Daniel Fernando Pavelec

Identificação da Autoria de Documentos: Análise Estilométrica da Língua Portuguesa usando SVM

Dissertação de Mestrado apresentado ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática.

Curitiba
2007

Daniel Fernando Pavelec

Identificação da Autoria de Documentos: Análise Estilométrica da Língua Portuguesa usando SVM

Dissertação de Mestrado apresentado ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática.

Área de Concentração: Computação Forense e Biometria

Orientador: Prof. Dr. Edson J. R. Justino
Co-orientador: Prof. Dr. Leonardo V. Batista

Curitiba
2007

Pavelec, Daniel Fernando
P337i Identificação da autoria de documentos : análise estilométrica
2007 da língua portuguesa usando SVM / Daniel Fernando Pavelec; orientador, Edson J. R. Justino; co-orientador, Leonardo V. Batista. - 2007.
xi, 98 f. : il. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Paraná, Curitiba, 2007
Bibliografia: f. 74-83

1. Reconhecimento de padrões. 2. Prova pericial - Processamento de dados. 3. Escrita - Identificação. I. Justino, Edson José Rodrigues. II. Batista, Leonardo V. III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. IV. Título.

CDD 21. ed. - 006.4
363.2565

Dedico este trabalho a meus pais Roberto e Climenes, pelo amor e constantes orações.
Dois exemplos de vida !

Agradecimentos

A Deus de todo coração, por tudo e por ser tudo.

A minha família pela compreensão e incentivo.

Ao Professor Edson Justino por sempre acreditar, auxiliar e incentivar. Ao Professor Leonardo Batista e Francisco, que mesmo à distância contribuíram muito com seus comentários e trabalho de referência.

Ao casal de amigos Marisa e Professor Luíz Soares, por guiarem-me nos primeiros passos (Prof. Luíz em vários passos mais...)

Ao Rodrigo Burgos, pelo apoio com a aplicação jurídica.

Aos amigos e colegas do HSBC, em especial Clésio Ramos, por ajudarem-me a conciliar os horários entre trabalho e estudo.

Ao grande amigo Márcio Fuckner, de longa caminhada acadêmica e profissional,... Um incentivador e co-orientador oculto.

A Cris, pelo carinho, palavras de apoio e por ouvir minhas ansiedades e preocupações.

A todos os Funcionários e Professores do PPGIA, em especial Professora Cinthia Freitas e ao Professor Niévola pelas colaborações.

Finalmente a meus tantos amigos, mexicanos e brasileiros, que entenderam minha ausência durante este trabalho.

Sumário

Agradecimentos ii
Sumário iii
Lista de Figuras vii
Lista de Tabelas viii
Lista de Abreviações ix
Resumo x
Abstract xi

Capítulo 1 - Introdução

1.1 Desafio . . . 2
1.2 Motivação . . . 2
1.3 Proposta . . . 2
1.4 Contribuições . . . 3
1.5 Organização . . . 4

Capítulo 2 - Fundamentação Teórica

2.1 Comunicação . . . 5
2.2 Princípios Básicos da Escrita . . . 5
2.3 Linguagem . . . 6
2.4 Língua Portuguesa . . . 7
2.4.1 Origem . . . 7
2.4.2 Padrões . . . 8
2.5 Língua Portuguesa Brasileira . . . 9
2.5.1 Variedade Lingüística . . . 9
2.5.2 Estrutura . . . 12
2.6 Linguagem Escrita . . . 13
2.7 Lingüística . . . 15
2.7.1 Lingüística Aplicada . . . 15
2.7.2 Variação Lingüística . . . 17
2.7.3 Lingüística Forense . . . 17

2.8 Estilística . . . 18
2.8.1 Estilística Lingüística . . . 18
2.8.2 Estilo . . . 18
2.8.3 Estilística Forense . . . 19
2.8.4 Estilometria . . . 20
2.9 Aplicação Jurídica . . . 21
2.9.1 Aplicação . . . 21
2.9.2 Prova . . . 21
2.9.3 Procedimento Probatório . . . 22
2.9.4 Prova Pericial . . . 23
2.9.5 Decisão do Juiz ou Tribunal . . . 24
2.10 Reconhecimento de Padrões . . . 25
2.10.1 Técnica de Aprendizado de Máquina - SVM . . . 26
2.10.1.1 SVM - Duas Classes . . . 26
2.10.1.2 $SVM^{multiclass}$ . . . 27
2.10.2 Atributos Estilométricos . . . 28
2.10.3 Vetores de Dissimilaridade . . . 30
2.10.4 Dicotomia . . . 31
2.10.5 Abordagens de Automatização . . . 32
2.10.6 Modelos de Classificação . . . 32
2.10.6.1 Modelo Global . . . 32
2.10.6.2 Modelo Pessoal . . . 32

Capítulo 3 - Estado da Arte

3.1 Histórico . . . 35
3.2 Identificação da Autoria - Estilística . . . 38
3.2.1 Probabilísticas e Estatísticas . . . 38
3.2.2 Computacionais . . . 39
3.2.2.1 PPM-C . . . 40
3.2.3 Aprendizado de Máquina . . . 41
3.3 Análise Crítica do Estado da Arte . . . 42

Capítulo 4 - Método Proposto

4.1 Abordagem . . . 43
4.2 Método de Identificação da Autoria . . . 43
4.3 Coleta e Formação da Base de Dados . . . 44

4.4 Extração das Características . . . 46
4.5 Geração dos Vetores de Dissimilaridade . . . 50
4.6 Geração dos Modelos . . . 51
4.6.1 Modelo Global . . . 51
4.6.1.1 SVM^{light} . . . 52
4.6.2 Modelo Pessoal . . . 52
4.7 Processo de Testes . . . 53
4.7.1 Modelo Global . . . 53
4.7.2 Modelo por autor - Multi-Classe . . . 54
4.8 Processo de Decisão . . . 55
4.8.1 Modelo Global . . . 55
4.8.2 Modelo por autor - Multi-Classe . . . 55

Capítulo 5 - Experimentos Realizados e Análise dos Resultados

5.1 Experimentos . . . 57
5.1.1 Modelo Global Utilizando SVM . . . 57
5.1.1.1 Divisão da Base de Dados . . . 57
5.1.1.2 Protocolo de Treinamento . . . 58
5.1.1.3 Protocolo de Testes . . . 59
5.1.1.4 Resultados . . . 60
5.1.2 Modelo por Autor Utilizando SVM Multiclasse . . . 62
5.1.2.1 Divisão da Base de Dados . . . 62
5.1.2.2 Protocolo de Treinamento . . . 62
5.1.2.3 Protocolo de Testes . . . 63
5.1.2.4 Resultados . . . 67
5.1.3 Modelo por Autor Utilizando PPM . . . 67
5.1.3.1 Divisão da Base de Dados . . . 67
5.1.3.2 Pré-processamento . . . 68
5.1.3.3 Protocolo de Treinamento . . . 68
5.1.3.4 Classificação . . . 68
5.1.3.5 Protocolo de Testes . . . 68
5.1.3.6 Resultados . . . 68
5.2 Comparativo entre os Experimentos . . . 69

Conclusão 72

Referências Bibliográficas 74

Apêndice A - Troca de E-mails sobre a autoria da "Uma Elegia Fúnebre"

Apêndice B - Tabela de Autores - Base de Dados

Apêndice C - Tabela de Atributos Estilométricos

C.1 Atributos da Língua Portuguesa 89

Apêndice D - Recentes Casos de Uso (Brasil)

D.1 Juíza se afasta de casos com o Opportunity 91

D.2 Caso Procurador Luiz Francisco de Souza 93

D.3 Exemplo de uma Coluna Inteira - Base de dados 96

Lista de Figuras

2.1 Evolução da Letra [Fab06] 6
2.2 Evolução do Alfabeto Fenício [Fra06] 6
2.3 Exemplo de Modelo Formal de Escrita [Mig04] 11
2.4 Exemplo de Modelo Informal - Webblogger 27/09/2004 00:14 12
2.5 Classificação entre duas classes W_1 e W_2 usando hiperplanos: (a) Hiperplanos arbitrários l_i e (b) hiperplano com separação ótima, máxima margem. 27
2.6 Agrupamentos e exemplos de características do estilo (Adaptado [AC05]) . 29
2.7 Exemplo de transformação: policotomia (a) → dicotomia (b) 31
2.8 Modelo Global de Identificação da Autoria de Textos 33

3.1 Exemplo de Divisão do campo da análise da autoria - Foco Estilometria. . 34
3.2 Modelo PPM-C depois do processamento da string abracadabra [CMRJB04] 41

4.1 Diagrama esquemático das etapas - Estilometria 44
4.2 Apresentação eletrônica de uma coluna 46
4.3 Exemplo de Vetor de Dissimilaridade 51
4.4 Fluxo com $SVM^{multiclass}$ 53
4.5 Fluxo do processo de testes 54
4.6 Exemplo de um arquivo resultado do SVM^{light} 55

5.1 Exemplo da combinação para vetores de autoria - Treinamento 58
5.2 Exemplo das combinações para vetores de não-autoria - Treinamento . . . 59
5.3 Exemplo das combinações para vetores de autoria - Testes 59
5.4 Exemplo das combinações para vetores de não-autoria - Testes 60
5.5 Exemplo dos vetores de autoria gerados para um modelo multi-classe . . . 62
5.6 Exemplo dos vetores de autoria gerados para um autor para treinamento . 63
5.7 Exemplo Fluxo Testes - Modelo Multi-classe 64

