmo outras bases de dados.

Além disso, a definição de um subcorpus pode ser feita por metadados do autor, título da obra ou ainda data de publicação, porém, nesta pesquisa, o subcorpus foi montado de acordo com o gênero, extraindo-se os textos de cada gênero manualmente. A Figura 3 demonstra o resultado de uma busca por dados bibliográficos, bem como a elaboração de um subcorpus.

[25] *Key Word in Context*

Figura 3: Resultado de busca por dados bibliográficos no Philologic

Após a familiarização com a ferramenta, partiu-se para exploração do corpus, com a finalidade de estabelecer procedimentos, de definir uma metodologia, em busca de uma possível sistematização para o desenvolvimento da pesquisa.

Inicialmente, foi realizada uma busca no programa com um dos traços sugeridos na tabela de traços contemporâneos, pois foi uma forma de analisar quais tipos de textos eram recuperados, bem como a frequência, além de ser realizada a leitura de alguns textos.

O primeiro traço a ser observado no corpus foi o **pronome na primeira pessoa singular (eu).** Como resultado, obtiveram-se 9.222 ocorrências e o programa recuperou uma série de gêneros textuais como cartas, diários, processo, sermão, certidão, sesmaria, instrumento, relação, biografia, roteiro, relatório, diálogo, ânua, arrematação, contrato, *termo*, registro, inventário, juramento, parecer, exortação, representação, edital, poesia, aviso, interrogatório, denúncia, autos, cartografia, inquirição, petição. Em decorrência dessa busca inicial, esse traço não foi considerado relevante no contexto do corpus DHPB, comparando-se os gêneros. Foi esse o procedimento adotado para todos os traços da tabela de Aires (2005).

Após ver quais textos e gêneros foram recuperados em cada traço e realizada a leitura de alguns deles, foi possível fazer algumas observações, tais como:

- textos cuja estrutura seja semelhante a uma ata e alguns pertencentes ao domínio jurídico (auto, assento, contrato, registro, juramento, arrematação, *termo*, processo) apresentam a seguinte forma: “Aos X dias de mês de X do ano de X“, como por exemplo:

1. “Aos vinteecinco dias do mez do mez de Agosto demil de mil seis centos etrinta eoito annos nesta Cidade do Salvador”
2. “Aos vinte e dois dias do mez dagosto de mil quinhentos e sessenta e quatro annos”

- em textos pertencentes ao domínio jurídico (inventário, testamento, petição, etc.), ocorria a seguinte expressão “Ano de nascimento de nosso senhor Jesus Cristo”:
  1. “ano do nacimento de Nosso Senhor Jesu Christo de mil e quinhentos e sesenta e tres anos, aos 27 dias do mes de Janeiro do dito ano”.
  2. “anno do nasim^to de noso Senhor Jezus Christo de mil e seis Sentos e sinCoenta e Coatro annos Aos uinte e noue dias do mes de Septembro do dito anno”.

Observado e registrado isso, foi realizada a busca com o **pronome na primeira pessoa do plural (nós)**, cujo resultado foi 1.927 ocorrências em diversos gêneros textuais: diário de navegação, sermão, cartas redigidas por padres, processos, relato, descrição, diálogo, instrução, exortação, representação, regimento, sesmarias, anais, atas, depoimento, auto, notícias, memórias entre outros. Seguem alguns exemplos da ocorrência do pronome no corpus:

a) “... que nós nom temos necessidade de casa ...”
b) “... foi a mor afronta que nesta viajem nós tínhamos visto.”
c) “... em nós sejam alumiados e n'elles enfatuados e confusos.”
d) “... quando vi encaminhar-se a nós um official militar,...”

Em seguida, foi realizada novamente a leitura de alguns desses textos em busca de hipóteses de traços recorrentes, observaram-se, então, algumas regularidades nos sermões:

- Iniciam-se com a expressão “Prègado na”, como nos exemplos:
  1. “Prègado na Igreja de Nsssa Senhora da Ajuda da Bahia, no anno de 1640”
  2. “Prègado na Igreja da Misericordia da Bahia, no Anno de 1637”
  3. “Prègado na Bahia, á Irmandade dos pretos de um Engenho em dia de S. João Evangelista, no anno de 1633 Maria de qua natus est Jesus, qui vocatur Christus”
- há também algumas expressões em latim presentes nos sermões, tais como: *“Quam mihi”, “Qui vocatur Christus” e “Ad quam nos*”, exemplificadas a seguir:

1. “... e encaminhadas á gloria: Quam mihi etc.”
2. “... penhor da gloria Quam mihi et vobis, etc.”
3. “... o Evangelista S. Matheus o de Christo: Qui vocatur Christus?”
4. “ De qua natus est Jesus; mas acrescentar: Qui vocatur Christus.”
5. “... na outra a gloria: Ad quam nos perducat, etc.”

A fim de averiguar se, de fato, essas expressões poderiam ser consideradas caracterizadores do domínio religioso, ou especificamente dos sermões, foram realizadas buscas no corpus com elas, e pôde-se considerar o resultado satisfatório, uma vez que foram recuperados apenas os sermões. Além disso, foram feitas também buscas com outras expressões em latim como “*Minimus Societatis Jesu*” e “*Vas electionis est mihi iste*”, porém, o resultado não foi semelhante ao anterior.

Embora o gênero não estivesse entre os utilizados na pesquisa, outra observação interessante é que as cartas religiosas, caracterizadas pela comunicação entre os padres, contêm as seguintes expressões recorrentes:

- *“Pax Christi”*
- “Graça de Nosso Senhor”
- “A graça e amor de N. Senhor Jesu Christo seja sempre em nosso favor e ajuda. Amen.”

Exemplos:

1. “A graça e amor de N. Senhor Jesu Christo seja sempre em nosso favor. Amen.”
2. “A graça e amor de N. Senhor Jesu Christo seja sempre em nosso favor e ajuda. Amen.”

Outra sugestão da tabela era quantificar a palavra “se”, porém, ela pode ocorrer no corpus das seguintes maneiras:

- Esedeterminasse
- que sedesocuparern ficaram
- Cazo que sedespeje

A princípio, pensou-se que a palavra *se* não poderia ser quantificada devido ao problema de junção, que é comum em textos históricos. Assim, o extrator de traços foi preparado também para lidar com esse tipo de problema.

Tudo o que foi realizado até este ponto da pesquisa, embora tomasse bastante tempo, foi importante porque permitiu uma rápida leitura e interpretação dos resultados

de busca e, sobretudo, um olhar sobre o corpus. Porém, sua maior contribuição está no fato de permitir pensar e definir uma metodologia de trabalho capaz de facilitar a descoberta de traços linguísticos de cada gênero. Diante disso e apoiando-se na tabela, passou-se a pesquisar especificamente cada gênero e levantar as características de cada subcorpus.

Pelo exposto e com base na tabela de traços contemporâneos, foram realizadas buscas no corpus DHPB, analisaram-se em quais gêneros os traços ocorriam, mas não eram consideradas a frequência e a predominância em cada gênero, o que é fundamental para descrevê-los. Assim, já direcionando a pesquisa para descrição do português para fins de classificação, foi definido que, por meio do *Philologic*, seria elaborado um subcorpus para cada gênero e, a partir daí, aplicar-se-ia a tabela, levantando expressões e unidades lexicais por gênero.

Foi nesse momento da pesquisa que foram determinados quais gêneros seriam utilizados para comparação. O critério de escolha foi selecionar os gêneros cuja quantidade de textos fosse equivalente, como já foi mencionado anteriormente; assim, montou-se um subcorpus com 20 sermões, 20 diários, 20 assentos e assim por diante.

Utilizando-se, pois, o *Philologic*, foram seguidas as seguintes etapas:

1) leitura técnica do texto, que "consiste na abordagem global dos itens informacionais, e tem por objetivo recolher os dados" (SILVEIRA e MOURA, 2007, p. 131);
2) leitura das palavras mais frequentes, para auxiliar na identificação de unidades lexicais, além de permitir fazer observações e comparar com outros gêneros.

A título de exemplo, somente com esse procedimento, foi possível observar a alta frequência e predominância das palavras *dia*, *léguas* e *é* no gênero diário. Contudo, ainda era necessária a comparação com os demais gêneros.

Dando continuidade à sequência de etapas, foram realizadas:

3) a identificação de expressões/unidades lexicais (UL);
4) a busca com a expressão ou ULs para verificar em quais gêneros ocorria, independentemente da frequência. Quando uma expressão é encontrada, realizava-se uma busca no corpus para ver sua frequência, quais os gêneros eram recuperados e se havia algum gênero predominante;
5) a busca no *Philologic* por similaridade com a UL (variação de grafia), para verificar outras formas de grafia e em quais gêneros ocorriam. Era nesse

momento então que se definia a expressão ou UL candidata a traço linguístico de determinado gênero ou domínio;

6) a observação da frequência da expressão ou UL no subcorpus, bem como a comparação da frequência entre os gêneros.

Essa sistematização do processo de identificação de traços linguísticos foi ao encontro da sugestão que Biber, Conrad e Reppen (1998) propuseram a respeito de comparar as características linguísticas de cada gênero, para identificar e descrever os padrões linguísticos de gêneros textuais. A partir do que foi feito, foram levantadas algumas hipóteses de traços, os quais versam sobre unidades lexicais e expressões, para os gêneros diário, registro e sermão, conforme se pode observar na Tabela 2.

Diários (Domínio Técnico administrativo/ oficial)
➢ **Dia** – apesar de ocorrer em diversos gêneros e domínios, a ocorrência desta unidade lexical é predominante em diários. ➢ **Léguas** – apesar de ocorrer em diversos gêneros e domínios, considerando suas variações de grafia, esta unidade lexical é predominante em diários e registro. ➢ **Vossa Magestade, senhoria, mercê, reverência** – optou-se por verificar se ocorriam no subcorpus e constatou-se que não.
Registro (Domínio Técnico administrativo/ oficial)
➢ **Deos goarde** – observou-se a ocorrência dessa grafia e foi confirmada sua predominância neste gênero. ➢ **suplicantes/ oSupplicantes/ supplicantes/ suplicantez** – observou-se que essa unidade lexical é predominante neste gênero, considerando sua variação. ➢ **Deconprido, deComprido –** observou-se a ocorrência dessa grafia e constatou-se que é predominante neste gênero. No entanto, a grafia "de comprido", além de ocorrer em carta de doação, ocorre nos demais gêneros, por isso as grafias "deconprido" e "decomprido" para fins de classificação automática podem ser consideradas traços linguísticos. ➢ **Delargo** – verificou-se que essa grafia é predominante neste gênero. ➢ **Legoa** – observou-se que, assim como nos diários, considerando suas variações de grafia, a unidade lexical *legoa* é predominante neste gênero. ➢ **Sua Magestade que deos goarde/ desua Magestade q. deos goarde** – observou-se que esse pronome de tratamento, considerando essas grafias, é predominante neste gênero. ➢ **Dito riacho/ ditoz riachos/ nodito riacho/dodito riacho –** a princípio, observou-se a ocorrência dessa expressão. Contudo, mesmo considerando suas variações de grafia, constatou-se que não é predominante neste gênero. ➢ **faço saber/ faso a saber/ Faso aSaber/faco asaber –** observou-se que essa expressão, considerando sua variação de grafia, é predominante neste gênero. ➢ **Qual/quais** – observou-se a ocorrência dessa unidade lexical, porém, é preciso comparar sua ocorrência e frequência com outros gêneros para afirmar que se trata de um traço linguístico, além de verificar qual a função morfossintática. ➢ **Capitão-mor/ Capitão mor/ Capitão-mór/ Capitáo-mór/ Capitão major/ capitam maior/**

oCapitam mor/ Capitio-mór – observou-se que as grafias *capitam mor, maior, mayor, major* são predominantes neste gênero. As demais grafias não devem ser consideradas como traço linguístico deste gênero. ➢ **Petição/ petisão** – a princípio, acreditou-se que a ocorrência dessa unidade lexical era um traço linguístico deste gênero. No entanto, verificou-se que ocorre em diversos gêneros na mesma proporção. ➢ **Officiaes da camara/ officiaes daCamara** – verificou-se que essa expressão, considerando as variações de grafia, ocorre em diversos gêneros, porém, sua ocorrência é predominante em registros. ➢ **Guarde Deos** – observou-se que essa expressão, considerando apenas essa grafia, é predominante neste gênero. Contudo, considerando a variação da grafia, no caso **Guarde Deus**, verificou-se que essa grafia não ocorre em registro, mas em outros gêneros. ➢ **Vossa ou Sua Magestade/majestade** – verificou-se que esse pronome de tratamento, considerando sua variação de grafia, é predominante em registros, embora apareça em outros gêneros, como nos sermões. ➢ **Vossa Alteza/altesa** – considerando a variação de grafia do pronome de tratamento, foi observada e verificada sua predominância neste gênero, apesar de ocorrer em outros gêneros, como nos sermões e nos *termos*. ➢ **Christandade –** verificou-se que essa unidade lexical pode ser um traço linguístico neste gênero e concluiu-se que ocorre em diversos gêneros na mesma proporção, por isso não é considerada um traço.
Sermão (Domínio Religioso)
➢ Percebeu-se que algumas expressões em latim estão presentes apenas em sermões, como: "*Quam mihi", "Qui vocatur Christus" e "Ad quam nos".* ➢ **Prègado na** – percebeu-se e contatou-se que todos os sermões são iniciados por essa expressão.

Tabela 2: Hipóteses de traços lingüísticos referente às unidades lexicais e expressões

Embora a quantificação dos traços nos gêneros ficasse por conta do extrator, este primeiro exercício de identificação e relação de algumas hipóteses permitiu fazer as seguintes constatações:

- a variação de grafia pode e deve ser levada em consideração ao definir um traço linguístico, uma vez que, no período em que os textos foram produzidos, não havia uma regra de grafia estabelecida, daí sua variação.
- o que pode determinar uma unidade linguística como traço linguístico não é apenas sua frequência, predominância e ocorrência, mas também sua posição dentro do texto, sua função em um gênero ou domínio, além da imprescindível comparação entre os gêneros.

Na sequência, ainda para levantar os traços linguísticos, bem como refiná-los e confirmar as hipóteses, foi iniciada outra fase, utilizando o processador de corpus *Unitex*. Neste momento, para investigar o maior número possível de textos

disponibilizados para a pesquisa, ampliou-se de 3 para 9 o número de gêneros, de modo que foi preciso encontrar um número comum de textos para cada gênero, o que foi aproximadamente entre 15 e 20 textos para cada um. Os demais gêneros não tinham o mínimo de 15 textos, por isso, não foram utilizados.

4.5 Segundo exercício com o corpus

Com base nas hipóteses de traços, para agregar rigor ao que havia sido levantado, principalmente no que se refere às unidades lexicais, foi gerada uma lista de palavras mais frequentes em cada gênero, utilizando o *Unitex*, de modo a observar quais seriam as ULs caracterizadoras de gêneros.

4.5.1 Unitex

O *Unitex* é um software desenvolvido na Universidade Marne-La-Vallée (França) por Sébastien Paumier (Paumier, 2002), consiste num conjunto de programas que permite o processamento de grandes quantidades de textos, em diversas línguas. Na versão 2.0, o *Unitex* tem módulos para o alemão, coreano, espanhol, finlandês, francês, grego antigo, grego moderno, inglês, italiano, norueguês, polonês, português do Brasil, português europeu, russo, sérvio (tanto com o alfabeto cirílico quanto com o latino) e tailandês. Para o projeto do DHPB, foi feito um dicionário histórico para ser acoplado ao software, de modo que o corpus histórico possa ser mais facilmente processado.

Uma característica que o diferencia de outros programas que trabalham com *corpus* (como, por exemplo, o *WordSmith Tools*) é o fato de o *Unitex* funcionar com base em dicionários eletrônicos de cada uma das línguas que o integram. Para o português do Brasil, o *Unitex* traz um dicionário eletrônico bastante extenso – cerca de 67.500 formas canônicas (ou lemas), 880 mil formas flexionadas e 4.500 formas compostas com hífen – que foi construído por Muniz (2004) a partir do léxico do NILC (ALMEIDA; VALE, 2008).

Além disso, o programa também permite que qualquer usuário crie seus próprios dicionários, integrando novas unidades lexicais ou, ainda, acrescentando novas informações morfológicas, sintáticas e semânticas ao léxico já existente ou ainda gerando novas formas a partir de uma forma canônica (ALMEIDA; VALE, 2008).

Esses dicionários possibilitam ao usuário do programa a realização de buscas pela forma exata, pela forma canônica e também pelas categorias gramaticais. Além disso, o programa permite a combinação desse tipo de busca com a busca por formantes. Essas características fazem com que o *Unitex* possa ser particularmente útil em buscas de construções complexas (ALMEIDA; VALE, 2008).

Na Figura 4, apresenta-se uma tela do *Unitex* contendo a lista de palavras do gênero escritura, ordenadas pela frequência.

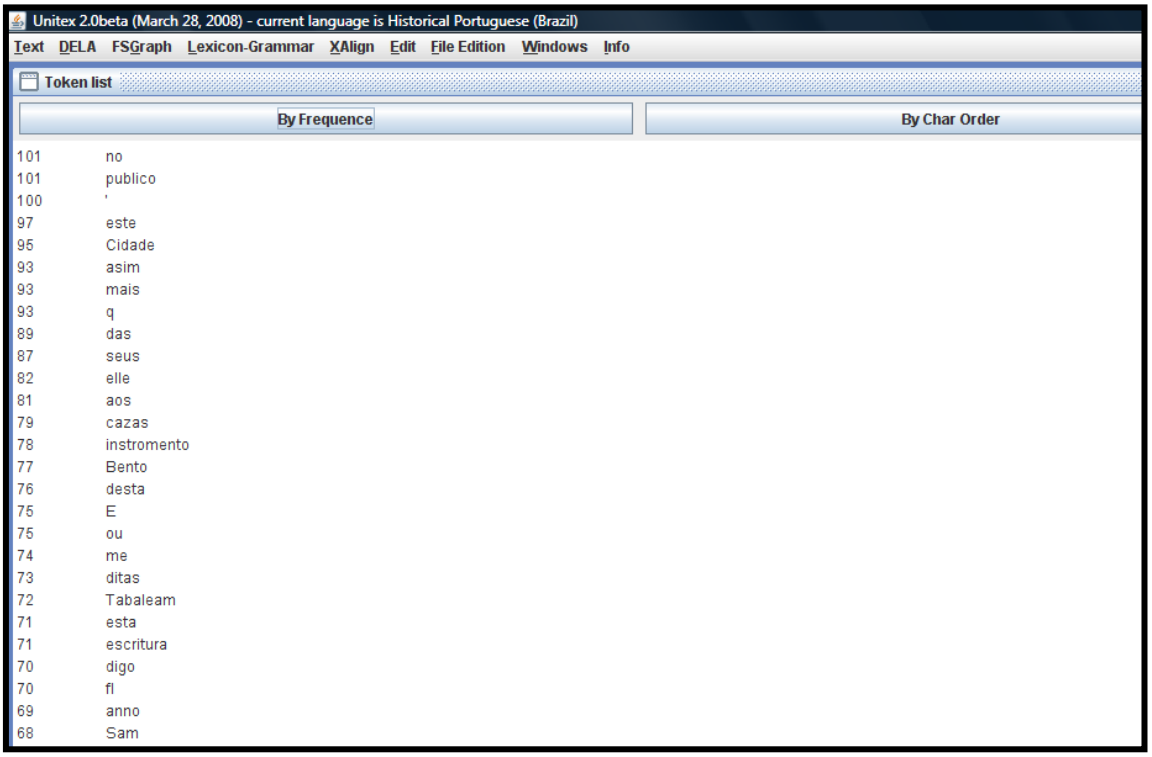

Figura 4: Interface com lista de palavras gerada pelo Unitex

Após gerar a lista, foram selecionadas as palavras mais frequentes, as quais são apresentadas na Tabela 3 a seguir, de acordo com o gênero.

GÊNEROS	UNIDADES LEXICAIS MAIS FREQUENTES
ASSENTO	Câmera, camara (41)[26]; cidade (54); estado (24); povo (24); oficiais (18); dias (18); prezentes (18), atas (15). Verbo: *dizer* na forma "digo" (18), Verbo: *ir* na forma "foi" (12).

[26] Dentro dos parênteses estão os números de ocorrência no subcorpus, considerando suas variações de grafia.

	Observação: as palavras mais frequentes são as que têm ocorrência superior a 15.
AUTO DE PROVIMENTO	Câmera e camara (63 + 134); villa (95); dito (69); termo (60); presente(60); escrivão (60); Doutor (58); ouvidor (39); officiais (35); comarca (29); corregedor (27); juiz (22); ouvidoria (23); certifico (17); corregedor (17); presente (60); presentes (24). Verbo: *Prover* na forma "Proveo" (52). Observação: as palavras mais frequentes são as que têm ocorrência superior a 20.

DIÁRIO	légua/legoas/legoa/leguas (679), dia/dias (567), Onde (519), grande (454), horas/hora (509), terra/terras (478), caminho (400), léguas (341), rio (313), vento (305), tempo (304), mar (156), serra/serras (219), marcha (143), Sudoeste (135), sol (131), água/aguas (191), nordeste (125), soldados/soldado (165), Noroeste (93), cidade (92), leste (84), latitude (71). Verbo: *Ter* nas formas "tem" (361), "tinha" (280), "ter" (188). Verbo: *ser* nas formas "era" (612), "são" (176) Verbo: *Ir* nas formas "foi (298), "fomos" (193) e "fui" (135) Verbo: dar na forma "dar" (130) Verbo: fazer nas formas "fazer" (235), "fez" (178)," faz" (146) Verbo: estar na forma "estava" (172) Observação: as palavras mais freqüentes são as que têm ocorrência superior a 70.
ESCRITURA	Adjetivo dito (318), dita (183), ditos (135), ditas (73), cidade (95), publico (101), instromento (81) (publico instromento), tabeliam (72), escritura (71), testemunha (48), dias (44), prezentes (38) Verbo: assinar na forma "asignaram"(30) Verbo: ir na forma "foi" (65) Verbo: dizer (digo 95) Observação: as palavras mais frequentes são as que têm ocorrência superior a 30.
NOTÍCIAS	Tamanho (224), rio/Rio (193+186), espécies (147), folhas (135), terra (132), carne (102), gentio(100), dia (89), arvore (82), onde (84), frutas (72), tempo (71), cor (66). Verbo: ser nas formas "são" (254) e "é" (679) Verbo: haver na forma "há" (288) Verbo: fazer na forma "fazem (90) Verbo: ter na forma "tem" (250) Observação: as palavras mais freqüentes são as que têm ocorrência superior a 60.
PARECER	Rio (60), guerra (54), anno (33), índios/Jndios (32+31), Brazil (29), testemunh/ testemunha/ testemunhas (74), asucar (26), tabaco (21) gentio (21), devassa/devaça (33), guanás (21), Canoas (20), devassa/devaça (33) Verbo: ter na forma "tem" (56) Verbo: depor na forma "depõem" (23) Observação: as palavras mais freqüentes são as que têm ocorrência superior a 20.
REGISTRO	Terras/terra (68+37), Villa (67), nome (64), goarde (58), magestade/Magestade (116), sismaria(56), senhor (46), capitania (43), data (42), escrivam (escrivão) (40), legoas/ leguas (56), capitão (35), petisam/petição (52), suplicante (47) Verbo: pedir nas formas "pedem" e " pede" (64) Observação: as palavras mais frequentes são as têm ocorrência superior a 30.
SERMÃO	Deus (988), Christo (632), terra (495), céo (345), homens/homem (503), Santo (270), Senhor (258), dia (255), verdade (207), alma (194), porque (826)misericórdia (189), razão (169), amor (159), pão (155), morte (149), palavras (148), rei (148), nome (147), filho (136), sacramento (131), corpo (128), graça (122) Verbo: dizer nas forma disse (402) Verbo: haver na forma "havia" (192) Verbo: fazer nas formas "fez" (186), "fazer" (164), "faz" (115) Verbo: poder nas formas "póde" (214) e "podia" (106) Verbo: ter nas formas "tinha"(170) e "tem" (212) Verbo: ser na forma "eram" (165) Verbo: ir nas formas "foram" (130) e " foi"(104) Verbo: dar nas formas "dar"(138) e "deu"(107) Observação: as palavras mais freqüentes são as que têm ocorrência superior a 120.
TERMO	camera (106), atas (32), ditos (55), reis (53), cidade (42) capitam (38) Verbo: fazer na forma "fazer" (38) Verbo: mandar na forma " mandarão" (18) Observação: as palavras mais freqüentes são as que têm ocorrência superior a 35.

Tabela 3: Unidades lexicais ocorridas em cada gênero

Com base nessa lista, elaborou-se uma lista de expressões que podem caracterizar um gênero. Elas são apresentadas a seguir:

- "Oficiais da camara": ocorre nos gêneros assento e auto de provimento.
- "Atas da camara": ocorre nos gêneros registro e *termo*.
- "Doutor ouvidor geral e corregedor/ ouvidor geral/corregedor/

ouvidor/doutor corregedor": ocorrem nos autos de provimento.

- "Termo e certifico" (certifico/termo): ocorre nos autos de provimento.
- "Deos goarde" e "capitão-mor": ocorrem nos registros.
- "público instrumento": ocorre nas escrituras.

Ainda assim, com base nas leituras realizadas de cada gênero, percebeu-se a ocorrência de determinadas expressões, as quais não eram possíveis recuperar por meio da lista de frequência, pois apareciam apenas uma vez no texto, seja no início ou final, configurando-se uma característica do gênero ou do domínio. Isso foi constatado durante o processo de leitura técnica. As expressões são apresentadas abaixo, por meio de exemplos extraídos dos subcorpora.

Expressão	Gêneros e frequência	Exemplos
Faço saber	Registro (42)	"... S. magestade que Deos goarde etc Faço saber aos que esta minha carta de Data esismaria Virem..."
Pregado	Sermão (18)	"Sermão de Santo Antonio prègado na cidade de S. Luiz do Maranhão" "... o tormento mortal de estar pregado e suspenso, derramando todo o sangue das veias até lhe faltar a vida"
Ano de nascimento	Escritura (27) Auto de provimento (13) Registro (10)	"que no anno do nascimento de nosso Senhor Jezus christo de mil e seis sentos e quarenta e nove annos aos trinta e hum dia do mes de Julho do dito anno nesta Cidade do Salvador Bahya" "Anno do Nascimanto de Nosso Senhor Jesus Christo de mil sete centos setenta e nove annos aos treze dias do mes de Fevereyro do dito anno
o escrevi/ assinei	Assento (12) Auto de provimento (26) Escritura (23)	"Eu Pascoal Teixeira Tabeliam o Escrevy por estar doente o Escrivam" "Provimento em que elles todos asinarão e eu Igancio Pereira de Azevedo escrivão que o escrevi." "...foi acabada e asignada em caza de mim Tabaleam sobredito o escreui Niculao Antunes"
Mihi[27] (47) non est (27) Domine/domini (36)	Sermão (47)	Unde hoc mihi non est rursus assumptus Domine, memento mei Nos autem in nomine Domini invocabimus

Tabela 4: Expressões identificadas em cada gênero

Com base naquela lista de palavras gerada pelo *Unitex*, dada a ocorrência de algumas palavras relacionadas ou ao meio ambiente, ao espaço ou território, foram criadas categorias mais amplas para contemplá-las, ao invés de quantificá-las isoladamente. É o caso de "cidade", "estado", "vila" e "comarca", que ocorrem nos gêneros assento, auto de provimento, diário, escritura e registro. A partir disso, criou-se a categoria para unidade lexical territorial (ULTr).

[27] As expressões correspondentes ao gênero Sermão foram alteradas após gerar o arquivo ARFF. Antes as expressões eram *Quam mihi", "Qui vocatur Christus" e "Ad quam nos".*

O mesmo ocorreu com as palavras mais frequentes no gênero sermão, consequentemente criou-se a categoria unidade lexical sacra (ULSa) “Deus”, apenas neste formato, como ocorrem nos sermões as palavras “santo”, “misericórdia”, “sacramento”, “graça”, “alma”, “almas”, “corpo” e “fé”.

Com base na lista de palavras do gênero notícias, criou-se uma categoria que abarcasse as características que descrevem o meio ambiente, como: “tamanho”, “folhas”, “árvores”, “rio”, “riacho”, “cor”, “espécies”, “vento”, “mar”, “serra”, “serras”, “águas”. Outras categorias foram formadas, como itens léxicos referentes a pessoas (povo, homem, índio, etc.) e itens referentes a pontos cardeais e colaterais (leste, oeste, norte, etc).

Referente aos verbos, a tabela de traços contemporâneos sugeria o levantamento dos verbos suasivos (*aderir, crer* e *dar*), privados (*ter* e *guardar*) e públicos (*abolir*, *promulgar* e *mencionar*), além do verbo *ser* nas formas *é* e *são*. O último foi mantido porque esteve entre os mais frequentes, mas os demais foram substituídos pelos verbos mais frequentes em cada gênero, pois se julgou mais coerente com a proposta da pesquisa e com as características do corpus. Assim, os verbos que integraram a tabela de traços foram: *dizer*, *fazer*, *haver*, *ir*, *pedir*, *poder*, *prover* e *ter*, nas suas formas mais frequentes no subcorpora.

Vale ressaltar que, da lista de palavras, foi selecionada apenas uma palavra de cada gênero, de modo a ser quantificada isoladamente, como no caso das palavras *dia, testemunha, devassa, suplicante, juiz*.

Poder-se-ia dizer que parte da tabela de traços linguísticos do português histórico estaria completa no que se referente a expressões e unidades lexicais. Contudo, junto a essa fase, era gerado o arquivo ARFF, que permitia refinar os traços, etapa essa que será descrita na Seção 4.7. Antes, serão apresentadas mais adaptações (de cunho morfossintático) feitas na tabela de traços contemporâneos.

4.6 Definição dos traços morfossintáticos

A tabela de traços contemporâneos sugere levantar os pronomes pessoais, interrogativos, demonstrativos, indefinidos, preposições, advérbios, interjeições, operadores argumentativos, marcadores discursivos, contrações, conjunções, além de

verbos, os quais já foram alterados, conforme apresentado na seção anterior.

Uma vez que o corpus não foi etiquetado, como seria possível quantificar cada uma das classes gramaticais? Quantos advérbios de tempo? Quantos pronomes interrogativos? Para fazer este levantamento, optou-se por recorrer à literatura, de modo a elencar as palavras que compõem as classes gramaticais. Assim, as características das categorias foram baseadas nas gramáticas de Rocha Lima (2000), Evanildo Bechara (2004) e Luiz Antonio Sacconi (1999).

A exemplo do que se pensou fazer, em princípio, pretendia-se escolher apenas algumas palavras capazes de representar cada advérbio (tempo, lugar, intensidade), baseando-se na frequência de cada uma. Contudo, isso seria desnecessário porque o próprio extrator quantifica cada classe, cada palavra. Dessa forma, com base nas gramáticas mencionadas, procurou-se incluir o máximo de palavras para cada classe.

Cabe esclarecer que o traço interjeição foi retirado da tabela, uma vez que os gêneros possuem características mais formais, ausência de frases exclamativas, não expressão de emoção ou sentimento repentino.

Além disso, é importante ressaltar que para cada palavra das classes gramaticais, era realizada a busca por variação de grafia no *Philologic*. A seguir, são apresentados os traços linguísticos morfossintáticos e algumas peculiaridades.

A. Pronome pessoal oblíquo

Todos os pronomes pessoais oblíquos confirmaram o que estava proposto na tabela, com exceção do *se, o, a* por assumirem diversas funções gramaticais: *o* e *a* podem figurar como pronome demonstrativo (quando se referem a aquele, aquela, aquilo) ou como artigo. Já o *se* foi quantificado como unidade lexical, como sugerido na tabela de traços contemporâneos. Os pronomes pessoais oblíquos, incluindo a variação de grafia, ficaram da seguinte forma:

- pronome pessoal oblíquo na primeira pessoa: *me, mim, mym nos, nós, comigo, commigo, comiguo, conosco, comnosco, connosco*;
- pronome pessoal oblíquo na segunda pessoa: *te, the, ti, contigo, comtigo, contiguo, vos, vós, convosco*;
- pronome pessoal oblíquo na terceira pessoa: *lhe, lhes, si, consigo.*

B. Pronome pessoal de tratamento

Os pronomes de tratamento são apresentados na tabela 5, relacionando-se o seu emprego ao contexto de uso:

1. Você	Tratamento familiar
2. Senhor, senhora	Tratamento de respeito
3. Vossa Senhoria	Tratamento cerimonioso
4. Vossa Excelência	Tratamento para altas autoridades
5. Vossa Eminência	Tratamento para cardeais
6. Vossa Santidade	Tratamento para o papa
7. Vossa Alteza	Tratamento para príncipes e duques
8. Vossa Magnificência	Tratamentos para reitores de universidades
9. Vossa Majestade	Tratamento para reis
10. Vossa Reverendíssima	Tratamento para sacerdotes

Tabela 5: Pronomes de tratamento e contexto de uso

O pronome *você*, tratamento familiar, com essa forma, não é um traço característico que se encaixe no português histórico, portanto, optou-se por utilizar *Vossa Mercê,* com as variações de grafia (*vosa mercê/merce, vossa mercê/merce*).

Vossa Eminência e *Vossa Reverendíssima* também não confirmam a tabela, porque a ocorrência no corpus DHPB é menor que 10. O pronome *Vossa Magnificência* também não foi considerado, já que não há um contexto de uso possível que se enquadre nos gêneros selecionados, por isso sua frequência no corpus DHPB é zero. Assim, os pronomes selecionados foram: *Senhor, Senhora, Vossa Senhoria, Vossa Excelência, Vossa Alteza, Vossa Santidade, Vossa Majestade, Vossa mercê/mercê.*

C. Pronomes demonstrativos

Os pronomes demonstrativos encontrados foram: *este, estes, deste, destes, d'estes, esta, estas, desta, destas, isto, disto, esse, esses, desse, desses, essa, essas, dessa, dessas, isso, disso, aquele, aqueles, aquelle, aquelles, daquele, daqueles, daquelle, daquelles, aquela, aquelas, aquella, aquellas, daquela, daquella, daquelas, daquellas, aquilo, aquillo, daquillo.*

D. Pronome relativo variável

Os pronomes relativos variáveis considerados foram: *a qual, aqual, os quais, os quaes, osquaes, cujo, cujos, cuja, cujas.*

E. Pronome interrogativo

Os pronomes interrogativos pertinentes foram: *quem, quanto, quando, qual, quais, quaes.*

F. Pronome indefinido

Nesta pesquisa, optou-se por classificar os pronomes indefinidos por referência a pessoa, lugar e coisa:

- referente à pessoa – *alguém, alguem, ninguém, ninguem, outrem, qualquer, quaesquer;*
- referente a lugar – *onde, aonde, donde, adonde*;
- referente a coisa e/ou pessoa – *algo, tudo, nada, todo, todos, toda, todas, vários, várias, certo, certa, pouco, poucos, pouca, poucas, muito, muitos, muyto, muytos, muita, tanto, tantos, tanta, tantas, cada, nenhum, nenhuma*.

G. Preposição

Todas as preposições foram consideradas, com as suas variações de grafia, tais como: *por, per, para, até, athé, athè, té, em, entre, contra, sem, sobre.*

H. Advérbios

Classe gramatical bastante abrangente, optou-se por classificar os advérbios de lugar, tempo e intensidade:

- lugar: *aqui, abaixo, acima, asima, cá, lá, ai, ahi, ali, alli, alem, além, alèm, dalém, dentro, longe, acolá, aquém, aquem, adiante, atrás, atras, atraz algures, alhures, perto, fora, defronte, embaixo, em baixo.*
- tempo: *hoje, oje, amanhã, ontem, hontem, manhã, já, sempre, nunca, ainda, antes, tarde, cedo, outrora, depois, depoiz, agora, logo, jamais, diariamente, anualmente, atualmente, sucessivamente, entrementes, imediatamente, immediatamente.*
- intensidade: *muito, pouco, bastante, mais, menos, bem, mal, muito, muyto, pouco, poco, bastante, bastantes, mais, maiz, menos, menoz, bem, mal, tão, tao, todo, assaz, tanto, quão, meio, meyo.*

I. Adjetivo

Trata-se de uma classe bastante abrangente. Porém, com base na leitura técnica, observou-se a ocorrência da palavra "dito" e optou-se por incluí-la na tabela como adjetivo, ficando da seguinte forma: *dito, ditto, odito, oditto*. Seguem exemplos:

- *(...) fizeram odito imprestimo eque aditta repartiçam sefazia (...)*
- *(...) vindo nodito tempo seobriga apagar (...)*
- *(...) emandaram que aportaria dodito Senhor Governador (...)*

J. Marcadores discursivos

Como se trata de um grupo abrangente, foram selecionados os seguintes itens: *pois, também, tambem, só, apenas, mesmo, mesma, nem, além de, alem de, tão, logo, portanto, então, porém, todavia, embora ainda, agora, consequentemente, assim, como, porque, perque, por que, ainda, depois, ou, portanto, por tanto, entretanto, entre tanto, a medida que, diante.*

Assim, após definir todos os itens acima (denominados traços) que seriam quantificados pelo extrator, já era possível gerar o arquivo ARFF e refinar os traços.

