TIPOLOGIA DE TRAÇOS LINGÜÍSTICOS DO PORTUGUÊS DO BRASIL DOS SÉCULOS XVI, XVII E XVIII: UMA PROPOSTA PARA FINS DE CLASSIFICAÇÃO AUTOMÁTICA DE GÊNEROS TEXTUAIS

Jacqueline A. SOUZA[1]

RESUMO: A partir de um corpus do português dos séculos XVI, XVII e XVIII, constituído por 2.459 textos e 7.5 milhões de palavras, pretende-se descrever os traços lingüísticos característicos desses textos, 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 isso, foi aplicada uma tabela de traços contemporâneos, como base para identificação de características do português histórico e utilizaram-se algumas ferramentas computacionais. Diante disso, este artigo relata os primeiros procedimentos para definir uma metodologia, assim como as etapas iniciais para identificação dos traços lingüísticos.

Palavras-chave: Descrição do português; Traços lingüísticos; Gêneros textuais; Classificação automática.

ABSTRACT: From a corpus of Portuguese language from the XVI, XVII and XVIII centuries, constituted of 2459 texts and 7,5 million words it is intended to describe the characteristic linguistic features of these texts, correlating them to their respective genres and to propose a typology of features so that it is possible to automatically identify the genre of each text. In order to do that, a table of contemporary features has been applied, as a basis for identification of the characteristics of the historic Portuguese and some computer tools have also been used. Therefore, this article relates the first procedures in order to define a methodology as well as the initial steps to identify the linguistic features.

Keywords: Description of Portuguese; Linguistic features; Textual genres; Automatic classification.

1. Introdução

Na atual Sociedade da Informação ou Sociedade do Conhecimento, trabalha-se com um volume de dados e informações muitas vezes exorbitantes, e facilitar a recuperação, o tratamento, o acesso à informação tornou-se imprescindível. Isso não é diferente no contexto da Lingüística de *Corpus*, definida por Berber Sardinha (2004) "como a área que se ocupa da coleta e exploração de *corpora*, ou conjuntos de dados lingüísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística, no que tange a manipulação de *corpora"*.

Como proposta para melhorar e facilitar a manipulação e exploração dos *corpora* sugere-se a descrição do português do Brasil para fins de classificação automática de gêneros textuais, e especificamente neste projeto, português histórico dos séculos XVI, XVII e XVIII. Dessa maneira, possibilita descrever-se-á a história de um povo, bem como o funcionamento

[1] Mestranda do Programa de Pós-graduação em Lingüística/ UFSCar

de uma sociedade em um contexto econômico, político e social muito distinto do atual, explicitando informações lingüísticas de importância cultural e histórica.

Neste sentido, a pesquisa tem como objetivo levantar os traços lingüísticos de textos históricos, correlacionando-os aos respectivos gêneros e, propor uma tipologia de traços lingüísticos de forma que seja possível identificar o gênero de cada texto de forma automática. Todavia, o foco do presente artigo é apresentar os procedimentos iniciais para definição de uma metodologia, as ferramentas computacionais utilizadas e resultados parciais referente apenas à descrição dos traços lingüísticos.

2. O Conceito de gênero

A expressão "gênero" circula por diversas áreas do conhecimento, entre vários estudiosos, sejam eles lingüistas, lingüistas computacionais, sociólogos, especialistas em ensino e aprendizagem, o que corrobora uma abordagem cada vez mais multidisciplinar dos estudos de gêneros. Assim, analisá-los pode implicar em analisar textos, discursos, descrição da língua, visão da sociedade em uma perspectiva histórica ou não, categorização, taxonomia e assim por diante. Dado o seu conceito variável, há diversas perspectivas teóricas, as quais, segundo Marcuschi (2008), 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. Nesse sentido, no âmbito desta pesquisa, que aborda questões relacionadas à caracterização, descrição, classificação, os conceitos de gênero basilares são os de Swales (1990) e Biber (1995).