Lista de Tabelas

2.1 Grafia - Português Brasileiro vs. Português Europeu 8
2.2 Níveis Lingüístico na língua falada e escrita 16

4.1 Colunas do autor Antônio Pietrobelli 45

5.1 Protocolo de Testes - Modelo Global - SVM 60
5.2 Execuções aleatórias para obtenção de uma boa "semente" de treinamento 61
5.3 Execuções aleatórias para obtenção de uma boa "semente" de referência . . 61
5.4 Protocolo Base Autores 1 a 10 - Documentos para treinamento 1 a 5 64
5.5 Protocolo Base Autores 1 a 10 - Documentos para treinamento 6 a 10 . . . 65
5.6 Protocolo Base Autores 1 a 10 - Documentos para treinamento 11 a 15 . . 65
5.7 Protocolo Base Autores 11 a 20 - Documentos para treinamento 1 a 5 . . . 65
5.8 Protocolo Base Autores 11 a 20 - Documentos para treinamento 6 a 10 . . 65
5.9 Protocolo Base Autores 11 a 20 - Documentos para treinamento 11 a 15 . . 65
5.10 Protocolo Base Autores 1 a 20 - Documentos para treinamento 1 a 5 66
5.11 Protocolo Base Autores 1 a 20 - Documentos para treinamento 6 a 10 . . . 66
5.12 Protocolo Base Autores 1 a 20 - Documentos para treinamento 11 a 15 . . 66
5.13 Resultados - Modelo por Autor - SVM Multi-classe 67
5.14 Resultados - Modelo por Autor - PPM-C 69
5.15 Taxa de Acerto - Comparativo PPM-C e SVM Multi-classe 70

B.1 Lista de Autores - Base de Dados 87

C.1 Características da Língua Portuguesa 89

Lista de Abreviações

PPM-C	*Variação do algoritmo de compressão PPM*
CPLP	*Comunidade dos Países de Língua Portuguesa*
IALA	*Association Internatiole de Linguistique Appliquée*
SVM	*Support Vector Machine*
MRS	*Minimização do Risco Estrutural*
QCS	*Quintus Curtius Snodgrass*
PPM	*Prediction by Partial Matching*
RBF	*Radial Basis Function*
RNA	*Redes Neurais Artificiais*

Resumo

Baseado nos estudos da formação da língua portuguesa brasileira, é constatada a riqueza em atributos estilométricos discrimatórios, e que se, cientificamente forem contextualizados e bem aplicados, podem auxiliar expressivamente na correta solução de casos forenses envolvendo textos de língua portuguesa. Este trabalho tem por finalidade propor uma metodologia científica na busca da identificação da autoria de documentos questionados, com conteúdo reduzido, baseado na análise de atributos estilométricos da língua portuguesa, utilizados pelo autor. As abordagens propostas baseiam-se em estilometria. São apresentados dois métodos. Ambos utilizam-se de *3-gram* de palavras-função (conjunções e advérbios) como características estilométricas. O primeiro método utiliza um modelo global com classificação através de SVM pacote SVM^{light}. Neste modelo somente duas classes são assumidas através de dicotomia: autoria e não autoria. O segundo método proposto utiliza modelo por autor com classificação multi-classe através do pacote $SVM^{multiclass}$. Nesta abordagem é gerado um único modelo, porém cada entrada é identificada através da classe a qual pertence. Para validar ambos modelos foi utilizada uma base de 30 autores e 15 documentos cada. As seguintes etapas constituem ambas abordagens: (1) formação da base de dados, (2) seleção dos atributos estilométricos que devem caracterizar o autor, (3) geração dos vetores de dissimilaridade, (4) produção de modelos (treinamento) e finalmente (5) o processo de decisão e voto analisando assim os resultados e desempenho dos métodos propostos bem como através de comparações com o método de compactação por PPM-C (*Variação do algoritmo de compressão PPM*). Analisando os tamanhos dos textos e os resultados obtidos pelo método de compactação PPM-C demonstra-se a potencialidade das características utilizadas na análise estilométrica de textos em português. As taxas de erros obtidas estão na faixa de 27,5% a 9% dependendo do modelo e protocolo de testes. Em paralelo ao trabalho científico de informática aplicada, é de fundamental importância a aceitação pelo Direito Brasileiro. O Capítulo 2.9 foi dedicado ao estudo das relações da análise estilométrica do autor em um processo de perícia no Direito Brasileiro.

Palavras-chave: Ciência Forense Computacional, Identificação da Autoria usando SVM, Estilometria, Estilística, Linguística Forense, Estilo Literário, Grafoscopia

Abstract

Based on the studies of the formation of the Brazilian Portuguese language, the richness is verified through stylistics attributes, and so, if scientifically properly applied, it can aid the correct solution forensic in cases involving texts of Portuguese language. This work presents a scientific methodology for the authorship identification of questioned documents, with short texts, based on the analysis of stylistics attributes of the Portuguese language, used by the author. The approaches are based on stylometry. Two methods are presented. Both are used *3-gram* of function words (conjunctions and adverbs) as stylistics features. The first method uses a global model with classification through SVM. In this model, only two classes are assumed through dichotomie: authorship and non-authorship. The second method uses model for author with classification SVM multi-class. In this approach only one model is generated, however each vector is identified through the class which pertence. To validate both models it was used a database of 30 authors and 15 documents each. These approaches consist of the following stages: (1) formation of the database, (2) selection of the stylistics features to characterize the author's profile, (3) generation dissimilarity vectors, (4) production of models (training) and finally (5) the process of decision and vote analyzing the performance of the methods proposed in relation to the results as well as through comparisons with the data compression method for PPM-C. Analyzing the sizes of the texts and the results for the compression method PPM-C the potentiality of the characteristics used in stylistics analysis of texts is demonstrated in Portuguese. Errors rate are about of 27,5 % to 9% depending on the protocol of tests. In parallel to the scientific work, its important the acceptance for the Brazilian Law. The Chapter 2.9 studies of the relationships of the author's stylistics analysis in an analysis process by Brazilian Law.

Keywords: Authorship Identification, Stylometry, Stylistic, Forensic Linguistics, Language-based Author Identification using SVM

Capítulo 1

Introdução

Notáveis são os avanços em pesquisas referentes ao estudo da individualidade da escrita manuscrita e de assinaturas no sentido de se identificar autores de documentos questionados[Jus02][Cha01a][BL03][Mor00][Bar05][Cre95][Gro06]. Várias características podem ser extraídas e analisadas computacionalmente, retirando a subjetividade do processo e assegurando a identificação da autoria de forma precisa quando submetidas a várias abordagens [SNSSHCHASL02] e em muitos casos com taxas de erros muito pequenas.

Contudo, existe uma outra abordagem, que também vem sendo estudada, em relação à identificação da autoria, que não está relacionada diretamente com a escrita manual do autor. Esta abordagem é principalmente direcionada aos casos em que a autoria de um documento eletrônico é questionada e não é possível utilizar-se ferramentas e processos que analisam a grafia. As características inseridas em um texto através do estilo literário muitas vezes são pistas únicas, pois independe da forma de como o documento está armazenado, se escrito à mão, em meio digital ou impresso. Em documentos eletrônicos (impressos ou em meio digital), diferentemente da análise de manuscritos, não é possível a extração de características da escrita manual do autor, levando a necessidade da análise de outros fatores inerentes ao que chamamos de estilo literário do autor (modo de expressão da escrita de um indivíduo em um texto). Muitos são os exemplos atuais de documentos eletrônicos de autoria questionada: e-mails, livros, notas de resgate, cartas de seqüestro e ameaçadoras, diários eletrônicos (blogs), cartas em meio magnético, colunas impressas em panfletos, jornais e revistas, etc. A tarefa desse estudo não é somente buscar recursos computacionais para a identificação da autoria em documentos questionados, através da análise do estilo literário do autor, mas também buscar o embasamento e aceitação jurídica através das leis vigentes e estudos de caso.

1.1 Desafio

Muito se tem estudado sobre características estruturais e análise do estilo literário em outros idiomas, principalmente a língua inglesa [McM02][Ols04][Cha01b][Cha97][Cha05] e somente nos últimos anos iniciaram-se pesquisas envolvendo identificação da autoria na língua portuguesa [CMRJB04]. O desafio deste trabalho está ligado a utilização de atributos estilométricos exclusivos da língua portuguesa, bem como, trabalhar com uma base de dados com amostras reduzidas (muitos estudos sugerem um mínimo de 1000 palavras por texto para se determinar um padrão, ver seção 3.2) para obter resultados significativos e aplicáveis ao Direito Brasileiro.

1.2 Motivação

O problema de identificação da autoria de documentos digitais por si só, poderia ser considerado como o principal fator motivacional, visto que uma solução aceitável deve passar por um rigoroso processo de avaliação, envolvendo não somente resultados comprobatórios, mas também compatíveis com os critérios aceitos pela comunidade jurídica internacional [Jus02]. Contudo, outros importantes fatores para realização deste trabalho foram:

- Demonstrar a relevância para estudo de características da língua portuguesa na identificação da autoria;
- Trabalhar com base de dados reduzidos, tanto em número de amostras quanto em conteúdo;
- Fomentar a área de pesquisa;
- Demonstrar conhecimento e embasamento científico e jurídico visando auxiliar nas limitações hoje enfrentadas pelo Direito Brasileiro no tratamento de documentos digitais e de autoria desconhecida.

1.3 Proposta

O objetivo geral deste trabalho é demonstrar a relevância e importância da utilização de características da língua portuguesa para identificação da autoria de documentos textuais. Através da extração de características de cada autor demonstra-se uma

metodologia utilizando uma abordagem quantitativa de simples atributos estilométricos da língua portuguesa.

Este trabalho assume a proposta de apresentar uma abordagem para identificação da autoria em documentos digitais, visando:

- Criar uma base de dados com poucas amostras (30 autores, 15 documentos cada) e armazená-la de modo a permitir uma análise do estilo literário do autor;
- Apresentar os conceitos da língua portuguesa e os possíveis atributos estilométricos a serem estudados;
- Selecionar e testar atributos estilométricos discriminantes da língua portuguesa aplicáveis ao foco desta pesquisa;
- Criar um processo automatizado para extração das características (atributos estilométricos) escolhidas;
- Comparar com outra metodologia, com outras características porém com a mesma base de dados;
- Apresentar um método que possa ser apreciado e eventualmente aceito pelo Direito Brasileiro;
- Contribuir para o trabalho atualmente realizado por peritos e lingüistas.
- Propor soluções computacionais que auxiliem na produção de laudos juridicamente aceitos, retirando desta maneira a subjetividade muitas vezes aplicada pelos peritos durante o processo de análise do documento.