4.7 Do extrator de traços ao arquivo ARFF

A partir das buscas no subcorpus, filtradas pela tabela de traços de Aires (2005), foi possível fazer uma nova versão da tabela de traços adaptada ao contexto histórico. Com essa nova tabela, utilizou-se o extrator de traços e gerou-se o arquivo ARFF.

Para essa tarefa, foi necessário selecionar texto a texto, nomeá-los e classificá-los de acordo com a tipologia de gêneros, separar os que seriam utilizados para treinamento (90 textos) dos que seriam utilizados para realizar os testes (70 textos) com os classificadores. Apresenta-se, na Figura 5, a organização dos textos já classificados em uma pasta para treinamento.

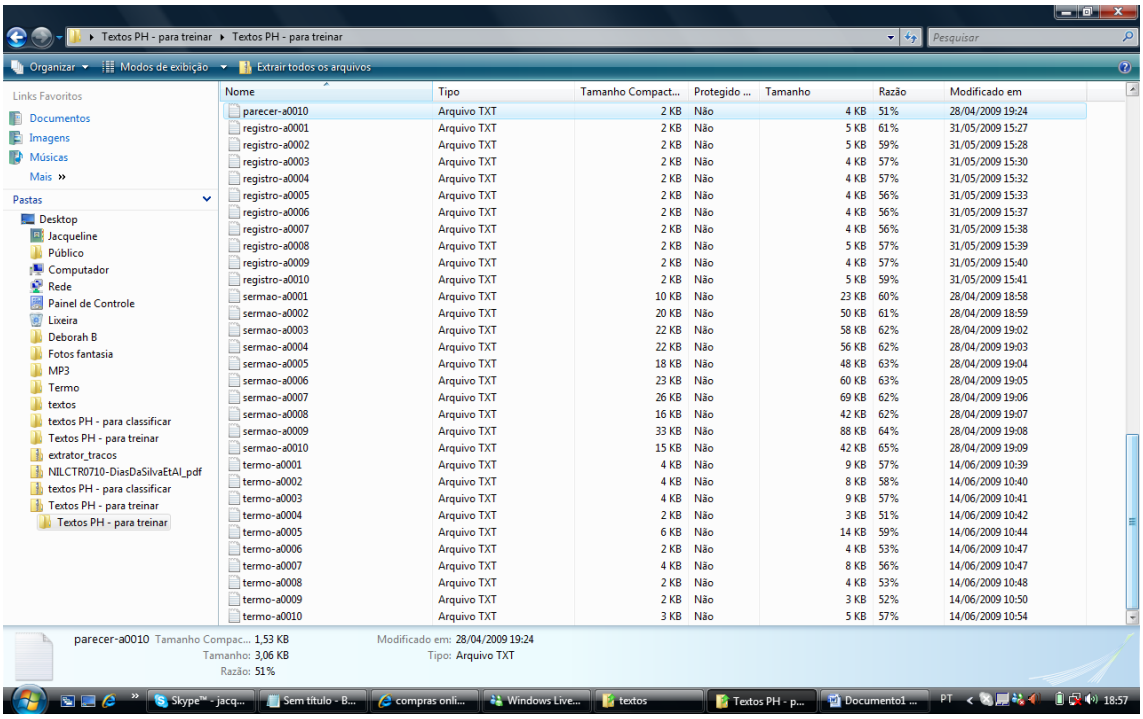

Figura 5: Textos classificados e organizados no arquivo para treinamento

Para utilizar os classificadores disponibilizados pelo Weka, os textos precisariam ser convertidos em arquivo ARFF, requisito do próprio programa de classificação, por isso foi criado o programa denominado *extrator de traços*. Primeiramente, explicar-se-á o extrator.

4.7.1 O Extrator de traços linguísticos

Elaborado por Cândido Junior[28], o objetivo do programa era quantificar os traços de cada texto e convertê-los em arquivo ARFF. Desenvolvido em Perl, linguagem de programação estável e multiplataforma, permite criar programas em vários ambientes operacionais.

A seguir, apresenta-se a Figura 6, a qual esboça o processamento do texto pelo extrator de traços para geração do arquivo ARFF.

[28] Doutorando no ICMC/USP e informata colaborador nesta pesquisa.

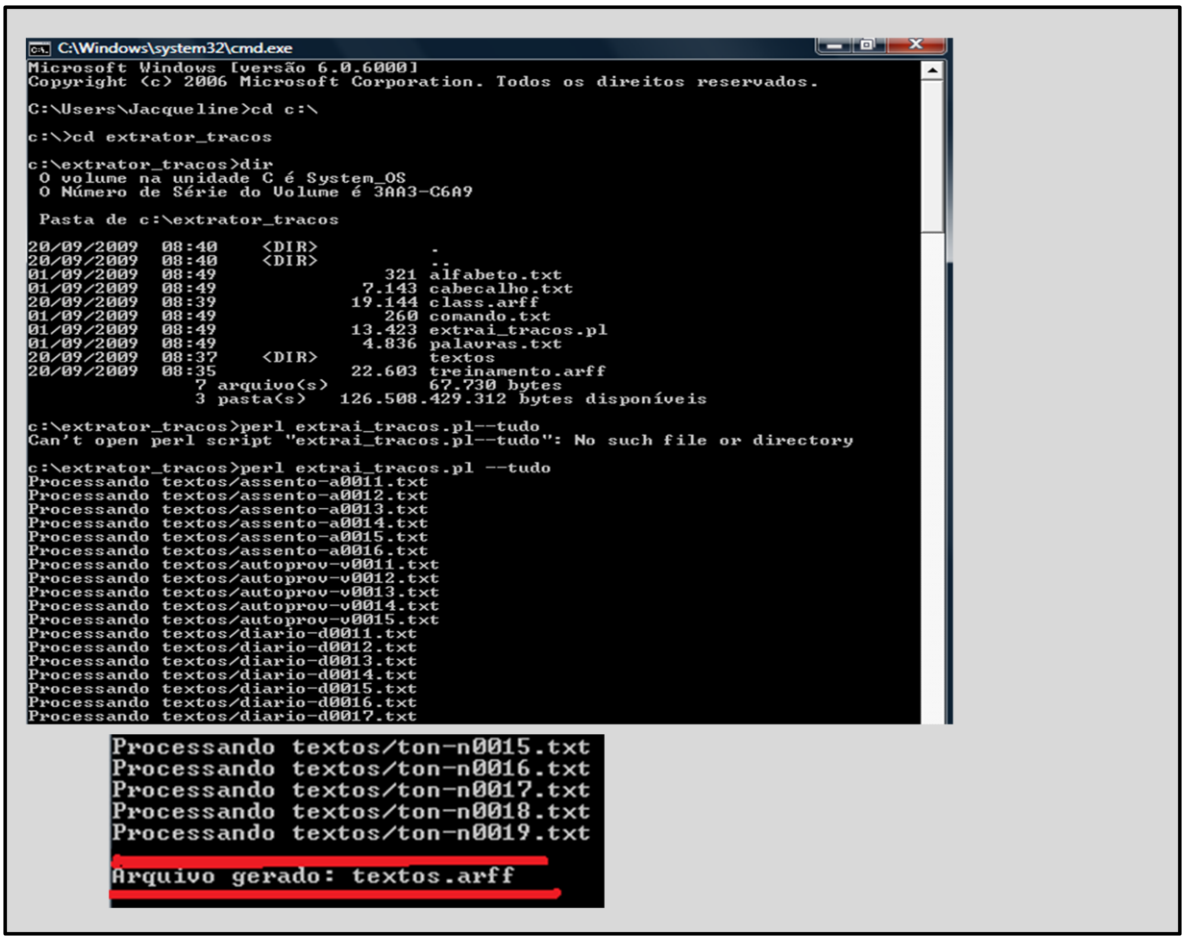

Figura 6: Processamento dos textos pelo extrator de traços

A Figura 6 ilustra o processamento dos textos pelo extrator de traços, o qual refere-se a quantificação dos traços de cada texto dos gêneros estudados, ou seja, em um texto do gênero assento há *n* pronomes pessoais de tratamento, há *n* expressões, *n* verbos e assim por diante, com base na tabela de traços adaptada ao contexto histórico. Feito isso, ele gera os dados que constitui o arquivo ARFF, descrito abaixo.

4.7.2 O Arquivo ARFF

Desenvolvido pelo Projeto de Aprendizado de Máquinas do Departamento de Ciência da Computação da Universidade de Waikato, o arquivo ARFF *(Attribute-Relation File Format)* é um arquivo de texto ASCII (*American Standard Code dor Information Interchange*)[29] que descreve uma lista de situações de partilha de um

[29] Código padrão americano para intercâmbio de informação.

conjunto de atributos de arquivos. Trata-se de um formato próprio do Weka, o qual exige descrever o domínio do atributo, uma vez que este não pode ser obtido automaticamente pelo seu valor. Ele é constituído em 3 partes: a relação, atributos e dados, explicados e exemplificados nas Figuras 7 e 8.

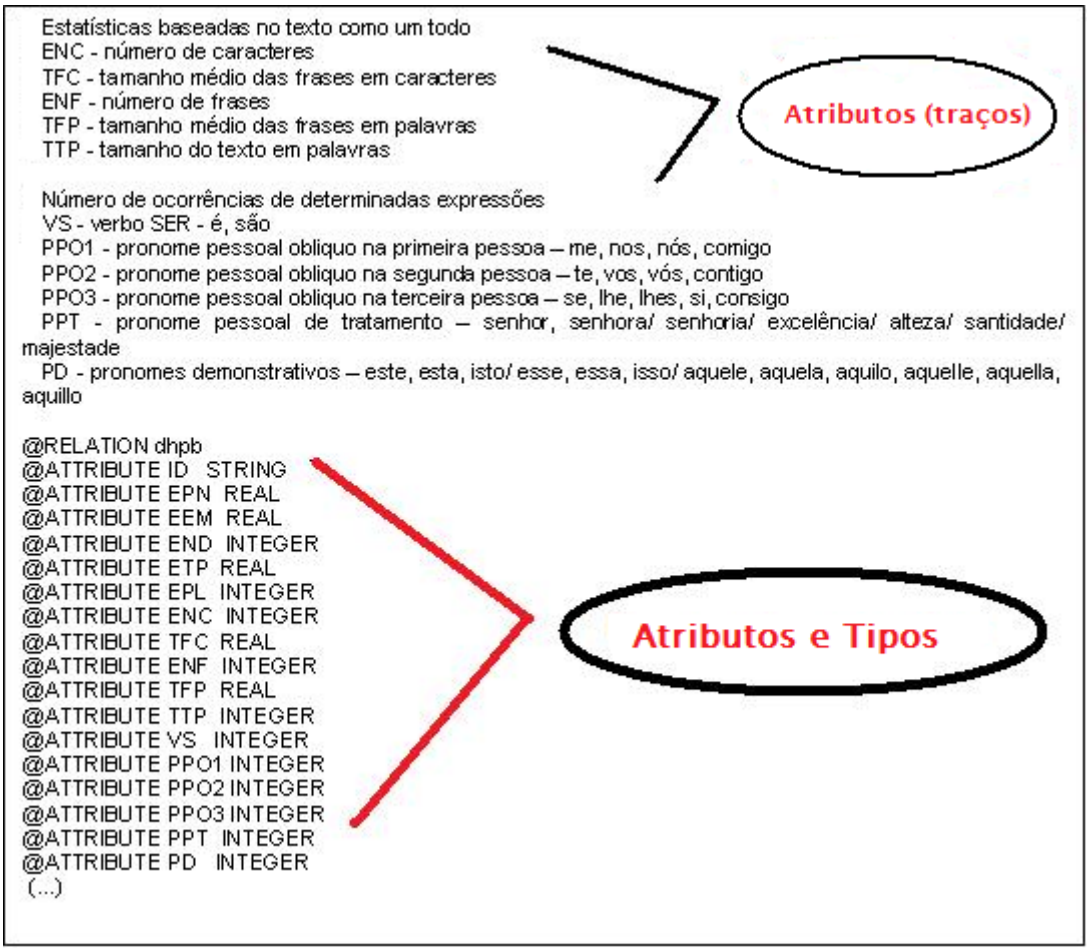

Figura 7: Cabeçalho do arquivo ARFF

A Figura 7 exemplifica o cabeçalho do arquivo ARFF, o qual contém o nome da relação (@RELATION dhpb), que identifica a tarefa estudada, seguida de uma lista de atributos (ID, EPN, EEM, END, ETP, EPL), e seus tipos (real, integral). O exemplo é de um conjunto de dados de textos do corpus DHPB.

Os dados são apresentados na Figura 8.

```
@DATA
a0001,0.54,0.65,27,5.51,274,4526,646.57,7,114.14,799,0,0,0,7,3,10,7,0,2,2,23,1,5,3,0,5,12,0,0,0,0,0,0,0,0,0,0,0,1,0,0,assento
a0002,0.53,0.65,27,5.27,148,2620,436.66,6,79.66,478,0,0,0,1,3,2,1,0,0,0,9,2,1,4,0,1,3,0,0,0,0,0,0,0,0,0,0,0,0,0,assento
a0003,0.45,0.63,58,5.07,476,9070,697.69,13,132.23,1719,1,1,0,14,4,10,2,1,1,2,35,3,2,18,0,3,28,0,0,0,0,0,0,0,0,0,1,0,0,assento
a0004,0.50,0.54,27,5.24,179,3137,522.83,6,96.66,580,0,0,0,2,5,6,4,0,1,2,18,1,0,4,0,0,6,0,0,0,0,0,0,0,0,0,0,0,0,assento
```

Figura 8: Formato dos dados do arquivo ARFF

De acordo com a Figura 8, cada linha representa os valores separados por vírgula correspondentes dos atributos (TFP, TTP e etc). Mais especificamente, representa os dados de um texto do gênero assento, com os traços elencados no cabeçalho, separados por vírgulas, como se estivessem em uma tabela. Para melhor elucidação, observe-se um texto do gênero assento na Tabela 6 a seguir:

	TFP	TTP	VS	PPO1	PPO2	PPO3	PPT	PD
a0001	799	0	0	0	7	3	10	7

LEGENDA
a0001 – código de identificação do texto.
TFP - Tamanho médio das frases em palavras
TTP – tamanho do texto em palavras
VS - Verbo SER
PPO1 - Pronome pessoal oblíquo na primeira pessoa.
PPO2 - Pronome pessoal oblíquo na segunda pessoa.
PPO3 - Pronome pessoal oblíquo na terceira pessoa.
PPT - Pronome pessoal de tratamento.
PD - Pronomes demonstrativos

Tabela 6: Explicação da leitura do arquivo ARFF

Após a geração do arquivo ARFF, era realizada a leitura para verificar se realmente os traços eram quantificados nos textos, quais os problemas ocorriam e para refinar a tabela. Vale ressaltar que este arquivo era gerado ao mesmo tempo em que era gerada a lista de palavras mais frequentes, de maneira que fosse possível a organização e inclusão de mais traços, se fosse o caso. No apêndice D encontra-se um exemplo do arquivo ARFF em Word.

A seguir, será relatado o modo como foi realizada a leitura do arquivo e o refinamento dos traços.

4.7.3 Refinamento dos traços

Em posse do arquivo ARFF, era feito o refinamento dos traços com base na leitura do arquivo. O refinamento consiste em verificar a ocorrência dos traços, em alteração na tabela, incluindo ou excluindo traços, inclusão de mais variações de grafia

e assim por diante. Dessa maneira, também era possível verificar algum erro do próprio extrator.

A verificação iniciou-se pelas expressões porque os resultados numéricos eram baixos: algumas formas deveriam ocorrer, mas não ocorriam, observe-se, por exemplo, o caso da palavra *suplicante*. Após imprimir o arquivo ARFF em documento Word e realizar a leitura, verificou-se que *suplicante*, hipótese de traço caracterizador de registro, não ocorreu em dois textos. Para verificar se realmente não ocorria, os textos foram acessados e verificou-se que havia o problema da grafia, pois a tabela não contemplara inicialmente a variação em decorrência da junção *osuplicante*, uma característica do corpus, e também a ocorrência de outra grafia *supplicante*. Assim, incluíram-se na tabela mais duas grafias possíveis: *osuplicante* e *supplicante*.

Uma unidade lexical observada foi *légua* que, de acordo com as hipóteses levantadas, seria um traço de diário e registro, mas a referida palavra não estava ocorrendo neste último, verificou-se também que o problema era devido à variação de grafia. Para solucionar a questão, acrescentaram-se outras grafias *leguas* e *legoas*.

Nos diários, o arquivo ARFF comprovou a predominância da unidade lexical *dia*. Essa palavra, assim como outras, pode ocorrer em qualquer outro gênero, mas em diferentes proporções. Por exemplo, enquanto nos diários a média de ocorrência de *dia* é de 52 ocorrências por texto, a maior ocorrência em outro gênero, o sermão, é de 12, sendo que a média do tamanho do texto em palavras nos dois gêneros é de aproximadamente 9 mil palavras. Esse é um exemplo bastante convincente de que a palavra *dia* pode ocorrer em qualquer gênero, mas é considerado um traço linguístico de diário devido à frequência.

Outro traço referente à unidade lexical são os pontos cardeais (norte, sul, leste oeste), os quais ocorreram no corpus, mas observou-se a necessidade de acrescentar alguns pontos colaterais e subcolaterais, como: *nordeste, sudeste, noroeste, sudoeste e nor-Nordeste, lés-Nordeste, lés-Sudeste, su-Sudeste, su-Sudoeste, oés-Sudoeste, oés-Noroeste, nor-noroeste,* para melhorar a recuperação dessas unidades lexicais em cada gênero.

De acordo com o primeiro arquivo ARFF, a expressão *atas da câmara* não ocorreu. Sendo assim, cada um dos textos foi analisado manualmente para tentar verificar se não se tratava mesmo de um traço ou a não ocorrência era referente à variação de grafia ou junção. Essa expressão ocorre nos gêneros do domínio jurídico

como nos assentos, em alguns *termos* e autos. Acessando os textos manualmente, verificou-se que a ausência era devido ao uso de letras maiúsculas e minúsculas.

A expressão *"o escrevi"* também não ocorreu, de acordo com o arquivo ARFF. Acessando manualmente cada texto, foi possível comprovar que ela ocorre no final de cada texto nos gêneros *termo,* escritura, assento, auto de provimento e registro. Apesar de encontrar variações de grafia que não contemplavam a tabela, ainda não é compreensível por que não ocorreu essa expressão. Assim, foi necessário comunicar o informata, para que ele identificasse algum problema com o extrator.

A expressão *"ano de nascimento de nosso senhor Jesus Cristo*" também não ocorreu, enquanto deveria ocorrer em alguns gêneros do domínio jurídico, principalmente nos autos. Constatou-se que devido à expressão ser extensa, com oito palavras e cada uma podendo ter variações de grafia, decidiu-se alterar o traço apenas para "*ano de nascimento*" e verificar o resultado seguinte gerado pelo arquivo ARFF. Assim, as variações que acabaram fazendo parte da tabela fora: *"anno de nascimento", "anno de nassimento", "anno do nascimento", "anno do nassimento".*

A expressão *"oficiais da câmara*" também não ocorreu nenhuma vez no corpus. Ao observar os gêneros nos quais deveriam ocorrer, como assento e auto de provimento, foi possível verificar que o problema poderia ser decorrente do uso de letras maiúsculas e minúsculas. Foram acrescentadas também mais algumas variações de grafia, como: *"officiaes daCamara", "officiaes desta câmera", "officiaes da câmera"*.

A expressão *"pregado em*", bem como determinadas expressões em latim, como "*quam mihi, qui vocatur christus, ad quam nos*", de acordo com o arquivo ARFF também não ocorreram nos sermões. Ao analisar os textos, observou-se a necessidade de incluir outras grafias referentes à expressão, como "*prègado, pregado".*

A expressão "*capitão mor*" também não ocorreu, constatou-se a necessidade de incluir mais variações de grafia, inclusive com o hífen, ficando da seguinte forma: *"capitão-mor", "capitão mor", "Capitão-mór", "Capitáo-mór", "Capitão major", "capitam maior", "capitam mor", "capitam major", "capitam Mayor*". O mesmo se deu com a expressão "*faço saber*".

A expressão "*em sua petição atrás escrita e declarada*" também não ocorreu. Assim, pelo fato de a expressão ser extensa e possibilitar muitas variações de grafia, além das que estão na tabela de traços, a expressão foi reduzida para *"escrita e declarada",* considerando as variações de grafia.

Por fim, a expressão "*Deus guarde*", hipótese de traço do gênero registro, não ocorreu. Manualmente, constatou-se a necessidade de verificar a questão do uso de letras maiúsculas e minúsculas, bem como acrescentar as seguintes variações de grafia: "*Deos gde*" e "*Deos g.*"

Além das unidades lexicais e expressões, de acordo com o primeiro arquivo ARFF gerado, outros traços chamaram atenção, por não ocorrerem em alguns gêneros, ou ainda, por terem baixa frequência, como:

1. o verbo ser, nas formas *é* e *são* (VS)
2. Pronome pessoal oblíquo na primeira pessoa (PPO1)
3. Pronome pessoal oblíquo na segunda pessoa (PPO2)
4. Pronome pessoal oblíquo na terceira pessoa (PPO3)

Para verificar se havia algum problema referente à variação de grafia, ou se realmente não ocorrem em determinados gêneros, em posse de uma lista de variantes de grafia, cedida pelo informata e elaborada no contexto do projeto DHPB, foi acrescentada à tabela outras variações de grafia. Seguem exemplos:

- VS – *"são, sam, é".*
- PPO1 – *"me, mim, mym, comigo, commigo, comiguo, conosco, comnosco, connosco"*
- PPO2 – "*te, the, ti, contigo, comtigo, contiguo, vos, vós, convosco"*
- PPO3 – "*lhe, lhes, si, consigo, comsigo, ele, eles, elle, elles, ela, elas, ella, ellas"*

Diante dessa análise preliminar, foi possível refinar os traços, acrescentando mais variações de grafia e mais traços na tabela, tais como:

- variações de grafia por junções, como: "*osuplicante"*;
- inserção de traços:
    - pronome de tratamento: *"vossa mercê"*;
    - unidades lexicais: *"nordeste, sudeste, noroeste, sudoeste e nor-Nordeste, lés-Nordeste, lés-Sudeste, su-Sudeste, su-Sudoeste, oés-Sudoeste, oés-Noroeste, nor-noroeste".*

Após refinar os traços e alterar a tabela, o extrator também precisou ser alterado, de modo que pudesse melhorar a recuperação e quantificação dos traços. Depois disso, gerou-se um segundo arquivo ARFF.

Segundo arquivo ARFF

Feitas as alterações na tabela e no extrator, foi gerado um segundo arquivo ARFF, que também foi analisado para outro refinamento. A expressão *"escrita e declarada"* não ocorreu no corpus, não foi recuperada, por isso foi excluída da tabela de traços.

Foi acrescentado na tabelo o adjetivo *"dito"*, incluindo suas variações de grafia, como *"odito"*. Verificou-se também que as expressões em latim não eram recuperadas. Neste momento da pesquisa, foi gerada uma lista de palavras mais frequentes para cada gênero, que deu subsídios para identificar mais unidades lexicais e expressões. Logo, as expressões em latim foram substituídas por outras, como: *"Mihi", "non est", "domine" "domini".*

Após a geração da lista de palavras mais frequentes obtida pelo *Unitex*, o extrator foi alterado e a tabela contemplou as seguintes unidades lexicais (UL), expressões (E) e verbos (V):

- ULT – *testemunha*
- ULJ – *juiz/juízes/juis*
- ULDe – *Deus*
- ULG – *guerra*
- ULTe – *terra/terras*
- ULI – *índios, indio*
- ULDv – *devassa, devaça*
- ULH – *homem*
- EXDo – *doutor ouvidor geral e corregedor/ ouvidor geral/corregedor/ ouvidor/doutor corregedor/corregedor da comarca*
- EXT – *termo e certifico/certifico/termo*
- EXPI – *público instrumento/instrumento/publico*
- VH – *verbo HAVER nas formas há, havia*
- VP – *verbo PEDIR nas formas pede, pedem*
- VPr – *verbo PROVER na forma proveu, proveo*
- VPo – *verbo PODER nas formas póde, podia*

A partir disso, foi alterada a tabela e gerado o terceiro arquivo ARFF, seguido de análise para refinamento dos traços.

Terceiro arquivo ARFF

No que se refere a unidades lexicais, todas foram recuperadas, mas em alguns casos como: *índio, homem,* criaram-se traços com unidades lexicais referentes a pessoas (ULP), ficando da seguinte forma:

- ULP – *índios, indio, gentio, povo, homem.*

Todas as ULs que obtidas foram elencadas entre as palavras mais frequentes. O mesmo ocorreu com a UL *Deus*, optando-se por criar uma UL referente ao discurso sacro, ficando da seguinte forma:

- ULSa – *deus, santo, misericórdia, sacramento, graça, alma, almas, corpo, fé.*

O mesmo ocorreu com outras ULs, criando-se os seguintes traços:

- ULMA – referente a aspectos descritivo do meio ambiente: *tamanho, grande, pequeno, folhas, arvores, árvores, rio, riacho, cor, espécies (terra está em territorial), vento, mar, serra, serras, águas, aguas.*
- ULTr – referente a aspectos territoriais: *cidade, estado, vila, Villa, comarca, terra, terras.*

As demais estimativas que confirmaram a tabela não apresentavam problemas referentes à recuperação, e a variação de grafia era levantada por meio do *Philologic*, a partir de uma lista que já existia.

Feitas todas as alterações, era possível descrever cada gênero, saber quais os principais traços, a média de ocorrência de cada um (Seção 4.5) e iniciar o treinamento e teste com os classificadores.

4.8 Do treinamento aos testes com os classificadores

Após gerar dois arquivos ARFF, um para treinamento outro para teste, inicia-se o treinamento com os classificadores. Uma vez que essa etapa encontra-se no âmbito do aprendizado de máquina supervisionado, foi fornecida aos classificadores uma amostra para o treinamento. Feito isso, foram seguidas as seguintes etapas ao acessar o Weka:

1) carrega o arquivo ARFF com o conjunto de treinamento;

2) selecionar o algoritmo de aprendizagem;

3) avaliar a classificação.

Neste último item também foi possível refinar os traços para melhorar o resultado. Apresenta-se na Figura 9 o resultado de um treinamento com o classificador J48.

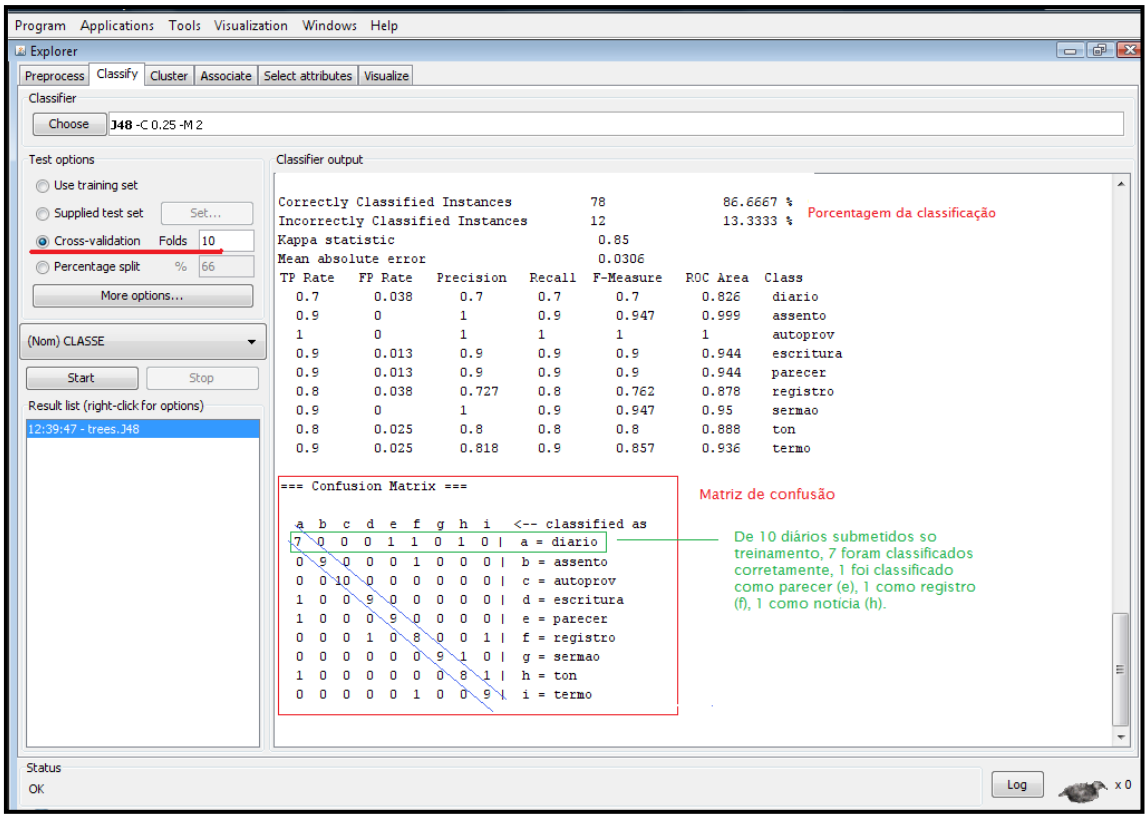

Figura 9: Resultado de treinamento com o classificador J48 do Weka

A Figura 9 mostra que, no treinamento, 86,6% dos textos foram classificados corretamente. Também apresenta a matriz de confusão, que fornece a qualidade da classificação, mostra o número de classificações corretas e as classificações preditas para casa classe, sobre determinado conjunto de exemplos. Referente à *Cross-validation* é uma forma de seleção de conjuntos de treino e teste, assim, um conjunto de dados é dividido em *n* subconjuntos com mais ou menos o mesmo número de amostras, e correspondentemente são feitos *n* pares para treino e teste.

O teste para classificação apresenta uma interface muito semelhante, mudando apenas os resultados. Os classificadores utilizados para testar a classificação foram selecionados com base no desempenho na fase de treinamento e correspondem aos seguintes conjuntos e subconjuntos de algoritmos, como está disponibilizado no Weka:

- Bayes: NaiveBayes (84,4% classificação correta), BayesNet (88,8%);

- Functions: SMO (86,6% classificação correta), MultilayerPerceptron (87,7%), RBFNetwork (91,1%);
- Tree: J48 (86,6% classificação correta), NBTree (86,6%).

Outros classificadores foram treinados, mas os resultados foram baixos, observem-se:

- Relus: OneR (32,2%), DecisionTable (53,3%), ZeroR (11,1%);
- Tree: DecisionStump (21,1%).

Os resultados referentes à descrição de traços do português histórico correlacionando-os a gêneros serão descritos no Capítulo 5, assim como o resultado dos testes com os classificadores.

5. Descrição dos traços linguísticos correlacionados aos gêneros

Neste capítulo são apresentados os resultados obtidos da classificação automática de gêneros. Uma vez que o objetivo da pesquisa é descrever os traços do português histórico correlacionando-os a gêneros e tendo como orientação a aplicação de uma tabela de traços contemporâneos adaptada ao contexto histórico, a apresentação dos resultados está organizada com base nos grupos sugeridos na própria tabela: estatísticas baseadas no texto como um todo e em palavras, outras estatísticas, as quais se referem a características morfossintáticas, expressões e unidades lexicais. Dentro de cada um desses grupos, cada gênero foi descrito pela média de ocorrência em cada texto, por exemplo: em cada texto pertencente ao gênero parecer, a média de ocorrência do verbo "fazer" nas formas "fez", "fazer" e "fazem" é 4,35. Ou ainda, os textos pertencentes ao gênero assento são formados por média de 114,6 frases. Além disso, a média está acompanhada do desvio padrão.

Após a descrição, serão apresentados os resultados da classificação automática, organizados pelos algoritmos de classificação utilizados na pesquisa.

5.1 Estatísticas baseadas no texto como um todo e em palavras

Gênero assento

Apresenta-se na Tabela 7 o gênero assento, a qual descreve a média dos traços referentes ao texto como um todo e em palavras, acompanhado de uma descrição das siglas elaboradas na pesquisa para melhor desempenho das tarefas computacionais.

Traços	Descrição	Média[30]
EPN	Estimativa de itens lexicais diferentes, dado pelo número de itens lexicais diferentes dividido pelo número de itens	0,59 *(0,09)*
EEM	Estimativa de itens lexicais diferentes, porém, considerando-se apenas os itens iniciados por letra maiúscula	0,71 *(0,10)*
END	Número de dígitos	13,62 *(36,46)*

[30] A média é seguida pelo desvio padrão, o número entre parênteses ( ).

ETP	Tamanho médio das palavras em caracteres	5 *(0,34)*
EPL	Número de palavras longas	170,75 (*103,15)*
ENC	Número de caracteres	2.783,56 *(1.961,41)*
TFC	Tamanho médio das frases em caracteres	660,8 *(293,67)*
ENF	Número de frases	4,81 *(3,39)*
TFP	Tamanho médio das frases em palavras	114,66 *(58,01)*
TTP	Tamanho do texto em palavras	438,56 *(380,06)*

Tabela 7: Média dos traços referentes ao texto como um todo: gênero assento

De acordo com a Tabela 7, é interessante ressaltar que se trata de um gênero cujos textos não são extensos, segundo o TTP, mas de riqueza lexical considerável, dada pelo EPN.

Seguem os dados referentes ao gênero auto de provimento.

Gênero auto de provimento

Traços	Descrição	Média
EPN	Estimativa de itens lexicais diferentes, dado pelo número de itens lexicais diferentes dividido pelo número de itens.	0,36 *(0,042)*
EEM	Estimativa de itens lexicais diferentes, porém considerando-se apenas os itens iniciados por letra maiúscula	0,58 *(0,06)*
END	Número de dígitos	23,26 *(20,93)*
ETP	Tamanho médio das palavras em caracteres	4,66 *(0,08)*
EPL	Número de palavras longas	296,26 *(114,07)*
ENC	Número de caracteres	5.754 *(2.195,07)*
TFC	Tamanho médio das frases em caracteres	203,3 *(137,75)*
ENF	Número de frases	83,73 *(119,22)*
TFP	Tamanho médio das frases em palavras	42,52 *(29,05)*
TTP	Tamanho do texto em palavras	1.195 *(475, 33)*

Tabela 8: Média dos traços referentes ao texto como um todo: gênero auto de provimento

Com base na tabela 8, pode-se considerar que os textos deste gênero possuem pouca riqueza lexical, dado pelo EPN, e os textos possuem tamanho médio.

Apresenta-se na Tabela 9, a seguir, os dados referentes ao gênero diário.

Gênero diário

Traços	Descrição	Média
EPN	Estimativa de itens lexicais diferentes, dado pelo número de itens lexicais diferentes dividido pelo número de itens.	0,16 *(0,04)*
EEM	Estimativa de itens lexicais diferentes, porém considerando-se apenas os itens iniciados por letra maiúscula	0,38 *(0,11)*
END	Número de dígitos	341,9 *(288,21)*
ETP	Tamanho médio das palavras em caracteres	4,42 *(0,19)*

EPL	Número de palavras longas	2.263,29 *(1.004,92)*
ENC	Número de caracteres	49.856,5 *(22.884,41)*
TFC	Tamanho médio das frases em caracteres	164,1 (59,79)
ENF	Número de frases	379,88 *(276,90)*
TFP	Tamanho médio das frases em palavras	35,45 *(13,14)*
TTP	Tamanho do texto em palavras	10.839 (*5.105,18)*

Tabela 9: Média dos traços referentes ao texto como todo: gênero diário

Com base na Tabela 9, verifica-se que textos pertencentes a esse gênero são extensos, com baixíssima riqueza lexical e com um número elevado de frases curtas, dado pelo TFP.

Apresenta-se na Tabela 10, a seguir, os dados referentes ao gênero escritura.

Gênero escritura

Traços	Descrição	Média
EPN	Estimativa de itens lexicais diferentes, dado pelo número de itens lexicais diferentes divididos pelo número de itens.	0,37 *(0,05)*
EEM	Estimativa de itens lexicais diferentes, porém considerando-se apenas os itens iniciados por letra maiúscula	0,55 *(0,06)*
END	Número de dígitos	7,85 *(3,70)*
ETP	Tamanho médio das palavras em caracteres	4.44 *(0,13)*
EPL	Número de palavras longas	333,6 *(180,76)*
ENC	Número de caracteres	6.793,3 *(3.655,66)*
TFC	Tamanho médio das frases em caracteres	367,4 *(157,58)*
ENF	Número de frases	20 *(9,07)*
TFP	Tamanho médio das frases em palavras	81,66 (35,50)
TTP	Tamanho do texto em palavras	1.502 *(799,99)*

Tabela 10: Média dos traços referentes ao texto como um todo: gênero escritura

Com base na Tabela 10, constata-se que o tamanho dos textos desse gênero em palavras não são extensos, mas não são tão pequenos quanto os textos do gênero assento. Possuem pouca riqueza lexical, com poucas frases, porém longas.

Apresenta-se na Tabela 11, a seguir, os dados referentes ao gênero parecer.