Swales 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 lingüísticas e retóricas. Por meio de seus pressupostos observa-se 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 através dos gêneros pertinentes a ela. Essa noção diz respeito àqueles que trabalham usualmente ou profissionalmente com um determinado gênero e que, desse 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 ressalta Bonini (2001) a cerca da perspectiva de Swales (1990), conhecer o padrão lingüístico 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.

Para Biber (1988, 1995) sob uma perspectiva histórico-cultural e sistêmica, sustenta que o gênero 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. Também, optou por utilizar o termo registro, como se preponderasse o viés sociolingüístico que a palavra registro carrega. Ele considera que registro/gênero como uma categoria mais ampla e abstrata, que congrega vários sub-registros, e são categorias de texto situacionalmente definidas. Embora seu conceito seja semelhante ao de Swales (1990), enfatiza o fato do "gênero/registro ser uma variedade definida por variáveis situacionais e não apenas lingüísticas" (BERBER SARDINHA, 2004). Portanto, 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.

3. A abordagem probabilística da linguagem

Ao estudar a relação entre o léxico e o texto, deve-se ressaltar a intensa ligação entre estes dois campos 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 é que sua atividade é expressa", segundo Hoey (1991).

Semelhante a essa perspectiva, encontra-se a abordagem teórica de Halliday (1994), em que a linguagem é um sistema probabilístico em que certos padrões são mais freqüentes que outros. Assim, o autor descreve a probabilidade dos sistemas lingüísticos, dados os contextos em que os falantes os empregam. Sua visão da linguagem enquanto sistema probabilístico pressupõe que, embora muitos padrões lingüísticos sejam possíveis teoricamente, eles não ocorrem com a mesma freqüência. O exemplo disso, no nível morfossintático, é a freqüência 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 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 freqüência.

Dessa maneira, o mais importante da diferença de freqüências 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 freqüências 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 freqüência maior ou menor de um padrão e um contexto de ocorrência. Ou, nas palavras de Biber (1988, 1995), há uma correlação entre características lingüísticas e situacionais (os contextos de uso). O conjunto da pesquisa desenvolvida por Biber apresenta evidências inequívocas de que conjuntos de padrões lingüísticos 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 freqüência 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 freqüência e as de menor freqüência, sendo que a diferença entre elas é relativa. Assim, algumas palavras têm freqüência 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 freqüência aparecerem.

No que tange a questão de investigar características lingüísticas, Biber, Conrad e Reppen (1998) enfatizam que as técnicas quantitativas não são suficientes e, as interpretações qualitativas são necessárias para examinar as bases funcionais que determinam os padrões de características lingüísticas. Os autores também propõem os seguintes requisitos para o estudo de características lingüísticas e gêneros, como: inclusão de um grande número de textos, consideração de uma ampla escala de características lingüísticas e a comparação das características que identificam os gêneros, pois é dessa comparação que se determina um traço lingüístico.

Com base nisso, foi necessária uma observação preliminar do *corpus* em um nível macrolingüístico, combinada a uma análise microlingüística de identificação de traços recorrentes, com o auxílio de uma metodologia de pesquisa de natureza essencialmente

quantitativa, para avaliação da freqüência das características e categorias lingüísticas co-ocorrentes. A seguir, apresenta-se a metodologia e os principais procedimentos para sua definição.

4. Metodologia

4.1. Desenho do corpus

O *corpus* utilizado foi desenvolvido no âmbito do projeto intitulado ***Dicionário Histórico do Português do Brasil*** – séculos XVI, XVII e XVIII, que integra o Programa Institutos do Milênio do CNPq. O projeto, coordenado inicialmente por Maria Tereza Camargo Biderman[2] (FCL, UNESP, Campus de Araraquara), tem como objetivo elaborar um dicionário do português do Brasil dos séculos XVI, XVII e XVIII a partir de *corpora*. 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, a seguir, apresenta mais algumas informações.