Esta pesquisa não tem como objetivo avanços científicos no campo de aprendizagem de máquina, contudo se faz uso de algoritmos demonstrando como técnicas de atribuição de autoria podem ser automatizadas através de modelos preditivos de autoria.

1.4 Contribuições

A principal contribuição deste trabalho é desenvolver uma metodologia de processos cientificamente embasados, tornando-a uma ferramenta de auxílio para peritos, advogados e juízes em situações onde existam documentos de autoria questionada, apresentando resultados eficazes e confiáveis o suficiente para que o documento possa ser utilizado como prova associativa ou dissociativa nos autos processuais.

Como contribuições indiretas pode-se citar:

- Inserir o Brasil e sua língua cada vez mais no contexto mundial de línguas analisadas com fins forenses através dos atributos particulares da língua;
- Fomentar a área de estudo no Brasil, tanto jurídica quanto áreas relacionadas a computação forense;
- Integração com a área jurídica buscando a fundamentação e aceitação do processo computacional;
- Auxiliar lingüistas a diminuir a subjetividade aplicada no processo de análise.

1.5 Organização

Este trabalho está organizado em cinco capítulos. O primeiro capítulo refere-se a introdução. O capítulo 2 detalha o processo de utilização da língua portuguesa desde sua origem, passando pelas variabilidades regionais e sociais, finalizando destacando conceitos importantes sobre estilo, estilística forense e aplicação jurídica. Ainda ao final deste capítulo são apresentados conceitos referente às análises, métodos e abordagens necessárias para realização deste trabalho. O capítulo 3 apresenta o estado da arte relacionado com identificação da autoria com estilometria. Nos dois últimos capítulos (4 e 5) são apresentados as metodologia propostas e os experimentos realizados para comprovação das mesmas com suas respectivas análises.

Este documento é finalizado com apêndices, seção 5.2, nos quais são encontradas tabelas com exemplos de atributos estilométricos de formatação e específicos da língua portuguesa. Estão descritos também, dois casos de uso, no Brasil, nos quais a análise do estilo literário pode ser aplicada.

Capítulo 2

Fundamentação Teórica

Este capítulo contém a fundamentação teórica, necessária para o desenvolvimento da pesquisa. A clarificação da teoria envolvendo as áreas de lingüística e direito são pilares importantes para o bom entendimento deste trabalho.

2.1 Comunicação

A humanidade demorou milhares de anos para que a linguagem escrita pudesse ser utilizada como meio de comunicação. Até então as formas de comunicação se limitavam a sons e sinais. A comunicação escrita tem seu papel essencial como disseminadora do conhecimento através dos séculos, ampliando relacionamentos e intercambiando experiências, documentos e processos. A comunicação escrita é matéria prima para este estudo como objeto de análise.

2.2 Princípios Básicos da Escrita

As letras manuscritas foram criadas pelos egípcios em 5000 a.C., não como um alfabeto organizado mas sim uma escrita com hieróglifos, pictogramas e ideogramas, representadas sem padrão e organização, Figura 2.1. A origem do primeiro alfabeto com símbolos abstratos distintos que representavam sons de consoantes e vogais (fonogramas) é atribuída aos Fenícios. Em seu quinto livro Histórias, o grego Herodotus[1], relata a utilização de um verdadeiro alfabeto pelos Fenícios [Fab06].

A descoberta do alfabeto fenício foi oficializada em 1929. O arqueólogo francês Claude Schaeffer encontrou várias tábuas escritas com língua cuneiforme desconhecida

[1]Herodotus viveu durante o quinto século antes de Cristo e é considerado um dos primeiros historiadores do mundo.

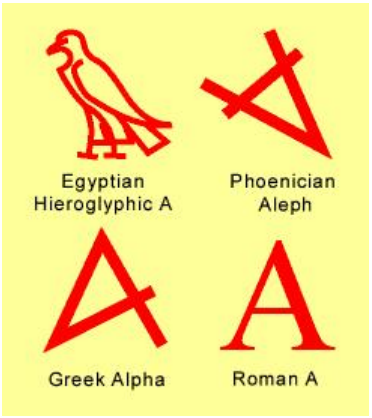

Figura 2.1: Evolução da Letra [Fab06]

enquanto escavava sobre as colinas de Ras Shamra, na antiga cidade de Ugarit[DOL81].

As tábuas de Ugarit foram intensamente analisadas e estudadas por um grande número de historiadores clássicos de historia antiga. Através da pesquisa árdua dos historiadores, em 1948, todos os vinte e oito caracteres do alfabeto cuneiforme Fenício foram corretamente identificados. Das 28 letras, 26 eram consoantes. Essas Tábuas de Ugarit continham o primeiro alfabeto da história humana.

Os Fenícios eram grandes comerciantes marítimos e por isso necessitavam de uma escrita ágil e eficiente para manter seus inventários, contabilidade, negociações, etc., diferente dos conhecidos hieróglifos egípcios que tornavam a escrita lenta e difícil pois era formado por ideogramas e pictogramas.

Desde então o alfabeto com suas variações é a mais importante, criativa e essencial ferramenta para comunicação de seres humanos entre gerações.

Figura 2.2: Evolução do Alfabeto Fenício [Fra06]

2.3 Linguagem

Linguagem é um sistema de comunicação através de um código[Inf01]. Esses códigos são utilizados tanto por seres humanos quanto pela natureza. Exemplos: mensagens sonoras entre golfinhos, diálogo entre surdo-mudos, linguagem corporal através dos

cinco sentidos,etc.

Entre os seres humanos o mais utilizado desses códigos é a língua. Língua é um conjunto de sons e ruídos (representados por um alfabeto ou pronunciados oralmente), utilizado para transmissão de mensagens a um receptor ou destinatário[Inf01]. Contudo, não basta receber a mensagem, é necessário entender o significado da mensagem que está sendo transmitida, por isso a língua representa um conjunto de signos e regras combinatórias entre estes signos, que são padronizados e utilizados por um grupo de indivíduos.

Falar e escrever possuem profundas diferenças na elaboração da mensagem a ser passada. Essa variação implica a criação de dois códigos distintos: língua falada e a língua escrita.

Tanto na língua falada como na língua escrita, cada indivíduo possui preferências a determinadas palavras (signos) ou construções (regras combinatórias). Essas preferências compõem o estilo literário do autor independentemente se escolhidas por hábito (inconsciente) ou por opção consciente.

2.4 Língua Portuguesa

O referente estudo está baseado na contrato social denominado Língua Portuguesa. A Língua Portuguesa possui uma variação particular para o Brasil: Língua Portuguesa Brasileira, a qual é base deste estudo.

2.4.1 Origem

A Língua Portuguesa desenvolveu-se no ocidente da Península Ibérica, que hoje é Portugal. Foi formada das transformações verificadas no latim e trazida por soldados romanos desde o século III a.C.. O Português começou a diferenciar-se das outras línguas românicas a partir do século V devido a queda do império romano e invasões bárbaras, que puseram fim as escolas romanas. Seus primeiros escritos datam do início do século IX e já no século XV, se tinha formado uma língua com uma literatura rica. A partir do século XV, com as navegações e descobrimentos portugueses, iniciou-se um longo processo de expansão lingüística em muitas regiões da Ásia, África e América, através de seus colonos e emigrantes. O alastramento da língua foi ajudado principalmente por casamentos mistos e pelos missionários católicos, recebendo o nome de cristã em muitos locais[Bue67].

Com mais de 200 milhões de falantes nativos, a língua de Camões (assim também conhecida devido a Luís de Camões, autor de Os Lusíadas) é a sexta língua materna mais popular no mundo e a segunda língua latina, só ultrapassada pelo Espanhol. É a terceira

língua mais falada no mundo ocidental [dPdLP06].

Atualmente, a língua portuguesa é oficial em alguns países tais como: Portugal, Brasil, Angola, Moçambique, Guiné-Bissau, Cabo Verde, São Tomé e Príncipe, Timor Leste. Em algumas regiões da Índia como Goa, Damão, Diu e Dadra e Nagar Haveli, o português é falado por uma parcela da população e em outras deu origem a alguns dialetos locais[dPdLP06].

A CPLP (*Comunidade dos Países de Língua Portuguesa*) é uma organização internacional que consiste nos oito países independentes, que têm o português como língua oficial. O português é também uma língua oficial da União Européia, Mercosul e uma das línguas de trabalho e oficiais da União Africana. A língua portuguesa tem ganho popularidade como língua de estudo na África, América do Sul e Ásia[dPdLP06].

2.4.2 Padrões

O português tem dois padrões reconhecidos internacionalmente:

- Português Europeu e Africano;
- Português do Brasil.

As diferenças entre o português da Europa e do Brasil estão no vocabulário, pronúncia e sintaxe. Na língua escrita nos textos formais as diferenças diminuem, já nos textos vernáculos as diferenças podem até impedir uma boa compreensão do texto.

Como principais diferenças, o português brasileiro tem na sua forma escrita a ausência da maioria dos primeiros "c", quando "cc", "cç" ou "ct"; e "p", quando "pc", "pç" ou "pt", porque não são pronunciados na forma culta da língua, Tabela 2.1. As diferentes normas de escrita do português foram padronizadas pelo Acordo Ortográfico da Língua Portuguesa de 1990, aprovados pelos países membros da CPLP.