Gênero parecer

Traços	Descrição	Média
EPN	Estimativa de itens lexicais diferentes, dado pelo número de itens lexicais diferentes divididos pelo número de itens.	0,43 *(0,11)*
EEM	Estimativa de itens lexicais diferentes, porém considerando-se apenas os itens iniciados por letra maiúscula	0,68 *(0,14)*
END	Número de dígitos	41,78 *(47,50)*
ETP	Tamanho médio das palavras em caracteres	4,77 *(0,19)*
EPL	Número de palavras longas	318,07 *(317,11)*
ENC	Número de caracteres	6.569,21

		(7.024,45)
TFC	Tamanho médio das frases em caracteres	216,7 *(167,42)*
ENF	Número de frases	51,78 *(78,56)*
TFP	Tamanho médio das frases em palavras	44 *(34,55)*
TTP	Tamanho do texto em palavras	1.258,3 *(1.517,98)*

Tabela 11: Média dos traços referentes ao texto como um todo: gênero parecer

De acordo com os traços da Tabela 11, verifica-se que o tamanho do texto em palavras é considerado médio, com pouca riqueza lexical e frases curtas.

Apresenta-se na Tabela 12, a seguir, os dados referentes ao gênero registro.

Gênero registro

Traços	Descrição	Média
EPN	Estimativa de itens lexicais diferentes, dado pelo número de itens lexicais diferentes dividido pelo número de itens.	0,49 *(0,04)*
EEM	Estimativa de itens lexicais diferentes, porém considerando-se apenas os itens iniciados por letra maiúscula	0,61 *(0,10)*
END	Número de dígitos	18,45 *(6,51)*
ETP	Tamanho médio das palavras em caracteres	5,1 *(0,16)*
EPL	Número de palavras longas	200,15 *(25,96)*
ENC	Número de caracteres	3.668 *(416,06)*
TFC	Tamanho médio das frases em caracteres	442,6 *(704,66)*
ENF	Número de frases	15,45 *(13,29)*
TFP	Tamanho médio das frases em palavras	83,97 *(131,30)*
TTP	Tamanho do texto em palavras	707,95 *(86,32)*

Tabela 12: Média dos traços referentes do texto como um todo: gênero registro

Com base na Tabela 12, constata-se que o tamanho dos textos desse gênero é pequeno, constituídos por poucas frases longas e com pouca riqueza lexical.

Apresenta-se na Tabela 13, a seguir, os dados referentes ao gênero sermão.

Gênero sermão

Traços	Descrição	Média
EPN	Estimativa de itens lexicais diferentes, dado pelo número de itens lexicais diferentes dividido pelo número de itens.	0,19 *(0,02)*
EEM	Estimativa de itens lexicais diferentes, porém considerando-se apenas os itens iniciados por letra maiúscula	0,37 *(0,06)*
END	Número de dígitos	4,25 *(3,69)*
ETP	Tamanho médio das palavras em caracteres	4,42 *(0,10)*
EPL	Número de palavras longas	1.976,4 *(715,53)*
ENC	Número de caracteres	42.525,9 *(14.725,38)*
TFC	Tamanho médio das frases em caracteres	159,73 *(22,76)*
ENF	Número de frases	266,45 *(81,46)*
TFP	Tamanho médio das frases em palavras	34,60 *(4,71)*

TTP	Tamanho do texto em palavras	9.227,8 *(3216,39)*

Tabela 13: Média dos traços referentes ao texto como um todo: gênero sermão

De acordo com a Tabela 13, verifica-se que os textos desse gênero são extensos, constituídos por muitas frases curtas e baixíssima riqueza lexical.

Apresenta-se na Tabela 14, a seguir, os dados referentes ao gênero *termo*.

Gênero *termo*

Traços	Descrição	Média
EPN	Estimativa de itens lexicais diferentes, dado pelo número de itens lexicais diferentes dividido pelo número de itens.	0,42 *(0,11)*
EEM	Estimativa de itens lexicais diferentes, porém considerando-se apenas os itens iniciados por letra maiúscula	0,59 *(0,12)*
END	Número de dígitos	15,58 *(11,52)*
ETP	Tamanho médio das palavras em caracteres	4,49 *(0,41)*
EPL	Número de palavras longas	210,7 *(106,93)*
ENC	Número de caracteres	4.698,23 *(2.585,69)*
TFC	Tamanho médio das frases em caracteres	522,6 *(180,52)*
ENF	Número de frases	9,41 *(5,02)*
TFP	Tamanho médio das frases em palavras	111,67 *(43,50)*
TTP	Tamanho do texto em palavras	1.018 *(611,28)*

Tabela 14: Média dos traços referentes ao texto como um todo: gênero *termo*

Com base na Tabela 14, conclui-se que os textos desse gênero têm tamanho mediano, é constituído por pouquíssimas frases longas e pouca riqueza lexical.

Apresenta-se na Tabela 15, a seguir, os dados referentes ao gênero notícia.

Gênero notícia

Traços	Descrição	Média
EPN	Estimativa de itens lexicais diferentes, dado pelo número de itens lexicais diferentes dividido pelo número de itens.	0,25 *(0,07)*
EEM	Estimativa de itens lexicais diferentes, porém considerando-se apenas os itens iniciados por letra maiúscula	0,56 *(0,11)*
END	Número de dígitos	98 *(72,89)*
ETP	Tamanho médio das palavras em caracteres	4,4 *(0,17)*
EPL	Número de palavras longas	777,85 *(598,65)*
ENC	Número de caracteres	16.891,6 *(11.232,31)*
TFC	Tamanho médio das frases em caracteres	194,51 *(57,89)*
ENF	Número de frases	90,4 *(58,98)*
TFP	Tamanho médio das frases em palavras	41,71 *(11,85)*
TTP	Tamanho do texto em palavras	3.604,35 *(2.334,14)*

Tabela 15: Média dos traços referentes ao texto como um todo: gênero notícia

Com base na Tabela 15, verifica-se que os textos desse gênero costumam ser extensos, com um número de frases curtas elevado e baixíssima riqueza lexical.

Foi realizada a descrição dos gêneros de acordo com características baseadas no texto como um todo e em palavras. A seguir, seguem gráficos de cada um dos traços da tabela referente a estatísticas baseadas no texto como um todo.

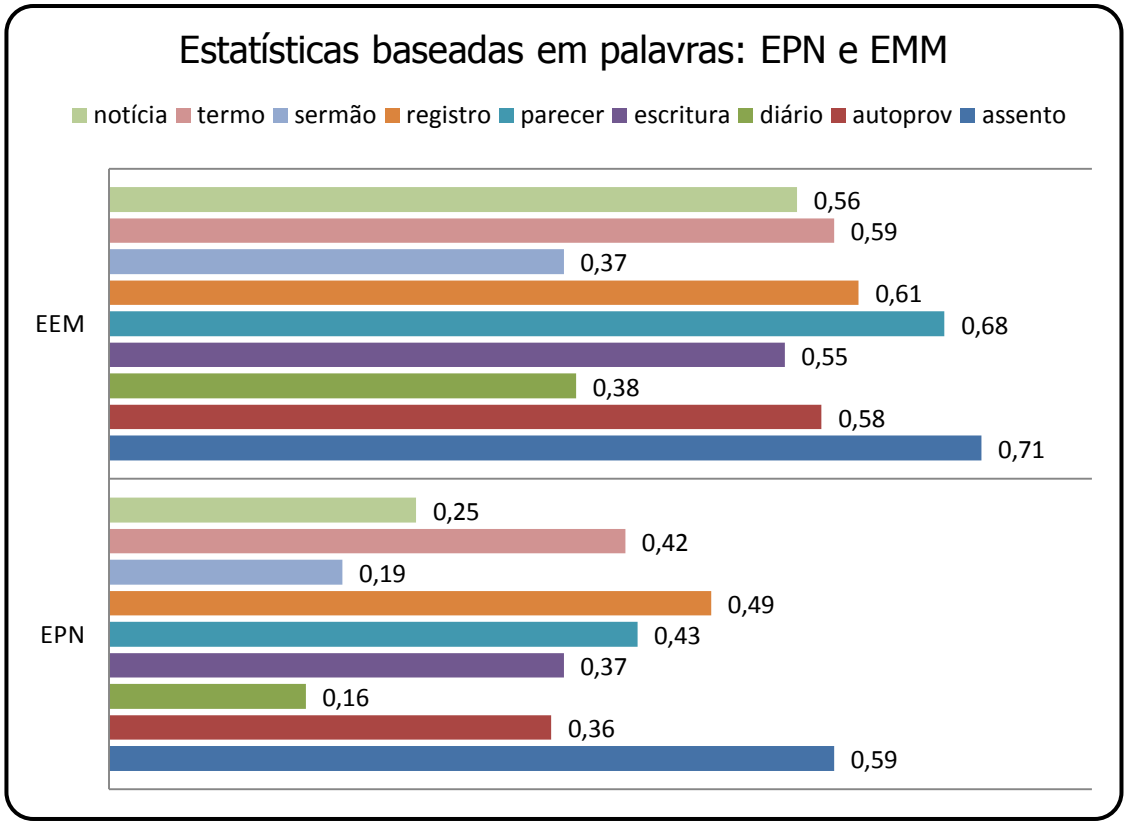

Gráfico 1: Estatísticas baseadas em palavras: EPN e EMM

De acordo com o gráfico 1, constata-se que o gênero que possui maior riqueza lexical (EPN) é o assento; e com menor riqueza lexical, o gênero diário. O mesmo ocorre com estimativas lexicais de itens diferentes iniciados por letras maiúsculas (EEM).

O gráfico 2 representa o tamanho médio das palavras em caracteres (ETP), o número de frases (TFP) e o número de dígitos (END) comparando os gêneros.

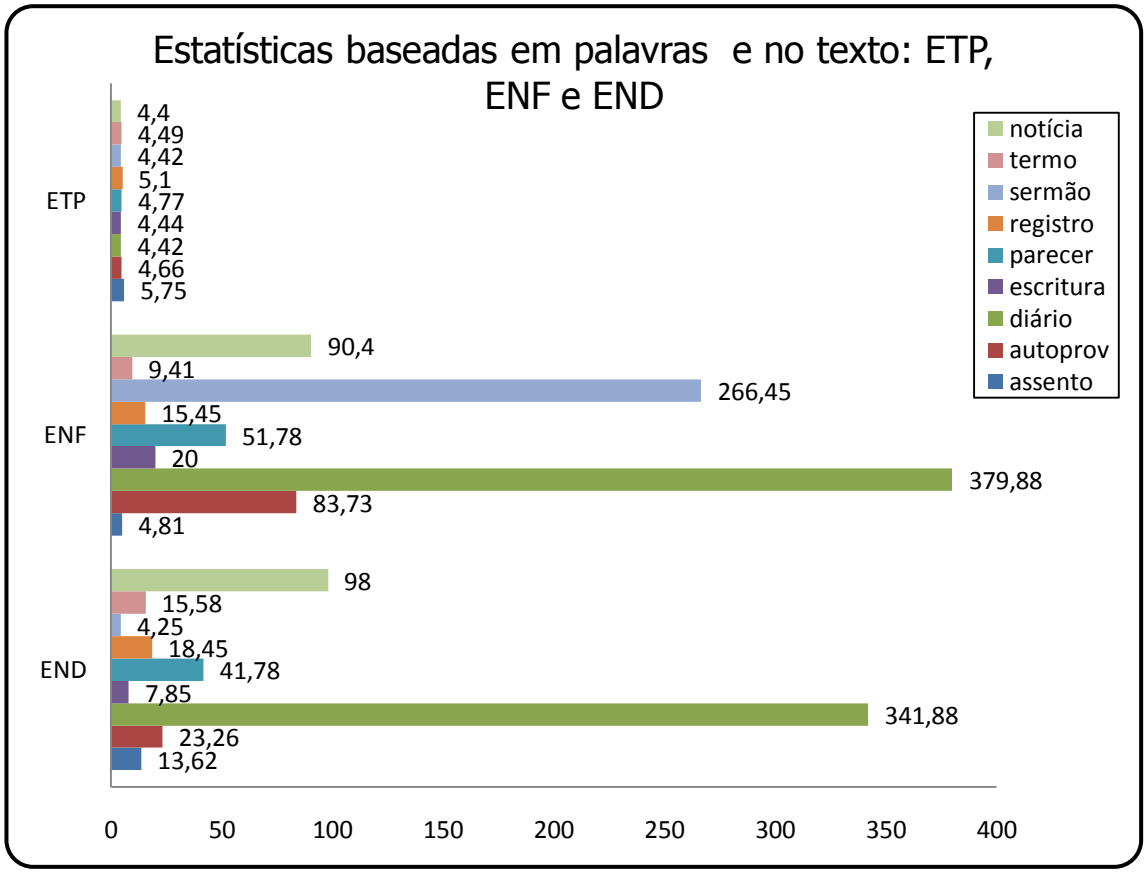

Gráfico 2: Estatísticas baseadas no texto e em palavras: ETP, ENF e END

De acordo com o gráfico 2, contata-se que o tamanho médio das palavras em caracteres (ETP) é semelhante em todos os gêneros. O mesmo não ocorre com o tamanho médio das frases em caracteres (ENF), no qual as maiores médias são do diário, sermão e notícia, o que é previsível por serem textos extensos. Já com o número de dígitos foi diferente, o que possui maior média e se destaca entre os demais gêneros é o diário.

A seguir, apresenta-se o gráfico 3 com número de palavras longas (EPL), tamanho médio das frases em palavras (TFP) e tamanho das frases em caracteres (TFC).

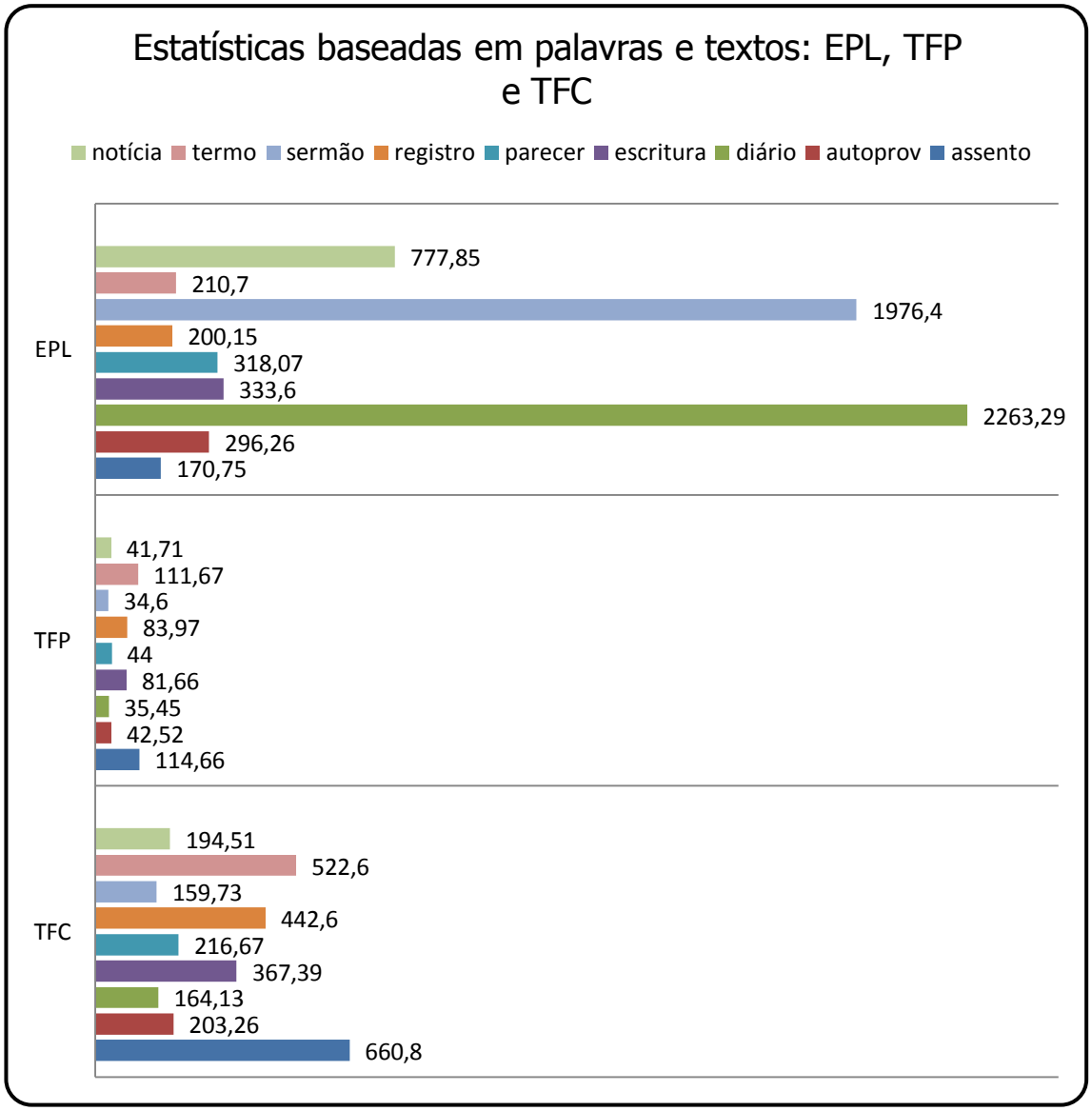

Gráfico 3: Estatísticas baseadas em palavras e textos: EPL, TFP e TFC

No gráfico 3, pode-se constatar que os gêneros que possuem maior número de palavras longas (EPL) são os diários, sermão e notícia. No entanto, referente ao tamanho médio das frases em palavras (TFP), têm destaque o gênero *termo*, escritura, assento e o mesmo ocorre com o tamanho das frases e caracteres (TFC).

A seguir, apresenta-se o último gráfico referente a estatísticas baseadas no texto e em palavras, mas especificamente número de caracteres (ENC) e tamanho do texto em palavras (TTP).

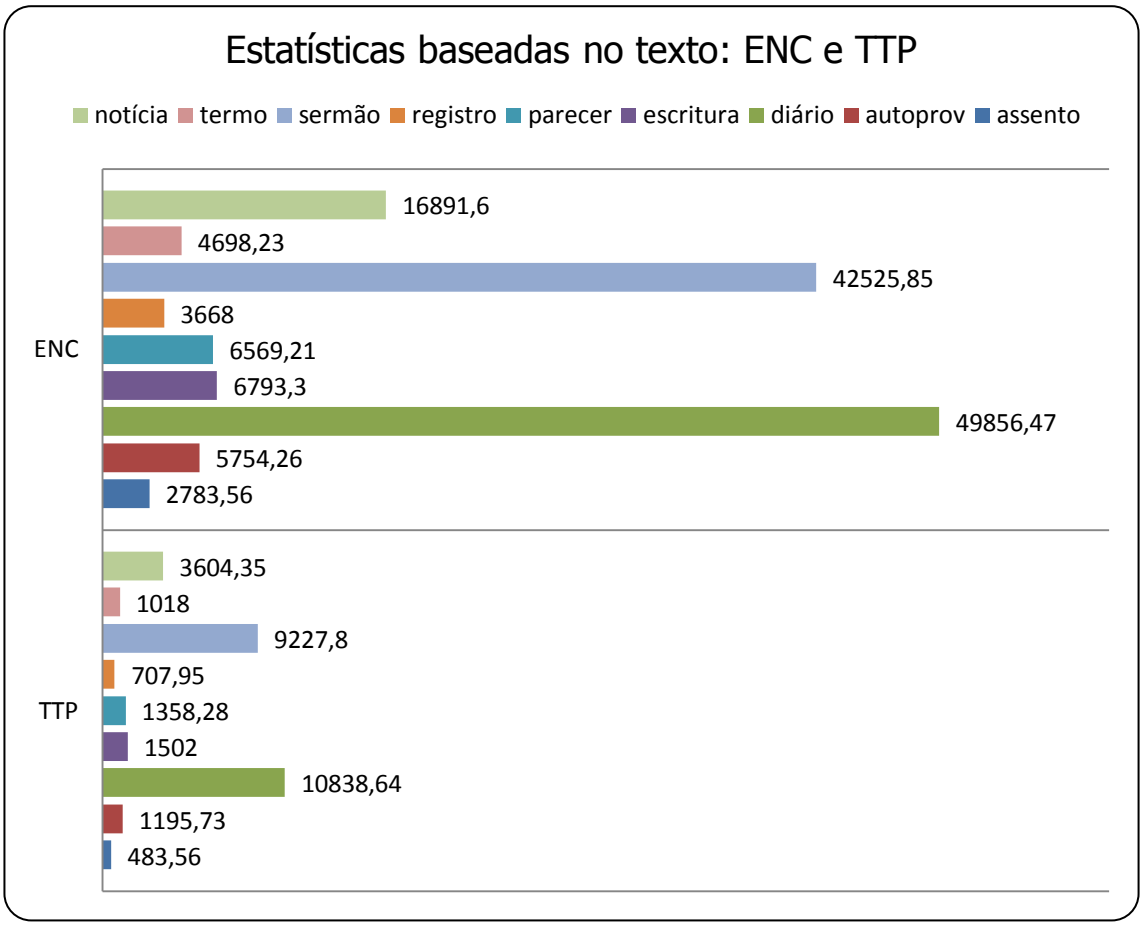

Gráfico 4: Estatísticas baseadas no texto: ENC e TTP

Com base no gráfico 4, verifica-se que os gêneros com textos mais extensos, dado pelo número de palavras, são o diário, sermão e notícia, e o menor é o assento. Como se pode deduzir, o mesmo ocorre com o número de caracteres.

Para finalizar o grupo de estimativas baseadas no texto como um todo e em palavras, apresenta-se a Tabela 16 com todos os traços por gênero, uma vez que facilita a comparação entre todos os gêneros e alguns traços, pois nos gráficos as escalas eram muito distintas.

	EPN	EEM	END	ETP	EPL	ENC	TFC	TPF	TTP	ENF
Assento	0,59	0,71	13,62	5,75	170,75	2783,56	660,8	114,66	483,56	4,81
Autoprov	0,36	0,58	23,26	4,66	296,26	5754,26	203,3	42,52	1195,7	83,73
Diário	0,16	0,38	341,9	4,42	2263,29	49856,5	164,1	35,45	10839	379.88
Escritura	0,37	0,55	7,85	4,44	333,6	6793,3	367,4	81,66	1502	20
Parecer	0,43	0,68	41,78	4,77	318,07	6569,21	216,7	44	1358,3	51,78
Registro	0,49	0,61	18,45	5,1	200,15	3668	442,6	83,97	707,95	15,45
Sermão	0,19	0,37	4,25	4,42	1976,4	42525,9	159,7	34,6	9227,8	266,5
Termo	0,42	0,59	15,58	4,49	210,7	4698,23	522,6	111,67	1018	9,41

Notícia	0,25	0,56	98	4,4	777,85	16891,6	194,5	41,71	3604,4	90,4

Tabela 16: Comparação da média dos traços referentes ao texto como um todo.

No gênero assento, os textos apresentam maior riqueza lexical (EPN=0,59), possuem menos palavras (TTP), menor número de palavras longas, menor número de frases (ENF), que por sua vez são mais longas (TPF) comparadas a outros gêneros.

Com base na tabela, constata-se que quanto menor o número de frases (ENF), maior o tamanho das frases (TPF) que compõem um texto, por exemplo: os textos do gênero sermão apresentam a maior média em números de frases (ENF=266) e a menor média no tamanho das frases (TPF = 34). Outro mesmo exemplo do corpus é o gênero assento, o qual tem a menor média em número de frases (ENF), mas a maior média em tamanho das frases (TPF).

Outra constatação é referente ao tamanho do texto em palavra (TTP) e a riqueza lexical dada pelo EPN, uma vez que, quanto maior o texto em palavras (TTP), menor a riqueza lexical. Um exemplo disso no corpus é o gênero assento, menor média em tamanho de texto e maior riqueza lexical.

Além disso, quanto maior o número de caracteres (ENC), menor o tamanho da frase em caracteres (TFC), assim como para o tamanho médio das palavras em caracteres.

5.2 Outras estatísticas

5.2.1 Outras estatísticas: verbos

Na sequência, são apresentadas tabelas para cada gênero, contendo a descrição de traços referente a verbos.

A Tabela 17, a seguir, refere-se ao gênero assento.

Gênero assento

Traços	Descrição	Média
VS	Verbo SER nas formas *é* e *são, sam*	0,37 *(0,71)*
VH	Verbo HAVER nas formas *há, havia*	0,37 *(0,5)*
VP	Verbo PEDIR nas formas *pede, pedem*	0 *(0)*
VPr	Verbo PROVER nas formas *proveu, proveo*	0 *(0)*
VPo	Verbo PODER nas formas *póde, podia*	0,37 *(0,5)*

VF	Verbo FAZER nas formas *fez, fazer* e *fazem*	0,68 *(1,07)*
VI	Verbo IR nas formas *foi, fomos* e *fui*	0,75*(0,68)*
VT	Verbo TER nas formas *tem* e *tinha*	0,5 *(0,81)*
VD	Verbo DIZER nas formas *dizer, disse* e *digo*	1,18 *(1,22)*

Tabela 17: Média de ocorrência dos verbos no gênero assento

De acordo com a Tabela 17 e como previsto na lista de palavras, o verbo que teria predominância em textos desse gênero é o *dizer*, o que ficou comprovado.

A Tabela 18, a seguir, refere-se ao gênero auto de provimento.

Gênero auto de provimento

VS	Verbo SER nas formas *é* e *são, sam*	2,33 *(1,79)*
VH	Verbo HAVER nas formas *há, havia*	0,33 *(0,48)*
VP	Verbo PEDIR nas formas *pede, pedem*	0 *(0)*
VPr	Verbo PROVER nas formas *proveu, proveo*	4,13 *(2,41)*
VPo	Verbo PODER nas formas *póde, podia*	0,2 *(0,41)*
VF	Verbo FAZER nas formas *fez, fazer* e *fazem*	5,6 *(1,35)*
VI	Verbo IR nas formas *foi, fomos* e *fui*	1,26 *(0,59)*
VT	Verbo TER nas formas *tem* e *tinha*	1,13 *(1,32)*
VD	Verbo DIZER nas formas *dizer, disse* e *digo*	0,93 *(0,79)*

Tabela 18: Média de ocorrência dos verbos no gênero auto de provimento

Com base na Tabela 18, pode-se constatar que o verbo *prover* é mais recorrente que os demais verbos, como previsto nas hipóteses. Os autos de provimento constituem textos redigidos pelo tabelião, a mando do Doutor ouvidor geral da comarca, tomando providências, regularizando diversas situações. Seguem trechos extraídos dos autos a título de elucidação: *"Proveu e determinou que se observassem e cumprissem exatamente todos os provimentos de seus antecessores"; "Proveu que ele dito Doutor Provedor e Ouvidor geral e corregedor da Comarca que atendendo a grave dano que resulta a esta vila".*

A Tabela 19, a seguir, refere-se ao gênero diário.

Gênero diário

VS	Verbo SER nas formas *é* e *são, sam*	52,29 *(49,23)*
VH	Verbo HAVER nas formas *há, havia*	18,52 *(16,20)*
VP	Verbo PEDIR nas formas *pede, pedem*	0,17 *(0,39)*
VPr	Verbo PROVER nas forma *proveu, proveo*	0,05 *(0,24)*
VPo	Verbo PODER nas formas *póde, podia*	6,05 *(6,26)*
VF	Verbo FAZER nas formas *fez, fazer* e *fazem*	28,88 *(28,43)*
VI	Verbo IR nas formas *foi, fomos* e *fui*	39,23 *(29,56)*
VT	Verbo TER nas formas *tem* e *tinha*	40,17 *(22,39)*
VD	Verbo DIZER nas formas *dizer, disse* e *digo*	9,17 *(8,83)*

Tabela 19: Média de ocorrência dos verbos no gênero diário

Esse gênero apresenta maior ocorrência dos verbos *ser, ir* e *ter*, com suas respectivas formas. A predominância do verbo *ser* é previsível, uma vez que os diários descrevem as viagens, as descobertas. Entre as passagens dos textos, encontram-se várias narrações do tipo: *são brancos, são índios, são outras muitas ilhas, as águas são férteis* e assim por diante; ou ainda, referente ao verbo *ir, fomos para a mesa, fomos para a cidade* e referente ao verbo *ter, tem muitos e belos edifícios, tem muitos comércios, a fortaleza tem dentro o quartel da tropa inglesa.* Esses exemplos elucidam o aspecto descritivo dos diários e justifica a predominância desses verbos.

A Tabela 20, a seguir, refere-se ao gênero escritura.

Gênero escritura

VS	Verbo SER nas formas *é* e *são, sam*	5,05 *(3,13)*
VH	Verbo HAVER nas formas *há, havia*	0,45 *(0,82)*
VP	Verbo PEDIR nas formas *pede, pedem*	0,25 *(0,71)*
VPr	Verbo PROVER nas formas *proveu, proveo*	0 *(0)*
VPo	Verbo PODER nas formas *póde, podia*	0,4 *(0,59)*
VF	Verbo FAZER nas formas *fez, fazer* e *fazem*	3,1 *(2,07)*
VI	Verbo IR nas formas *foi, fomos* e *fui*	3,95 *(2,21)*
VT	Verbo TER nas formas *tem* e *tinha*	2,45 *(1,57)*
VD	Verbo DIZER nas formas *dizer, disse* e *digo*	4.05 *(3,53)*

Tabela 20: Média de ocorrência dos verbos no gênero escritura

Com base na tabela, verifica-se que os verbos que mais ocorrem nos textos pertencentes a esse gênero são os verbos *ser, dizer* e *ir*, como previsto e mencionado na metodologia.

A Tabela 21, a seguir, refere-se ao gênero parecer.

Gênero parecer

VS	Verbo SER nas formas *é* e *são, sam*	3,92 *(9,07)*
VH	Verbo HAVER nas formas *há, havia*	0,09 *(0,24)*
VP	Verbo PEDIR nas formas *pede, pedem*	0,07 *(0,26)*
VPr	Verbo PROVER nas formas *proveu, proveo*	0 *(0)*
VPo	Verbo PODER nas formas *póde, podia*	0,07 *(0,26)*
VF	Verbo FAZER nas formas *fez, fazer* e *fazem*	4,35 *(4,18)*
VI	Verbo IR nas formas *foi, fomos* e *fui*	0,78 *(1,84)*
VT	Verbo TER nas formas *tem* e *tinha*	4,64 *(6,95)*
VD	Verbo DIZER nas formas *dizer, disse* e *digo*	1,18 *(2,40)*

Tabela 21: Média de ocorrência dos verbos no gênero parecer

De acordo com a tabela e o que foi mencionado na metodologia, especificamente na lista de palavras mais frequentes, era previsível a predominância do verbo *ter*. Os textos desse gênero costumam descrever fatos, menciona testemunhas, devassa etc.

Seguem exemplos: *"a Nação dos Muras, gentio Bravo do Rio da Madeira, tem acometido a muitas Canoas, e nelas tem feito as hostilidades".*

A Tabela 22, a seguir, refere-se ao gênero registro.

Gênero registro

VS	Verbo SER nas formas *é* e *são, sam*	0,45 *(0,60)*
VH	Verbo HAVER nas formas *há, havia*	0 *(0)*
VP	Verbo PEDIR nas formas *pede, pedem*	3,9 *(2,07)*
VPr	Verbo PROVER nas formas *proveu, proveo*	0 *(0)*
VPo	Verbo PODER nas formas *póde, podia*	0,05*(0,22)*
VF	Verbo FAZER nas formas *fez, fazer* e *fazem*	0,3 *(0,57)*
VI	Verbo IR nas formas *foi, fomos* e *fui*	0,6 *(0,75)*
VT	Verbo TER nas formas *tem* e *tinha*	1,35 *(1,18)*
VD	Verbo DIZER nas formas *dizer, disse* e *digo*	0,55 *(0,68)*

Tabela 22: Média de ocorrência dos verbos no gênero registro

Com base na Tabela 22, conclui-se que o verbo que mais ocorre nesse gênero é o *pedir*, o que foi previsto na lista de palavras. O registro trata de uma carta oficial de concessão de sesmaria, compreende-se como pedidos oficiais dos "suplicantes", por exemplo: *"(...), portanto pedem a Vossa Mercê cada um deles suplicantes em nome de sua majestade que Deus guarde três Léguas de terra de comprido".*

A Tabela 23, a seguir, refere-se ao gênero sermão.

Gênero sermão

VS	Verbo SER nas formas *é* e *são, sam*	102 *(37,04)*
VH	Verbo HAVER nas formas *há, havia*	10,05 *(5,71)*
VP	Verbo PEDIR nas formas *pede, pedem*	1 *(1,52)*
VPr	Verbo PROVER nas formas *proveu, proveo*	0,1 *(0,30)*
VPo	Verbo PODER nas formas *póde, podia*	16,55 *(8,21)*
VF	Verbo FAZER nas formas *fez, fazer* e *fazem*	20,85 *(11,12)*
VI	Verbo IR nas formas *foi, fomos* e *fui*	24,8 *(16,97)*
VT	Verbo TER nas formas *tem* e *tinha*	19,35 *(10,18)*
VD	Verbo DIZER nas formas *dizer, disse* e *digo*	24,35 *(12,16)*

Tabela 23: Média de ocorrência dos verbos no gênero sermão

De acordo com a tabela, o verbo que mais ocorre nos textos desse gênero é o verbo *ser*, o que não estava previsto na lista de palavras, e também os verbos *dizer*, *ir* e *fazer*, como o esperado.

A Tabela 24, a seguir, refere-se ao gênero *termo*.

Gênero *termo*

VS	Verbo SER nas formas *é* e *são, sam*	0,58 *(0,71)*
VH	Verbo HAVER nas formas *há, havia*	0 *(0)*
VP	Verbo PEDIR nas formas *pede, pedem*	0,17 *(0,72)*
VPr	Verbo PROVER nas formas *proveu, proveo*	0 *(0)*

VPo	Verbo PODER nas formas *póde, podia*	0,47 *(0,79)*
VF	Verbo FAZER nas formas *fez, fazer* e *fazem*	3,29 *(2,11)*
VI	Verbo IR nas formas *foi, fomos* e *fui*	0,82 *(1,81)*
VT	Verbo TER nas formas *tem* e *tinha*	1,47 *(1,28)*
VD	Verbo DIZER nas formas *dizer, disse* e *digo*	0,64 *(1,16)*

Tabela 24: Média de ocorrência dos verbos no gênero *termo*

Com base na Tabela 24, verifica-se que a ocorrência do verbo *fazer* é distinta das demais, como previsto. Com função semelhante ao assento, pode-se compreender que *termo* é uma menção escrita, integrante dos autos, escrita por um escrivão ou tabelião com vistas a regularizar algo, como no exemplo a seguir: *"Termo de Resolução que os oficiais da câmara tomaram sobre a repartição do sal", "oficiais da câmara mandarão fazer este termo".*

A Tabela 25, a seguir, refere-se ao gênero notícia.

Gênero notícia

VS	Verbo SER nas formas *é* e *são, sam*	30,95 *(30,72)*
VH	Verbo HAVER nas formas *há, havia*	17,4 *(16,69)*
VP	Verbo PEDIR nas formas *pede, pedem*	0,05 *(0,22)*
VPr	Verbo PROVER nas formas *proveu, proveo*	0 *(0)*
VPo	Verbo PODER nas formas *póde, podia*	0,8 *(1,43)*
VF	Verbo FAZER nas formas *fez, fazer* e *fazem*	9,5 *(9,63)*
VI	Verbo IR nas formas *foi, fomos* e *fui*	3,85 *(5,44)*
VT	Verbo TER nas formas *tem* e *tinha*	19,2 *(13,46)*
VD	Verbo DIZER nas formas *dizer, disse* e *digo*	1,25 *(1,86)*

Tabela 25: Média de ocorrência dos verbos no gênero notícia

De acordo com a tabela, constata-se que nos textos do gênero notícia, dentre os verbos listados, o que ocorreu com predominância foi o verbo *ser*, depois o verbo *ter*, como já estava previsto. As notícias descrevem tudo o que encontravam: índios, fauna e flora, cidades, como apresentam os exemplos a seguir: *"são todos peixes de coiro"; "Codornizes: todas tem as carnes duras, e secas."; "as Cornudas são mais capazes de se comerem"; "maracujá merim, tem os cipós roliços, as folhas do tamanho de uma mão com 3 pontas como uma meia estrela"; "como as galinhas, todas cantam, e de verão tem o canto muito triste".*

A seguir, apresentam-se gráficos para visualizar os traços por gêneros, permitindo, assim, compará-los.