Dados	Valores
Tokens	16.505.808
Types	368.850
Formas simples	7.492.473
Formas simples únicas	368.529
Sentenças	287.570
Textos	2.458

Tabela 1: Desenho do *corpus*

4.2. Philologic

A principal ferramenta utilizada até a etapa de definição de uma metodologia é o Philologic, ferramenta Web desenvolvida 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). Ele foi adaptado por pesquisadores no NILC[3], com o objetivo de manipular o corpus do projeto DHPB, e foi escolhido por atender às necessidades e peculiaridades deste tipo de pesquisa, de modo a auxiliar na busca, extração e recuperação de textos e fragmentos de textos. Além de ser acessível via Web, os

[2] 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.
[3] Núcleo Interinstitucional de Lingüística Computacional - http://www.nilc.icmc.usp.br/nilc/

recursos que a ferramenta oferece são: um concordanciador, um gerador de colocações, um contador de freqüências e um buscador de dados de cabeçalho, que é capaz de listar os documentos do corpus e formar *subcorpus*, pesquisas por metadados do autor, título da obra, data de publicação, busca por similaridade, o que permite abarcar as variações de grafia, como por exemplo, *magestade e majestade.*

4.3 A descrição diacrônica e a tabela de traços lingüísticos

Por se tratar de uma descrição diacrônica, é importante mencionar que o ponto de partida para identificação de traços lingüísticos não está explícito na literatura. Dessa maneira, e de acordo com Berber Sardinha (2004), 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. Para isso, aplicar-se-á a tabela de traços lingüísticos de Aires (2005)[4] que sugere um levantamento estatístico baseado em palavras, como itens lexicais diferentes, palavras longas, assim como estatísticas baseadas no texto como um todo, como número de frases e de caracteres, tamanho do texto em palavras e outros elementos como marcadores discursivos, advérbios de lugar e tempo, pronomes, preposição, artigo, determinadas expressões recorrentes, entre outras.

A partir disso, seguem os procedimentos iniciais para definição de uma metodologia.

4.4 Procedimentos para definição de uma metodologia

4.4.1 Procedimento inicial

Inicialmente, utilizando o Philologic, realizaram-se buscas com as características sugeridas na tabela de Aires (2005), pois seria uma forma de analisar quais tipos de textos eram recuperados, bem como a freqüência e, sobretudo, explorar o *corpus*, por meio da leitura dos textos.

A primeira busca foi com o **pronome na primeira pessoa singular (eu).** Como resultado, obteve-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,

[4] Essa tabela está explicitada no Anexo A.

cartografia, inquirição, petição. Em decorrência dessa busca inicial, esse traço não foi pode ser considerado relevante, no que se refere ao corpus, não em relação a um gênero específico, comparado a outro.

Feito isso, foi realizada a leitura de alguns dos textos, a partir da qual foi possível levantar algumas hipóteses: 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 estrutura: "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"

Outra hipótese é que em textos pertencentes ao domínio jurídico (inventário, testamento, petição, etc.) ocorre a seguinte expressão "Ano de nascimento de nosso senhor Jesus Cristo", como por exemplo:

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".

Além disso, os sermões são iniciados com a expressão "Prègado na", como no exemplo:

1. "Prègado na Igreja de Nsssa Senhora da Ajuda da Bahia, no anno de 1640"

A fim de averiguar se, de fato, essas expressões poderiam ser consideradas caracterizadoras do domínio religioso, ou especificamente dos sermões, foram realizadas buscas com elas, de modo que foi possível verificar em quais gêneros elas ocorriam e se seriam candidatas a traço lingüístico. Assim foi feito com cada hipótese extraída da leitura de cada texto, analisando em quais gêneros os possíveis traços ocorriam, sem considerar sua freqüência, conseqüentemente a predominância ou não em cada gênero, o que é fundamental para descrevê-lo.