Europa e África	**Brasil**
acção	ação
contacto	contato
direcção	direção
eléctrico	elétrico
óptimo	ótimo

Tabela 2.1: Grafia - Português Brasileiro vs. Português Europeu

2.5 Língua Portuguesa Brasileira

A colonização portuguesa no Brasil, iniciou-se lentamente a partir de 1532, com a criação das capitanias hereditárias. Nesse período existiam muitas comunidades indígenas, principalmente Tupi e Guarani, habitando a região do litoral brasileiro entre Rio de Janeiro e Bahia. Para melhorar a comunicação com os nativos, os portugueses aprenderam os dialetos e idiomas indígenas. A partir do tupinambá criou-se uma língua geral falada por índios e não-índios, documentada pelos jesuítas para catequização dos índios. Essa língua geral, derivada do tupinambá, foi a primeira influência do idioma dos portugueses no Brasil.

Outro fator que influenciou a língua portuguesa brasileira, foi a língua dos negros africanos trazidos como escravos para o país. Os escravos acabaram aprendendo o português para se comunicar com seus senhores, juntamente com a língua geral utilizada pelos colonos.

Com aproximadamente dois séculos da utilização minoritária do português, sua predominância começa a se dar a partir do século XVIII, com uma maior imigração de portugueses para exploração de minas de ouro e diamante, tornando assim o biligüismo cada vez menor. Em 1758, através do Marques de Pombal, a língua portuguesa se torna o idioma oficial no Brasil e proíbe o uso da língua geral. Contudo, a língua geral falada pelos colonos, juntamente com a influência africana, já eram maioria e com o vocabulário repleto de palavras de origem indígena e africana.

Além do vocabulário, as mudanças fonéticas influenciadas pelos africanos e indígenas foram muito fortes, originando uma fonética bem diferenciada do português europeu.

Após a independência do Brasil, houve a chegada de mais imigrantes europeus, como italianos e alemães. O contato da língua portuguesa com outras línguas incrementou o vocabulário, com palavras oriundas dos países europeus, originando as diversas variedades regionais existentes hoje no Brasil.

2.5.1 Variedade Lingüística

As diferentes derivações da língua portuguesa no Brasil, tanto na língua falada quanto na escrita, são derivadas dos seguintes fatores:

- Geográficos

  Com os diferentes pontos de concentração dos imigrantes no Brasil a língua portuguesa sofreu adequações dando origem ao que chamamos de variações regionais que constituem os dialetos, sotaques. Maiores acentos dos Brasil são:

  - Caipira: Interior do estado de São Paulo, norte do Paraná e sul de Minas Gerais;
  - Cearense: Ceará;
  - Baiano: região da Bahia;
  - Fluminense: Estados do Rio de Janeiro e Espírito Santo, com exceção da cidade do Rio de Janeiro que tem um dialeto bastante próprio;
  - Gaúcho: Rio Grande do Sul;
  - Mineiro: Minas Gerais;
  - Nordestino: Estados do nordeste brasileiro. O interior nordestino e Recife tem um dialeto particular;
  - Nortista: Estados da bacia do Amazonas;
  - Paulistano: cidade de São Paulo;
  - Sertanejo: Estados de Goiás e Mato Grosso;
  - Sulista: Estados do Paraná e Santa Catarina. Curitiba tem um falar próprio e ainda existe um pequeno dialeto no litoral catarinense próximo do açoriano.

- Sociais

  Existem muitas diferenças entre o português utilizado por indivíduos que tiveram acesso a escola e por indivíduos privados de instrução. Com isso, é estabelecida uma segregação de língua. A norma culta é instrumento de ascensão profissional e social. Neste sentido, a língua torna-se uma ferramenta de dominação e discriminação social.

- Profissionais

  Para exercícios de certas atividades profissionais muitas vezes é necessária a utilização de uma linguagem técnica. Esta linguagem é repleta de conceitos técnicos específicos da área e importante para a comunicação entre os especialistas.

- Situacionais

  A capacidade do ser humano em adaptar-se ao meio em que está, faz com que a língua utilizada seja aplicada em formas diferentes, de modo a adequar o nível vocabular e lingüístico ao ambiente ou situação que se encontra. O fator situacional está relacionado diretamente com os resultados da análise quantitativa de atributos estilométricos. Uma análise grafoscópica entre textos informais e formais manuscritos, pode ser feita com uma precisão similar a uma análise grafoscópica entre

documentos manuscritos de mesma linguagem. Contudo, na análise quantitativa de atributos estilométricos, baseado no estilo literário, a linguagem utilizada pelo autor tem alta relevância nos resultados. A escrita tem claramente dois modelos a serem comentados:

– Modelo formal

No modelo formal, exemplificado na Figura 2.3 o autor demanda um esforço maior de criação, pois preocupa-se com: aplicação das corretas regras da língua, tornar a mensagem precisa e independente do lugar e do tempo em que é produzida.

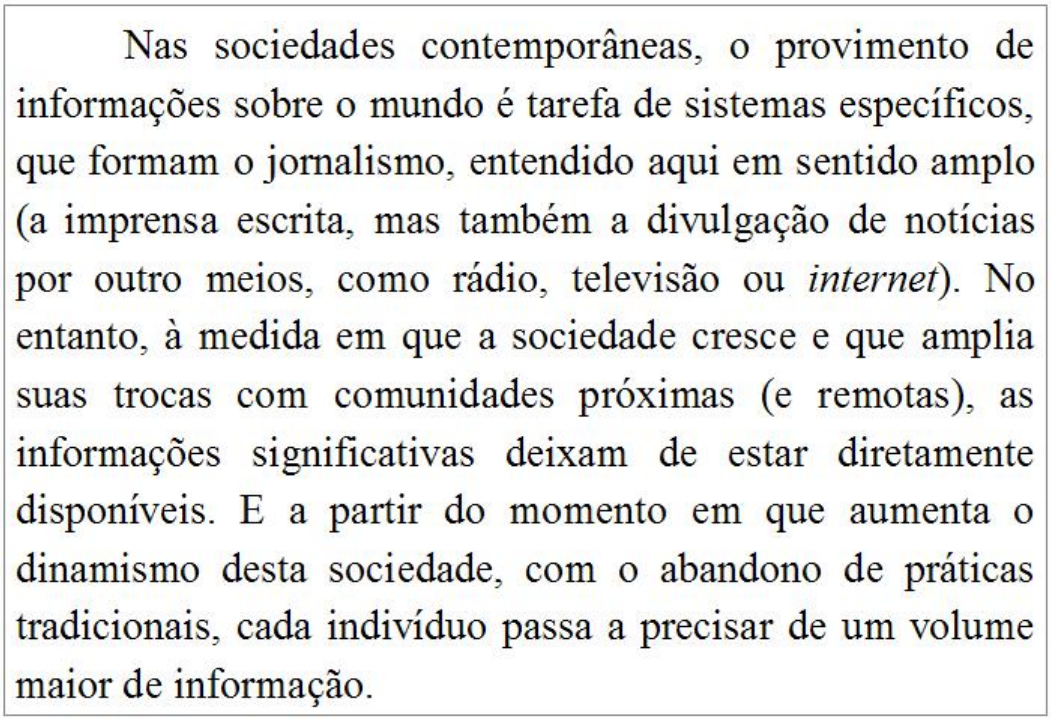
Nas sociedades contemporâneas, o provimento de informações sobre o mundo é tarefa de sistemas específicos, que formam o jornalismo, entendido aqui em sentido amplo (a imprensa escrita, mas também a divulgação de notícias por outro meios, como rádio, televisão ou *internet*). No entanto, à medida em que a sociedade cresce e que amplia suas trocas com comunidades próximas (e remotas), as informações significativas deixam de estar diretamente disponíveis. E a partir do momento em que aumenta o dinamismo desta sociedade, com o abandono de práticas tradicionais, cada indivíduo passa a precisar de um volume maior de informação.

Figura 2.3: Exemplo de Modelo Formal de Escrita [Mig04]

– Modelo informal

Na escrita informal o indivíduo, na maioria dos casos, está direcionando seu texto a um leitor específico ou a um grupo de leitores com características afins. Em cartas pessoais e com o aumento da utilização de e-mails, blogs, programas de trocas de mensagens instantâneas (celulares ou internet), uma "nova" linguagem informal, criativa e personaliza é utilizada, ver Figura 2.4, fazendo com que documentos formais e informais, produzidos pelo mesmo autor, apresentem grandes diferenças inviabilizando assim o estudo comparativo de alguns atributos estilométricos.