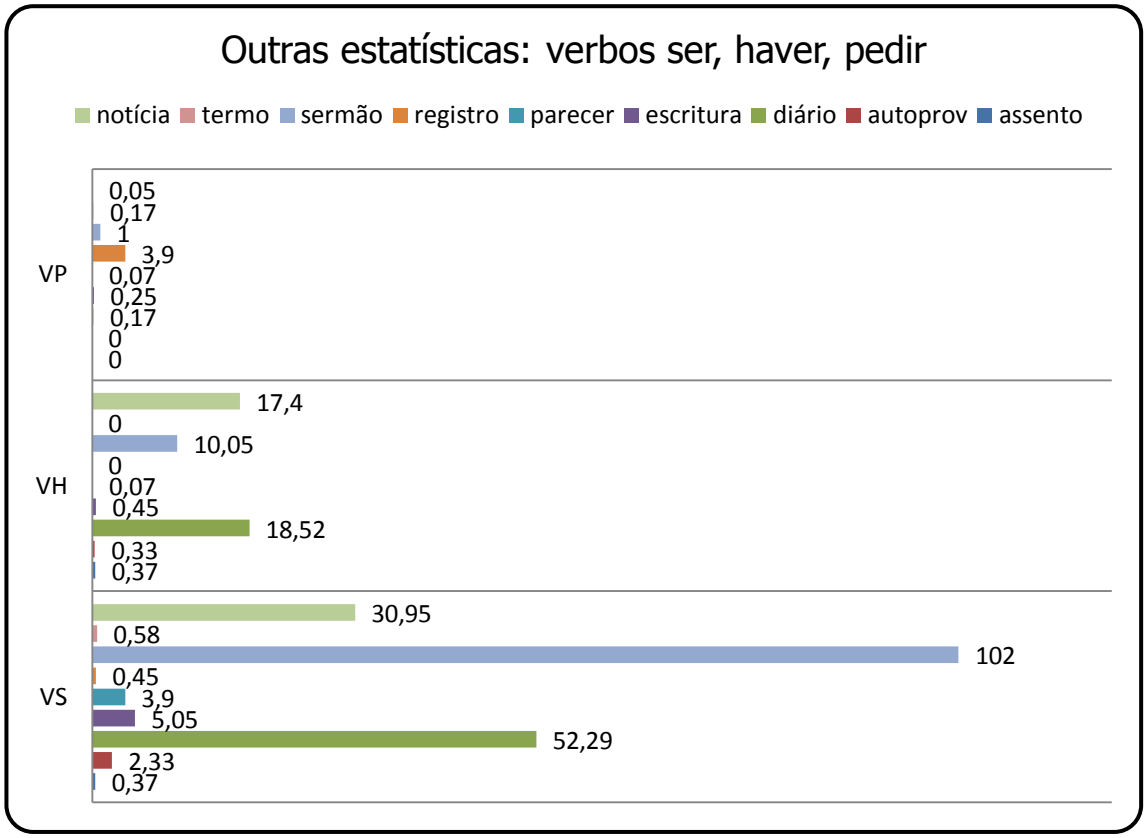

Gráfico 5: Outras estatísticas: verbos ser, haver, pedir

O gráfico 5, que contempla os verbos *ser*, *haver* e *pedir*, apresenta em quais gêneros a média de ocorrência é maior. Assim, constata-se, como previsto, a predominância do verbo *pedir* em textos do gênero registro, a predominância do verbo *haver* nos gêneros notícia, sermão e diário, e por fim, a predominância do verbo *ser* nos textos dos gêneros diário, sermão e notícia.

A seguir, o gráfico 6 apresenta a média de ocorrência dos verbos *prover*, *poder* e *ficar* em todos os gêneros.

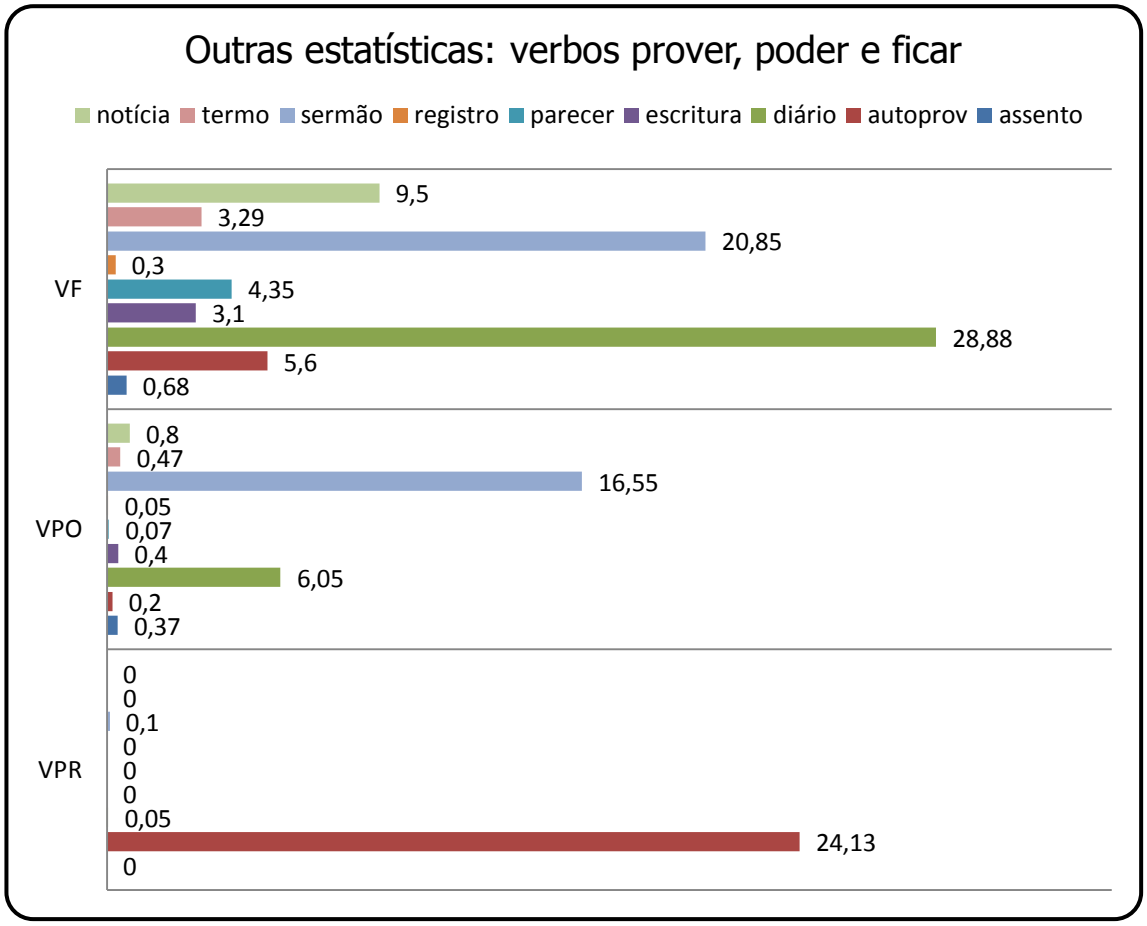

Gráfico 6: Outras estatísticas: verbos prover, poder e ficar

Com base no gráfico 6, comprova-se a predominância do verbo *prover* nos autos de provimento, a predominância do verbo *poder* nos sermões e o verbo *fazer* nos diários e nos sermões. A seguir, apresenta-se o gráfico 7 que traz dados sobre os verbos *ir*, *ter* e *dizer*.

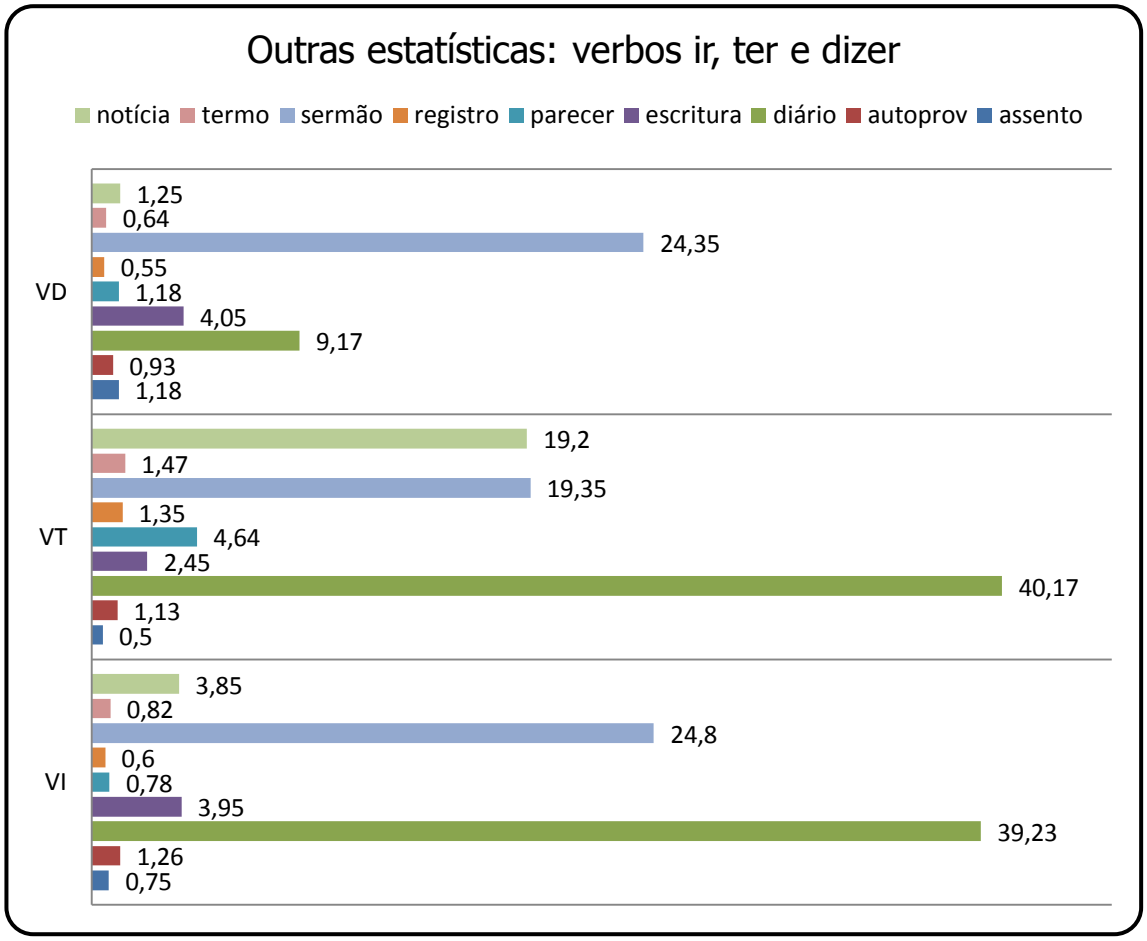

Gráfico 7: Outras estatísticas: verbos ir, ter e dizer

Por meio do gráfico 7, comprova-se a predominância do verbo *ir* nos diários, a predominância do verbo *ter* nos sermões, notícias e no diáro, bem como a predominância do verbo *dizer* nos sermões.

Com base nos resultados acerca dos verbos, pode-se concluir que, de modo geral, a ocorrência e predominância dos verbos também varia de gênero para gênero. O maior exemplo disso, baseado no corpus e nos gêneros utilizados, é o verbo *prover*, que praticamente só ocorre nos autos de provimento. Os textos dos gêneros diário e notícia são extensos e, embora compartilhem algumas características, como no caso da predominância do verbo *haver*, com outros verbos não ocorre o mesmo, como o verbo *ir*. Isso leva a compreender que, o fato de os textos do gêneros serem extensos, nem sempre os traços serão predominates. Essa comparação pode ser melhor percebida na tabela 26 a seguir.

	VS	VH	VP	VPR	VPO	VF	VI	VT	VD
Assento	0,37	0,37	0	0	0,37	0,68	0,75	0,5	1,18
Autoprov	2,33	0,33	0	24,13	0,2	5,6	1,26	1,13	0,93
Diário	52,29	18,52	0,17	0,05	6,05	28,88	39,23	40,17	9,17
Escritura	5,05	0,45	0,25	0	0,4	3,1	3,95	2,45	4,05
Parecer	3,9	0,07	0,07	0	0,07	4,35	0,78	4,64	1,18
Registro	0,45	0	3,9	0	0,05	0,3	0,6	1,35	0,55
Sermão	102	10,05	1	0,1	16,55	20,85	24,8	19,35	24,35
Termo	0,58	0	0,17	0	0,47	3,29	0,82	1,47	0,64
Notícia	30,95	17,4	0,05	0	0,8	9,5	3,85	19,2	1,25

Tabela 26: Comparação da média de ocorrência dos verbos nos gêneros

5.2.2 Outras estatísticas: pronomes

A seguir, serão apresentadas tabelas para cada gênero, contendo a descrição de traços referente a pronomes.

A Tabela 27, a seguir, refere-se ao gênero assento.

Gênero assento

Traço	Descrição	Média
PPO1	Pronome pessoal oblíquo na primeira pessoa	0,87 *(1,02)*
PPO2	Pronome pessoal oblíquo na segunda pessoa	0 *(0)*
PPO3	Pronome pessoal oblíquo na terceira pessoa	1,81 *(2,85)*
PPT	Pronome pessoal de tratamento	1,18 *(1,68)*
PD	Pronomes demonstrativos	8,06 *(5,85)*
PRV	Pronomes relativos variáveis	0,68 *(0,87)*
PINT	Pronomes interrogativos	1,5 *(1,89)*
PIP	Pronomes indefinidos referente à pessoa	0,06 *(0,25)*
PIL	Pronomes indefinidos referente a lugar	0,25 *(0,57)*
PIC	Pronomes indefinidos referente a coisas	1,68 *(1,49)*

Tabela 27: Média de ocorrência de pronomes no gênero assento

Pelo exposto na Tabela 27, verifica-se que o pronome que mais ocorre neste gênero é o pronome demonstrativo.

A Tabela 28, a seguir, refere-se ao gênero auto de provimento.

Gênero auto de provimento

PPO1	Pronome pessoal oblíquo na primeira pessoa	2,4 *(1,76)*
PPO2	Pronome pessoal oblíquo na segunda pessoa	0,13 *(0,13)*
PPO3	Pronome pessoal oblíquo na terceira pessoa	8,4 *(4,54)*
PPT	Pronome pessoal de tratamento	2 *(1,06)*
PD	Pronomes demonstrativos	19,73 *(9,26)*
PRV	Pronomes relativos variáveis	3,53 *(2,38)*

PINT	Pronomes interrogativos	5,06 *(5,28)*
PIP	Pronomes indefinidos referente à pessoa	0,8 *(1,20)*
PIL	Pronomes indefinidos referente a lugar	2,13 *(1,24)*
PIC	Pronomes indefinidos referente a coisas	9,6 *(5,32)*

Tabela 28: Média de ocorrência dos pronomes no gênero auto de provimento

Pelo exposto acima, verifica-se que os pronomes que mais ocorrem nesse gênero são o demonstrativo, o pessoal oblíquo em terceira pessoa e os interrogativos. O pronome oblíquo em primeira pessoa, pessoal de tratamento, relativos variáveis, indefinidos referente a lugar ocorrem com frequência semelhante.

A Tabela 29, a seguir, refere-se ao gênero diário.

Gênero diário

PPO1	Pronome pessoal oblíquo na primeira pessoa	113,41 *(111,41)*
PPO2	Pronome pessoal oblíquo na segunda pessoa	1,47 *(3,33)*
PPO3	Pronome pessoal oblíquo na terceira pessoa	51,76 *(41,38)*
PPT	Pronome pessoal de tratamento	11,2 *(13,81)*
PD	Pronomes demonstrativos	130,5 *(70,20)*
PRV	Pronomes relativos variáveis	20,47 *(16,31)*
PINT	Pronomes interrogativos	54,05 *(41,33)*
PIP	Pronomes indefinidos referente à pessoa	4,11 *(4,18)*
PIL	Pronomes indefinidos referente a lugar	37,9 *(26,63)*
PIC	Pronomes indefinidos referente a coisas	91,47 *(57,66)*

Tabela 29: Média de ocorrência de pronomes no gênero diário

De acordo com a Tabela 29, verifica-se que os pronomes que mais ocorrem são o pessoal oblíquo em primeira pessoa, o demonstrativo e o indefinido referente a coisas. Os pronomes pessoais oblíquos em terceira pessoa, interrogativos e indefinidos referente a lugar ocorrem em proporções semelhantes. Os demais apresentam frequências baixas.

A Tabela 30, a seguir, refere-se ao gênero escritura.

Gênero escritura

PPO1	Pronome pessoal oblíquo na primeira pessoa	9,55 *(4,69)*
PPO2	Pronome pessoal oblíquo na segunda pessoa	0,2 *(0,41)*
PPO3	Pronome pessoal oblíquo na terceira pessoa	17,65 *(7,47)*
PPT	Pronome pessoal de tratamento	3,1 *(3,24)*
PD	Pronomes demonstrativos	15,55 *(7,38)*
PRV	Pronomes relativos variáveis	1,75 *(2,24)*
PINT	Pronomes interrogativos	9,55 *(5,67)*
PIP	Pronomes indefinidos referente à pessoa	1,5 *(1,96)*
PIL	Pronomes indefinidos referente a lugar	2,2 *(1,93)*
PIC	Pronomes indefinidos referente a coisas	15,05 *(7,81)*

Tabela 30: Média de ocorrência de pronomes no gênero escritura

Com base na Tabela 30, constata-se que os pronomes que mais ocorrem nesse gênero são pronome pessoal oblíquo na terceira pessoa, demonstrativo, indefinido referente a coisas, interrogativo e pessoal oblíquo na primeira pessoa.

A Tabela 31, a seguir, refere-se ao gênero parecer.

Gênero parecer

PPO1	Pronome pessoal oblíquo na primeira pessoa	6,21 *(6,57)*
PPO2	Pronome pessoal oblíquo na segunda pessoa	0 *(0)*
PPO3	Pronome pessoal oblíquo na terceira pessoa	10,28 *(11,33)*
PPT	Pronome pessoal de tratamento	0,78 *(0,69)*
PD	Pronomes demonstrativos	22,5 *(27,47)*
PRV	Pronomes relativos variáveis	2,42 *(4,51)*
PINT	Pronomes interrogativos	5,5 *(7,90)*
PIP	Pronomes indefinidos referente à pessoa	0,64 *(0,92)*
PIL	Pronomes indefinidos referente a lugar	1,21 *(1,57)*
PIC	Pronomes indefinidos referente a coisas	13,57 *(16,63)*

Tabela 31: Média de ocorrência de pronomes no gênero parecer

Com base na Tabela 31, os pronomes que mais ocorrem nesse gênero são os demonstrativos, pessoal oblíquo em terceira pessoa e indefinido referente a coisas.

A Tabela 32, a seguir, refere-se ao gênero registro.

Gênero registro

PPO1	Pronome pessoal oblíquo na primeira pessoa	4,8 *(1,60)*
PPO2	Pronome pessoal oblíquo na segunda pessoa	0,2 *(0,41)*
PPO3	Pronome pessoal oblíquo na terceira pessoa	5,9 *(1,74)*
PPT	Pronome pessoal de tratamento	4,15 *(1,53)*
PD	Pronomes demonstrativos	7,35 *(3,29)*
PRV	Pronomes relativos variáveis	1,85 *(1,22)*
PINT	Pronomes interrogativos	3,6 *(1,18)*
PIP	Pronomes indefinidos referente à pessoa	0 *(0)*
PIL	Pronomes indefinidos referente a lugar	0,65 *(1,26)*
PIC	Pronomes indefinidos referente a coisas	2,9 *(1,71)*

Tabela 32: Média de ocorrência de pronomes no gênero registro

Conforme a Tabela 32, os pronomes que mais ocorrem nesse gênero são o demonstrativo, o pessoal de tratamento e o pessoal oblíquo na primeira pessoa.

A Tabela 33, a seguir, refere-se ao gênero sermão.

Gênero sermão

PPO1	Pronome pessoal oblíquo na primeira pessoa	67,3 *(33,68)*
PPO2	Pronome pessoal oblíquo na segunda pessoa	44,8 *(39,67)*
PPO3	Pronome pessoal oblíquo na terceira pessoa	74,95 *(32,37)*
PPT	Pronome pessoal de tratamento	18,7 *(10,84)*
PD	Pronomes demonstrativos	152,7 *(58,10)*
PRV	Pronomes relativos variáveis	5,55 *(4,89)*
PINT	Pronomes interrogativos	63,35 *(26,58)*

PIP	Pronomes indefinidos referente à pessoa	6,55 *(5,93)*
PIL	Pronomes indefinidos referente a lugar	17,05 *(11,13)*
PIC	Pronomes indefinidos referente a coisas	108,3 *(39,93)*

Tabela 33: Média de ocorrência de pronomes no gênero sermão

Com base na Tabela 33, observa-se que os pronomes que mais ocorrem são o demonstrativo e o indefinido referente a coisas.

A Tabela 34, a seguir, refere-se ao gênero *termo*.

Gênero *termo*

PPO1	Pronome pessoal oblíquo na primeira pessoa	2,47 *(2,62)*
PPO2	Pronome pessoal oblíquo na segunda pessoa	1,64 *(2,28)*
PPO3	Pronome pessoal oblíquo na terceira pessoa	7,29 *(7,08)*
PPT	Pronome pessoal de tratamento	2,64 *(3,55)*
PD	Pronomes demonstrativos	14,64 *(9,60)*
PRV	Pronomes relativos variáveis	0,47 *(0,87)*
PINT	Pronomes interrogativos	1,64 *(1,80)*
PIP	Pronomes indefinidos referente à pessoa	0,23 *(0,56)*
PIL	Pronomes indefinidos referente a lugar	1,22 *(2,52)*
PIC	Pronomes indefinidos referente a coisas	9,29 *(7,63)*

Tabela 34: Média de ocorrência de pronomes no gênero *termo*

De acordo com a Tabela 34, os pronomes que mais ocorrem são o demonstrativo, indefinido referente a coisas e pessoal oblíquo na terceira pessoa.

A Tabela 35, a seguir, refere-se ao gênero notícia.

Gênero notícia

PPO1	Pronome pessoal oblíquo na primeira pessoa	13,9 *(17,91)*
PPO2	Pronome pessoal oblíquo na segunda pessoa	0,2 *(0,41)*
PPO3	Pronome pessoal oblíquo na terceira pessoa	9,9 *(9,34)*
PPT	Pronome pessoal de tratamento	4,2 *(11,93)*
PD	Pronomes demonstrativos	38,2 *(39,28)*
PRV	Pronomes relativos variáveis	4,05 *(9,11)*
PINT	Pronomes interrogativos	6,05 *(8,50)*
PIP	Pronomes indefinidos referente à pessoa	0,75 *(1,48)*
PIL	Pronomes indefinidos referente a lugar	9,2 *(8,78)*
PIC	Pronomes indefinidos referente a coisas	31,15 *(26,99)*

Tabela 35: Média de ocorrência de pronomes no gênero notícia

Com base na Tabela 35, os pronomes que mais ocorrem nesse gênero são os demonstrativos, indefinidos referente a coisas e pessoal oblíquo em primeira pessoa.

A seguir, apresenta-se o gráfico 8 que compara a ocorrência de todos os pronomes em cada gênero.

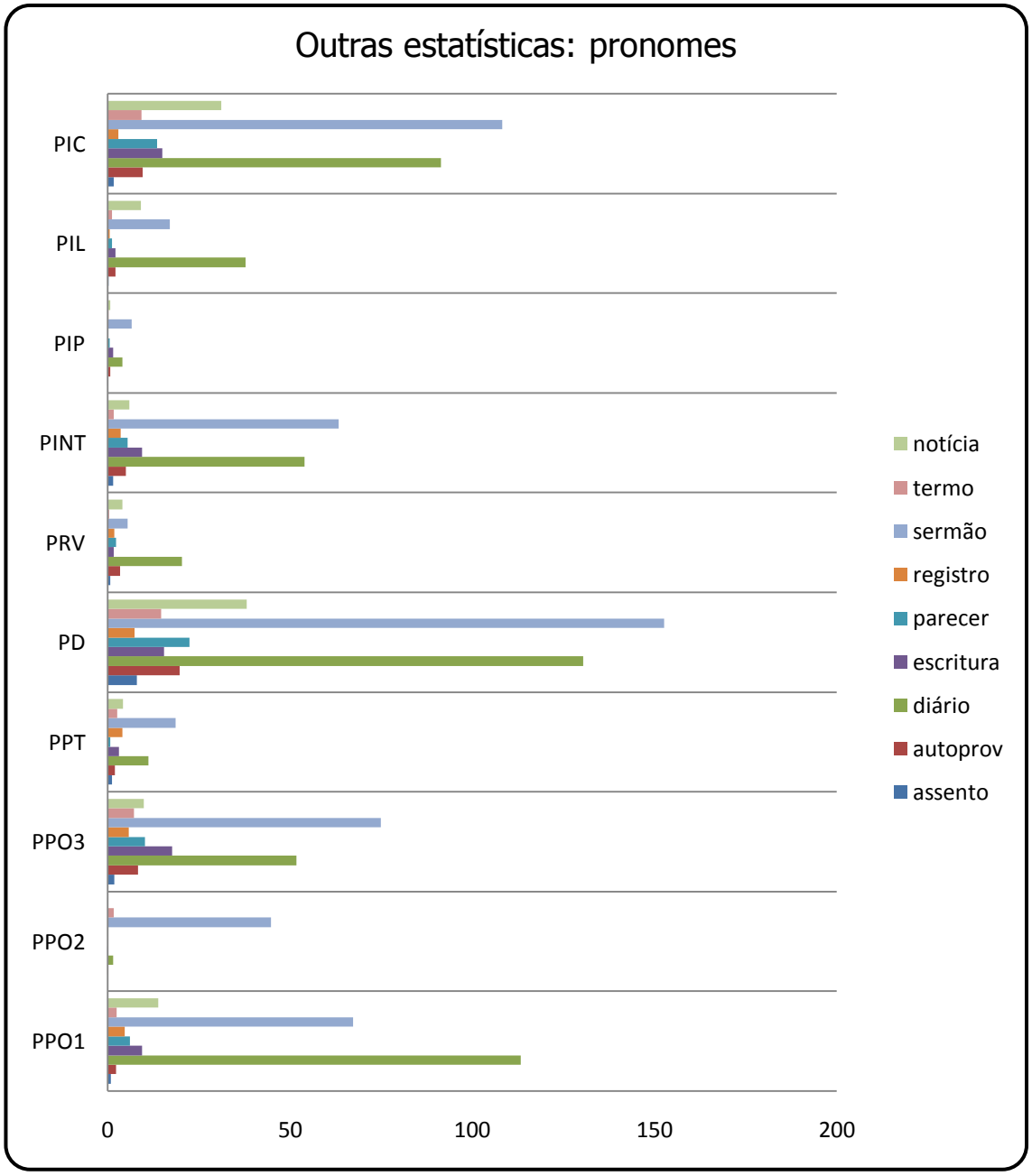

Gráfico 8: Outras estatísticas: pronomes

Conforme o gráfico 8, constata-se que os pronomes ocorrem com maior frequência nos diários, sermão e notícia. Além disso, os pronomes que mais ocorrem são: pronome demonstrativo, indefinido referente a coisas e pessoal em primeira e terceira pessoas.

A Tabela 36, a seguir, apresenta uma comparação entre os gêneros e os pronomes.

	PPO1	**PPO2**	**PPO3**	**PPT**	**PD**	**PRV**	**PINT**	**PIP**	**PIL**	**PIC**
Assento	0,87	0	1,81	1,18	8,06	0,68	1,5	0,06	0,25	1,68

Autoprov	2,4	0,13	8,4	2	19,73	3,53	5,06	0,8	2,13	9,6
Diário	113,4	1,47	51,76	11,2	130,5	20,47	54,05	4,11	37,9	91,47
Escritura	9,55	0,2	17,65	3,1	15,55	1,75	9,55	1,5	2,2	15,05
Parecer	6,21	0	10,28	0,78	22,5	2,42	5,5	0,64	1,21	13,57
Registro	4,8	0,2	5,9	4,15	7,35	1,85	3,6	0	0,65	2,9
Sermão	67,3	44,8	74,95	18,7	152,7	5,55	63,35	6,55	17,1	108,3
Termo	2,47	1,64	7,29	2,64	14,64	0,47	1,64	0,23	1,22	9,29
Notícia	13,9	0,2	9,9	4,2	38,2	4,05	6,05	0,75	9,2	31,15

Tabela 36: Comparação da média de ocorrência de pronomes nos gêneros

5.2.3 Outras estatísticas: advérbios

Na Tabela 37, a seguir, são apresentados os dados para cada gênero, contendo a descrição de traços referente a advérbios.

ASSENTO		
Traços	Descrição	Média
ADVL	Advérbio de lugar	1,68 *(1,25)*
ADVT	Advérbio de tempo	1,81 *(2,19)*
ADVI	Advérbio de intensidade	5,06 *(5,32)*
AUTO DE PROVIMENTO		
Traços	Descrição	Média
ADVL	Advérbio de lugar	2,93 *(3,36)*
ADVT	Advérbio de tempo	1,86 *(1,24)*
ADVI	Advérbio de intensidade	20,1 *(8,57)*
DIÁRIO		
Traços	Descrição	Média
ADVL	Advérbio de lugar	80,64 *(43,43)*
ADVT	Advérbio de tempo	82,70 *(41,96)*
ADVI	Advérbio de intensidade	201,47 *(104,23)*
ESCRITURA		
Traços	Descrição	Média
ADVL	Advérbio de lugar	7,1 *(4,45)*
ADVT	Advérbio de tempo	8,9 *(4,02)*
ADVI	Advérbio de intensidade	16,1 *(10,66)*
PARECER		
Traços	Descrição	Média
ADVL	Advérbio de lugar	2,07 *(3,40)*
ADVT	Advérbio de tempo	7,57 *(12,15)*
ADVI	Advérbio de intensidade	18,5 *(21,27)*
REGISTRO		
Traços	Descrição	Média
ADVL	Advérbio de lugar	3,65 *(1,30)*

ADVT	Advérbio de tempo	0,4 *(0,75)*
ADVI	Advérbio de intensidade	6,75 *(2,24)*
SERMÃO		
Traços	Descrição	Média
ADVL	Advérbio de lugar	31,85 *(15,33)*
ADVT	Advérbio de tempo	60,55 *(21,79)*
ADVI	Advérbio de intensidade	225,9 *(80,78)*
TERMO		
Traços	Descrição	Média
ADVL	Advérbio de lugar	2,35 *(1,86)*
ADVT	Advérbio de tempo	1,58 *(1,69)*
ADVI	Advérbio de intensidade	11,70 *(9,31)*
NOTÍCIA		
Traços	Descrição	Média
ADVL	Advérbio de lugar	15,25 *(17,79)*
ADVT	Advérbio de tempo	9,85 *(7,22)*
ADVI	Advérbio de intensidade	52,7 *(43,44)*

Tabela 37: Média de ocorrência de advérbios nos gêneros

De acordo com a Tabela 37, constata-se que em todos os gêneros, a frequência do advérbio de intensidade é muito maior que os outros tipos. Nos gêneros parecer e sermão, o advérbio de tempo (ADVT) ainda é mais frequente que o advérbio de lugar (ADVL).

A Tabela 38, a seguir, compara os gêneros e a frequência dos respectivos advérbios. Os mesmos dados estão explícitos no gráfico 9 para melhor visualização.

	ADVL	**ADVT**	**ADVI**
Assento	1,68	1,81	5,06
Autoprov	2,93	1,86	20,06
Diário	80,64	82,7	201,5
Escritura	7,1	8,9	16,1
Parecer	2,07	7,57	18,5
Registro	3,65	0,4	6,75
Sermão	31,85	60,6	225,9
Termo	2,35	1,58	11,7
Notícia	15,25	9,85	52,7

Tabela 38: Comparação da média de ocorrência de advérbios nos gêneros

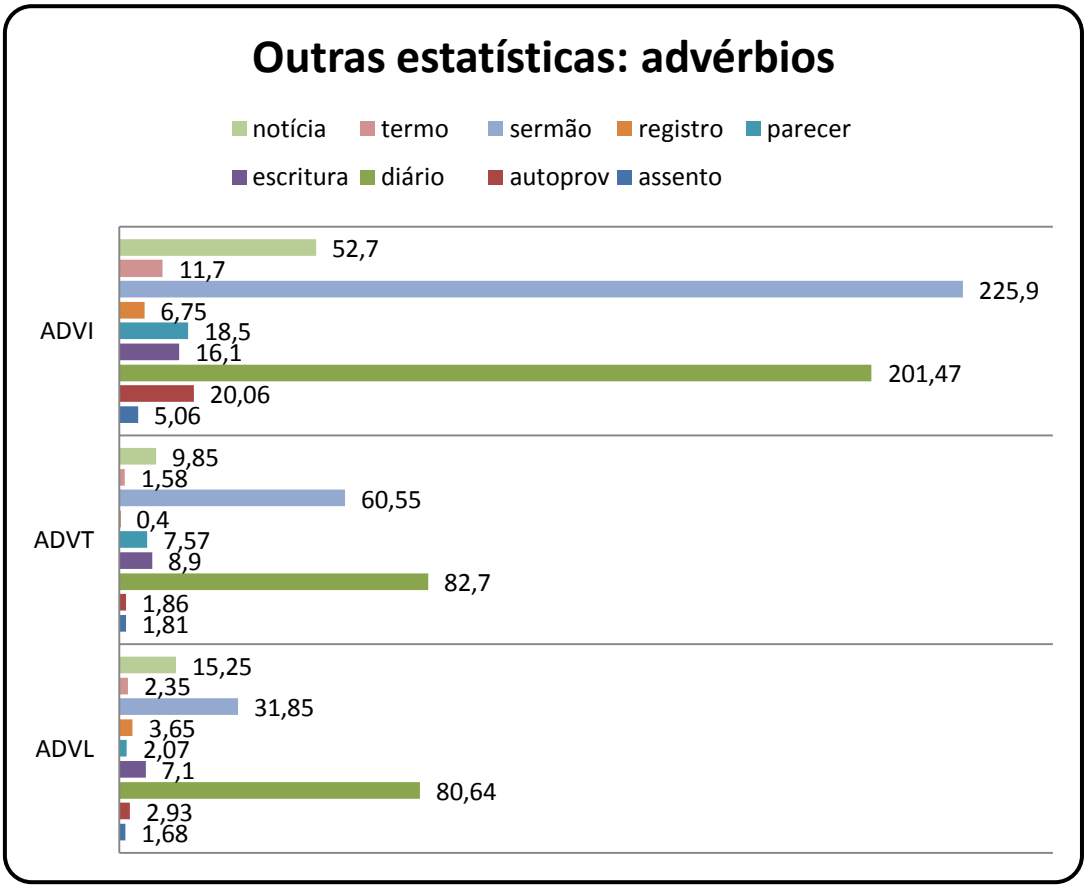

Gráfico 9: Outras estatísticas: advérbios

De acordo com a tabela 38 e o gráfico 9, verifica-se que os advérbio de lugar (ADVL) ocorrem com maior frequencia nos diários, assim como o advérbio de tempo (ADVT). Já o advérbio de intesidade (ADVI) ocorre mais no sermão.

5.2.3 Outras estatísticas: preposições, marcadores discursivos e adjetivos

Com base na lista de palavras apresentada na metodologia, os adjetivos deveriam ser predominantes em escritura e auto de provimento. Contudo, foi predominante primeiramente nos diários e em seguida nas escrituras, como se pode observar na Tabela 39 e no gráfico 10.

Gêneros	PREP	MD	ADJ
Assento	11,75	6,56	3
Autoprov.	44	19,5	5,66
Diário	412,6	257	26,47
Escritura	54,25	19,2	15,9

Parecer	36,92	35,8	0,71
Registro	20,75	7,1	1,75
Sermão	273	424	2,55
Termo	23,64	11,8	5,35
Notícia	108,2	70,5	3,2

Tabela 39: Comparação: preposições, marcadores discursivos e adjetivos

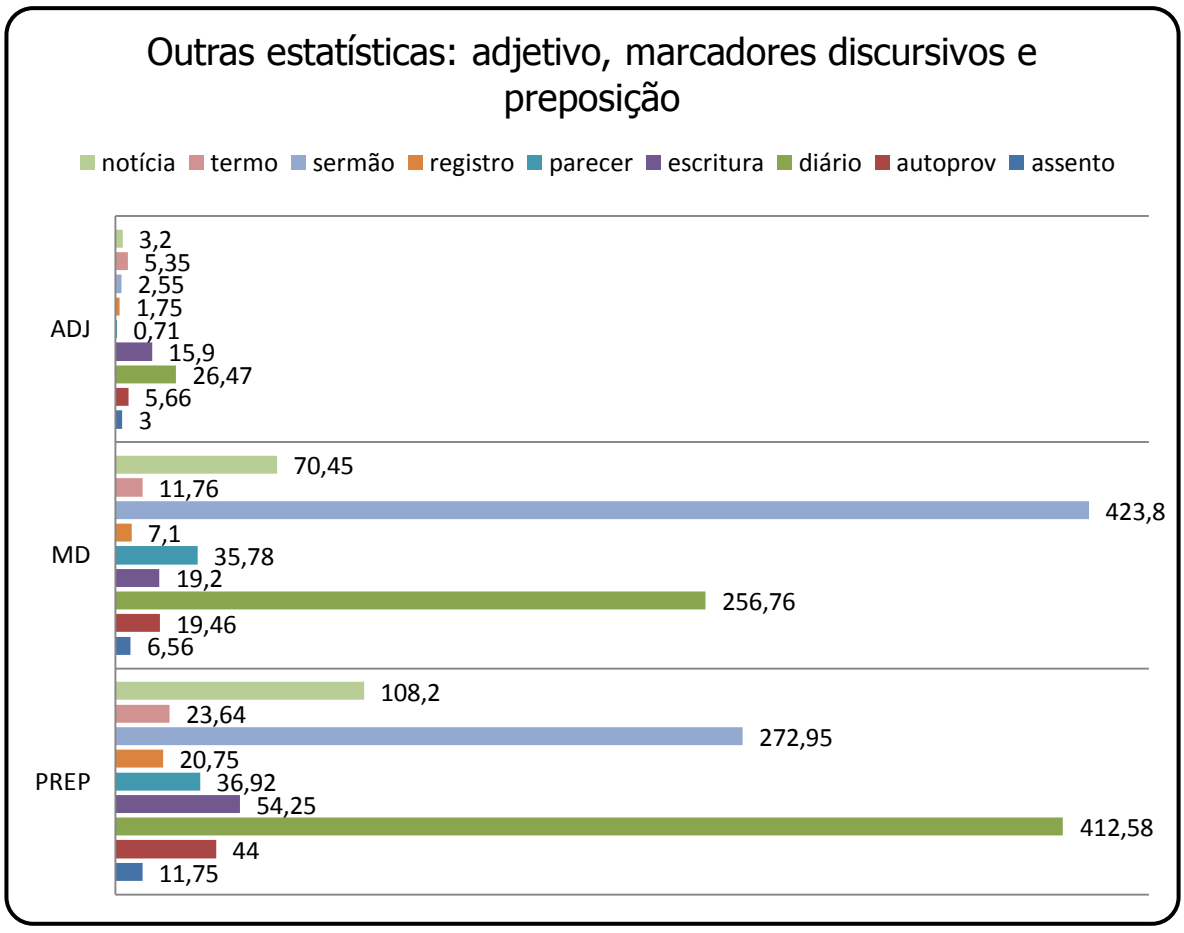

Gráfico 10: Outras estatísticas: adjetivo, marcadores discursos e preposição

De acordo com o gráfico 10, verifica-se a ocorrência das preposições é maior nos gêneros cujos textos são mais extensos, como nos diários, sermões e notícias. O mesmo ocorre com os marcadores discursivos.