Dessa forma, esta primeira etapa foi importante por que permitiu uma rápida leitura e interpretação dos resultados e, sobretudo um olhar sobre o *corpus* e quais seriam as principais dificuldades. Sua maior contribuição está no fato de permitir pensar e definir uma metodologia de trabalho capaz de facilitar a descoberta de traços lingüísticos de cada gênero e, apoiando-se na tabela, passou-se a pesquisar especificamente cada gênero e a levantar suas características.

4.4.2 Sistematização dos procedimentos para identificação dos traços

Diante do que foi exposto acima, foi possível sistematizar os procedimentos para levantar os traços lingüísticos, as características que deverão ser analisadas. A seguir, apresenta-se a sistematização, utilizando o programa Philologic:

1) leitura técnica do texto – "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 freqüentes – auxilia na identificação de unidades lexicais, bem como permite fazer observações e comparar com outros gêneros.

A título de exemplo, na primeira etapa, observou-se no gênero diário a freqüência das palavras *dia, léguas* e *é*, como também verbos na primeira pessoa do plural. Deve-se observar a ocorrência e freqüência dessas unidades lexicais em outros gêneros, bem como trabalhar a hipótese de que há predominância de verbos conjugados na primeira pessoa do plural em diários. Continuando a seqüência de passos, ainda há:

3) identificação de expressões/ unidades lexicais (UL);
4) busca no Philologic com a expressão ou UL, para verificar em quais gêneros ocorria, independentemente da freqüência. Quando uma expressão é encontrada, realiza-se uma busca no *corpus* para ver sua freqüência, quais os gêneros recuperados e se há algum gênero predominante.
5) busca no Philologic por similaridade com a UL (variação de grafia), para verificar outras formas de grafia e em quais gêneros ocorrem. É só então que se define a expressão ou UL candidata a traço lingüístico de determinado gênero ou domínio;
6) observação da freqüência da expressão ou UL no *subcorpus*, bem como a comparação da freqüência entre os gêneros.

A sistematização do processo de identificação de traços lingüísticos foi ao encontro da sugestão que Biber, Conrad e Reppen (1998) propuseram a respeito de comparar as características lingüísticas de cada gênero, para identificar e descrever os traços lingüísticos de gêneros textuais. A partir do que foi feito, foram obtidos alguns resultados referente à identificação dos traços lingüísticos.

5. Análise dos resultados parciais

Referente às expressões que podem ser consideradas um traço lingüístico, seguem alguns exemplos:

- Ano de nascimento de nosso senhor Jesus cristo – domínio jurídico
- faço saber – gênero Carta de doação

- *Pregado na, pregado em – gênero Sermão*
- o escrevy, o escrevi, o escreuy – gênero Assento
- Deos goarde – gênero Registro
- Atas da Câmara – gênero Termo e Assento

Quanto às unidades lexicais que deverão ser analisadas, algumas delas são:dia , léguas, leste, oeste, norte, sul, suplicante, petição, considerando suas variações de grafia.

Assim, no contexto de uma descrição diacrônica, também deverão ser analisados o uso dos pronomes de tratamento, bem como pessoal obliquo, que não foram sugeridos na tabela de traços contemporâneos.

Desse modo, no âmbito da identificação de traços lingüísticos recorrentes em um corpus histórico, esses resultados parciais permitem, preliminarmente, concluir que:

- a variação de grafia pode e deve ser levada em consideração ao definir um traço lingüístico, 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 lingüística como traço lingüístico não é apenas sua freqüência, predominância e ocorrência, mas também sua posição dentro do texto, sua função em um gênero ou domínio, além da imprescindível comparação entre os gêneros.

REFEERÊNCIAS

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.

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

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.

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.

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.

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.

ANEXO - Traços lingüísticos

(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
Os marcadores discursivos “agora”, “da mesma forma”, “de qualquer forma”, “de qualquer maneira” e “desse modo”