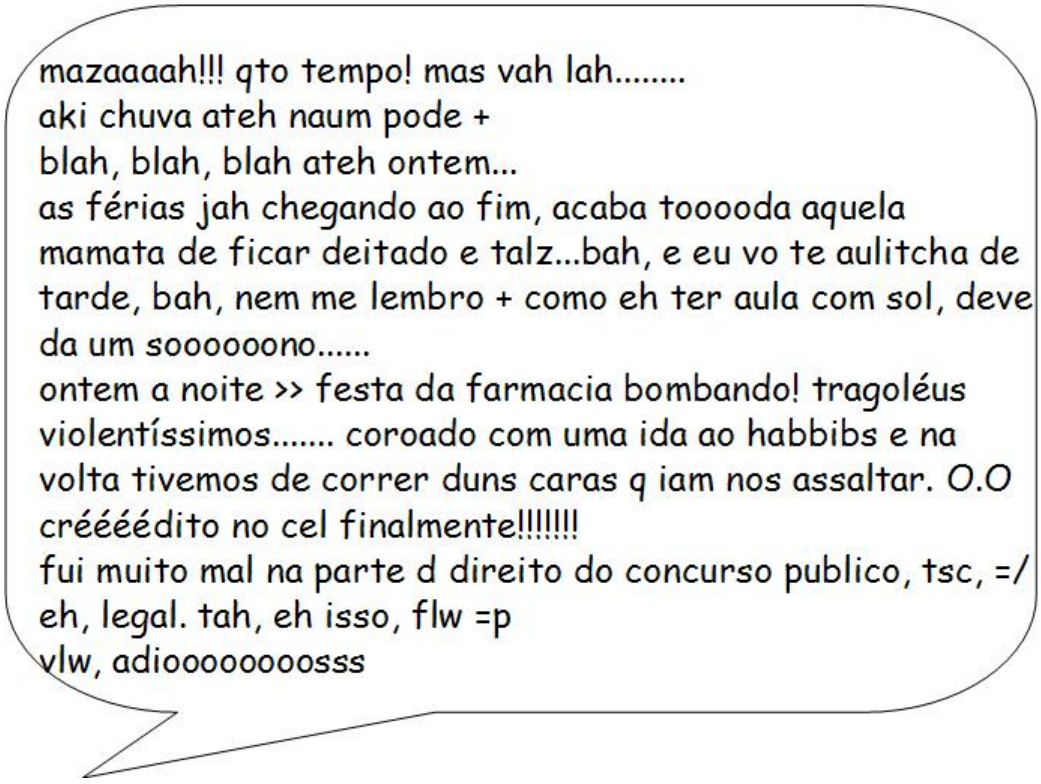

Figura 2.4: Exemplo de Modelo Informal - Webblogger 27/09/2004 00:14

- Literária

  A literatura portuguesa é muita rica em seus escritos. Quando um escritor utiliza a língua, não só para as necessidade práticas do cotidiano, e passa a preocupar-se em incorporar preocupações estéticas, combinando e criando elementos lingüísticos, surge a língua literária. A leitura sensível a língua literária é um processo gradual e contínuo [Inf01].

A língua falada e escrita possuem uma ligação estreita. Fatores já vistos como regionalismos, sociais, etc. refletem no estilo de comunicação do indivíduo, tanto falado como escrito.

2.5.2 Estrutura

- Gramática;

  Gramática é um conjunto de regras e exemplos para um correto e apropriado uso da língua [Inf01]. Pode ser tradicionalmente dividida em:

  – Fonologia;

Estudo dos sons da língua ou fonemas e sua organização silábica. Inclui acentuação, ortografia e pronúncia dos vocábulos.

– Morfologia.

Estudo das formas da língua, da estrutura e formação de palavras[Fer04]. Inclui:

* Formação das Palavras;
* Estudo dos Verbos;
* Estudo dos Substantivos;
* Estudo dos Artigos;
* Estudo dos Adjetivos;
* Estudo dos Advérbios;
* Estudo dos Pronomes;
* Estudo dos Numerais;
* Estudo das Preposições;
* Estudo das Conjunções;
* Estudo das Interjeições.

– Sintaxe.

Estudo das relações estabelecidas entre palavras nas orações ou entre as orações nos períodos.

2.6 Linguagem Escrita

Escrita é um sistema gráfico de símbolos lingüísticos para representação e comunicação da informação. É um processo de efetuar marcar visíveis (símbolos) em uma superfície, seja ela um papel, papiro ou em uma tela de computador. Escrita representa diretamente informações sobre coisas e eventos.

Escrita também é formular um texto coerente, como um discurso, é criar um poema, uma carta comercial, um relatório, etc. É também o processo de produzir um texto da forma como o escritor quer e deseja passar a informação.

Por muito tempo a escrita foi considerada a representação da linguagem falada. Nos últimos anos, diversos estudos [Cha86][GC98][Ols97][WN88] mostram que a escrita e a linguagem falada são diferentes em diversos aspectos. A escrita está relacionada com a linguagem falada, mas não é derivada do ato de falar[McM02].

A linguagem falada é efêmera, social, rápida e requer um ouvinte. A escrita é duradoura, visual, lenta, solitária e permite ser alterada durante e depois. A língua falada possui muito mais subjetividade, atributos estilométricos não lingüísticos para ligar idéias, gírias, sotaque. Língua falada tem suas idéias contextualizadas e acontece no momento. A escrita emprega mais formalidade, letras maiúsculas, sublinhado e possui menor variabilidade. O contexto e os detalhes devem estar nas palavras e nas pontuações, juntamente com uma estrutura textual coesa e vocabulário ajustado.

O processo de aprendizado da escrita requer muito mais do que apenas conhecimento de palavras. Aprender a escrever requer desenvolvimento motor, coordenação (olhos e mão), segurar a ferramenta utilizada para escrita, traçados, orientação, simetria, espaçamento, continuidade, movimentação esquerda para direita e cima para baixo, etc.

O aprendizado da escrita requer uma instrução formal e habilidade de ler o alfabeto relacionando as letras aos sons correspondentes, copiando as letras, repetindo palavras difíceis de serem pronunciadas e então escrevendo pequenas histórias e fatos do cotidiano.

As dificuldades e habilidades naturais de cada indivíduo são determinantes na formação da linguagem, por isso alguns fatores são importantes para o aprendizado da escrita por uma criança[GC98]:

- Devem estar ativamente envolvida no processo (interessada);
- Conhecimento adquirido antes de iniciar a escola;
- Oportunidades e encorajamento para escrever;
- Escrever o que ela diz e não o que quem ensina diz;
- Compreender o alfabeto, letras representam sons e estas são usadas para formar palavras;
- Oportunidade de ter contato com outros que ela pode ver escrevendo.

Na aplicação ou não dos fatores acima citados, é iniciado a distinção da escrita entre indivíduos. Mesmo crianças ensinadas por um mesmo processo de aprendizado, com o passar dos anos, apresentarão diferenças em suas escritas. Isto se deve a alguns fatores:

- Habilidade no aprendizado da linguagem falada;

  Como o processo de aprendizado da linguagem é interrelacionado entre língua falada e escrita a habilidade em assimilar a língua falada resultará em diferenças no processo de aprendizado da língua escrita.

- Situação Contextual;
- Dificuldade da Língua;

  Línguas tonais apresentam dificuldade na associação rápida e correta das letras a sons, pois apresentam mesmas letras porém com pronúncias diferentes como por exemplo a língua chinesa.

- Intenção da Escrita;
- Habilidades Naturais;
- Entendimento e Interpretação da Linguagem.

2.7 Lingüística

Lingüística é uma ciência que estuda as regras e padrões de uma língua. A lingüística tem o seu papel teórico e prático. Teórico pois estuda características presentes em uma língua específica ou em um grupo similar de línguas. Prático pois tenta utilizar este conhecimento para melhorar a comunicação através da língua, seja ela aplicada a alguma determinada área, por exemplo, aplicações forenses, ou simplesmente no ensino básico da língua. Um foco importante do estudo lingüístico está em pesquisar o que o indivíduo internaliza, durante o curto período que leva para adquirir sua primeira língua[McM02].

Existem quatro canais utilizados pelos usuários de uma linguagem. Dois estão relacionados com habilidades de expressão: falar e escrever; e outros dois relacionados com a capacidade de receptividade e compreensão: ouvir e ler. A técnica de uma língua é descrita por níveis lingüísticos apresentados na Tabela 2.2:

2.7.1 Lingüística Aplicada

Lingüística aplicada é a utilização do conhecimento e de técnicas lingüísticas focadas nas necessidades humanas. Os primeiros relatos de lingüística aplicada eram somente relacionados com a língua ensinada e adquirida. Atualmente existem não menos que 25 comissões científicas de acordo com IALA (*Association Internatiole de Linguistique Appliquée*) aplicadas às mais diversas áreas:

- Ensino da língua para adultos;
- Linguagem infantil;

Tabela 2.2: Níveis Lingüístico na língua falada e escrita

Nível Lingüístico	**Língua Falada**	**Língua Escrita**
FORMA DA LÍNGUA		
Fonética	Sons (fonemas)	Letras (grafemas)
Fonologia	Padrões de Sons e Combinações	Letras e Dígrafos
Morfologia	Formação das Palavras (morfemas)	Parte das Palavras (raízes e anexos)
Léxico	Palavras	Vocabulário (Dicionário)
Sintaxe	Formação da Sentença	Sentenças Escritas (gramática)
Semântica	Significado Expresso	Significado das palavras e sentenças
FUNÇÃO DA LÍNGUA		
Discursiva	Conversas e Narrativas	Histórias
Pragmática	Jogando com palavras	Equivalente em escrita

- Comunicação e as profissões;
- Lingüística Distinta e análise de erros;
- Análise de Discursos;
- Tecnologia Educacional e Aprendizado da Linguagem;
- Metodologia de ensino de língua estrangeira e educação do professor;
- *Lingüística Forense*;
- Educação por Imersão;
- Interpretação e Tradução;
- Linguagem e ecologia;
- Educação da linguagem em um cenário multi-língue;
- Educação e Gênero;
- Sociolingüística;
- Linguagem e a mídia;
- Linguagem para propósitos especiais;
- Planejamento da linguagem;
- Autonomia da língua e aprendizado da língua;

- Lexicografia e Lexicologia;
- Alfabetização;
- Educação da língua mãe;
- Psicolinguística;
- Retórica e Estilística;
- Aquisição da Segunda Linguagem;
- Linguagem de Sinais.

2.7.2 Variação Lingüística

A análise da variação lingüística aplicada a lingüística forense é de fundamental importância. As variações deixam evidências no traçado e estilo, com as quais é possível associar indivíduos ou uma classe de indivíduos que possuam as mesmas características, existentes em um documento questionado.

Existem dois tipos de variação: A variação intrapessoal, do indivíduo com ele mesmo e interpessoal, do indivíduo em relação a uma comunidade de outros indivíduos.

A observação e análise da variação lingüística utilizada por qualquer indivíduo pode ser mensurada, visto que é influenciada sistematicamente por fatores internos (lingüísticos) e externos (não lingüísticos). A mensuração da variação depende do entendimento preciso do que cada variação significa no contexto analisado. Este entendimento está em saber como a utilização da língua se altera entre indivíduos.