5.2.5 Outras estatísticas: expressões

A seguir, são apresentadas tabelas para cada gênero, contendo a descrição de traços referente a expressões.

A Tabela 40, a seguir, refere-se ao gênero assento.

Gênero assento

Siglas	Descrição	Média
EXD	Expressão *"Deus guarde"*	0,06 *(0,25)*
EXFS	Expressão *"faço saber"*	0 *(0)*
EXCM	Expressão *"capitão-mor"*	0 *(0)*
EXL	Expressões em latim	0 *(0)*
EXPr	Expressão *"pregado"*	0 *(0)*
EXO	Expressão *"oficiais da câmara"*	0,56 *(1,03)*
EXA	Expressão *"ano de nascimento"*	0 *(0)*
EXEs	Expressão *"o escrevi"*	0,56 *(1,36)*
EXAt	Expressão *"atas da câmara"*	0,43 *(0,26)*
EXDo	Expressão *"Doutor ouvidor geral"*	0 *(0)*
EXT	Expressão *"termo e certifico"*	1,75 *(0,85)*
EXPI	Expressão *"público instrumento"*	0,93 *(0,77)*

Tabela 40: Média de ocorrência das expressões no gênero assento

De acordo com a Tabela 40, a única expressão que deve ocorrer em textos desse gênero é a expressão *"termo e certifico"* (EXT), cuja média é maior que um.

A Tabela 41, a seguir, refere-se ao auto de provimento.

Gênero auto de provimento

EXD	Expressão *"Deus guarde"*	0,33 *(0,61)*
EXFS	Expressão *"faço saber"*	0 *(0)*
EXCM	Expressão *"capitão-mor"*	0 *(0)*
EXL	Expressões em latim	0 *(0)*
EXPr	Expressão *"pregado"*	0 *(0)*
EXO	Expressão *"oficiais da câmara"*	2,06 *(1,90)*
EXA	Expressão *"ano de nascimento"*	0,93 *(0,25)*
EXEs	Expressão *"o escrevi"*	17,5 *(7,37)*
EXAt	Expressão *"atas da câmara"*	5,86 *(3,96)*
EXDo	Expressão *"Doutor ouvidor geral"*	1,53 *(1,18)*
EXT	Expressão *"termo e certifico"*	9,2 *(3,25)*
EXPI	Expressão *"público instrumento"*	0 *(0)*

Tabela 41: Média de ocorrência das expressões no gênero auto de provimento

Na Tabela 41, é possível constatar a predominância de algumas expressões, tais como: *"o escrevi"* (EXEs), *"termo e certifico"* (EXT) e *"atas da câmara"* (EXAt).

A Tabela 42, a seguir, refere-se ao gênero diário.

Gênero diário

EXD	Expressão *"Deus guarde"*	0,82 *(1,66)*
EXFS	Expressão *"faço saber"*	0,11 *(0,33)*
EXCM	Expressão *"capitão-mor"*	0,64 *(2,42)*
EXL	Expressões em latim	0 *(0)*
EXPr	Expressão *"pregado"*	0 *(0)*
EXO	Expressão *"oficiais da câmara"*	0 *(0)*
EXA	Expressão *"ano de nascimento"*	0 *(0)*
EXEs	Expressão *"o escrevi"*	1,05 *(1,81)*
EXAt	Expressão *"atas da câmara"*	0,17 *(0,39)*
EXDo	Expressão *"Doutor ouvidor geral"*	0,41 *(0,79)*
EXT	Expressão *"termo e certifico"*	0,64 *(1,32)*
EXPI	Expressão *"público instrumento"*	0 *(0)*

Tabela 42: Média de ocorrência das expressões no gênero diário

Com base na Tabela 42, nesse gênero pode ocorrer apenas uma das expressões elencadas, a expressão "o escrevi" (EXEs).

A Tabela 43, a seguir, refere-se ao gênero escritura.

Gênero escritura

EXD	Expressão *"Deus guarde"*	0 *(0)*
EXFS	Expressão *"faço saber"*	0,1 *(0,30)*
EXCM	Expressão *"capitão-mor"*	0 *(0)*
EXL	Expressões em latim	0 *(0)*
EXPr	Expressão *"pregado"*	0 *(0)*
EXO	Expressão *"oficiais da câmara"*	0 *(0)*
EXA	Expressão *"ano de nascimento"*	0,15 *(0,48)*
EXEs	Expressão *"o escrevi"*	0,45 *(0,99)*
EXAt	Expressão *"atas da câmara"*	1,25 *(1,25)*
EXDo	Expressão *"Doutor ouvidor geral"*	5,2 *(2,50)*
EXT	Expressão *"termo e certifico"*	5,8 *(3,15)*
EXPI	Expressão *"público instrumento"*	0 *(0)*

Tabela 43: Média de ocorrência das expressões no gênero escritura

De acordo com a Tabela 43, nesse gênero as expressões que mais ocorrem são *"termo e certifico"* (EXT) e *"Doutor ouvidor geral"* (EXDo). Também deve ocorrer pelo menos uma vez em cada texto desse gênero a expressão *"atas da câmara"* (EXAt).

A Tabela 44, a seguir, refere-se ao gênero parecer.

Gênero parecer

EXD	Expressão *"Deus guarde"*	0,14 *(0,36)*
EXFS	Expressão *"faço saber"*	0 *(0)*
EXCM	Expressão *"capitão-mor"*	0 *(0)*
EXL	Expressões em latim	0 *(0)*
EXPr	Expressão *"pregado"*	0 *(0)*

EXO	Expressão *"oficiais da câmara"*	0 *(0)*
EXA	Expressão *"ano de nascimento"*	0 *(0)*
EXEs	Expressão *"o escrevi"*	0,21 *(0,57)*
EXAt	Expressão *"atas da câmara"*	0 *(0)*
EXDo	Expressão *"Doutor ouvidor geral"*	0,21 *(0,57)*
EXT	Expressão *"termo e certifico"*	0 *(0)*
EXPI	Expressão *"público instrumento"*	0 *(0)*

Tabela 44: Média de ocorrência das expressões no gênero parecer

De acordo com a Tabela 44, nenhuma expressão é recorrente nesse gênero.

A Tabela 45 refere-se ao gênero registro.

Gênero registro

EXD	Expressão *"Deus guarde"*	2,45 *(2,35)*
EXFS	Expressão *"faço saber"*	0,7 *(0,47)*
EXCM	Expressão *"capitão-mor"*	2,5 *(1,39)*
EXL	Expressões em latim	0 *(0)*
EXPr	Expressão *"pregado"*	0 *(0)*
EXO	Expressão *"oficiais da câmara"*	0,75 *(1,06)*
EXA	Expressão *"ano de nascimento"*	0,45 *(0,51)*
EXEs	Expressão *"o escrevi"*	0,15 *(0,67)*
EXAt	Expressão *"atas da câmara"*	0,95 *(0,88)*
EXDo	Expressão *"Doutor ouvidor geral"*	0 *(0)*
EXT	Expressão *"termo e certifico"*	2,75 *(2,04)*
EXPI	Expressão *"público instrumento"*	0 *(0)*

Tabela 45: Média de ocorrência das expressões no gênero registro

A Tabela 45 permite verificar que as expressões que ocorrem em textos desse gênero são *"Deus guarde"* (EXD), *"capitão-mor"* (EXCM) e *"termo e certifico"* (EXT).

A Tabela 46, a seguir, refere-se ao gênero sermão.

Gênero sermão

EXD	Expressão *"Deus guarde"*	0,15 *(0,36)*
EXFS	Expressão *"faço saber"*	0 *(0)*
EXCM	Expressão *"capitão-mor"*	0 *(0)*
EXL	Expressões em latim	5,5 *(5,19)*
EXPr	Expressão *"pregado"*	0,9 *(0,78)*
EXO	Expressão *"oficiais da câmara"*	0 *(0)*
EXA	Expressão *"ano de nascimento"*	0 *(0)*
EXEs	Expressão *"o escrevi"*	0,5 *(0,82)*
EXAt	Expressão *"atas da câmara"*	0,4 *(0,82)*
EXDo	Expressão *"Doutor ouvidor geral"*	0,6 *(0,68)*
EXT	Expressão *"termo e certifico"*	0,1 *(0,30)*
EXPI	Expressão *"público instrumento"*	0 *(0)*

Tabela 46: Média de ocorrência das expressões no gênero sermão

Com base na Tabela 46, conclui-se que o traço mais recorrente em textos do gênero sermão são as expressões em latim. A expressão *"pregado"* (EXPr) deve ocorrer ao menos uma vez em cada texto desse gênero, contudo, talvez pela variação da expressão, a média não foi obtida.

A Tabela 47, a seguir, refere-se ao gênero *termo*.

Gênero *termo*

EXD	Expressão *"Deus guarde"*	0 *(0)*
EXFS	Expressão *"faço saber"*	0 *(0)*
EXCM	Expressão *"capitão-mor"*	0 *(0)*
EXL	Expressões em latim	0 *(0)*
EXPr	Expressão *"pregado"*	0 *(0)*
EXO	Expressão *"oficiais da câmara"*	0 *(0)*
EXA	Expressão *"ano de nascimento"*	0 *(0)*
EXEs	Expressão *"o escrevi"*	0 *(0)*
EXAt	Expressão *"atas da câmara"*	2,35 *(1,16)*
EXDo	Expressão *"Doutor ouvidor geral"*	0,11 *(0,33)*
EXT	Expressão *"termo e certifico"*	0,7 *(0,77)*
EXPI	Expressão *"público instrumento"*	1,76 *(0,66)*

Tabela 47: Média de ocorrência das expressões no gênero *termo*

De acordo com a Tabela 47, as expressões que ocorrem pelo menos uma vez nos textos desse gênero são *"atas da câmara"* (EXAt) e *"público instrumento"* (EXPI).

A Tabela 48, a seguir, refere-se ao gênero notícia.

Gênero notícia

EXD	Expressão *"Deus guarde"*	0,05 *(0,22)*
EXFS	Expressão *"faço saber"*	0 *(0)*
EXCM	Expressão *"capitão-mor"*	0,3 *(1,34)*
EXL	Expressões em latim	0 *(0)*
EXPr	Expressão *"pregado"*	0 *(0)*
EXO	Expressão *"oficiais da câmara"*	0 *(0)*
EXA	Expressão *"ano de nascimento"*	0 *(0)*
EXEs	Expressão *"o escrevi"*	0,9 *(3,21)*
EXAt	Expressão *"atas da câmara"*	0,25 *(0,63)*
EXDo	Expressão *"Doutor ouvidor geral"*	0,4 *(0,94)*
EXT	Expressão *"termo e certifico"*	0 *(0)*
EXPI	Expressão *"público instrumento"*	0 *(0)*

Tabela 48: Média de ocorrência das expressões no gênero notícia

Nesse gênero, poucas expressões foram recuperáveis. Ainda assim, a que ocorreu com mais predominância foi a expressão *"o escrevi"* (EXEs), talvez por se tratar de uma notícia, descrevendo os fatos e tudo que se descobria no país, portanto, cabia uma assinatura.

O gráfico 11 esclarece a diferença na ocorrência e predominância das expressões em cada gênero. A expressão *"Deus guarde"* (EXD) é predominante em registro, tal como a expressão *"capitão-mor"* (EXCM). Além disso, expressões em latim (EXL), como previsto, ocorre apenas em textos do gênero sermão.

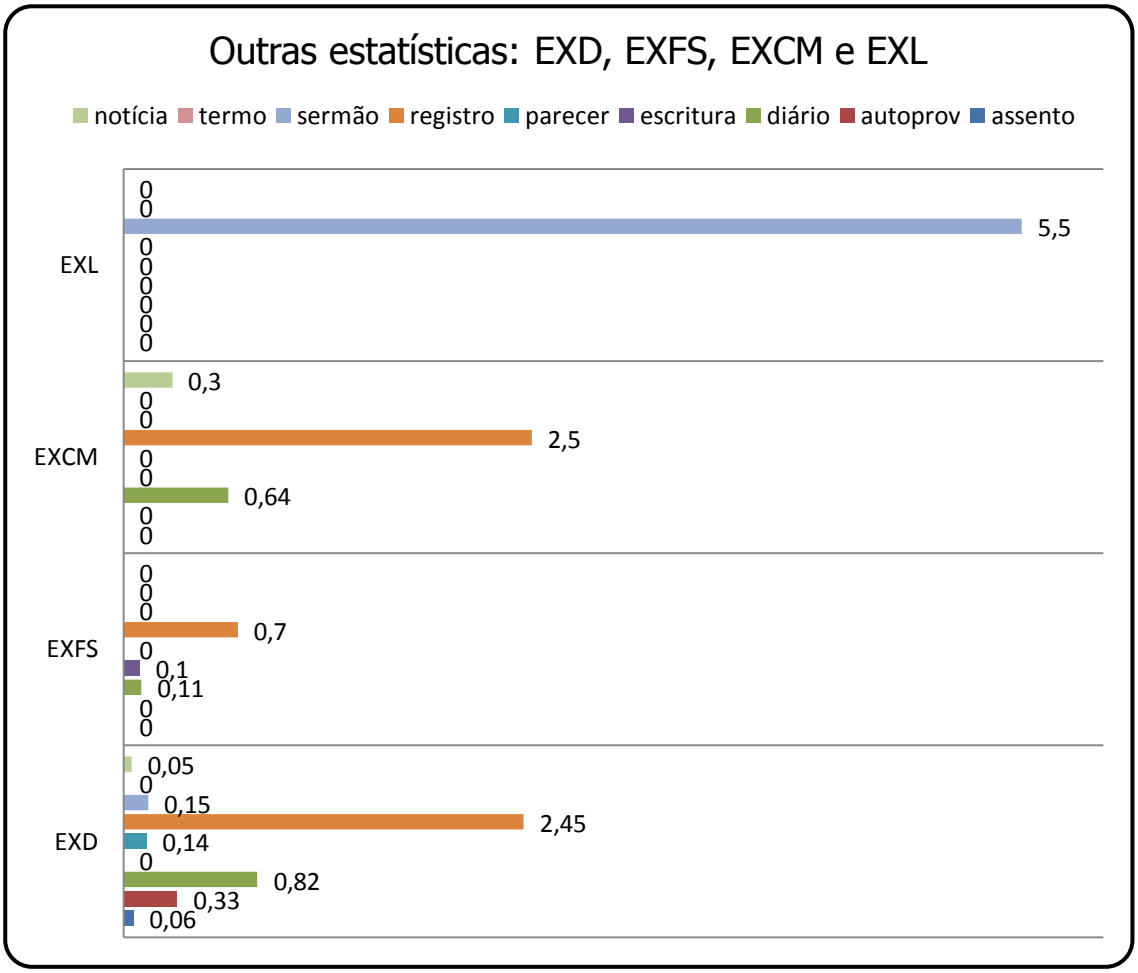

Gráfico 11: Outras estatísticas: EXD, EXFS, EXCM e EXL

Na sequência, apresenta-se o gráfico 12, contendo outras expressões.

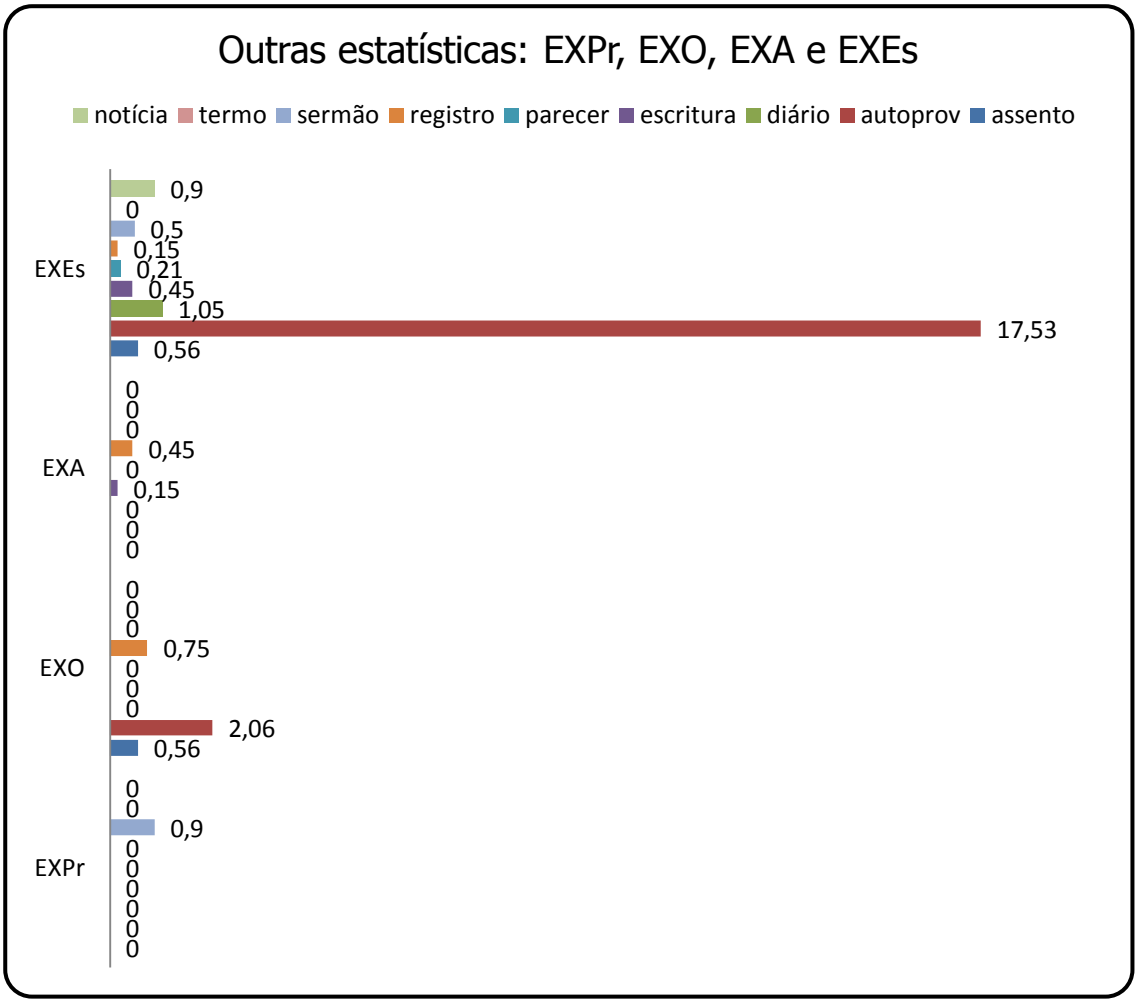

Gráfico 12: Outras estatísticas: EXPr, EXO, EXA e EXEs

De acordo com o gráfico 12, verifica-se que a expressão *"pregado"* (EXPr) ocorria apenas em textos do gênero sermão. Já a expressão *"oficiais da câmara"* (EXO) ocorre mais em textos do gênero auto de provimento, assim como a expressão *"o escrevi"* (EXEs). A expressão "ano de nascimento" (EXA) não apresentou bons resultados, dado sua baixa ocorrência.

O gráfico 13, a seguir, apresenta as expressões *"atas da câmara"* (EXAt), *"doutor ouvidor"* (EXDo), *"termo e certifico"* (EXT) e *"público instrumento"* (EXPI).

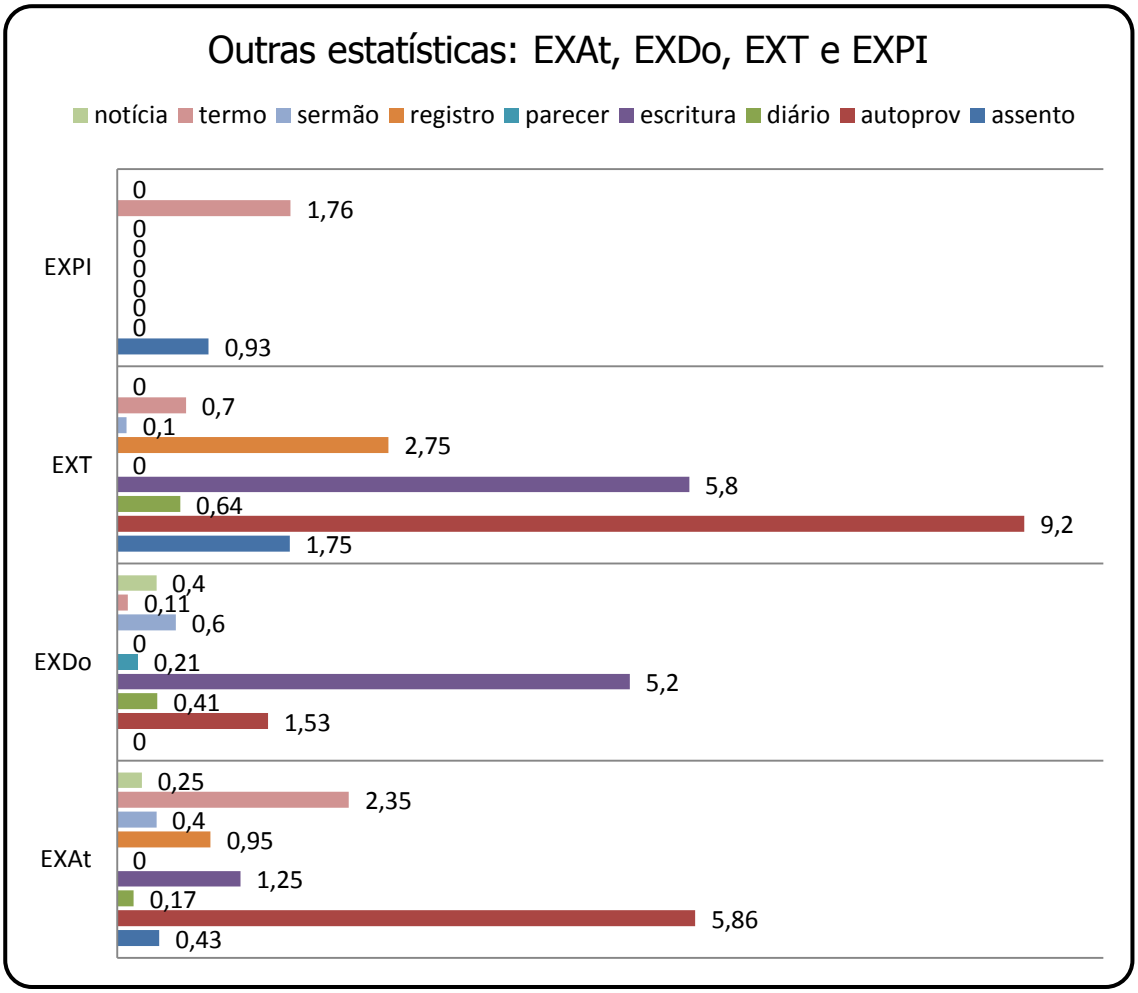

Gráfico 13: Outras estatísticas: EXAt, EXDo, EXT e EXPI

No gráfico 13, constata-se que a expressão *"atas da câmara"* (EXAt) é predominante em textos do gênero auto de provimento, assim como a expressão *"termo e certifico"* (EXT). Já a expressão *"doutor ouvidor"* (EXDo) é predominante em textos do gênero escritura, e a expressão *"público instrumento"* (EXPI) deve ocorrer ao menos uma vez em textos do gênero *termo*.

A seguir, a Tabela 49 apresenta a ocorrência dos traços em cada gênero, de modo a facilitar a visualização e comparação entre os gêneros.

	EXD	**EXFS**	**EXCM**	**EXL**	**EXPr**	**EXO**	**EXA**	**EXEs**	**EXAt**	**EXDo**	**EXT**	**EXPI**
Assento	0,06	0	0	0	0	0,56	0	0,56	0,43	0	1,75	0,93
Autoprov	0,33	0	0	0	0	2,06	0.93	17,5	5,86	1,53	9,2	0
Diário	0,82	0,11	0,64	0	0	0	0	1,05	0,17	0,41	0,64	0
Escritura	0	0,1	0	0	0	0	0,15	0,45	1,25	5,2	5,8	0
Parecer	0,14	0	0	0	0	0	0	0,21	0	0,21	0	0
Registro	2,45	0,7	2,5	0	0	0,75	0,45	0,15	0,95	0	2,75	0
Sermão	0,15	0	0	5,5	0,9	0	0	0,5	0,4	0,6	0,1	0

Termo	0	0	0	0	0	0	0	0	2,35	0,11	0,7	1,76
Notícia	0,05	0	0,3	0	0	0	0	0,9	0,25	0,4	0	0

Tabela 49: Comparação da média de ocorrência das expressões nos gêneros

De acordo com a Tabela 49, verifica-se que a recuperação dessas características foi baixa (0>1), o que é justificável pela variação de grafia e outras implicações próprias do tipo de corpus. Também como previsto, muitas dessas expressões ocorrem uma única vez em um texto de cada gênero, como no caso da expressão *"pregado"* (EXPr) nos sermões, ou a expressão *"publico instrumento"* em escrituras.

Conforme mostrado na metodologia, a expressão *"deus guarde"* (EXD) ocorreu e foi predominante em registro. A expressão *"faço saber"* (EXFs), hipótese de traço do registro, também ocorreu e foi predominante no gênero, assim como a expressão *"capitão-mor"* (EXCM).

As expressões em latim (EXL) também ocorreram unicamente no sermão, como previsto e apontado durante a metodologia. A expressão *"oficiais da câmara"* (EXO), de acordo com o que foi apresentado na metodologia, deveria ocorrer e ser predominante em assento e auto de provimento, contudo, de acordo com a Tabela 49, foi recorrente e com predominância apenas em auto de provimento.

A expressão *"ano de nascimento"* (EXA) deveria no mínimo ocorrer uma única vez em escritura, registro e autos de provimento, porém, nos primeiros não ocorreu e foi predominante apenas nos autos de provimento.

A expressão *"o escrevi"* (EXEs) era para ocorrer em assento, autos de provimento e escrituras. De acordo com a Tabela 49, ela ocorreu em todos os gêneros, mas com predominância nos autos de provimento. Já a expressão *"atas da câmara"* (EXAt), segundo o que foi apresentado na metodologia, deveria ocorrer em registro e *termo*. Ela ocorreu, mas com predominância nos autos de provimento, depois nos *termos*. O mesmo aconteceu com a expressão *"doutor ouvidor geral"* (EXDo), que além de ocorrer nos autos de provimento, ocorreu com predominância em escrituras.

Referente à expressão *"termo e certifico"*, de acordo com a tabela, constata-se que, como previsto, ocorreu com predominância nos autos de provimento, mas também em escrituras e nos demais gêneros, mas com baixa proporção.

Por último, a expressão *"público instrumento"* (EXPI) estava prevista para ocorrer em escrituras, o não aconteceu. Ela ocorreu em assento e *termo*, sendo que neste último ela foi predominante.

A seguir, apresentam-se os resultados da descrição referentes a unidades lexicais.

5.2.6 Outras estatísticas: unidades lexicais

A seguir, apresenta-se a Tabela 50 do gênero assento, contendo a média de ocorrência dos traços nos textos do respectivo gênero.

Gênero assento

Traços	Descrição	Média
ULE	Unidades lexicais referente a pontos cardeais	0 *(0)*
ULD	Unidade lexical *"dia"*	0,12 *(0,34)*
ULL	Unidade lexical *"légua"*	0 *(0)*
ULS	Unidade lexical *"suplicante"*	0 *(0)*
ULQ	Unidades lexicais *"que"* e *"se"*	20,18 *(12,91)*
ULT	Unidade lexical *"testemunha"*	0 *(0)*
ULJ	Unidade lexical *"juiz"*	0,18 *(0,40)*
ULSa	Unidades lexicais referente ao contexto sacro	0,18 *(0,75)*
ULMA	Unidades lexicais referente ao meio ambiente	0,31 *(0,87)*
ULTr	Unidades lexicais referente a territórios	5,68 *(3,77)*
ULP	Unidades lexicais referente a pessoas	2,37 *(1,31)*
ULDv	Unidade lexical *"devassa"*	0 *(0)*

Tabela 50: Média de ocorrência das unidades lexicais no gênero assento

Nesse gênero, as unidades lexicais mais recorrentes são referentes a territórios (ULTr) e pessoas (ULP).

A seguir, apresenta-se a Tabela 51 contendo dados do gênero auto de provimento.

Gênero auto de provimento

Traços	Descrição	Média
ULE	Unidades lexicais referente a pontos cardeais	0 (0)
ULD	Unidade lexical *"dia"*	0,4 (0,91)
ULL	Unidade lexical *"légua"*	0 (0)
ULS	Unidade lexical *"suplicante"*	0 (0)
ULQ	Unidades lexicais *"que"* e *"se"*	67,13 (31,97)
ULT	Unidade lexical *"testemunha"*	0,06 (0,25)
ULJ	Unidade lexical *"juiz"*	1,5 (1,88)
ULSa	Unidades lexicais referente ao contexto sacro	2 (1,73)
ULMA	Unidades lexicais referente ao meio ambiente	1,73 (1,70)
ULTr	Unidades lexicais referente a territórios	12,9 (7,25)
ULP	Unidades lexicais referente a pessoas	1,8 (1,61)
ULDv	Unidade lexical *"devassa"*	0,26 (0,70)

Tabela 51: Média de ocorrência das unidades lexicais no gênero auto de provimento

A partir da Tabela 51 acima, verifica-se que as unidades lexicais recorrentes em textos desse gênero são os referente a territórios (ULTr).

A seguir, apresenta-se a Tabela 52, contendo dados do gênero diário.

Gênero diário

Traços	Descrição	Média
ULE	Unidades lexicais referente a pontos cardeais	62,5 *(120,50)*
ULD	Unidade lexical *"dia"*	73,2 *(66,84)*
ULL	Unidade lexical *"légua"*	38,94 *(52,92)*
ULS	Unidade lexical *"suplicante"*	0 *(0)*
ULQ	Unidades lexicais *"que"* e *"se"*	583,41 *(258,72)*
ULT	Unidade lexical *"testemunha"*	0,11 *(0,33)*
ULJ	Unidade lexical *"juiz"*	0,6 *(1,36)*
ULSa	Unidades lexicais referente ao contexto sacro	6,76 *(6,08)*
ULMA	Unidades lexicais referente ao meio ambiente	170, 23 *(137,52)*
ULTr	Unidades lexicais referente a territórios	47,47 *(50,52)*
ULP	Unidades lexicais referente a pessoas	13,05 *(11,16)*
ULDv	Unidade lexical *"devassa"*	0 *(0)*

Tabela 52: Média de ocorrência das unidades lexicais no gênero diário

De acordo com a Tabela 52, as unidades lexicais mais recorrentes em textos desse gênero são: unidades lexicais referente ao meio ambiente (ULMA), a unidade lexical *"dia"* (ULD), unidades referentes a território (ULTr) e pontos cardeais (ULE) e a unidade lexical *"légua"*.

A seguir, apresenta-se a Tabela 53, contendo dados do gênero escritura.

Gênero escritura

Traços	Descrição	Média
ULE	Unidades lexicais referente a pontos cardeais	0,5 *(0,94)*
ULD	Unidade lexical *"dia"*	1,1 *(1,41)*
ULL	Unidade lexical *"légua"*	0,0,5 *(0,22)*
ULS	Unidade lexical *"suplicante"*	0,3 *(0,80)*
ULQ	Unidades lexicais *"que"* e *"se"*	60,55 *(34,50)*
ULT	Unidade lexical *"testemunha"*	0,15 *(0,48)*
ULJ	Unidade lexical *"juiz"*	1,15 *(2,03)*
ULSa	Unidades lexicais referente ao contexto sacro	1,15 *(1,95)*
ULMA	Unidades lexicais referente ao meio ambiente	0,65 *(1,03)*
ULTr	Unidades lexicais referente a territórios	12,9 *(5,79)*
ULP	Unidades lexicais referente a pessoas	0,1 *(0,44)*
ULDv	Unidade lexical *"devassa"*	0 *(0)*

Tabela 53: Média de ocorrência das unidades lexicais no gênero escritura

Com base na Tabela 53, a unidade lexical mais ocorrente em textos desse gênero são as unidades lexicais referentes a territórios (ULTr).

A seguir, apresenta-se a Tabela 54, contendo dados do gênero parecer.

Gênero parecer

Traços	Descrição	Média
ULE	Unidades lexicais referente a pontos cardeais	0,64 *(1,64)*

ULD	Unidade lexical *"dia"*	0,35 *(0,63)*
ULL	Unidade lexical *"légua"*	0,28 *(1,06)*
ULS	Unidade lexical *"suplicante"*	0 *(0)*
ULQ	Unidades lexicais *"que"* e *"se"*	78,21 *(70, 05)*
ULT	Unidade lexical *"testemunha"*	1,85 *(3,79)*
ULJ	Unidade lexical *"juiz"*	0 *(0)*
ULSa	Unidades lexicais referente ao contexto sacro	1,57 *(2,40)*
ULMA	Unidades lexicais referente ao meio ambiente	7,07 *(5,18)*
ULTr	Unidades lexicais referente a territórios	3,71 *(5,39)*
ULP	Unidades lexicais referente a pessoas	2,85 *(1,46)*
ULDv	Unidade lexical *"devassa"*	2,35 *(2,46)*

Tabela 54: Média de ocorrência das unidades lexicais no gênero parecer

De acordo com a Tabela 54, verifica-se que as unidades lexicais mais frequentes em textos do gênero parecer são as referente ao meio ambiente (ULMA) e a territórios (ULTr).

A seguir, apresenta-se a Tabela 55, contendo dados do gênero registro.

Gênero registro

Traços	Descrição	Média
ULE	Unidades lexicais referente a pontos cardeais	0,25 *(0,71)*
ULD	Unidade lexical *"dia"*	0,15 *(0,36)*
ULL	Unidade lexical *"légua"*	3,05 *(1,60)*
ULS	Unidade lexical *"suplicante"*	4 *(1,52)*
ULQ	Unidades lexicais *"que"* e *"se"*	23,5 *(8,46)*
ULT	Unidade lexical *"testemunha"*	0 *(0)*
ULJ	Unidade lexical *"juiz"*	0,05 *(0,22)*
ULSa	Unidades lexicais referente ao contexto sacro	2,15 *(2,64)*
ULMA	Unidades lexicais referente ao meio ambiente	6,55 *(2,72)*
ULTr	Unidades lexicais referente a territórios	9,55 *(3,92)*
ULP	Unidades lexicais referente a pessoas	0,2 *(0,41)*
ULDv	Unidade lexical *"devassa"*	0 *(0)*

Tabela 55: Média de ocorrência das unidades lexicais no gênero registro

A partir dos dados apresentados na Tabela 55, constata-se que as unidades lexicais mais recorrentes em textos desse gênero são referentes a territórios (ULTr), a pessoas (ULP) e ao meio ambiente (ULMA).

A seguir, apresenta-se a Tabela 56, contendo dados do gênero sermão.

Gênero sermão

Traços	Descrição	Média
ULE	Unidades lexicais referente a pontos cardeais	0,35 *(0,81)*
ULD	Unidade lexical *"dia"*	12,9 *(10,06)*
ULL	Unidade lexical *"légua"*	0,3 *(0,57)*
ULS	Unidade lexical *"suplicante"*	0 *(0)*
ULQ	Unidades lexicais *"que"* e *"se"*	577,95 *(217,37)*
ULT	Unidade lexical *"testemunha"*	0,1 *(0,30)*

ULJ	Unidade lexical *"juiz"*	0,3 *(0,65)*
ULSa	Unidades lexicais referente ao contexto sacro	107,05 *(56,96)*
ULMA	Unidades lexicais referente ao meio ambiente	25,25 *(20,44)*
ULTr	Unidades lexicais referente a territórios	36,7 *(27,36)*
ULP	Unidades lexicais referente a pessoas	12,8 *(7,82)*
ULDv	Unidade lexical *"devassa"*	0 *(0)*

Tabela 56: Média de ocorrência das unidades lexicais no gênero sermão

Com base na Tabela 56, verifica-se que as unidades lexicais mais recorrentes nos textos do gênero sermão são as referente ao contexto sacro (ULSa).

A seguir, apresenta-se a Tabela 57, contendo dados do gênero *termo*.

Gênero *termo*

Traços	Descrição	Média
ULE	Unidades lexicais referente a pontos cardeais	0 *(0)*
ULD	Unidade lexical *"dia"*	0,52 *(1,17)*
ULL	Unidade lexical *"légua"*	0 *(0)*
ULS	Unidade lexical *"suplicante"*	0,17 *(0,72)*
ULQ	Unidades lexicais *"que"* e *"se"*	61,05 *(45,27)*
ULT	Unidade lexical *"testemunha"*	0,05 *(0,24)*
ULJ	Unidade lexical *"juiz"*	0,5 *(1,00)*
ULSa	Unidades lexicais referente ao contexto sacro	0,17 *(0,39)*
ULMA	Unidades lexicais referente ao meio ambiente	3,17 *(3,33)*
ULTr	Unidades lexicais referente a territórios	5,7 *(4,19)*
ULP	Unidades lexicais referente a pessoas	0,11 *(0,48)*
ULDv	Unidade lexical *"devassa"*	0 *(0)*

Tabela 57: Média de ocorrência das unidades lexicais no gênero *termo*

De acordo com a Tabela 57, verifica-se que das unidade lexicais elencadas, as mais ocorrentes em textos desse gênero são referentes a territórios (ULTr) e ao meio ambiente (ULMA).

A seguir, apresenta-se a Tabela 58, contendo dados do gênero notícia.

Gênero notícia

Traços	Descrição	Média
ULE	Unidades lexicais referente a pontos cardeais	3 *(5,11)*
ULD	Unidade lexical *"dia"*	4,55 *(5,99)*
ULL	Unidade lexical *"légua"*	1,25 *(1,91)*
ULS	Unidade lexical *"suplicante"*	0 *(0)*
ULQ	Unidades lexicais *"que"* e *"se"*	121,7 *(120,35)*
ULT	Unidade lexical *"testemunha"*	0 *(0)*
ULJ	Unidade lexical *"juiz"*	0,05 *(0,22)*
ULSa	Unidades lexicais referente ao contexto sacro	4,65 *(6,49)*
ULMA	Unidades lexicais referente ao meio ambiente	59,6 *(50,58)*
ULTr	Unidades lexicais referente a territórios	13,15 *(8,22)*
ULP	Unidades lexicais referente a pessoas	8,75 *(13,33)*
ULDv	Unidade lexical *"devassa"*	0 *(0)*

Tabela 58: Média de ocorrência de determinadas unidades lexicais no gênero notícia

Com base na Tabela 58, constata-se que as unidades lexicais mais recorrentes em textos desse gênero são referentes a territórios (ULTr) e a pessoas (ULP).

A seguir, o gráfico 14 apresenta a ocorrência das unidades lexicais referentes a pontos cardeais e às unidades lexicais *"dia", "léguas"* e *"suplicante"*.

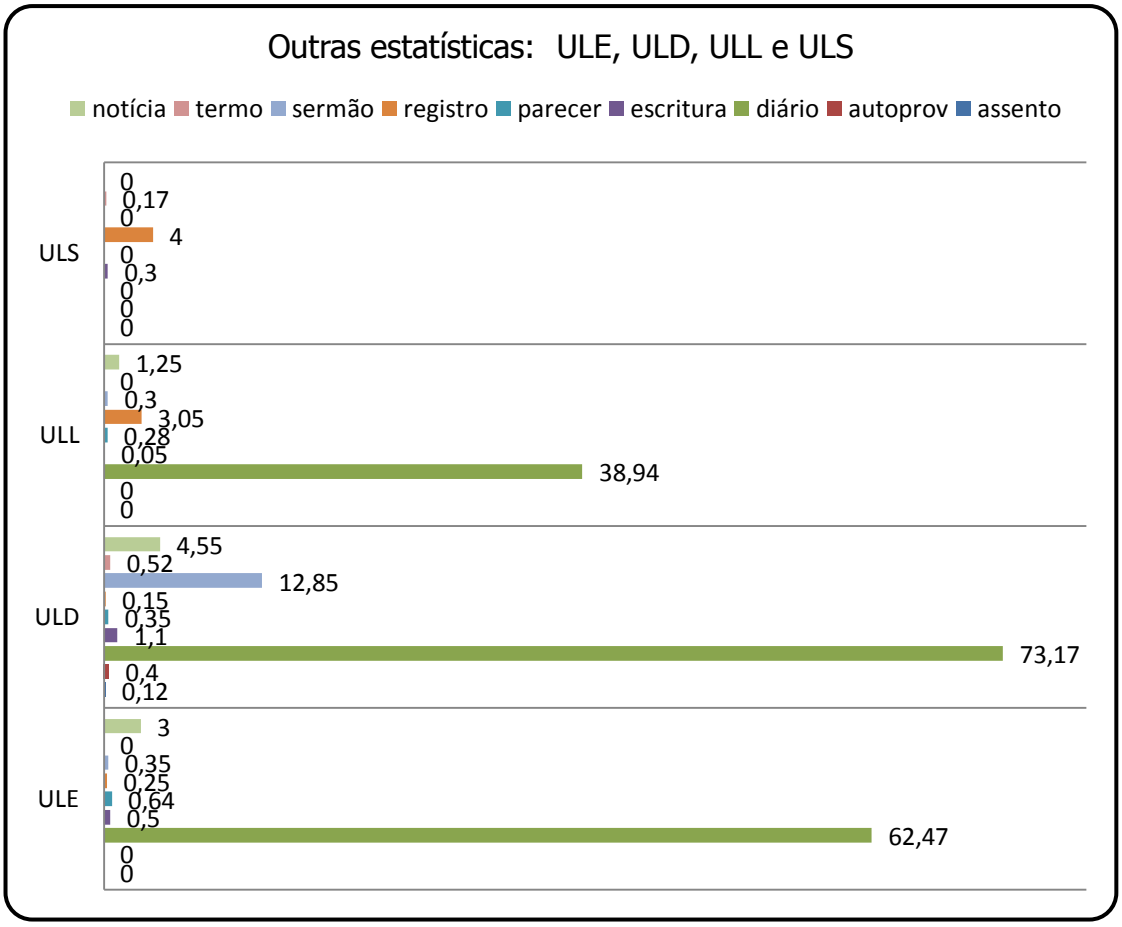

Gráfico 14: Outras estatísticas: ULE, ULD, ULL e ULS

Com base no gráfico 14, constata-se que as unidades lexicais referente a pontos cardeais e às unidades *"dia"* e *"leguas"* ocorrem com predominância em textos do gênero diário. A unidade lexical *"suplicante"* é predominante apenas em registro.

O gráfico 15 apresenta as unidades lexicais *"testemunha", "juiz"* e do contexto sacro.

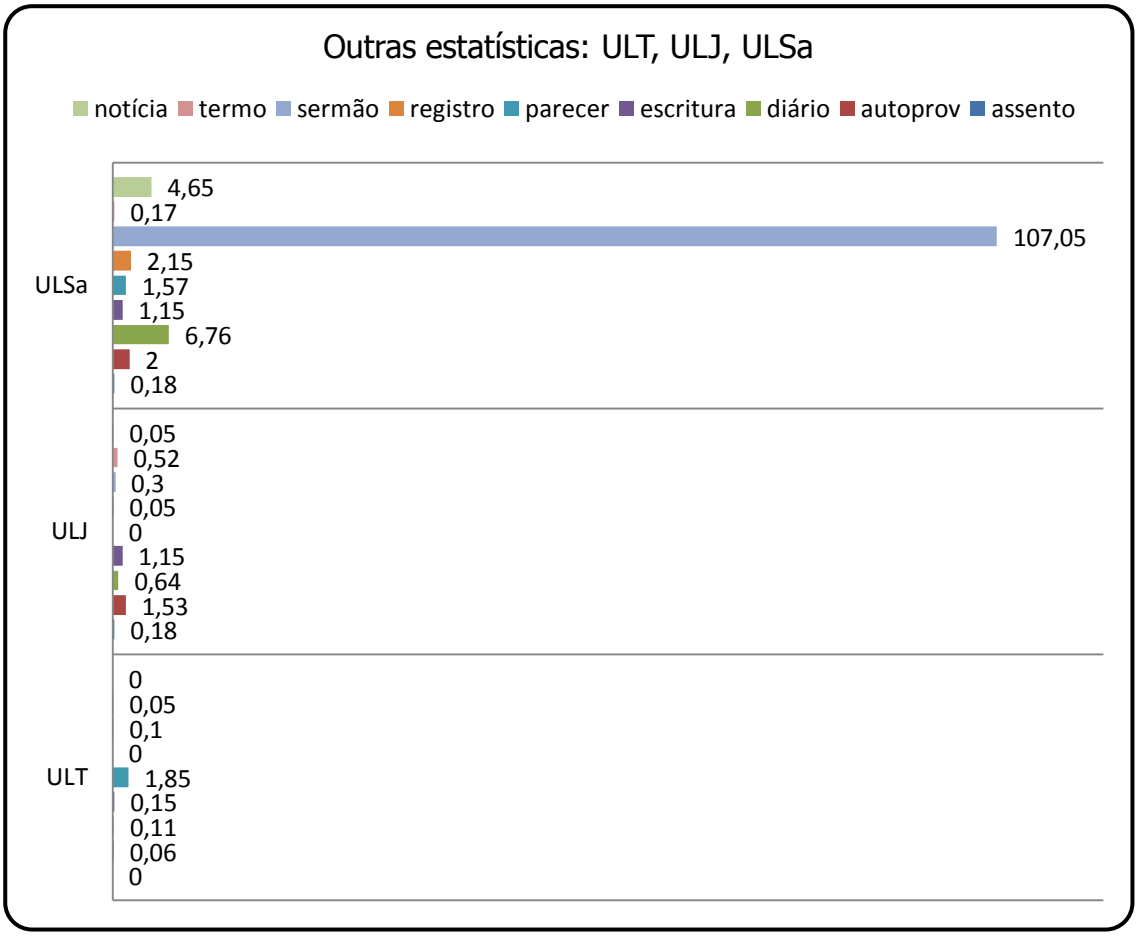

Gráfico 15: Outras estatísticas: ULT, ULJ, ULSa

O gráfico 15 demonstra que as unidades lexicais *"testemunha"* e *"juiz"* possuem baixa frequência em todos os gêneros, como previsto, mas são, respectivamente, predominantes em parecer e auto de provimento.

Adiante, encontra-se o gráfico 16, apresentando as unidades lexicais referente ao meio ambiente (ULMA), a territórios (ULTr), a pessoas (ULP) e à unidade lexical *"devassa"* (ULDv).

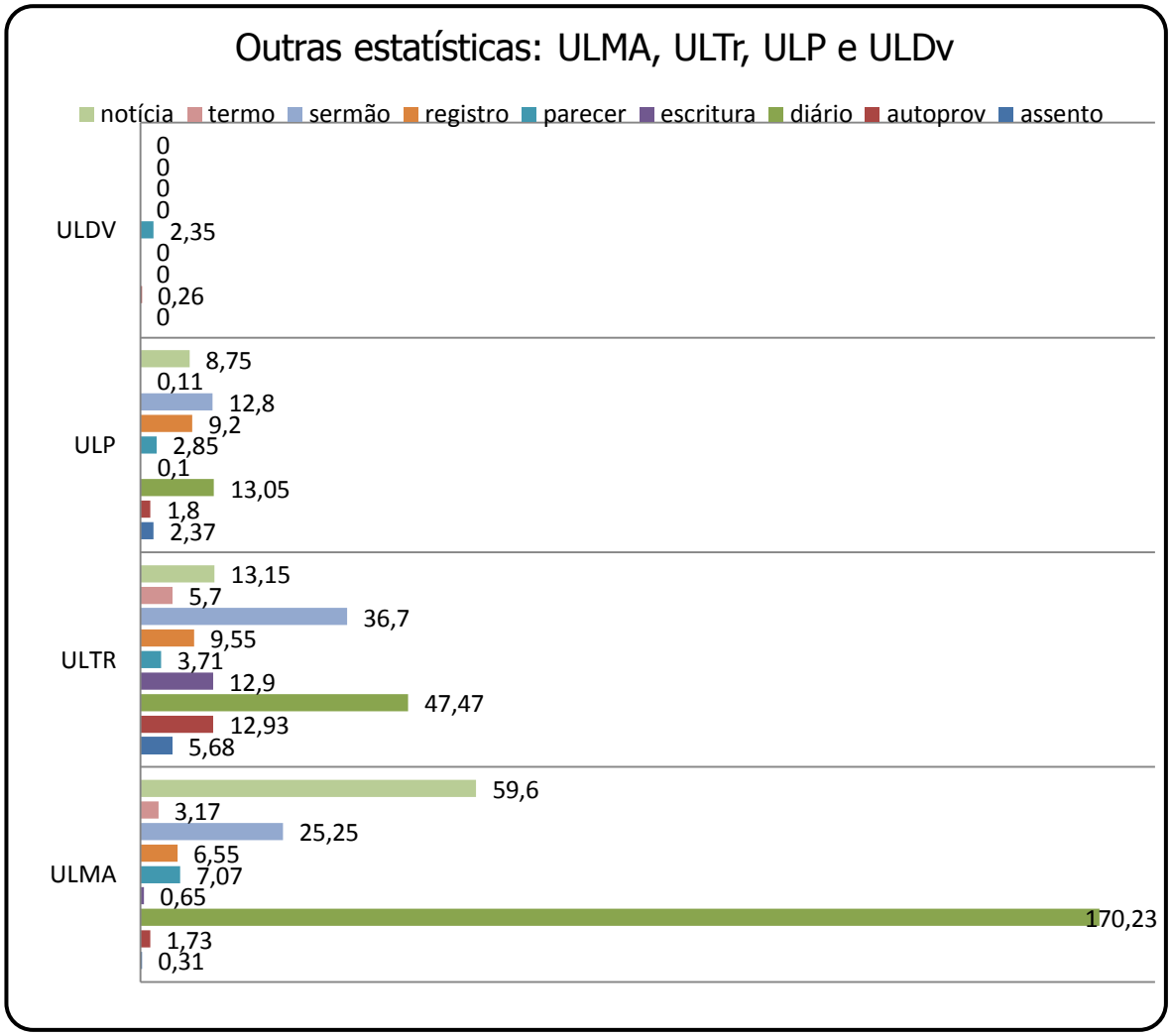

Gráfico 16: Outras estatísticas: ULMA, ULTr, ULP e ULDv

Com base no gráfico 16, conclui-se que as unidades lexicais referentes ao meio ambiente (ULMA), territórios (ULTr) e pessoas (ULP) ocorrem predominantemente em textos do gênero diário. Já a unidade lexical *"devassa"* é predominante apenas em parecer, como esperado.

A seguir, o gráfico 17 apresenta a unidade lexical *"que"* e *"se"*, cuja predominância está nos texto do gênero diário e sermão.

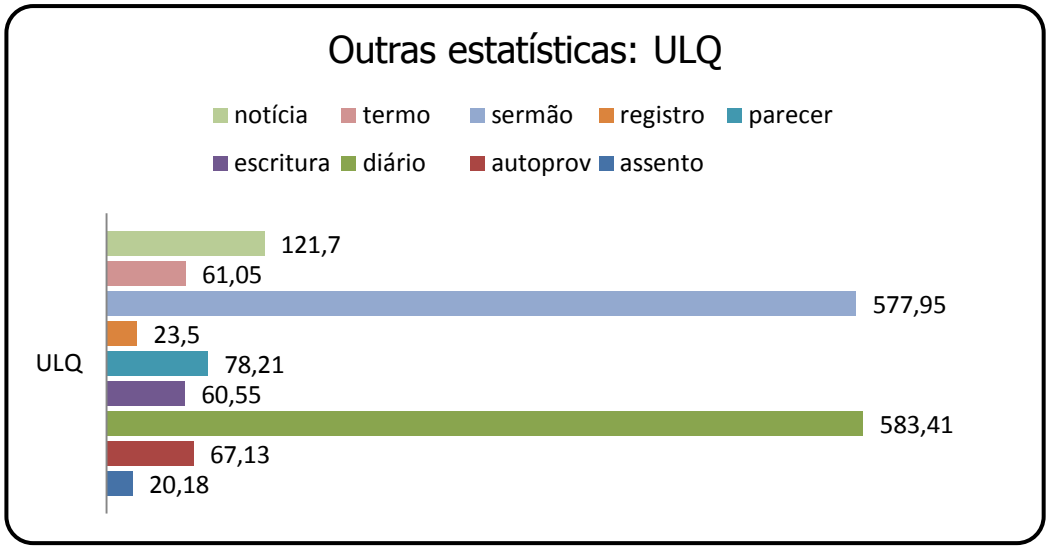

Gráfico 17: Outras estatísticas: ULQ

Na sequência, a tabela 59 apresenta os dados de forma comparativa. De acordo com ela, constata-se que cada uma das unidades lexicais é predominante em determinados gêneros.

A unidade lexical referente a pontos cardeais é predominante em diário, assim como a unidade lexical *"dia"* (ULD) e *"léguas"* (ULL) e que essas unidades lexicais ocorrem e podem ocorrer nos demais textos dos outros gêneros, mas dada a comparação, verifica-se que com predominância e frequência distintas.

A unidade lexical *"suplicante"* (ULS) correspondeu ao esperado, é traço característico de textos do gênero registro, tal como *"testemunha"* (ULT). O mesmo ocorreu com *"juiz"* (ULJ) que ocorre e é predominante em auto de provimento e parecer.

A unidade lexical referente ao contexto sacro (ULSa), como previsto, também foi predominante em textos do gênero sermão, embora ocorra com baixa frequência em outros textos.

Já a unidade lexical referente ao meio ambiente (ULMA), como previsto, é predominante em diário e notícia. A unidade lexical referente a territórios (ULTr) ocorre em todos os gêneros com predominância no diário e sermão, inclusive é relevante mencionar que se trata da unidade mais compartilhada entre os gêneros. Enquanto que a unidade lexical *"devassa"* ocorreu em textos do gênero parecer, como previsto.

As unidades lexicais *"que"* e *"se"* ocorrem em todos os gêneros, como previsto, com maior predominância em textos cuja extensão em palavras e frases é maior.

	ULE	**ULD**	**ULL**	**ULS**	**ULT**	**ULJ**	**ULSa**	**ULMA**	**ULTR**	**ULP**	**ULDV**	**ULQ**
Assento	0	0,12	0	0	0	0,18	0,18	0,31	5,68	2,37	0	20,18
Autoprov	0	0,4	0	0	0,06	1,53	2	1,73	12,93	1,8	0,26	67,13
Diário	62,47	73,2	38,9	0	0,11	0,64	6,76	170	47,47	13,05	0	583,4
Escritura	0,5	1,1	0,05	0,3	0,15	1,15	1,15	0,65	12,9	0,1	0	60,55
Parecer	0,64	0,35	0,28	0	1,85	0	1,57	7,07	3,71	2,85	2,35	78,21
Registro	0,25	0,15	3,05	4	0	0,05	2,15	6,55	9,55	9,2	0	23,5
Sermão	0,35	12,9	0,3	0	0,1	0,3	107	25,3	36,7	12,8	0	578
Termo	0	0,52	0	0,17	0,05	0,52	0,17	3,17	5,7	0,11	0	61,05
Notícia	3	4,55	1,25	0	0	0,05	4,65	59,6	13,15	8,75	0	121,7

Tabela 59: Comparação da média de ocorrência das unidades lexicais nos gêneros

6. Classificação automática

Com base na tabela de traços linguísticos adaptada ao contexto histórico e a geração do arquivo ARFF, foram realizados os testes com os classificadores citados no Capítulo 3. A seguir, a Tabela 60 apresenta os resultados de acordo com os classificadores bem como as medidas: acurácia, precisão, revocação, medida-F.

Acurácia é uma medida de precisão ou validade. Trata-se da correlação entre o valor estimado e os valores das fontes de informação, ou seja, mede o quanto a estimativa obtida está relacionada com o valor real do parâmetro. Dessa maneira, pode-se avaliar em que grau os dados medem o que eles deveriam medir ou quanto os resultados de uma aferição correspondem ao estado verdadeiro do fenômeno aferido.

CLASSIFICADOR	**NaiveBayes**		
ACURÁCIA	84%		
GÊNERO	PRECISÃO	REVOCAÇÃO	MEDIDA-F
Assento	0,6	1	0,75
AutoProv	1	1	1
Diário	0.583	1	0,737
Escritura	1	1	1
Parecer	0,5	0,25	0,333
Registro	1	1	1
Sermão	1	1	1
Termo	1	0,714	0,833
Notícia	0,8	0,4	0,533
CLASSIFICADOR	**BayesNet**		
ACURÁCIA	85,5%		
GÊNERO	PRECISÃO	REVOCAÇÃO	MEDIDA-F
Assento	0,66	1	0,8
AutoProv	0,83	1	0,8
Diário	0,7	1	0,82
Escritura	1	0,8	0,88
Parecer	0,75	0,75	0,75
Registro	1	1	1
Sermão	1	1	1
Termo	0,83	0,71	0,76
Notícia	0,83	0,5	0,62
CLASSIFICADOR	**Multilayer Perceptron**		
ACURÁCIA	91,30%		
GÊNERO	PRECISÃO	REVOCAÇÃO	MEDIDA-F
Assento	1	1	1

AutoProv	1	1	1
Diário	0,75	0,85	0,8
Escritura	1	0,9	0,94
Parecer	0,66	1	0,8
Registro	0,90	1	0,95
Sermão	1	1	1
Termo	1	1	1
Notícia	0,85	0,6	0,70
CLASSIFICADOR	**RBF Network**		
ACURÁCIA	81,15%		
GÊNERO	PRECISÃO	REVOCAÇÃO	MEDIDA-F
Assento	0,75	1	0,85
AutoProv	1	1	1
Diário	0,58	1	0,73
Escritura	0,6	1	0,8
Parecer	0,8	1	0,88
Registro	1	0,5	0,66
Sermão	1	1	1
Termo	1	0,4	0,57
Notícia	1	0,71	0,83
CLASSIFICADOR	**SMO**		
ACURÁCIA	89,85%		
GÊNERO	PRECISÃO	REVOCAÇÃO	MEDIDA-F
Assento	0,85	1	0,92
AutoProv	1	1	1
Diário	0,77	1	0,87
Escritura	1	0,9	0,94
Parecer	0,5	1	0,66
Registro	1	1	1
Sermão	1	1	1
Termo	0,85	0,85	0,92
Notícia	0,5	0,5	0,66
CLASSIFICADOR	**J48**		
ACURÁCIA	81,15%		
GÊNERO	PRECISÃO	REVOCAÇÃO	MEDIDA-F
Assento	0,6	1	0,8
AutoProv	1	0,85	0,88
Diário	0,66	0,85	0,70
Escritura	1	0,8	0,88
Parecer	0,5	0,5	0,5
Registro	0,81	0,9	0,85
Sermão	1	0,9	0,94
Termo	1	0,6	0,8
Notícia	0,75	0,85	0,75

CLASSIFICADOR	**NBTree**		
ACURÁCIA	85,5%		
GÊNERO	PRECISÃO	REVOCAÇÃO	MEDIDA-F
Assento	0,85	1	0,92
AutoProv	0,71	1	0,83
Diário	0,77	1	0,87
Escritura	1	0,6	0,75
Parecer	0,5	1	0,66
Registro	1	1	1
Sermão	1	1	1
Termo	1	1	1
Notícia	0,8	0,4	0,53

Tabela 60: Resultado da classificação automática de gêneros

De acordo com a Tabela 60, conclui-se que a classificação foi satisfatória, entre 84% e 92%, com os índices de precisão e revocação bons, com excelente equilíbrio entre eles, dado pela medida-F.

Os classificadores que apresentam a estrutura arbórea são o J48 e NB Tree. A Figura 10 ilustra visualmente quais traços eles utilizaram para classificar, desde os mais discriminativos (na raiz e níveis seguintes) até os menos expressivos (nas folhas da árvore).

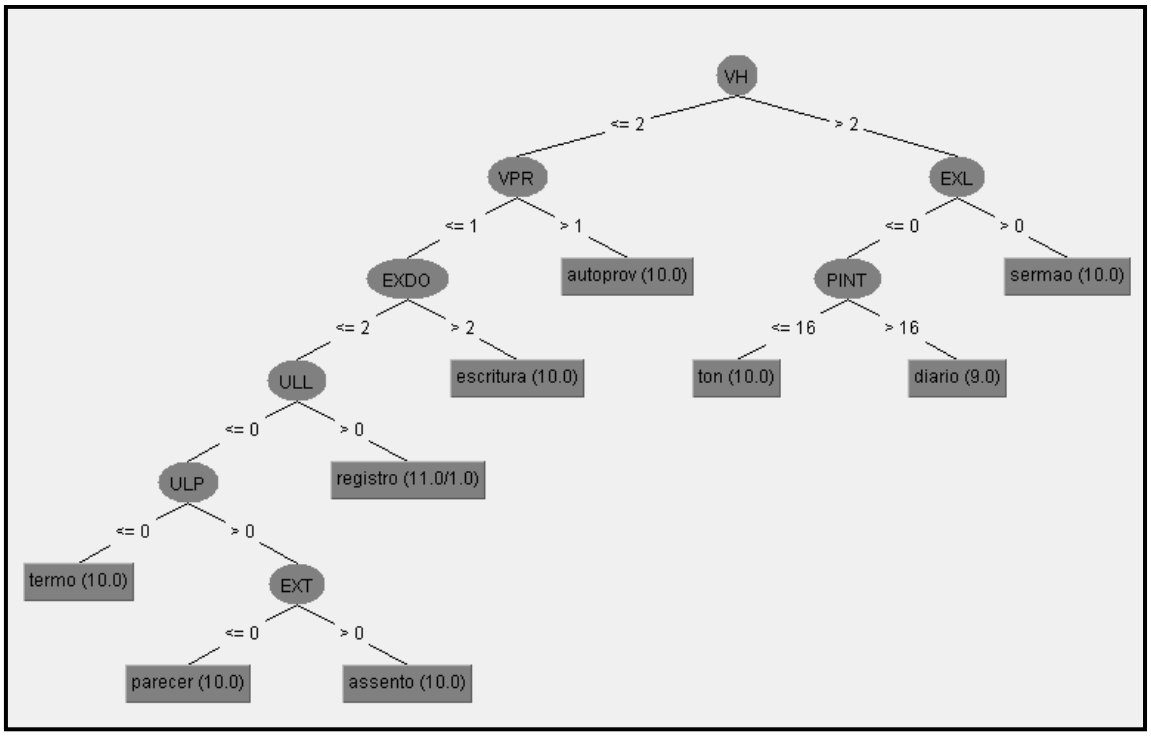

Figura 10: Estrutura arbórea do classificador J48

A ilustração referente ao J48 mostra que o classificador utilizou os seguintes traços: verbos *haver*, *prover*, expressões (EXs), unidades lexicais (ULs) e pronome interrogativo. Por exemplo, sermão: se contém o verbo *haver* (VH) maior que 2 e

expressões em latim (EXL) maior que 0 é sermão. Mas se contém o verbo *haver* maior que 2, expressão em latim menor ou igual a 0 e pronome interrogativo (PINT) menor ou igual a 16 é notícia (ton), ou ainda pronome interrogativo maior que 16 é diário.

Com base na estrutura, observa-se que não é preciso ter uma recuperação elevada de cada traço, quando é feita sua extração, afinal, alguns dos gêneros foram classificados com base entre menor, maior ou igual a 0, 1, 2, como no caso de parecer e assento, *termo*, registro, auto de provimento.

A Figura 11, a seguir, apresenta a estrutura arbórea do classificador NB Tree.

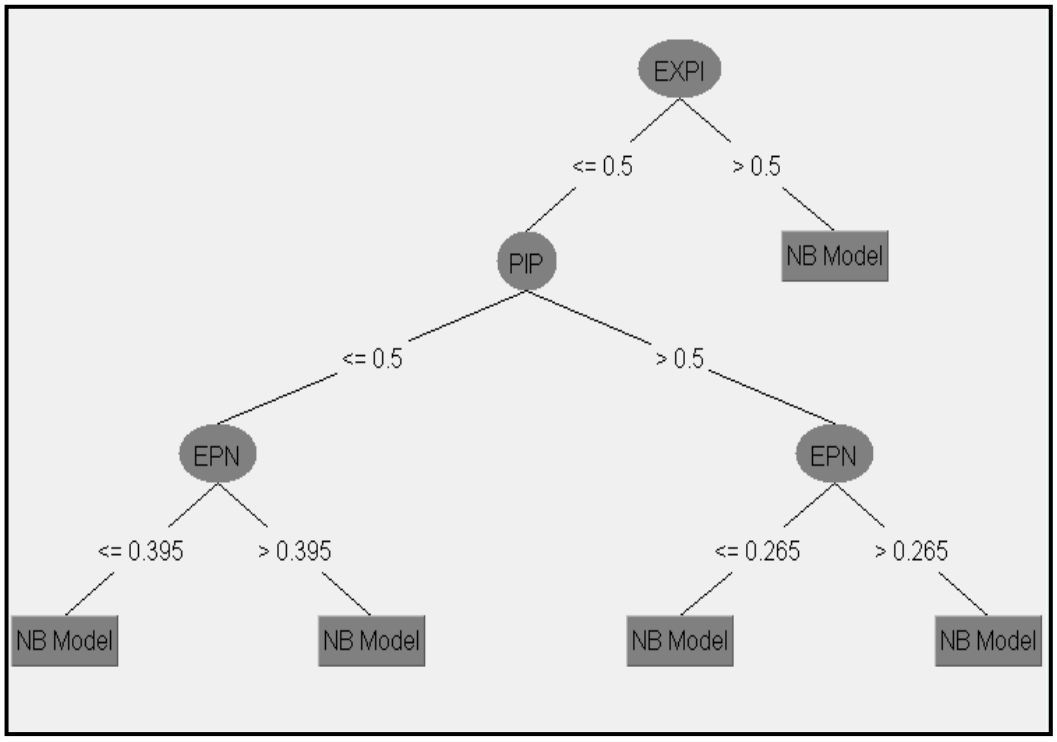

Figura 11: estrutura arbórea do classificador NB Tree

Essa estrutura apresentada da Figura 11 não indicou a classificação dos gêneros, mas diferentemente da outra estrutura, utilizou estimativas baseadas no texto e em palavras (EPN), apenas uma expressão e um pronome. Também se valeu de baixos números para classificar, menor, maior ou igual a 0.

Apresenta-se o Gráfico 18 os trinta atributos mais relevantes utilizados para o treinamento com os classificadores. Para isso, aplicou-se o algoritmo de ganho de informação, InformationGainAtributteEval, disponibilizado também no Weka.

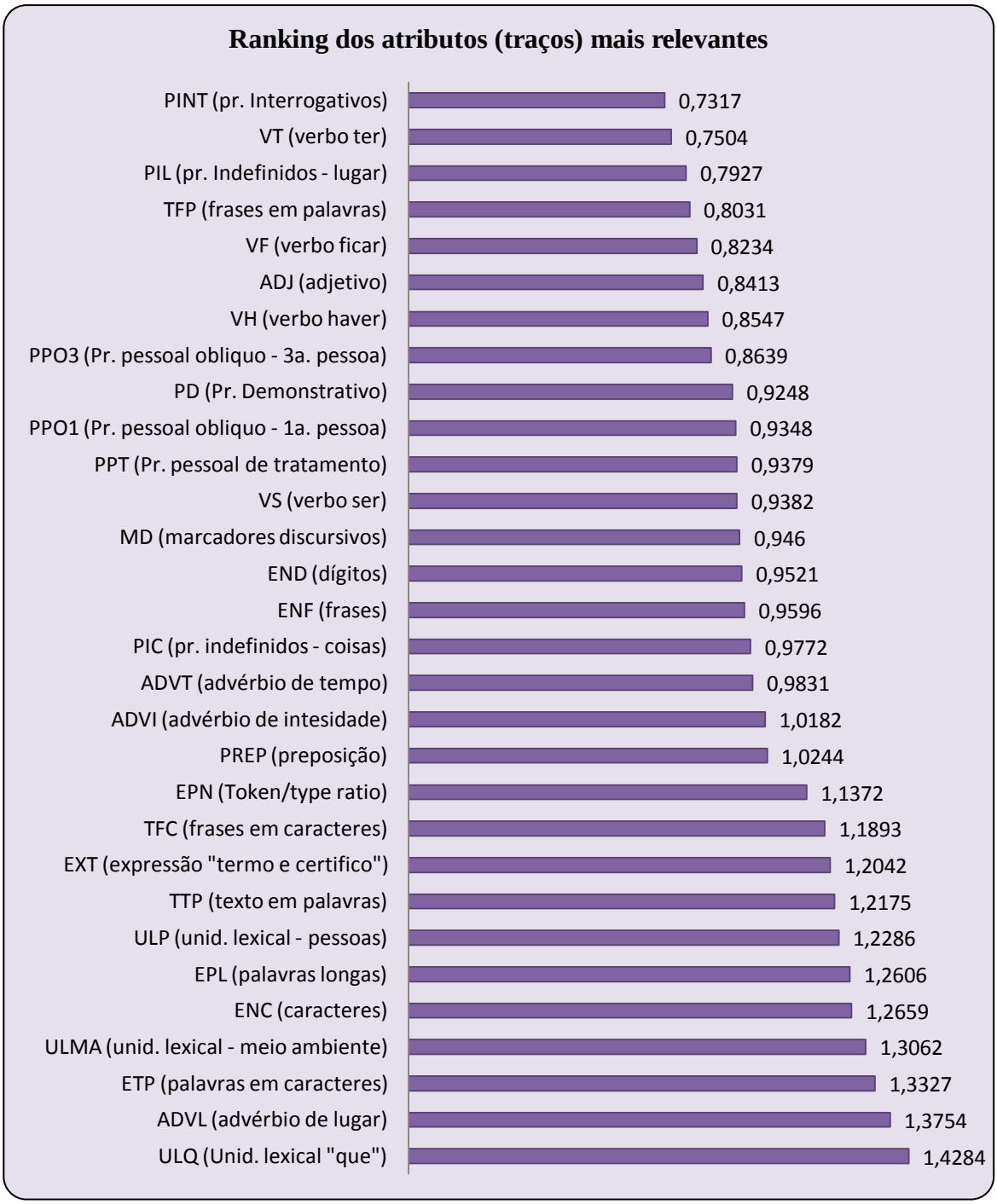

Gráfico 18: *Ranking* dos atributos mais relevantes

Com base no Gráfico 18, observa-se que do conjunto de traços utilizados para treinar os classificadores, 61 no total, os mais relevantes foram a unidade lexical “que” (ULQ), seguida do advérbio de lugar (ADVL) e tamanho médio das palavras em caracteres (ETP). Já os traços menos relevantes, dentre os trinta apresentados no gráfico, foram pronome interrogativo (PINT), seguido do verbo ter (VT) e pronome indefinido referente a lugar (PIL). Contudo, vale ressaltar que do conjunto total de traços, os menos relevantes foram: unidade lexical “juiz” (ULJ), unidade lexical

“testemunha” (ULT), expressão “faço saber” (EXFS) e expressão “pregado” (EXPr), os quais não estão esboçados no gráfico.

7. Conclusões

Na introdução foram elencadas algumas questões vinculadas ao propósito da pesquisa, tais como:

- as características dos gêneros comutam de um contexto para outro?
- é possível a convergência de todas as características determinantes dos gêneros de forma a se obter uma classificação coerente de gêneros textuais?
- como lidar com textos históricos, uma vez que a referência de mundo do classificador humano é a do contexto atual?
- quais características do português sobressaem em textos de um corpus histórico; são as mesmas características dos textos de um corpus contemporâneo?

Toda a realização da pesquisa permitiu respondê-las. Julgou-se relevante iniciar as teorias acerca do conceito de gênero por Bakhtin (2000) apenas em decorrência do pioneirismo de sua abordagem. Contudo, a perspectiva sistêmico funcional de Halliday e Hasan (1989), seguidos por Eggins e Martin (1997) ao apresentar como os textos se estruturam para construírem significados, o conceito de Estrutura Genérica Potencial (EGP), constituída pelas escolhas léxico-gramáticas dos interlocutores, pela função interpessoal e função textual, juntamente com a perspectiva de Swales (1990), nos quais os gêneros são sócio-retoricamente construídos, são eventos codificados, inseridos em processos sociais comunicativos compartilhados pelas comunidades em que ocorrem e reconhecidos por seus membros como legítimos permitiu fundamentar a ocorrência de padrões sociocomunicativos.

A pesquisa permitiu constatar que os gêneros possuem suas características específicas, seus traços, o que não se pode afirmar que nenhum dos traços irá ocorrer em outros gêneros, como por exemplo, a unidade lexical *"dia"*, traço característico de diário, mas é uma palavra que pode ocorrer em qualquer outro gênero. O que o torna traço é a frequência, ocorrência e o fato de ser um padrão em textos do gênero diário.

Além disso, é o conjunto de traços de determinado gênero que o diferencia dos demais, como demonstra a figura 11, se contém o verbo *haver*, pronome interrogativo (maior que 16) e não tem expressão em latim é o gênero diário.