2.7.3 Lingüística Forense

Lingüística Forense é o estudo científico da linguagem aplicada aos propósitos e contextos forenses. Lingüística forense está relacionada com a evolução das características do texto, incluindo gramática, sintaxe, pronúncia, vocabulário e frasologia. Esta evolução é realizada através da comparação de material textual de autoria conhecida e não conhecida, na tentativa de revelar idiossincrasias peculiares a autoria para determinar se os autores poderiam ser idênticos [BNNH90].

A lingüística forense é dividida nas seguintes áreas:

- Fonética (Audível e Acústica)

  Interpretação e significado expressado.

- Pragmática e Discurso

  Interpretação e significado deduzido.

- *Estilística e Autoria Questionada de Documentos;*

  Esta é a área de aplicação deste trabalho dentro da lingüística forense;

- Linguagem do Direito e Tribunal;

- Interpretação e Tradução de Textos.

2.8 Estilística

Estilística é a disciplina que estuda a expressividade de uma língua[Fer04]. A expressividade de uma língua é demonstrada através dos estilos utilizados. A escolha e análise destes estilos é foco de estudo desse trabalho e abordado neste capítulo.

2.8.1 Estilística Lingüística

Estilística lingüística é o estudo científico dos atributos estilométricos de um indivíduo para caracterização de um idioleto[2], bem como atributos estilométricos de classes de acordo com o idioma e grupos de dialeto.[McM02]

2.8.2 Estilo

Estilo é a reflexão do indivíduo, variação do grupo em uma linguagem escrita. Maneira de escrever caracterizada pelo emprego de expressões e características próprias de um indivíduo, classe, profissão, ou grupo [Fer04].

A construção do estilo de um indivíduo na linguagem escrita é resultando de recorrentes processos de escolhas de um indivíduo. Essas escolhas podem ser dentro de norma, corretas, por exemplo, cinqüenta/cincoenta. Podem ferir alguma norma referente à língua escrita e, portanto, considerada incorreta, por exemplo a frase, *Me Liga.* ou simplesmente tratar-se de idiossincrasias, isto é, forma específica do autor.

Em relação as normas, podem haver dois tipos: prescritiva e descritiva. Norma prescritiva é relacionada a aceitação social e a descritiva reflete o que é aceitável e correto gramaticalmente.

[2]Idioleto é a linguagem utilizada por um único indivíduo [Fer04]

2.8.3 Estilística Forense

Estilística forense é a aplicação da ciência da estilística lingüística para contextos e propósitos forenses. A aplicação principal da estilística forense é a área de autoria questionada.

A identificação da autoria forense é realizada através da análise do estilo da linguagem escrita. A estilística explora duas premissas de variabilidade da linguagem:

- Dois escritores de uma língua não escrevem exatamente do mesmo modo;
- Um mesmo escritor não escreve exatamente do mesmo modo todo o tempo.

De um modo amplo a análise do estilo pode ser classificada em qualitativa e quantitativa. O estudo qualitativo da escrita consiste na análise das formas usadas pelo autor e em como e porque elas foram utilizadas [Joh00]. Apesar de algumas linhas de pesquisa questionarem a cientificidade da análise qualitativa, existem algumas razões forenses importantes pelas quais deve ser considerada e contextualizada:

- Descrição qualitativa é o passo inicial: A “medida” depende das descrições e da categorização dos elementos lingüísticos analisados;
- Evidências qualitativas são mais demonstráveis que evidências quantitativas, ou seja, quando se refere ao impacto a um corpo de jurados, características de como ou porque foi inserido ou usado uma determinado estilo (mesmo que seja somente um e uma ocorrência) tem um maior ”senso de convencimento”do que relatórios estatíscos com métodos não comprovados cientificamente;
- Resultados qualitativos demonstram um senso de probabilidade. Exemplo: Um determinado autor nunca utilizou uma determinada palavra em toda suas obras conhecidas.
- Casos onde não é possível contar ou medir a evidência.

Na análise quantitativa, usa-se a medida da variação da língua escrita como poderosa ferramenta para discriminação entre autores e portanto, importante para o sucesso da análise e interpretação do estilo. O foco da análise quantitativa da escrita está em quanto e como formas comuns são utilizadas por um autor[Joh00].

Mesmo com muitos avanços tecnológicos, ainda são poucas as ferramentas confiáveis de análise para auxiliar o perito na identificação e mensuração dos atributos estilométricos

desejados. Esta carência faz com que a tarefa do perito seja lenta, trabalhosa e nem sempre exata. Outra limitação a ser considerada é que, em muitos casos, o material a ser analisado não fornece dados suficientes ou significativos para que possam ser utilizados.

Apesar das limitações e dificuldades, o processo de análise quantitativa tem sua importância forense por encontrar requisitos internos (metodológicos) e externos (judiciais) para demonstração, ou seja, mesmo não sendo possível visualizar as diferenças entre autores, através de técnicas e modelos comprovados pela comunidade científica (requisitos internos) suportado por leis (requisitos externos) demonstram a atribuição da autoria.

Este trabalho está focado na análise quantitativa dos atributos estilométricos da língua portuguesa contextualizados à base de dados gerada.

Estilística forense faz uso da análise estilística para obter conclusões e opiniões relacionados a autoria de um documento questionado, dentro de um contexto jurídico. O objeto de estudo é linguagem de um indivíduo (idioleto), cujas características são extraídas resultando em uma identificação pessoal.

Na identificação da autoria em documentos questionados basicamente existem dois modelos de análise:

- Modelo de Identificação:

  Dado o documento questionado tentar identificar o escritor dado um conjunto de n possíveis escritores. Este modelo tem a vantagem de ser capaz de identificar o escritor diretamente, entretanto depende do conhecimento de todos os escritores possíveis antecipadamente.

- Modelo de Verificação:

  Dado dois documentos de quaisquer escritores, busca-se determinar se os documentos foram escritos pela mesma pessoa.

O resultado da análise pode ser:

- Determinar a semelhança da escrita questionada;
- Identificar um ou mais autores suspeitos;
- Inconclusiva em relação aos dados fornecidos para identificação ou eliminação.

2.8.4 Estilometria

Estilometria tenta definir as características do estilo de um autor e determinar modelos (estatísticos, probabilísticos) para medir as características discrimatórias entre

dois ou mais documentos. Com base nas premissas de variabilidade da linguagem vistas na seção 2.8.3 e nas premissas demonstradas por Holmes [Hol98] é que esta pesquisa é norteada.

2.9 Aplicação Jurídica

Este capítulo tem por finalidade analisar a legislação existente acerca do tema, e a utilização pelo Direito como ferramenta para auxiliar nos processos e na elucidação de eventuais questionamentos de autoria de documentos eletrônicos ou impressos.

2.9.1 Aplicação

O questionamento da autoria de documentos impressos e digitais, muitas vezes, está diretamente relacionado com o mérito de processos judiciais.

Como exemplo de documentos eletrônicos ou impressos de autoria questionada pode-se citar: e-mails, livros, notas de resgate, cartas de seqüestro, cartas ameaçadoras, testamentos, diários eletrônicos (blogs), mensagens em meio magnético, colunas impressas em panfletos, jornais e revistas e demais documentos notariais cuja autoria é desconhecida e a análise da grafia não se aplica para identificação do autor.

No entanto, a utilização desses documentos como prova fica ainda sujeita a algumas considerações, pelo fato de ser praticamente desconhecida e pouco utilizada no Brasil.

2.9.2 Prova

A finalidade da prova é o convencimento do juiz, que é o seu destinatário final. No processo, a prova não tem um fim em si mesma, fim moral ou filosófico, mas sim formar a convicção do magistrado/julgador. Na prova não se busca uma certeza absoluta, a qual, é na maioria dos casos impossível, mas sim a certeza relativa suficiente para convicção do juiz [GF03].

O conceito de prova, no sentido jurídico, pode ser a demonstração que se faz pelos meios legais, da existências ou veracidade de um ato material ou de um ato jurídico, em virtude da qual se conclui por sua existência ou se firma a certeza a respeito da existência do fato ou do ato demonstrado [SSFC04].

No processo judicial, a verdade deve ser sempre buscada pelo juiz, mas o legislador não a coloca como um fim absoluto. Ou seja, o que é suficiente, muitas vezes, para a validade e a eficácia da sentença é a verossimilhança dos fatos. O que se pretende

alcançar com a análise da autoria de um documento, relacionando-a a um determinado autor é convencer o julgador, no sentido de que possa ele fazer a correta aplicação da lei no caso concreto. De acordo com o artigo 332 do Código de Processo Civil: "todos os meios legais, bem como os moralmente legítimos, ainda que não especificados neste Código, são hábeis para provar a verdade dos fatos controversos, em que se funda a ação ou a defesa". Os meios hábeis para provar a verdade dos fatos especificados no ordenamento brasileiro são:

- Depoimento pessoal;
- Confissão;
- Exibição de documento ou coisa;
- Prova documental;
- Prova testemunhal;
- Prova pericial;
- Inspeção judicial.

2.9.3 Procedimento Probatório

O procedimento probatório é o espaço reservado à coleta das provas e compreende três estágios:

- Proposição;

  Ocorre quando a autoria de um documento é questionado requerendo sua prova.

- Deferimento;

  Momento em que o juiz acolhe a necessidade da prova.

- Produção.

  Momento da efetivação para que a prova seja incorporada aos autos do processo.

Quando em um processo judicial a autoria de um documento é questionada, a parte que questiona a autoria deve requerer ao juiz a produção de prova de autoria ou não autoria do documento. Este tipo de prova requer conhecimento específico que o juiz não possui e só pode ser realizada por especialistas na área da identificação de autoria por meio

de prova pericial. Nesse contexto, os lingüistas são os profissionais com o conhecimento necessário para efetuar a análise.

Em documentos eletrônicos existem inúmeros questionamentos na doutrina quanto a possibilidade de sua utilização, devido a sua fragilidade, possibilidade de alteração e na confiabilidade do processo de determinação da autoria.

A utilização do estilo literário pretende melhorar a confiabilidade do processo, nos casos em que é possível a análise grafoscópica do documento (documentos manuscritos) e viabilizar um mecanismo científico, quando única e exclusivamente for necessária a análise contextual (documentos eletrônicos), na busca da identificação da autoria.

Caberá ao juiz o exame no caso concreto, com o objetivo de garantir os direitos das pessoas envolvidas, não impedir a aceitação da modernização dos meios de produção de provas, desde que estas provas sejam mais significativas e úteis do que as tradicionais para a verificação do fato.

A aceitação do Poder Judiciário de que peritos utilizem novas ferramentas baseadas em processos cientificamente automatizados, certamente trará mais clareza e confiabilidade aos processos envolvendo autoria questionada.

2.9.4 Prova Pericial

Perícia, do latim *peritia*, significa habilidade, saber. Consiste no meio pelo qual pessoas entendidas verificam fatos interessantes à causa, transmitindo ao juiz o respectivo parecer [Fer04]. A perícia é feita pelo perito oficial nomeado pelo juiz. O perito deve possuir o conhecimento técnico especializado. Junto ao perito pode figurar o assistente técnico, também conhecedor da área mas que atua como auxiliar de cada parte, e que tem por obrigação, concordar, criticar ou complementar o laudo do perito oficial, através de seu parecer [Sil91].

Na produção da perícia, o juiz nomeia o perito e fixa de imediato um prazo para a entrega do laudo pericial. Às partes, cabe indicar seus assistentes técnicos e apresentar os quesitos ao perito. Estes quesitos são os questionamentos que o perito deverá responder com análise no laudo pericial.

Os tipos de investigação pericial do Código de Processo Civil no seu artigo 420 são: Exame, Vistoria e Avaliação. Para a identificação da autoria o tipo de perícia utilizada é o exame, que consiste na inspeção feita por perito sobre pessoas e coisas móveis, livros comerciais, documentos e papéis de um modo geral, para a verificação de circunstâncias e fatos.

Na análise do estilo literário se faz necessária uma certa quantidade de documentos

do(s) autor(es) questionado(s), para fazer a análise do estilo. O perito deverá, nestes casos, solicitar ao juiz documentos que estejam em poder das partes ou em repartições públicas para que ele os requisite. Tal situação fica clara no artigo 429 do Código de Processo Civil: "O perito e os assistentes técnicos no desempenho de sua função, podem utilizar-se de todos os meios necessários, ouvindo testemunhas, obtendo informações, solicitando documentos que estejam em poder de parte ou em repartições públicas, bem como instruir o laudo com plantas, desenhos, fotografias e outras quaisquer peças". Em suma, os peritos e os assistentes possuem livre acesso a recorrer a todas as informações que visem o esclarecimento dos quesitos apresentados em seu laudo.

Encerrada as diligências e a análise, o perito deverá apresentar seu laudo em cartório para que sejam intimadas as partes a se pronunciarem sobre o laudo, no cumprimento do contraditório e para que os assistentes técnicos apresentem pareceres a respeito, aceitando, discutindo e criticando o laudo pericial. O laudo pericial consiste na designação da peça escrita pelo perito, na qual faz relatório da perícia realizada. Compõe-se de duas partes: Expositiva, na qual descreve os objetos da perícia e a metodologia adotada e a parte conclusiva, onde responde os quesitos formulados pelas partes.

Embora prova técnica, científica, a perícia é uma prova como qualquer outra no que diz respeito à possibilidade de conter erros, imperfeições e até vícios que a tornem imprestável. Por isso ela está sujeita a esclarecimentos, que serão dados em audiência com a intimação do perito pelo juiz [Sil91].

2.9.5 Decisão do Juiz ou Tribunal

Posteriormente a fase da produção de provas, na continuação do processo, ocorrerá a sua apreciação e avaliação pelo órgão julgador que deverá proferir a decisão. Nas legislações processuais contemporâneas, incluindo a brasileira, o sistema de apreciação e valoração da prova admitido é o sistema da persuasão racional, que dá liberdade ao juiz para a valoração das provas, ou seja, não há valor ou peso determinado para cada prova, mas ao mesmo tempo vincula a decisão às provas apresentadas, decidindo de acordo com o que lhe foi apresentado e obrigando o juiz a fundamentar sua decisão de modo que se possa conhecer as provas em que se embasa sua conclusão e as razões de seu convencimento. O sistema da persuasão racional está contido no artigo 131 do Código de Processo Civil: "O juiz apreciará livremente a prova atendendo aos fatos e circunstâncias constantes nos autos, ainda que não alegados pelas partes; mas deverá indicar, na sentença, os motivos que lhe formaram o convencimento". A justificação presta-se a dúplice papel, o de convencer as partes e o público da justiça da decisão e de possibilitar o controle do ato

decisório.

De modo diferente não dispõe o Código de Processo Penal, ao determinar, no art. 381, que, entre outros requisitos, a sentença conterá a indicação dos motivos de fato e de direito em que se fundar a decisão, não se confundindo esta com a simples enumeração dos meios de prova de que se utilizou o julgador para decidir.

Em sua decisão, o juiz não é obrigado a acolher nem o laudo nem os pareceres dos assistentes técnicos, podendo inclusive, determinar uma nova perícia caso esta não seja suficientemente elucidativa. Com isso, o juiz pode fundamentar sua decisão de acordo com o seu livre convencimento dentro dos limites do processo. Conclui-se então, que o juiz não está obrigado a valorar com maior importância qualquer prova, incluindo-se a prova pericial da análise do estilo literário para identificação da autoria, foco deste estudo. Reza o Código de Processo Civil em seu artigo 436: "O juiz não está adstrito ao laudo pericial, podendo formar a sua convicção com outros elementos ou fatos provados nos autos". A prova pericial aqui referida deve, então, ser muito clara de modo a que realmente possa influenciar na decisão do juiz, para que não seja simplesmente uma mera prova e sim que cumpra seu papel de auxiliar o juiz na busca da verdade dos fatos.

Um problema enfrentado pelo direito processual é o do surgimento de novos meios de provas, devido aos progressos tecnológicos e ainda não disciplinados de forma expressa pela lei. A lei, em muitos casos, não consegue acompanhar a evolução da tecnologia, fazendo com que não haja embasamento jurídico correto em casos envolvendo documentos eletrônicos como meio de prova[Cal99].

Com embasamento científico é perfeitamente possível a realização de uma análise baseado no estilo literário do autor pois, "com o surgimento de novas tecnologias certamente novos meios de prova podem ser apresentados no âmbito do direito processual e que futuramente virão a se consolidar. Exemplo disso são: a prova judicial via satélite, o interrogatório do acusado no processo penal via satélite, atos cometidos no meio eletrônico com o uso de documentos eletrônicos ou via internet"[Cal99].

2.10 Reconhecimento de Padrões

Nesta seção são apresentadas as abordagens, técnicas e modelos que foram necessários para o correto entendimento e abrangência do campo de pequisa, bem como obter conhecimento para proposição dos métodos de identificação da autoria apresentados no capítulo 4.

2.10.1 Técnica de Aprendizado de Máquina - SVM

Técnicas de aprendizado de máquina tem sido utilizadas exitosamente nos últimos anos em pesquisas para classificação de padrões de autoria [DKLP03] [vHBT+05]. Isto deve-se ao fato de que algoritmos de aprendizado de máquina tem a capacidade de predizer (através da generalização) dados não conhecidos. Os algoritmos devem produzir um modelo com os dados aprendidos. Esses modelos serão testados através da classificação correta de dados não conhecidos [Cor03]. Existem vários tipos de algoritmos de aprendizado de máquina: Aprendizado Baseado em Regras, Árvores de Decisão, Redes Neurais, SVM (*Support Vector Machine*), etc. Nas próximas subseções aborda-se brevemente SVM (Duas Classes) e SVM Multi-Classe por se tratarem dos classificadores utilizados nesta pesquisa (Capítulo 5). Não é foco deste trabalho comparações com outros algoritmos de aprendizado de máquina.

2.10.1.1 SVM - Duas Classes

Support Vector Machine (SVM) foi desenvolvido por V. Vapnik [Vap98] e é uma técnica de aprendizado estatístico. A técnica se baseia no princípio da MRS (*Minimização do Risco Estrutural*). A MRS possui dois objetivos. O primeiro é controlar o risco empírico no conjunto de treinamento. O segundo é controlar a capacidade da função de decisão f usada para obter esse valor de risco. A Função de decisão do *SVM* linear é descrito por um vetor de peso $\overrightarrow{w}$, um bias $\overrightarrow{b}$ e um padrão de entrada $\overrightarrow{x}$ (Equação 2.1).

$$f(\overrightarrow{x}) = sign(\overrightarrow{w}.\overrightarrow{x} + b) \tag{2.1}$$

Dado um conjunto de vetores de treinamento S_l (Equação 2.2) pertencente a duas classes separáveis, $W_1(y_i = +1)$ e $W_2(y_i = -1)$, o *SVM* encontra o hiperplano com a máxima distância Euclidiana do conjunto de treinamento. De acordo com o princípio da MRS, existirá somente um hiperplano com a margem máxima δ, definida como a soma das distâncias do hiperplano até o ponto mais próximo das classes. Esse limiar do classificador linear é a separação ótima do hiperplano (Equação 2.2):

$$S_l = ((\overrightarrow{x}_1, y_1), ...(\overrightarrow{x}_l, y_l)), \overrightarrow{x}_i \in \Re^n, y_i \in \{-1, 1\} \tag{2.2}$$

No caso de conjuntos de treinamento não separáveis, o i-ésimo ponto possui uma variável de folga ξ_i, que representa a magnitude do erro de classificação. A função de