No âmbito das teorias acerca do Aprendizado de Máquina (AM) fez-se necessário compreender suas estratégias e modelos, utilizando-se o aprendizado de máquina supervisionado por indução, o qual tem como objetivo induzir conceitos a partir de exemplos que foram pré-classificados (gêneros diário, termo, assento etc.).

Mesmo com todas as dificuldades em lidar com um corpus histórico, inexistência de ortografia, problemas de junção, trabalhar com expressões foi mais difícil que o esperado, poucas eram quantificadas, dada as variações que cada palavra tinha, mesmo tentando identificá-las. Um exemplo disso foi com a expressão *ano de nascimento*, que possui no mínimo 15 possibilidades, considerando a variação de cada palavra.

É possível observar que os traços contemporâneos descrevem textos históricos, com exceção dos pronomes de tratamento, do adjetivo *dito*, de expressões e unidades lexicais, que foram acrescentados à tabela.

Por fim, esta dissertação consistiu em identificar os traços linguísticos do português histórico do Brasil e utilizá-los como insumo para obter uma classificação automática de gêneros. Para isso, partiu-se de um conceito operacional de gênero, uma tabela de traços contemporâneos e uma abordagem computacional, mais especificamente, o aprendizado de máquina supervisionado. Os traços constantes da tabela foram refinados a partir de um processo incremental e iterativo de buscas no corpus (*bootstrapping method*), finalizando com a geração do arquivo ARFF e treinamento com os algoritmos classificadores. Concluiu-se que a classificação foi satisfatória, entre 84% e 92%, com os índices de precisão e revocação bons, com excelente equilíbrio entre eles, dado pela medida-F. Vale citar que a taxa de acerto da classificação de gêneros com textos contemporâneos verificados na pesquisa de Aires (2005) foi entre 88,42% e 97,21%, ou seja, em comparação com a classificação de gêneros utilizando corpus histórico foi semelhante.

Assumiu-se a premissa de que é possível classificar gêneros pertencentes ao contexto histórico, uma vez que eles possuem características próprias.

Para a área de PLN, Linguística e Ciência da Informação, a contribuição desta pesquisa refere-se à própria metodologia: como levantar informações linguísticas de um corpus histórico com todas as suas peculiaridades e como utilizá-las para implementação de sistemas computacionais.

Como trabalhos futuros, sugerem-se:

- Aumentar pesquisas sobre expressões classificadoras de um domínio.

- Verificar se as expressões de determinados domínios em um corpus histórico ocorrem na mesma proporção em um corpus contemporâneo.
- Comparar os traços de um gênero ou domínio de um corpus histórico e um corpus contemporâneo.
- Investigar informações lingüísticas que podem indicar tendências e informações para aplicar no contexto de inteligência competitiva.

Referências

ALMEIDA, G. M. B.; VALE, O. A. Do texto ao termo: interação entre Terminologia, Morfologia e Linguística de Corpus na extração semi-automática de termos. In: ISQUERDO, Aparecida Negri & FINATTO, Maria José Bocorny. (Org.). **As ciências do Léxico: Lexicologia, Lexicografia e Terminologia.** 1 ed. Campo Grande: Editora da UFMS, 2008, v. IV, p. 483-499.

AIRES, R.V.X. **Uso de marcadores estilísticos para busca na *Web* em português**. 2005.185 f. Tese (Doutorado em Ciência da Computação) – Curso de Pós-graduação e, Ciências da Computação e Matemática Computacional, Universidade de São Paulo, 2005.

AUGER, A.; BARRIÈRE, C. Pattern-based approaches to semantic relation extraction. A state-of-art. **Terminology**, vol 14, n. 1, 2008. p. 1-19.

BAKHTIN, M. Os gêneros do discurso. *In* **Estética da criação verbal**, 3ª. ed., São Paulo: Martins Fontes, 2000.

BAZERMAN, C.; A. P. DIONÍSIO & J. C. HOFFNAGEL. (Orgs.). *Gêneros textuais, tipificação e interação*. São Paulo: Cortez, 2005.

BECHARA, Evanildo. Moderna gramática portuguesa. 37ª. ed., Rio de Janeiro: Lucerna, 2004.

BERBER SARDINHA, T. **Linguística de Corpus**. Barueri, SP: Manole, 2004.

BERBER SARDINHA, T. **Linguística de Corpus**: histórico e problemática. **DELTA** , São Paulo, v. 16, n. 2, 2000 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502000000200005&lng=en&nrm=iso>. Acesso em: 29 July 2008. doi: 10.1590/S0102-44502000000200005

BHATIA, Vijay K. **Analysing genre***:* language use in professional settings. New York: Longman, 1993.

BIBER, D. **Variation across speech and writing**. New York: Cambridge University Press, 1988.

BIBER, D. e FINEGAN, E. (1989). Drift and the evolution of English style: a history of three genres. **Language** 65 (3): 487-517.

BIBER, D. Representativeness in corpus design. **Literary and Linguistic Computing**, v. 114 8, n. 4, p. 1.15, 1993.

BIBER, D. Using register-diversified corpora for general language studies. **Computational Linguistics**, v. 19, n. 2, p. 219.241, 1993.

BIBER, D. **Sociolinguistic Perspective on Register**. New York: Oxford University Press, 1994.

BIBER, D. **Dimensions of Register Variation**: A cross-linguistic comparison. 1. ed. Cambridge: Cambridge University Press, 1995.

BIBER, D., CONRAD, S., REPPEN, R. **Corpus Linguistics***: Investigating Language Structure and Use*. Cambridge: Cambridge University Press. 1998.

BONINI, A. O conhecimento de jornalistas sobre gêneros textuais: um estudo introdutório. In: **Linguagem em (dis)curso on line** , v. 2, n. 1, 2001.

BONINI, A., MOTTA-ROTH, D. (Orgs.). **Gêneros:** teorias, métodos e debates. São Paulo: Parábola Editorial, 2005, p.81-106.

BRONCKART, J. **Atividades de Linguagem, Textos e Discursos. Por um Interacionismo Sócio-discursivo.** São Paulo: Editora da PUC-SP, EDUC. 1999.

BRUCKSCHEN, M.; MUNIZ, F.; SOUZA, J. G. C.; FUCHS, J. T.; INFANTE, K.; MUNIZ, M.; GONÇALVES, P. N.; VIEIRA, R.; ALUÍSIO, S. M. (2008). **Anotação Lingüística em XML do Corpus PLN-BR**. Série de Relatórios do NILC (NILC-TR-09-08). São Carlos - SP, Junho 2008, 39 p.

CÂNDIDO JR. A. **Criação de um ambiente para o processamento de corpus de Português histórico**. Dissertação (Mestrado em Ciência da Computação) – Curso de Pós-graduação e, Ciências da Computação e Matemática Computacional, Universidade de São Paulo, 2008.

CRISTOVÃO, V.L.L. Gêneros ensinados em inglês como língua estrangeira: uma problemática de transposição. In: **III Conferência de Pesquisa Sócio-cultural**, PUC São Paulo e Universidade Estadual de Londrina. Julho de 2000.

DE CONTO, J. M. . A Carta de Apresentação na Perspectiva Sistêmico-Funcional. In: **VII Seminário Internacional em Letras:** Linguagem, Cultura e Identidade, 2007, Santa Maria. Anais - VII Seminário Internacional em Letras: linguagem, cultura e identidade. Santa Maria : editora da UNIFRA, 2007. v. único.

DOLZ, J. & B. SCHNEUWLY . ***Pour un enseignement de l'oral. Initiation aux genres formels à l'école.*** Paris: ESF ÉDITEUR. 1998.

EGGINS, S. 1994. **An introduction to systemic functional linguistics**. London: Printer Publishers.

EGGINS, S. e J. R. MARTIN 1997. Genres and Registers of Discourse: In: T. A. van Dijk (org). **Discourse as structure and process – Discourse Studies**: a multidisciplinary introduction, Vol. I. Londres: SAGE Publ.

FIORIN, J.L. **Introdução ao pensamento de Bakhtin.** São Paulo: Ática, 2006.

GALHO, Thaís S.; MORAES, Sílvia M. W. **Categorização Automática de Documentos de Texto Utilizando Lógica Difusa**. Gravataí: ULBRA. Trabalho de Conclusão de Curso. 2003.

HALLIDAY, M. A. K. , HASAN, R. ***Language, context and text: aspects of language in a social perspective***. Oxford: Oxford University Press, 1989.

HALLIDAY, M. A. K. ***An introduction to functional grammar***. London: Edward Arnold, 1994.

HOEY, M. **Patterns of lexis in text.** Oxford: Oxford University Press, 1991.

KAUFFMANN, C. H. O Corpus do Jornal: A variação linguística, gêneros e dimensões da imprensa diária escrita. Dissertação (Mestrado) — LAEL - PUC, São Paulo, 2005.

KENNEDY, G. ***An Introduction to Corpus Linguistics*.** New York: Longman, 1998.

MACHADO, Irene. Gêneros discursivos. *In*: **Bakhtin:** conceitos-chave. Beth Brait (org). São Paulo: Contexto, 2005.

MARCUSCHI, L. A. **Gêneros textuais: definição de funcionalidade**. In: DIONÍSIO, A. P.; MACHADO, A. R.; BEZERRA, M. Gêneros Textuais & Ensino. Rio de Janeiro: Lucerna, 2002.

MARCUSCHI, L. A.; XAVIER, A. C. (orgs.) **Hipertexto e gêneros digitais**. 2a. ed. Rio de Janeiro: Lucerna, 2005.

MARCUSCHI, L. A. **Produção textual, análise de gêneros e compreensão**. São Paulo: Parábola, 2008. 296 p.

MARTINS, C. A. **Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado**. 2003. 208 f. Tese (Doutorado em Ciências de Computação e Matemática Computacional) – Instituto de Ciência Matemáticas e de Computação, Universidade de São Paulo, São Carlos 2004.

MATSUBARA, E.T. **o algoritmo de aprendizado semi-supervisionado co-training e sua aplicação na rotulação de documentos**. 2004. 105 f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de computação, Universidade de São Paulo, São Carlos, 2004.

METZ, J. Interpretação de clusters gerados por algoritmos de *clustering* hierárquico. 2006. 152 f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de computação, Universidade de São Paulo, São Carlos, 2006.

MILLER, C. “Genre as social Action”. **Quartely Journal of Speech 70**. (1984) p.151-167.

MOENS, Marie-Francine. **Automatic indexing and abstract of document texts**. Massachusetts: Kluwer Academic Publishers. 2000.

MONARD, M. C.; BARANAUSKAS, J.A. (2003). Conceitos sobre aprendizado de máquina. In S. O. Rezende (Ed.), **Sistemas Inteligentes***:* Fundamentos e Aplicações, p. 89–114. Manole.

MONARD, M. C. ; BARANAUSKAS, J. A. **Conceitos Sobre Aprendizado de Máquina.** In: Solange O. Rezende. (Org.). Sistemas Inteligentes Fundamentos e Aplicações. 1 ed. Barueri-SP: Manole Ltda, 2003, v. 1, p. 89-114.

MONARD, M.C.;BATISTA, G.E.A.P.; KAWAMOTO, S.; PUGLIESI, J.B. **Uma introdução ao aprendizado Simbólico de Máquina por exemplos**. São Carlos: ICMC/USP, 1997. Notas didáticas.

MORAES, S.M.W. STRUBE DE LIMA, V.L. Um Estudo sobre Categorização Hierárquica de uma Grande Coleção de Textos em Língua Portuguesa. In: **V Workshop em Tecnologia da Informação e Linguagem Humana**, XXVII Congresso da SBC, 5-6 julho, SBC, Rio de Janeiro, 2007.

MOTTA-ROTH, D., HEBERLE, V. O conceito de "estrutura potencial do gênero" de Ruqayia Hasan. In: MEURER, J. L., BONINI, A., MOTTA-ROTH, D. (Orgs.) **Gêneros, teorias, métodos e debates**. São Paulo: Parábola Editorial, 2005. p. 12 – 28.

MOTTA-ROTH, D. Questões de metodologia em análise de gêneros. In: KARWOSKI, A. M.; GAYDECZKA, B.; BRITO, K. S. (Orgs.). **Gêneros textuais:** reflexões e ensino. Palmas e União da Vitória, PR: Kaygangue, 2005. p.179 – 202

MUNIZ, M. C. M. **A construção de recursos linguístico-computacionais para o português do Brasil: o projeto de *Unitex*-PB**. Dissertação de Mestrado. Instituto de Ciências Matemáticas de São Carlos, USP. 72p. 2004.

NOGUEIRA, B.M. **Avaliação de métodos não supervisionados de seleção de atributos para mineração de textos**. 2009. 104 f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de computação, Universidade de São Paulo, São Carlos, 2009.

OLIVEIRA, F. M. **A investigação de gêneros textuais no contexto digital:** uma análise de sites educacionais para professores de língua em formação ou em serviços. 2005.

OLIVEIRA, L. P. **Compilação de um corpus representativo do português do Brasil e análise multidimensional da variação entre gêneros discursivos. [2003]**

PAUMIER, S. (2002). ***Unitex user manual.*** disponível em: http://www-igm.univ-mlv.fr/~*Unitex*

PEREIRA, J. S., ALMEIDA, M. B. "Sabe tudo sobre tudo": análise da seção de cartas-pergunta em revistas femininas para adolescentes. In: MEURER, J. L. & MOTTA-ROTH, D. (orgs.). **Gêneros textuais e práticas discursivas:** subsídios para o ensino da linguagem. Bauru, SP: EDUSC, 2002. p. 239 – 258.

PEIXOTO, M.D.F. BATISTA, M.G.T.R.H. CAPELO, M.J.T.S.P. **Categorização de textos**. Disponível em: http://www.di.ubi.pt/~api/text_categorization.pdf acessado 10 de outubro de 2009.

REZENDE, S. O. (Org.). **Sistemas Inteligentes**: Fundamentos e Aplicações. 1. ed. Barueri, SP: Editora Manole Ltda, 2003. v. 1. 560 p.

RIGO,S., OLIVEIRA, J.P. e BARBIERI, C. **Classificação de Textos baseada em ontologias de domínio**. TIL 2007.

ROCHA LIMA, C. H. **Gramática normativa portuguesa**. 38ª. ed., Rio de Janeiro: José Olympio, 2000.

SACCONI, L. A. **Nossa gramática: teoria e prática**. 25ª. ed., São Paulo: Atual Editora, 1999.

SANCHES, M.K. **Aprendizado de máquina semi-supervisionado**: proposta de um algoritmo para rotular exemplos a partir de poucos exemplos rotulados. 2003. 142 f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de computação, Universidade de São Paulo, São Carlos, 2003.

SILVA, V. L. Paredes. Forma e função nos gêneros de discurso. *Alfa*, São Paulo, 41, 79-98, 1997. Número especial.

SILVA, M. C. A noção de gênero em Swales: revisitando conceitos. **Recorte** - Revista de Linguagem, Cultura e Discurso. Ano 2, n. 3. jul-dez/2005.

SILVEIRA, F. J. N., MOURA, M. A. A estética da recepção e as práticas de leitura do bibliotecário-indexador. *Perspect. ciênc. inf.*, Jan./Apr. 2007, vol.12, no.1, p.123-135. ISSN 1413-9936.

SOUZA, J.A, S.M. ALUÍSIO, G.M.B. ALMEIDA. Tipologia de gêneros textuais do português do Brasil dos séculos XVI, XVII e XVIII. In: Workshop do projeto Dicionário Histórico do Português do Brasil, II, 2006, Araraquara, SP. **Preparação do**

Córpus de documentos em Português dos séculos XVI, XVII e XIII do Projeto *Dicionário Histórico do Português do Brasil* para ser utilizado com as ferramentas *UNITEX* e *Philologic* e ser disponibilizado para outras pesquisas.

SWALES, John M. ***Genre analysis****:* English in academic and research settings. Cambridge: Cambridge University Press, 1990.

VIEIRA, R., STRUBE DE LIMA, V.L. **Linguística computacional: princípios e aplicações.**

WITTEN, Ian H., FRANK, Eibe. **Data Mining**: Practical machine learning tools and techniques, 2a. ed., 2005.

ZANOTTO, N. **E-mail e carta comercial**: estudo contrastivo de gênero textual. Rio de Janeiro: Lucerna; Caxias do Sul: Educs, 2005.

APÊNDICE A - Traços linguísticos para classificação automática

(Lista baseada em Aires (2005))

Estatísticas baseadas em palavras

Estimativa de itens lexicais diferentes, dado pelo número de itens lexicais diferentes dividido pelo número de itens (*type/token ratio*)

Estimativa de itens lexicais diferentes, porém considerando-se apenas os itens iniciados por letra maiúscula (*capital type token ratio*)
Número de dígitos
Tamanho médio das palavras em caracteres
Número de palavras longas (com mais de 6 caracteres)

Estatísticas baseadas no texto como um todo
Número de caracteres
Tamanho médio das frases em caracteres
Número de frases
Tamanho médio das frases em palavras
Tamanho do texto em palavras

Outras estatísticas
Número de ocorrências das expressões "acho", "acredito que", "parece que", e "tenho impressão (de) que"
Verbo SER (nas formas "é" e "são")
Pronomes na primeira pessoa
Pronomes na segunda pessoa
Pronomes na terceira pessoa
Freqüência e tipo de pronomes demonstrativos
Freqüência e tipo de Pronomes indefinidos
Freqüência e tipo de pronomes interrogativos
Freqüência e tipo de preposições
Advérbios (lugar, tempo e terminados em -mente)
Freqüência e tipo de interjeições
(operadores argumentativos) (livro do Platão e Fiorim e em gramáticas do português)
Os marcadores discursivos "agora", "da mesma forma", "de qualquer forma", "de qualquer maneira" e "desse modo"

APÊNDICE B – Tabela de traços adaptada ao contexto histórico

Estatísticas baseadas em palavras
EPN - Estimativa de itens lexicais diferentes, dado pelo número de itens lexicais diferentes dividido pelo número de itens (*type/token ratio*)
EEM - Estimativa de itens lexicais diferentes, porém considerando-se apenas os itens iniciados por letra maiúscula (*capital type token ratio*)
END - Número de dígitos
ETP - Tamanho médio das palavras em caracteres
EPL - Número de palavras longas (com mais de 6 caracteres)

Estatísticas baseadas no texto como um todo
ENC - Número de caracteres
TFC - Tamanho médio das frases em caracteres
ENF - Número de frases
TFP - Tamanho médio das frases em palavras
TTP - Tamanho do texto em palavras

Outras estatísticas - Verbos
Número de ocorrências de determinadas expressões
V**S** - Verbo SER nas formas é e são, sam.
VH - Verbo HAVER nos formar há, havia
VP - Verbo PEDIR nas formas pede, pedem
VPr - Verbo PROVER na forma proveu, proveo
VPo - Verbo PODER nas formas póde, podia
VF – Verbo FAZER nas formas fez, fazer e fazem
VI – Verbo IR nas formas foi, fomos e fui
VT – Verbo TER nas formas tem e tinha
VD – Verbo DIZER nas formas dizer, disse e digo

Outras estatísticas – Pronomes, adjetivo, preposição, advérbios, marcadores discursivos.
PPO1 - Pronome pessoal obliquo na primeira pessoa – me, mim, mym nos, nós, comigo, commigo, comiguo, conosco, comnosco, connosco.
PPO2 - Pronome pessoal obliquo na segunda pessoa – te, the, ti, contigo, comtigo, contiguo, vos, vós, convosco
PPO3 - Pronome pessoal obliquo na terceira pessoa – lhe, lhes, si, consigo.
PPT - Pronome pessoal de tratamento – Senhor, Senhora/ Vossa Senhoria/ Vossa Excelência/ Vossa Alteza/ Vossa Santidade/ Vossa Majestade. (neste caso, substituir vossa por sua, como por exemplo: Sua Majestade, Sua Alteza etc).Vossa mercê/merce
PD - Pronomes demonstrativos – este, estes, deste, destes, d'estes, esta, estas, desta, destas, isto, disto, esse, esses, desse, desses, essa, essas, dessa, dessas, isso, disso, aquele, aqueles, aquelle, aquelles, daquele, daqueles, daquelle, daquelles, aquela, aquelas, aquella, aquellas, daquela, daquella, daquelas, daquellas, aquilo, aquillo, daquillo.
ADJ - Adjetivo - dito, ditto, odito, oditto
PRV - Pronome relativo variável – a qual, aqual, os quais, os quaes, osquaes, cujo, cujos, cuja, cujas.
PINT - Pronomes interrogativos – quem, quanto, quando, qual, quais, quaes.
PIP - Pronomes indefinidos referente à pessoa – alguém, alguem, ninguém, ninguem, outrem,
PIL – Pronomes indefinidos referente a lugar – onde, aonde, donde, adonde
PIC – Pronome indefinido referente a pessoas, lugares e coisas- algo, tudo, nada, todo, todos, toda, todas, vários, várias, certo, certa, pouco, poucos, pouca, poucas, muito, muitos, muyto, muytos, muita, tanto, tantos, tanta, tantas, cada, nenhum, nenhuma, qualquer, quaesquer.
PREP - Tipo e frequência de preposições – por, per, para, até, athé, athè, té, em, entre, contra, sem, sobre

ADVL - Advérbio de lugar - aqui, abaixo, acima, asima, cá, lá, ai, ahi, ali, alli, alem, além, alèm, dalém, dentro, longe, acolá, aquém, aquem, adiante, atrás, atras, atraz algures, alhures, perto, fora, defronte, embaixo, em baixo.
ADVT - de tempo – hoje, oje, amanhã, ontem, hontem, manhã, já, sempre, nunca, ainda, antes, tarde, cedo, outrora, depois, depoiz, agora, logo, jamais, diariamente, anualmente, atualmente, sucessivamente, entrementes, imediatamente, immediatamente.
ADVI - de intensidade - muito, pouco, bastante, mais, menos, bem, mal
muito, muyto, pouco, poco, bastante, bastantes, mais, maiz, menos, menoz, bem, mal, tão, tao, todo, assaz, tanto, quão, meio, meyo.
MD - Marcadores discursivos - mas, porém, assim, como, porque, perque, por que, ainda, depois, ou, então, portanto, por tanto, entretanto, entre tanto, a medida que, diante, pois, também, tambem, só, apenas, mesmo, mesma, nem, além de, alem de.

Outras estatísticas: expressões
EXD - Deos goarde
EXFS - faço saber, faso a saber, Faso aSaber, faco asaber, faco saber, faso saber.
EXCM - Capitão-mor, Capitão mor, Capitão-mór, Capitáo-mór, Capitão major, capitam maior, capitam mor, capitam major, **capitam** Mayor (manter o hífen)
EXL - Mihi, non est, domine/domini
EXPr – *prègado, pregado.*
EXO - Officiaes da camara, officiaes daCamara, officiaes desta câmera, officiaes da camera.
EXA - anno de nascimento, anno do nascimento, anno do nassimento, anno do nascimanto
EXEs - o escrevy, o escrevi, o escreuy, oescrevi, oescrevy, sobescrevy, sobescrevi, escreuy, escreui, escreuj, sobescreui, sobescreuy, asignou, asignei, asigney, tabeliam/ tabaliam/ escrivam/ escrivão
EXAt - Atas da câmara
EXDo - Doutor ouvidor geral e corregedor/ ouvidor geral/corregedor/ ouvidor/doutor corregedor/corregedor da comarca
EXT - Termo e certifico/certifico/termo
EXPI - público instrumento/instrumento/publico

Outras estatísticas: unidades lexicais
ULE - leste, oeste, norte, sul, nordeste, sudeste, noroeste, sudoeste e nor-Nordeste, lés-Nordeste, lés-Sudeste, su-Sudeste, su-Sudoeste, oés-Sudoeste, oés-Noroeste, nor-noroeste.
ULD - Dia
ULL – léguas, leguas, legoas, llegoas, leguoas.
ULS - Suplicantes, suplicante, supplicantes, supplicante, osuplicante, osupplicante
ULQ – que, qui e se (juntei que e se como sugere a tabela de Aires)
ULT – testemunha (retirar a palavra escritura)
ULJ - juiz/juízes/juis
ULSa - deus, santo, misericórdia, sacramento, graça, alma, almas, corpo, fé.
ULMA - tamanho, grande, pequeno, carne, folhas, arvores, árvores, rio, riacho, cor, espécies (terra está em territorial), vento, mar, serra, serras, águas, aguas.
ULTr - cidade, estado, vila, Villa, comarca, terra, terras
ULP - índios, indio, gentio, povo, homem.
ULDv – devassa, devaça

APÊNDICE C - Tipologia de gêneros do português do Brasil

Domínio **Discursivo** 1 - Religioso

1.1. Auto de confissão
1.2. Breve (Carta pontifícia, menos solene que a bula papal, tratando de assuntos menos abrangentes que esta e deferindo em alguns de seus caracteres externos e na disposição do texto. Belloto, H.L.Como fazer análise diplomática e análise tipológica de documento de arquivo.2002.)
1.3. Carta pastoral
1.4. Epístola
1.5. Moções
1.6. Oração
1.7. Sermão
1.8. Voto
1.9. Cartas
1.10. Capítulo

Domínio Discursivo 2 – Jurídico

2.1. Legislativo
2.1.1. Códigos
2.1.2. Constituição
2.1.3. Decreto
2.1.4. Decreto lei
2.1.5. Ementas constitucionais
2.1.6. Lei
- 2.1.6.1. Leis complementares
- 2.1.6.2. Leis ordinárias
- 2.1.6.3. Leis delegadas

2.1.7. Medida provisória
2.1.8. Portaria
2.1.9. Resoluções
2.1.10. Tratados internacionais

2.2. Jurisprudência
2.2.1. Acordão
2.2.2. aresto
2.2.3. minuta
2.2.4. petição
2.2.5. sentença
2.2.6. súmula

2.3. Jurídico-Administrativo
2.3.1. Atos administrativos
- 2.3.1.1. Alvará
- 2.3.1.2. Autorização
- 2.3.1.3. Licença
- 2.3.1.4. Admissão

2.3.1.5. Permissão
2.3.1.6. Aprovação
2.3.1.7. Homologação
2.3.1.8. Parecer
2.3.1.9. Visto
2.3.1.10. Exposição de motivos (1.Na linguagem burocrática, ofício dirigido por Ministro de Estado ao Presidente da República. In *Aurélio Eletrônico*, 2004)
2.3.1.11. Decreto
2.3.1.11.1. Bando (Ocorre apenas na administração colonial. É ordem ou o decreto, em geral, dos governadores e capitães generais proclamada oralmente em pregão público ou afixada em lugar ou veículo de circulação pública. Belloto, H.L. Como fazer análise diplomática e análise tipológica de documento de arquivo. 2002.)
2.3.1.12. Resolução
2.3.1.12.1. Assento [Interpretação de um documento legal; resolução. (segundo Morais, 1813]
2.3.1.13. Portaria
2.3.1.14. Circular
1.1.1.1. Despacho
2.3.1.16. Depoimento

2.3.2. Licitação
2.3.2.1. Edital
2.3.2.1.1. Édito (1.Ordem judicial publicada por anúncios ou editais. In *Aurélio Eletrônico*, 2004)
2.3.2.2. Concorrência
2.3.2.3. Convite
2.3.2.4. Pregão
2.3.2.5. Tomada de preços

2.3.3. Contrato
2.3.3.1. Concessão
2.3.3.2. Contrato de obra pública e de prestação de serviço
2.3.3.3. De fornecimento
2.3.3.4. De gestão
2.3.3.5. De convênio
2.3.3.6. De consórcio

2.3.4. Administrativo Público Notarial
2.3.4.1. Testamento
2.3.4.2. Devassa
2.3.4.3. Inventário
2.3.4.4. Certidão
2.3.4.4.1. Certidão de justificação
2.3.4.5. Certificado
2.3.4.6. Treslado
2.3.4.7. Sumário de testemunha
2.3.4.8. Procuração

2.3.5. Abaixo-assinado
2.3.6. ata
2.3.7. Atestação (Declaração escrita e assinada sobre a verdade de um fato, para servir a outrem de documento; atestado, testemunho In *Aurélio Eletrônico*, 2004)
2.3.8. atestado
2.3.9. auto (Registro escrito e autenticado de qualquer ato In *Aurélio Eletrônico*, 2004)
2.3.9.1. auto de abertura
2.3.9.2. auto de anulação
2.3.9.3. auto de assento
2.3.9.4. auto de averiguação
2.3.9.5. auto de arrematação
2.3.9.6. auto de demarcação
2.3.9.7. auto de diligência
2.3.9.8. auto de inquirição
2.3.9.9. auto de justificação

2.3.9.10. auto de posse
2.3.9.10. auto de vereação

2.3.8. foral (1.Carta de lei que regulava a administração duma localidade ou concedia privilégio a indivíduos ou corporações. 2.Carta de aforamento de terras; foro. In *Aurélio Eletrônico*, 2004)
2.3.9. Lançamento (Jur. Ato pelo qual, em certos casos, o juiz afasta da ação penal pública o acusador privado (querelante), por não haver ele apresentado o libelo no devido prazo, declarando-a perempta ou devolvendo-a ao Ministério Público. In *Aurélio Eletrônico*, 2004)
2.3.10. Notificação
2.3.11. Ofício
2.3.12. Carta patente
2.3.13. Carta precatória
2.3.14. Carta de sesmaria
2.3.15. Carta régia
2.3.16. Provimento (5.Jur. Manifestação dos tribunais superiores ao receberem e julgarem favoravelmente o recurso interposto contra decisões dos juízes inferiores. 6.Jur. Instruções ou determinações administrativas baixadas pelo corregedor ao realizar as correições. In *Aurélio Eletrônico*, 2004)
2.3.17. Termo (13.Jur. Peça em que se formaliza determinado ato processual. In *Aurélio Eletrônico*, 2004)
2.3.17.1. Termo de declaração
2.3.17.2. Termo de vereação
2.3.17.3. Termo em junta
2.3.17.4 Termo de testemunha

2.4. Jurídico Comercial
2.4.1. Escritura (ou compromisso)
2.4.2. Registro
2.4.3. Representação (13.Jur. Contrato remunerado, firmado entre dois comerciantes ou empresas comerciais, para que uma parte promova a venda de produtos da outra, efetuando negócios em nome dela, ou realize aproximação de fregueses, etc., mediante condições variáveis em cada caso.
15.Bras. Jur. Pedido que a vítima de certos delitos — ou seus representantes legais — formula à autoridade policial ou judiciária, e bem assim ao órgão do Ministério Público, para que se proceda contra o delinquente, sem o que será nula a ação penal que se intentar na espécie. In *Aurélio Eletrônico*, 2004)

Domínio Discursivo 3 – Científico

3.1. Divulgação
3.1.1. artigo
3.1.2. ensaio
3.1.3. resenha
3.1.4. resumo

3.2. Pesquisa
3.2.1. dissertação
3.2.2. monografia
3.2.3. projeto
3.2.4. tese

Domínio Discursivo 4 – Informativo

4.1. Jornalístico
4.1.1. editorial
4.1.2. entrevista
4.1.3. reportagem

4.2. Informe
4.2.1. aviso
4.2.2. boletim
4.2.3. comunicado

Domínio Discursivo 5 – Referencial

5.1. catálogo
5.2. dicionário
5.3. glossário
5.2. índex
5.5. verbete

Domínio Discursivo 6 – Instrucional

6.1. Didático
6.1.1. apostila
6.1.2. livro-texto

6.2. Procedimental
6.2.1. bula
6.2.2. manual
6.2.3. receita

Domínio Discursivo 7 – Técnico administrativo e/ou oficial

7.1. Comunicacional – entre 2 ou mais pessoas, informando, solicitando, exigindo....
7.1.1. ato
- 7.1.1.1. ato de nomeação
- 7.1.1.2. ato de sujeição e obediência e vassalagem
- 7.1.1.3. aviso
  - 7.1.1.3.1. aviso público

7.1.2. carta
- 7.1.2.1.carta de apresentação
- 7.1.2.2. carta régia
- 7.1.2.3.carta de abrasão de armas de nobreza e fidalguia
- 7.1.2.4.carta de confirmação
- 7.1.2.5.carta de conta
- 7.1.2.6.carta de diligência
- 7.1.2.7.carta de doação
- 7.1.2.8.carta de examinação
- 7.1.2.9.carta de mercê
- 7.1.2.10.carta de nomeação
- 7.1.2.11.carta de ofício
- 7.1.2.12.carta de ordenança
- 7.1.2.13.carta de prego (Carta fechada, na qual se determina o que o comandante de um navio deve fazer, e que ele só deve abrir quando fora da barra.In *Aurélio Eletrônico*, 2004)
- 7.1.2.14.carta de privilégio
- 7.1.2.15.carta de propriedade
- 7.1.2.16.carta de sentença
- 7.1.2.17.carta oficial
- 7.1.2.18. carta-relatório [ver: Cartas Jesuíticas].
- 7.1.2.19. carta de alforria
- 7.1.2.22. carta de sesmaria

7.1.3. circular
7.1.4. declaração
7.1.5. despacho
7.1.6. informação de serviço
7.1.7. memorando
7.1.8. ofício
7.1.9. provisão (5.Documento oficial em que o governo confere cargo, mercê, dignidade, ofício, etc., autoriza o exercício de uma profissão ou expede instruções. In *Aurélio Eletrônico*, 2004)
7.1.10. requerimento
7.1.11. solicitação

7.2. Descritivo

7.2.2. ata

7.2.3. auto de exame médico

7.2.4. balanço (financeiro)

(Livro onde se registram, em ordem cronológica, todas as operações contabilizáveis de uma empresa. In *Aurélio Eletrônico*, 2004)

7.2.4.1. diário (de viagem)

7.2.4.2. diário do governo

7.2.5. informe (Bras. Mil. Qualquer documento, fotografia, mapa, relatório ou observação, relativos ao inimigo ou a uma conjuntura complexa, e que pode contribuir para esclarecer a situação dele ou dela. In *Aurélio Eletrônico*, 2004)

7.2.6. levantamento

7.2.7. ordem do dia (2.Mil. Conjunto de determinações e instruções divulgadas diariamente por comandante militar. In *Aurélio Eletrônico*, 2004)

7.2.8. planejamento

7.2.9. regimento

7.2.9.1. regimento das fronteiras

7.2.10. relatório (relação, lista, quadros demonstrativos...)

7.2.11. memorial descritivo [acompanha desenhos, projetos, plantas de engenharia, arquitetura]

7.3. Comercial

7.3.1. conhecimento (Econ. Documento representativo de mercadoria depositada ou entregue para transporte, e que, se endossado, pode ser negociado como título de crédito. In *Aurélio Eletrônico*, 2004)

7.3.2. contrato

7.3.3. nota

7.3.4. recibo

Domínio Discursivo 8 – Literário

8.1. Prosa

8.1.1. apólogo

8.1.2 biografia

8.1.3. conto

8.1.4. crítica

8.1.5. crônica

8.1.6. ensaio

8.1.7. lenda

8.1.8. libelo (artigo ou escrito de caráter satírico ou difamatório; panfleto In *Aurélio Eletrônico*, 2004)

8.1.9. novela

8.1.10. resenha

8.1.11. romance

8.2. Poesia/poema

8.2. 1. acróstico

8.2. 2.balada

8.2. 3.canção

8.2. 4.cantata

8.2. 5. écloga

8.2. 6.elegia

8.2. 7. epitalâmio

8.2. 8. hino

8.2. 9. idílio

8.2. 10. madrigal

8.2. 11.ode

8.2. 12. quadra (ou trova)

8.2. 13. rondó

8.2. 14. sátira (ou poema satírico),

8.2. 15. sextilha
8.2. 16. soneto
8.2. 17. terceto
8.2. 18. vilancete

8.3. Teatro
8.3.1. ato (Cada uma das maiores partes em que se divide a peça, e cujo número pode variar, ger., de um a cinco. In *Aurélio Eletrônico*, 2004)
8.3.1.1. peça (Texto e/ou representação teatral. In *Aurélio*, 2004)
8.3.2. Auto
8.3.3. Milagre
8.3.4. Farsa
8.3.5. Tragédia
8.3.6. Comédia
8.3.7.Tragicomédia

Domínio Discursivo 9 – Pessoal

9.1. Correspondência
9.1.1. anotação
9.1.2. bilhete
9.1.3. carta (ou missiva)
9.1.4. dedicatória
9.1.5. diário de memórias pessoais
9.1.6. memento (2.Marca que serve para lembrar qualquer coisa. 3.Papel ou caderneta onde se anotam coisas que devem ser lembradas; memorial, memorando, memória. 4.Essa anotação; apontamento, memória. 5.Livrinho onde se acham resumidas as partes essenciais de uma questão. In *Aurélio Eletrônico*, 2004)
9.1.7. elogio

9.2. Documento Pessoal
9.2.1. certidão
9.2.2. certificado
9.2.3. diário
9.2.4. diploma
9.2.5. memorial [de concursos acadêmicos]

APÊNDICE D – Arquivo ARFF em Word

% Este arquivo contém informações sobre os textos do córpus DHPB para a classificação por domínio e gênero
%Estatísticas baseadas em palavras
%EPN - estimativa de itens lexicais diferentes, dado pelo número de itens lexicais diferentes dividido pelo número de itens (type/token ratio)
%EEM - estimativa de itens lexicais diferentes, porém considerando-se apenas os itens iniciados por letra maiúscula (capital type token ratio)
%END - número de dígitos
%ETP - tamanho médio das palavras em caracteres
%EPL - número de palavras longas (com mais de 6 caracteres)

%Estatísticas baseadas no texto como um todo
%ENC - número de caracteres
%TFC - tamanho médio das frases em caracteres
%ENF - número de frases
%TFP - tamanho médio das frases em palavras
%TTP - tamanho do texto em palavras

%Número de ocorrências de determinadas expressões
%VS - verbo SER - é, são, sam
%VH - Verbo HAVER - há, havia
%VP - Verbo PEDIR - pede, pedem
%VPR - Verbo PROVER - proveu, proveo
%VPO - Verbo PODER - póde, podia
%VF - Verbo FAZER - fez, fazer, fazem
%VI - Verbo IR - foi, fomos, fui
%VT - Verbo TER - tem, tinha
%VD - Verbo DIZER - dizer, disse, digo

%PPO1 - pronome pessoal obliquo na primeira pessoa - me, mim, mym, nos, nós, comigo, conosco, comnosco, connosco
%PPO2 - pronome pessoal obliquo na segunda pessoa - te, ti, the, contigo, comtigo, contiguo, vos, vós, convosco
%PPO3 - pronome pessoal obliquo na terceira pessoa - se, lhe, lhes, si, consigo
%PPT - pronome pessoal de tratamento - senhor, senhora, senhoria, excelência, alteza, santidade, majestade, vossa mercê, vossa merce
%PD - pronomes demonstrativos - este, estes, deste, destes, esta, estas, desta, destas, isto, disto, esse, esses, desse, desses, essa, essas, dessa, dessas, isso, disso, aquele, aqueles, aquelle, aquelles, daquele, daqueles, daquelle, daquelles, aquela, aquelas, aquella, aquellas, daquela, daquella, daquelas, daquellas, aquilo, aquillo, daquillo
%ADJ - adjetivo - dito, ditto, odito, oditto
%PRV - pronome relativo variável - a qual, aqual, os quais, os quaes, osquaes, cujo, cujos, cuja, cujas.
%PINT - pronomes interrogativos - quem, quanto, quando, qual, quais, quaes
%PIP - pronomes indefinidos referente à pessoa - alguém, alguem, ninguém, ninguem, outrem, qualquer, quaesquer
%PIL - referente a lugar - onde, aonde, donde, adonde
%PIC - referente a coisas ou pessoas - algo, tudo, nada, todo, todos, toda, todas, vários, várias, certo, certa, pouco, poucos, pouca, poucas, muito, muitos, muyto, muytos, muita, tanto, tantos, tanta, tantas, cada, nenhum, nenhuma
%PREP - tipo e freqüência de preposições - por, per, para, até, athé, athè, té, em, entre, contra, sem, sobre
%ADVL - advérbio de lugar - aqui, abaixo, acima, asima, cá, lá, ai, ahi, ali, alli, além, alèm, dalém, dentro, longe, acolá, aquém, aquem, adiante, atrás, atras, atraz algures, alhures, perto, fora, defronte, embaixo, em baixo

%ADVT - de tempo - hoje, amanhã, ontem, hontem, manhã, já, sempre, nunca, ainda, antes, tarde, cedo, outrora, outrossim, outrosim, depois, depoiz, agora, logo, jamais, diariamente, anualmente, atualmente, sucessivamente, entrementes, imediatamente, immediatamente.
%ADVI - de intensidade - muito, muyto, pouco, poco, bastante, bastantes, mais, maiz, menos, menoz, bem, mal, tão, tao, todo, assaz, tanto, quão, meio, meyo, e os terminados em mente
%OPA - operadores argumentativos - pois, também, tambem, só, apenas, mesmo, mesma, nem, além de, alem de
%MD - marcadores discursivos - mas, porém, assim, como, porque, perque, por que, ainda, depois, ou, então, portanto, por tanto, entretanto, entre tanto, diante

%Expressões
%EXD - deos goarde, deos guarde, deus goarde, deus guarde, guarde deos, guarde deus
%EXFS - faço saber, faso a saber, faso asaber, faco saber, faso saber
%EXCM - capitão-mor, capitão mor, capitáo-mór, capitão major, capitam maior, capitam mor, capitam major, capitam mayor
%EXL - mihi, non est, domine, domini
%EXPR - prègado, pregado
%EXO - officiaes da camara, officiaes dacamara, officiaes desta câmera, officiaes da camera
%EXA - anno de nascimento, anno do nascimento, anno do nassimento, anno do nascimanto
%EXES - o escrevy, o escrevi, o escreuy, oescrevi, oescrevy, sobescrevy, sobescrevi, escreuy, escreui, escreuj, sobescreui, sobescreuy, asignou, asignei, asigney, tabeliam, tabaliam, escrivam, escrivão
%EXAT - atas da câmara
%EXDO - doutor ouvidor geral e corregedor, ouvidor geral, corregedor, ouvidor, doutor, corregedor, corregedor da comarca
%EXT - termo e certifico, certifico, termo
%EXPI - público instrumento, instrumento, publico

%Unidades lexicais
%ULE - leste, oeste, norte, sul, nordeste, sudeste, noroeste, sudoeste
%ULD - dia
%ULL - léguas, leguas, legoas, llegoas, leguoas
%ULS - suplicantes, suplicante, supplicantes, supplicante, osuplicante, osupplicante
%ULQ - que, qui, se
%ULT - testemunha
%ULJ - juiz, juízes, juis
%ULSA - deus, santo, misericórdia, sacramento, graça, alma, almas, corpo, fé
%ULG - guerra
%ULMA - tamanho, grande, pequeno, carne, folhas, arvores, árvores, rio, riacho, cor, espécies, vento, mar, serra, serras, águas, aguas
%ULTR - terra, terras, cidade, estado, vila, villa, comarca
%ULP - índios, indio, gentio, povo, homem
%ULDV - devassa, devaça
@ATTRIBUTE CLASSE {diario, assento, autoprov, escritura, parecer, registro, sermao, ton, termo}

0.56,0.66,0,5.53,274,4478,1119.5,4,199,796,1,1,0,0,0,4,1,1,3,0,0,3,3,14,1,1,7,0,2,2,23,2,5,4,6,12,0,0,0,0,0,1,0,0,2,0,3,1,0,1,0,0,36,0,0,0,0,2,11,2,0,assento
0.56,0.66,0,5.30,148,2572,857.33,3,158.33,475,1,0,0,0,0,1,1,0,0,2,0,1,3,4,5,1,1,0,0,3,9,2,4,4,1,3,0,0,0,0,0,1,0,0,0,0,2,1,0,0,0,0,20,0,0,0,0,0,4,4,0,assento
0.47,0.64,4,5.09,476,8974,1282,7,244.71,1713,2,0,0,0,0,2,2,1,1,2,0,12,4,24,16,1,4,1,1,5,35,3,7,23,4,29,1,0,0,0,0,0,0,0,1,0,1,3,0,1,0,0,57,0,0,3,2,2,13,1,0,assento
0.53,0.55,0,5.26,179,3089,1029.66,3,192.33,577,0,0,0,0,0,0,2,0,4,0,0,1,5,12,5,0,4,0,1,3,18,3,0,4,0,6,0,0,0,0,0,4,0,5,0,0,3,2,0,0,0,0,30,0,1,0,0,0,12,4,0,assento
0.64,0.79,0,5.75,59,850,850,1,144,144,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,3,1,0,3,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,9,0,0,0,0,0,1,1,0,assento
0.63,0.63,0,6.07,150,2020,505,4,81.5,326,0,1,0,0,0,0,0,0,2,0,0,1,0,6,4,2,0,0,0,0,4,4,0,2,1,1,0,0,0,0,0,1,0,0,1,0,1,0,0,0,0,0,0,14,0,0,0,0,0,6,3,0,assento
0.45,0.70,107,5.94,181,2819,234.91,12,36.91,443,0,0,0,0,1,0,0,0,0,0,0,2,0,5,2,2,1,0,0,0,13,1,2,4,1,0,0,0,0,0,0,0,0,2,1,0,2,1,0,0,0,0,11,0,0,0,0,0,4,1,0,assento
0.65,0.70,0,6.19,155,2372,593,4,94,376,0,0,0,0,1,0,1,2,1,2,0,2,1,11,0,0,1,0,0,1,7,1,3,6,0,2,0,0,0,0,0,0,0,0,0,0,3,1,0,0,0,0,19,0,0,0,0,0,5,4,0,assento

0.69,0.84,0,6.00,97,1572,524,3,85.33,256,0,1,0,0,1,1,1,0,2,2,0,1,0,6,0,0,0,0,0,1,6,1,0,2,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,10,0,1,0,0,0,3,2,0,assento
0.70,0.89,0,5.62,75,1211,605.5,2,105.5,211,0,0,0,0,0,1,1,0,0,0,0,0,0,3,3,0,1,0,0,3,10,0,0,2,1,4,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,15,0,0,0,0,0,2,1,0,assento
0.25,0.29,218,5.59,2732,44552,578.59,77,100.48,7737,6,6,0,0,6,11,12,8,19,14,0,29,19,129,48,11,24,1,4,27,188,27,29,81,21,84,1,0,0,0,0,9,0,9,7,0,28,15,0,2,0,0,323,0,3,3,5,5,91,38,0,assento
0.37,0.64,35,4.83,348,5975,271.59,22,54,1188,3,0,0,2,0,5,1,1,0,5,0,5,2,21,3,3,1,0,2,15,39,1,0,21,14,15,0,0,0,0,3,1,19,4,0,10,0,0,0,0,0,57,0,2,4,0,0,8,1,0,autoprov
0.29,0.70,7,4.64,168,3444,7.99,431,1.48,640,1,0,0,3,0,5,0,0,0,1,0,3,1,11,6,2,2,0,2,5,25,1,0,15,6,7,0,0,0,0,0,1,1,11,1,2,6,0,0,0,0,0,0,34,0,2,0,0,1,5,0,0,autoprov
0.33,0.67,11,4.60,150,3207,11.96,268,2.35,630,1,0,0,2,0,8,2,2,1,2,0,8,1,11,5,2,1,0,0,5,27,2,2,8,3,7,0,0,0,0,0,1,1,18,4,0,5,0,0,0,0,0,0,38,0,0,4,0,6,4,0,0,autoprov
0.35,0.59,42,4.73,524,9826,81.20,121,16.72,2024,2,1,0,2,0,5,1,1,0,6,0,17,3,34,3,5,8,4,2,18,81,14,3,26,9,22,0,0,0,0,0,7,1,28,10,1,10,0,0,3,0,0,114,1,0,4,0,2,24,1,2,autoprov
0.40,0.57,21,4.75,252,4449,78.05,57,16.01,913,1,0,0,3,0,5,1,0,0,1,0,4,3,14,5,2,3,0,1,7,33,2,1,10,2,2,1,0,0,0,0,5,1,26,5,0,8,0,0,0,0,0,0,45,0,2,2,0,2,12,1,2,autoprov
0.40,0.60,16,4.71,249,4683,234.15,20,48.8,976,5,1,0,2,0,5,2,1,1,2,0,9,3,13,3,2,2,0,2,11,35,1,1,26,5,12,0,0,0,0,0,0,1,25,3,2,9,0,0,0,0,0,0,54,0,1,1,0,0,6,1,0,autoprov
0.43,0.63,6,4.58,149,2961,370.12,8,78.62,629,1,0,0,3,0,6,2,0,0,0,0,5,1,8,2,2,2,0,2,1,23,3,2,11,1,3,0,0,0,0,0,1,1,10,1,1,4,0,0,0,0,0,35,0,1,1,0,1,7,1,0,autoprov
0.44,0.60,22,4.78,239,4442,317.28,14,64.57,904,1,0,0,6,0,6,2,3,1,3,0,4,1,13,1,3,3,0,4,7,30,0,1,12,6,13,0,0,0,0,0,1,1,16,1,0,5,0,0,1,0,0,50,0,1,0,0,1,9,1,0,autoprov
0.31,0.49,81,4.50,434,8355,214.23,39,46.10,1798,0,1,0,10,0,6,1,0,1,2,0,8,1,25,1,3,7,1,4,20,56,3,3,14,13,30,0,0,0,0,0,1,1,5,13,3,7,0,0,2,0,0,115,0,1,0,0,3,19,5,0,autoprov
0.34,0.51,3,4.65,423,8486,353.58,24,75.04,1801,3,0,0,6,1,6,1,1,2,4,1,16,3,32,18,9,15,2,3,9,72,3,3,34,10,18,1,0,0,0,0,1,1,23,8,2,12,0,0,0,0,0,0,114,0,0,3,0,0,20,4,0,autoprov
0.14,0.23,349,4.66,4444,86328,68.73,1256,14.28,17936,35,5,0,62,3,84,19,17,14,36,2,126,30,296,85,53,76,12,32,144,660,44,28,301,101,191,5,0,0,0,0,31,14,263,88,23,138,0,0,6,0,0,1007,1,23,30,0,26,194,27,4,autoprov
0.21,0.54,241,4.70,2114,39660,177.84,223,36.18,8069,10,10,0,1,2,21,19,17,4,16,5,42,41,70,11,20,37,1,32,79,307,46,63,206,71,88,0,0,0,0,0,0,0,0,0,0,1,0,0,25,60,6,0,633,0,0,16,3,58,7,33,0,diario
0.13,0.32,366,4.43,1581,34183,196.45,174,42.42,7382,7,5,0,0,1,15,37,34,3,134,0,19,17,14,16,13,32,0,72,49,344,60,119,122,53,95,0,0,10,0,0,0,0,0,0,0,0,0,0,5,179,41,0,389,0,0,1,1,109,8,13,0,diario
0.15,0.33,544,4.43,2044,46265,89.48,517,19.05,9850,30,15,0,0,3,28,58,22,21,85,0,24,40,106,33,3,48,2,21,34,449,86,60,181,113,180,2,0,0,0,0,0,0,0,0,0,0,64,62,42,0,580,0,0,18,0,92,21,13,0,diario
0.15,0.39,69,4.26,1551,37360,232.04,161,52.45,8445,78,25,0,0,2,23,22,50,3,43,0,29,4,142,11,11,31,3,15,120,449,98,89,220,66,183,0,0,0,0,0,0,0,1,0,0,4,0,21,27,19,0,478,0,4,4,0,140,16,1,0,diario
0.22,0.52,69,4.41,763,15936,143.56,111,31.25,3469,16,5,0,0,2,12,20,10,1,33,0,16,2,60,6,13,17,7,16,28,125,27,54,52,36,75,1,0,0,0,0,0,0,0,0,0,0,23,19,0,0,155,0,0,1,2,83,6,0,0,diario
0.15,0.43,193,4.58,3546,73342,139.43,526,29.36,15447,71,30,0,0,14,24,82,78,27,273,0,142,3,214,54,67,165,4,90,122,601,108,164,285,174,368,0,0,0,0,0,0,0,2,0,0,1,0,6,42,2,0,822,0,0,5,4,163,114,19,0,diario
0.18,0.34,240,4.59,3515,69044,146.90,470,30.81,14483,78,18,1,0,24,35,33,66,17,247,1,132,10,240,25,33,101,13,31,171,567,94,139,334,167,370,5,1,0,0,0,0,0,6,0,0,0,0,2,29,2,0,763,1,0,17,9,68,80,10,0,diario
0.24,0.46,121,4.61,840,17027,205.14,83,42.75,3549,6,0,0,0,1,4,13,16,4,53,0,16,0,58,10,9,19,0,16,18,155,15,46,59,8,61,0,0,0,0,0,0,0,5,0,0,0,0,27,25,25,0,165,0,0,0,0,55,3,10,0,diario
0.11,0.27,704,4.33,2700,63255,237.80,266,52.10,13860,134,6,0,0,11,20,53,43,3,53,2,42,3,261,38,28,39,6,34,87,615,107,67,217,119,187,0,0,0,0,0,0,0,0,0,1,1,0,0,210,81,173,0,635,0,0,4,0,339,60,6,0,diario
0.22,0.60,83,4.54,1621,31858,212.38,150,44.96,6744,90,3,0,0,5,5,9,23,1,51,0,19,2,110,7,13,30,0,29,58,256,48,49,152,57,110,0,0,0,0,0,0,0,0,1,1,3,0,0,27,11,5,0,290,0,2,4,0,109,49,11,0,diario
0.06,0.20,5812,4.40,38476,847576,131.24,6458,28.53,184257,889,315,3,1,103,491,667,683,156,1928,25,880,191,2219,450,348,919,70,644,1555,7014,1371,1406,3425,1422,2943,14,2,11,0,0,0,0,18,3,7,11,0,1062,1244,662,0,9918,2,11,115,49,2894,807,222,0,diario
0.36,0.62,6,4.35,275,5865,325.83,18,73.27,1319,0,0,0,0,0,4,3,1,3,7,1,18,3,9,12,1,5,2,0,12,42,2,6,12,1,8,0,0,0,0,0,0,2,0,0,7,5,0,2,0,0,0,40,0,3,7,0,1,12,0,0,escritura
0.34,0.59,8,4.58,449,8498,293.03,29,63.31,1836,6,0,0,0,1,6,6,4,1,13,0,18,3,17,12,1,14,3,2,12,73,4,11,16,9,26,0,0,0,0,0,0,0,0,0,7,10,0,0,1,0,0,66,0,2,1,0,0,12,0,0,escritura
0.34,0.50,7,4.56,318,6196,364.47,17,78.82,1340,10,0,0,0,0,2,4,0,0,12,0,6,1,18,30,0,8,0,4,10,45,5,3,6,2,6,0,0,0,0,0,0,1,0,3,5,4,0,0,1,0,0,56,0,7,0,0,0,7,0,0,escritura
0.38,0.59,8,4.44,311,6143,767.87,8,170.62,1365,2,1,0,0,1,2,6,1,2,9,0,15,1,11,14,3,15,1,2,14,56,13,12,21,5,20,0,0,0,0,0,0,0,0,0,1,4,4,0,2,0,0,0,52,0,0,1,0,0,9,0,0,escritura

0.39,0.62,9,4.54,394,7647,246.67,31,53.64,1663,6,0,0,0,0,3,4,3,3,9,0,22,1,16,10,1,10,1,1,13,49,11,7,24,8,17,0,0,0,0,0,0,0,0,1,6,4,0,0,0,1,0,92,2,1,1,0,4,16,0,0,escritura
0.27,0.43,17,4.50,726,14655,444.09,33,97.12,3205,5,2,0,0,0,7,8,5,13,24,0,33,15,38,30,2,20,6,7,23,108,11,13,46,3,29,0,1,0,0,0,0,0,0,0,3,12,13,0,0,4,0,1,126,0,6,1,0,0,20,0,0,escritura
0.42,0.53,11,4.70,230,4327,309.07,14,64.64,905,6,0,0,0,0,2,1,1,3,4,0,11,2,8,9,0,3,3,1,8,24,4,8,9,2,11,0,0,0,0,0,0,0,1,3,4,0,0,0,0,0,39,0,0,4,0,0,12,0,0,escritura
0.36,0.51,4,4.48,295,5865,217.22,27,47.77,1290,6,1,0,0,0,1,2,2,2,6,0,16,3,9,23,3,8,0,2,14,44,9,5,7,5,18,0,0,0,0,0,0,0,0,1,3,4,0,0,2,0,0,57,0,1,2,0,2,20,0,0,escritura
0.39,0.61,6,4.37,277,5863,651.44,9,147.22,1325,1,0,1,0,0,4,2,3,2,7,1,17,1,15,14,1,10,1,2,17,55,5,12,13,3,13,0,0,0,0,0,0,0,0,2,4,6,0,0,1,0,3,52,0,1,0,0,0,10,0,0,escritura
0.36,0.55,8,4.40,305,6304,274.08,23,61.60,1417,6,0,3,0,0,4,3,3,4,11,0,16,3,19,19,1,5,1,1,8,46,5,6,21,1,11,0,0,0,0,0,0,0,0,1,6,8,0,3,2,0,0,44,0,0,0,0,0,12,0,0,escritura
0.11,0.23,157,4.46,6672,135885,339.71,400,75.1,30040,101,9,5,0,8,62,79,49,81,191,4,353,62,310,318,35,191,30,44,301,1085,142,178,322,65,319,0,2,0,0,0,0,3,9,25,104,116,0,10,22,1,6,1211,3,23,23,1,13,258,2,0,escritura
0.56,0.84,9,4.72,81,1653,551,3,112.66,338,0,0,0,0,0,0,0,1,2,2,0,2,0,5,0,0,1,0,2,2,8,0,0,6,2,9,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,21,0,0,0,0,3,1,3,1,parecer
0.57,0.78,14,4.79,119,2209,157.78,14,32.14,450,0,0,0,0,0,1,0,2,0,1,0,3,2,9,0,1,0,0,0,1,11,1,0,3,1,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,28,0,0,1,2,2,0,3,2,parecer
0.46,0.80,14,4.57,135,2798,466.33,6,98,588,0,0,0,0,0,4,0,6,2,3,0,5,2,11,0,0,0,0,0,9,16,0,1,17,5,8,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,54,0,0,1,2,5,2,3,3,parecer
0.27,0.49,138,4.62,505,11796,40.39,292,8.19,2392,1,0,0,0,0,6,0,6,5,4,0,25,1,18,0,3,5,2,5,30,36,2,6,16,23,52,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,169,6,0,2,10,11,2,4,2,parecer
0.47,0.74,27,4.93,155,2724,75.66,36,14.69,529,1,0,0,0,0,2,0,1,0,2,0,3,1,3,1,0,2,0,1,6,8,1,2,5,1,7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,34,10,0,1,2,3,1,3,2,parecer
0.47,0.74,27,4.93,155,2724,75.66,36,14.69,529,1,0,0,0,0,2,0,1,0,2,0,3,1,3,1,0,2,0,1,6,8,1,2,5,1,7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,34,10,0,1,2,3,1,3,2,parecer
0.30,0.38,41,4.86,623,12335,513.95,24,103.37,2481,5,1,0,0,0,11,1,5,1,14,0,21,1,23,0,16,8,2,1,27,61,0,11,29,36,45,1,0,0,0,0,0,0,0,0,2,0,0,0,1,0,0,205,0,0,2,21,17,7,5,10,parecer
0.35,0.70,10,4.70,424,8376,194.79,43,39.93,1717,13,0,1,0,0,4,4,2,3,16,0,15,0,37,2,4,12,2,1,9,63,3,17,49,22,41,1,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,88,0,0,4,1,5,5,3,0,parecer
0.28,0.52,148,4.30,1030,23694,164.54,144,36.83,5304,0,0,0,0,1,9,6,28,8,11,0,25,1,94,0,1,16,2,2,51,125,7,37,39,46,62,0,0,0,0,0,0,0,0,0,1,0,0,4,1,0,0,177,0,0,0,4,7,20,2,0,parecer
0.40,0.62,36,4.79,125,2469,129.94,19,25.05,476,0,0,0,0,0,2,0,3,0,5,0,1,1,16,1,0,2,0,0,2,15,1,0,6,3,6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,40,0,0,0,4,9,2,5,3,parecer
0.20,0.35,585,4.64,4453,91969,126.85,725,26.22,19016,55,1,1,0,1,61,11,65,21,87,0,144,11,315,10,34,77,9,17,190,517,29,106,259,178,323,2,0,0,0,0,0,0,3,0,3,0,0,9,5,4,0,1095,26,0,22,60,99,52,40,33,parecer
0.40,0.41,13,4.78,175,3314,165.7,20,34.1,682,2,0,2,0,0,1,0,0,0,2,0,7,7,6,1,1,2,0,5,7,18,1,0,1,0,4,0,1,1,0,0,0,0,0,0,0,1,0,3,1,4,2,23,0,0,0,0,12,10,0,0,registro
0.49,0.66,13,5.16,197,3690,194.21,19,37.05,704,1,0,5,0,1,0,3,1,1,8,0,9,2,5,3,1,4,0,3,1,14,5,1,6,3,8,6,0,3,0,0,0,0,0,3,0,5,0,0,0,5,4,32,0,0,6,0,2,11,1,0,registro
0.56,0.78,14,5.36,169,3191,1063.66,3,196,588,0,0,4,0,0,0,0,0,2,5,0,5,3,3,1,1,4,0,1,2,13,5,0,6,2,2,6,1,2,0,0,0,0,0,0,0,4,0,0,0,4,3,26,0,0,6,0,3,3,0,0,registro
0.54,0.73,14,5.29,171,3073,0,0,0,573,1,0,6,0,0,0,1,2,0,6,0,8,4,4,1,1,3,0,0,3,14,4,0,6,1,3,5,0,2,0,0,0,0,0,0,0,0,4,0,0,1,2,4,22,0,0,5,0,6,5,1,0,registro
0.54,0.78,14,5.18,171,3335,0,0,0,635,0,0,6,0,0,0,1,3,0,6,0,8,4,5,2,1,4,0,0,2,17,5,1,6,1,5,5,1,1,0,0,0,0,0,0,1,0,4,0,0,1,2,4,28,0,0,5,0,11,6,1,0,registro
0.54,0.71,14,5.21,187,3382,563.66,6,107,642,0,0,8,0,0,0,1,1,1,5,1,4,5,4,0,1,3,0,0,4,14,4,1,10,4,10,4,0,1,0,0,0,0,0,0,5,0,0,0,2,1,28,0,0,4,0,6,4,0,0,registro
0.57,0.76,14,5.28,178,3217,3217,1,601,601,0,0,5,0,0,0,1,3,0,6,0,7,4,3,1,1,3,0,0,6,16,4,0,9,2,5,5,0,2,0,0,0,0,0,1,0,4,0,0,0,2,4,24,0,0,5,0,4,8,0,0,registro
0.52,0.73,14,5.17,186,3442,573.66,6,109.5,657,0,0,8,0,0,0,1,0,1,6,0,6,6,4,3,4,4,0,0,3,25,6,1,6,2,2,4,0,0,0,0,0,0,0,0,0,0,0,1,5,20,0,1,5,0,6,5,0,0,registro
0.53,0.72,14,5.14,171,3158,789.5,4,151.5,606,0,0,6,0,0,0,1,3,0,5,0,7,4,5,2,1,4,0,1,2,15,5,0,6,1,2,6,0,1,0,0,0,0,0,0,0,4,0,0,0,2,3,27,0,0,6,0,7,6,0,0,registro
0.43,0.47,13,4.66,191,3806,158.58,24,33.54,805,1,0,2,0,0,2,1,1,1,3,0,8,5,9,2,0,0,0,0,6,19,4,0,10,3,4,0,1,2,0,0,0,0,0,1,0,4,0,0,0,1,4,30,0,0,0,0,7,9,0,0,registro
0.13,0.21,369,5.09,4003,73379,237.47,309,45.82,14159,9,0,78,0,1,6,12,27,11,96,4,118,83,147,35,37,72,0,13,58,415,73,8,135,45,97,49,14,41,0,0,15,9,3,19,0,55,0,5,3,61,80,470,0,1,43,0,131,191,4,0,registro

0.26,0.57,0,4.46,895,18418,167.43,110,35.93,3953,42,12,1,0,10,6,7,11,15,14,3,26,9,49,2,3,27,1,9,53,108,9,26,119,47,133,0,0,0,3,0,0,0,0,0,0,0,0,1,0,0,232,0,0,17,10,6,5,4,0,sermao
0.22,0.41,2,4.50,1898,39885,189.02,211,40.28,8501,78,14,0,0,6,26,21,3,20,69,19,55,20,134,0,2,63,3,9,108,270,22,80,237,95,242,1,0,0,2,0,0,0,0,0,1,0,0,0,7,2,0,650,0,0,42,17,28,37,12,0,sermao
0.21,0.33,12,4.51,2248,46352,176.24,263,37.59,9887,79,7,1,1,19,8,40,26,13,91,15,93,14,132,0,6,46,2,18,118,307,33,49,247,171,324,1,0,0,7,1,0,0,0,1,1,0,0,0,11,0,0,557,0,0,125,18,30,87,7,0,sermao
0.20,0.42,2,4.42,2099,44494,158.90,280,34.56,9678,115,10,0,0,18,32,15,19,11,45,19,106,9,129,1,8,83,12,22,110,289,45,55,266,181,323,1,0,0,2,0,0,0,0,0,1,0,0,0,3,1,0,686,0,0,81,1,28,39,17,0,sermao
0.18,0.42,5,4.37,1707,38231,108.92,351,23.79,8352,136,7,0,0,16,21,15,15,54,55,69,85,14,133,5,1,55,6,15,114,226,47,53,206,92,258,0,0,0,1,1,0,0,0,0,0,0,0,1,22,0,0,581,0,0,50,2,33,49,14,0,sermao
0.21,0.38,6,4.53,2419,47813,180.42,265,38.29,10149,98,5,0,0,6,15,31,16,26,91,43,95,23,177,3,15,68,4,28,78,257,30,79,214,177,288,0,0,0,16,1,0,0,0,0,0,0,0,0,5,0,0,550,1,0,99,14,15,18,7,0,sermao
0.20,0.41,0,4.46,2535,54875,196.68,279,42.36,11820,99,26,1,0,14,29,38,25,28,63,67,107,21,201,7,19,80,4,25,130,303,59,73,271,191,315,0,0,0,2,0,0,0,0,1,0,0,0,3,8,0,0,795,0,1,149,3,18,108,14,0,sermao
0.20,0.43,4,4.25,1327,33129,118.74,279,26.62,7429,110,4,3,1,13,17,18,12,22,54,72,60,7,101,4,2,40,1,5,102,228,25,51,186,80,197,0,0,0,8,0,0,0,0,0,0,1,0,0,12,0,0,440,0,1,139,1,16,51,6,0,sermao
0.16,0.24,7,4.46,3510,71212,189.89,375,40.89,15335,129,17,1,0,19,46,63,31,46,66,36,107,21,219,3,11,106,14,21,215,463,35,78,335,328,468,0,0,0,16,1,0,0,2,2,2,0,0,0,31,0,0,808,0,2,247,2,48,30,28,0,sermao
0.18,0.28,4,4.45,1599,33819,150.30,225,32.35,7280,91,5,1,0,16,7,15,36,21,37,24,47,4,143,1,3,41,3,14,90,200,15,40,130,123,215,0,0,0,1,1,0,0,0,0,0,0,0,0,25,0,0,425,0,0,171,0,34,12,12,0,sermao
0.07,0.15,85,4.42,39528,850536,159.60,5329,34.63,184556,2040,201,20,2,331,417,496,387,487,1346,896,1499,374,3053,51,111,1267,131,341,2165,5459,637,1211,4518,2988,5488,3,0,0,110,18,0,0,10,8,12,2,0,7,257,6,0,11559,2,6,2141,77,505,734,256,0,sermao
0.34,0.51,19,4.28,302,6659,605.36,11,133.81,1472,0,0,0,0,2,3,1,2,0,2,1,18,0,33,3,0,3,2,4,11,45,4,2,25,7,17,0,0,0,0,0,0,0,0,2,0,2,2,0,0,0,0,135,0,1,0,0,2,11,0,0,termo
0.30,0.50,19,3.96,241,6047,403.13,15,95.86,1438,1,0,0,0,1,5,0,4,0,2,7,10,0,24,3,1,7,0,0,25,36,2,0,21,3,20,0,0,0,0,0,0,0,0,5,0,1,2,0,1,0,0,126,0,0,0,0,11,4,0,0,termo
0.32,0.58,23,4.19,342,6707,558.91,12,125.16,1502,0,0,0,0,2,3,0,2,1,7,2,9,2,27,5,0,0,0,1,10,33,1,3,17,2,11,0,0,0,0,0,0,0,0,3,0,2,3,0,0,0,0,89,0,3,0,0,2,7,0,0,termo
0.43,0.64,2,4.32,84,2100,420,5,92.2,461,0,0,0,0,0,2,0,1,0,4,1,8,0,10,0,0,0,0,0,6,8,1,0,9,1,5,0,0,0,0,0,0,0,0,2,0,1,1,0,0,0,0,29,0,3,0,0,3,4,0,0,termo
0.26,0.47,37,3.90,436,11264,1024,11,242.27,2665,2,0,3,0,0,9,7,2,0,6,8,27,0,34,19,0,3,0,0,23,56,2,4,27,5,27,0,0,0,0,0,0,0,0,1,0,0,2,0,3,0,0,141,0,1,0,1,5,14,0,0,termo
0.38,0.51,5,3.95,109,2952,738,4,176.75,707,0,0,0,0,0,2,0,2,0,0,1,2,0,6,10,1,2,1,10,6,21,5,5,6,4,1,0,0,0,0,0,0,0,0,2,0,0,2,0,0,0,0,26,1,0,0,0,3,4,0,0,termo
0.51,0.66,9,5.19,321,6017,752.12,8,139.75,1118,1,0,0,0,2,4,0,2,4,0,0,10,12,12,5,1,3,0,2,3,20,4,3,13,0,11,0,0,0,0,0,0,0,0,1,0,0,2,0,0,0,0,44,0,0,0,0,2,7,0,0,termo
0.58,0.75,8,5.04,152,2981,496.83,6,95.33,572,1,0,0,0,0,1,0,1,0,1,1,3,1,9,1,3,1,0,0,3,17,1,1,6,0,2,0,0,0,0,0,0,0,0,1,0,0,1,0,1,0,3,20,0,0,0,0,0,1,0,0,termo
0.46,0.67,7,4.30,93,2033,338.83,6,75.5,453,0,0,0,0,0,4,0,0,1,0,0,0,3,8,3,0,2,1,0,5,16,1,0,5,2,3,0,0,0,0,0,0,0,0,3,0,1,1,0,0,0,0,30,0,0,0,0,0,2,0,0,termo
0.44,0.56,6,4.95,233,4006,445.11,9,86.33,777,1,0,0,0,0,3,0,1,1,3,0,2,1,5,5,0,0,0,0,4,6,0,2,3,2,1,0,0,0,0,0,0,0,0,4,0,1,2,0,0,0,0,25,0,0,1,0,5,3,0,0,termo
0.18,0.30,265,4.36,3582,79886,499.28,160,108.16,17306,10,0,3,0,8,56,14,25,11,42,28,124,45,249,91,8,28,4,19,158,402,40,27,199,54,146,0,0,0,0,0,0,0,0,0,40,2,12,30,0,9,0,3,1038,1,9,3,2,54,97,2,0,termo
0.16,0.28,177,4.61,2050,38728,207.10,187,43.24,8087,6,5,0,0,5,10,19,31,3,25,0,28,53,57,19,8,15,0,32,63,332,61,27,135,82,60,1,0,0,0,0,0,0,0,0,0,0,0,8,25,7,0,419,0,0,23,5,99,26,35,0,ton
0.39,0.70,32,4.33,134,3077,146.52,21,32.28,678,4,4,0,0,0,2,5,2,0,1,0,0,3,4,0,0,1,0,2,9,22,6,7,17,13,12,0,0,0,0,0,0,0,0,0,0,0,0,0,7,0,0,36,0,0,0,0,12,9,0,0,ton
0.29,0.68,53,4.26,299,6946,210.48,33,46.12,1522,10,4,0,0,0,5,0,9,0,3,0,3,0,9,0,0,0,0,1,4,21,6,1,25,6,25,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,35,0,0,0,0,25,3,1,0,ton
0.20,0.60,164,4.51,1025,21140,173.27,122,35.77,4364,66,58,0,0,0,11,3,25,1,14,0,5,0,27,0,0,0,1,4,23,68,22,9,40,14,45,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,72,0,0,0,1,90,17,5,0,ton
0.27,0.50,58,4.19,446,10408,150.84,69,33.85,2336,2,5,0,0,0,3,0,14,1,2,0,2,0,34,0,0,1,0,12,24,95,6,3,19,14,23,0,0,0,0,0,0,0,0,0,0,0,0,0,0,6,3,3,0,125,0,0,0,0,35,15,4,0,ton
0.24,0.55,117,4.27,500,11704,114.74,102,24.91,2541,40,17,0,0,0,2,0,20,0,10,0,5,0,9,0,0,0,0,9,29,50,6,1,24,13,32,0,0,0,0,0,0,0,0,0,0,0,0,0,7,0,1,0,47,0,0,5,0,62,4,4,0,ton
0.24,0.64,68,4.43,357,8191,163.82,50,34.56,1728,22,16,0,0,0,4,1,10,0,8,0,1,0,9,0,1,2,0,3,8,45,2,5,29,6,22,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,11,0,0,0,0,48,9,2,0,ton

0.19,0.60,124,4.25,920,24347,172.67,141,37.80,5331,56,45,0,0,0,27,0,41,1,19,1,19,0,26,0,0,2,0,9,59,115,
8,14,61,21,56,0,0,0,0,0,0,0,0,0,0,0,0,1,3,1,0,58,0,0,14,0,124,8,7,0,ton
0.19,0.60,124,4.25,920,24347,172.67,141,37.80,5331,56,45,0,0,0,27,0,41,1,19,1,19,0,26,0,0,2,0,9,59,115,
8,14,61,21,56,0,0,0,0,0,0,0,0,0,0,0,0,1,3,1,0,58,0,0,14,0,124,8,7,0,ton
0.24,0.64,211,4.31,598,14396,191.94,75,41.4,3105,26,27,0,0,0,6,0,33,0,11,0,10,0,25,0,1,1,0,13,20,93,7,3,
33,22,51,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,33,0,0,2,0,39,23,0,0,ton