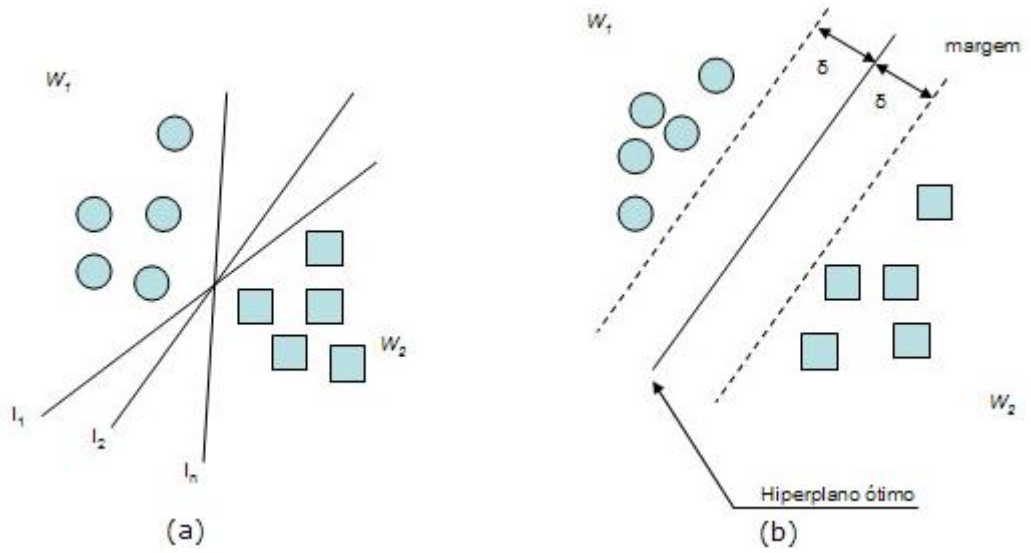

Figura 2.5: Classificação entre duas classes W_1 e W_2 usando hiperplanos: (a) Hiperplanos arbitrários l_i e (b) hiperplano com separação ótima, máxima margem.

penalidade $f'(\xi)$ representa a soma dos erros de classificação (Equação 2.3):

$$f'(\xi) = \sum_{i=1}^{l} \xi_i \tag{2.3}$$

A solução do SVM pode ser encontrada através da minimização dos erros de treinamento com a seguinte minimização (Equação 2.4):

$$\min_{\overrightarrow{w},b,\overrightarrow{\xi}} = \frac{1}{2}\overrightarrow{w}.\overrightarrow{w} + C\sum_{i=1}^{n} \xi_i \tag{2.4}$$

sendo que C ¿ 0, determina uma negociação entre o erro empírico e o termo de complexidade. O parâmetro C é escolhido livremente. Um grande valor para C corresponde a uma associação de uma penalidade mais alta para os erros [SJBS04].

2.10.1.2 $SVM^{multiclass}$

$SVM^{multiclass}$ é uma implementação do algoritmo multi-classe do SVM proposto por Crammer [CS01]. Para o conjunto de treinamento $(x_1, y_1)...(x_n, y_n)$ com $y_1 \in [1..k]$ este algoritmo encontra a solução para o problema de otimização durante o treinamento

demonstrado na Equação 2.5:

$$\begin{aligned} &\min \textstyle\sum_{i=1..k} w_i * w_i + C/n \sum_{i=1..n} \xi_i \\ &\quad s.t.\ for\ all\ y\ in\ [1..k] : [x_1 \bullet w_{yi}] >= [x_1 \bullet w_y] + 100 * \Delta(y_1, y) - \xi_1 \\ &\qquad \ldots \\ &\qquad for\ all\ y\ in\ [1..k] : [x_n \bullet w_{yn}] >= [x_n \bullet w_y] + 100 * \Delta(y_n, y) - \xi_n \end{aligned} \tag{2.5}$$

na qual C é o parâmetro de regularização comum que "negocia" com o tamanho da margem e o erro de treinamento. $\Delta(y_n, y)$ é a função de perda que retorna 0 se y_n igual a y e 1 caso contrário.

Para resolver esse problema de otimização, $SVM^{multiclass}$ implementa um algoritmo diferente do SVM original [CS01]. O algoritmo é baseado em SVM Estrutural [THJA04] e em SVM^{struct} apresentando para kernel linear (utilizado nesta pesquisa) $SVM^{multiclass}V2.12$ um excelente tempo de processamento para a classificação.

2.10.2 Atributos Estilométricos

A tarefa de atribuição de autoria pode ser vista como uma tarefa de classificação, na qual documentos de autoria conhecida são utilizados como treinamento com o objetivo de identificar os autores corretos de documentos desconhecidos baseado nos modelos gerados. Em nesta caso, como em vários outras classificações, o principal problema é não se ter certeza de quais características devem ser utilizadas para se classificar, ou seja, distinguir os autores.

De acordo com Rudman [Rud98] "pelo menos 1000 características existem para pesquisa estilométrica". Muitos avanços ocorreram nos últimos anos com a seleção de novas características e combinações entre elas. Contudo analisando os últimos trabalhos (Capítulo 3) verifica-se que não há consenso entre os pesquisadores e suas linhas de pesquisa sobre quais características, de fato, são as melhores para atribuição de autoria [Cor03].

Características estilométricas que auxiliam na atribuição da autoria podem ser classificadas em 4 grupos [ZQHC06] (A Figura 2.6, apresenta esta classificação [AC05]):

- Léxicas

  As características léxicas podem ser baseadas em palavras ou em caracteres. Características baseadas em palavra incluem tamanho de palavras, quantidade de palavras por frase, distribuição dos tamanhos das palavras, e diversidade do vocabulário. Como diversidade do vocabulário inclui-se o número de palavras que aparecem uma

única vez (*hapax legomena*) e duas (*hapax dislegomena*) como também as medidas estatísticas utilizadas em vários trabalhos (ver Abordagens Estatísticas 3.2.1). Características baseadas em caracter incluem total de caracteres, caracteres por frase, caracteres por palavra e freqüência das letras isoladamente [ZQHC06] [AC05]. Em alguns trabalhos nos quais foram utilizadas unicamente estas características os resultados demonstraram que as mesmas não são por si só suficientes como elemento discrimatório. Isto se deve a variações dependendo da categoria do texto, além de perderem muito de sua discriminatoriedade em se tratando de textos pequenos [Mal04] [Hol95].

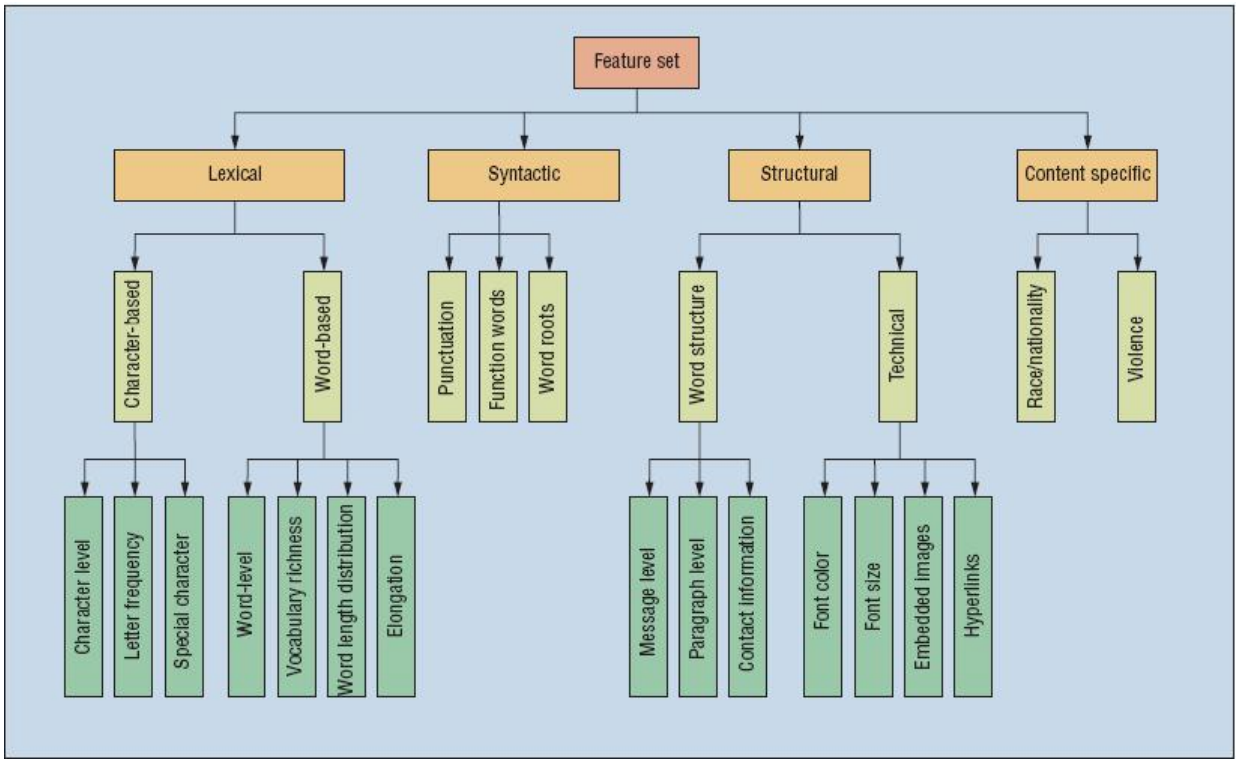

Figura 2.6: Agrupamentos e exemplos de características do estilo (Adaptado [AC05])

- Sintáticas

  Segundo dicionário Aurélio [Fer04] sintaxe é "Parte da gramática que estuda a estrutura da frase...". Esta categoria inclui os padrões utilizados para formar as frases, como por exemplo, pontuações, palavras gramaticais. Palavras gramaticais, palavras-função, *function words* ou *grammatical words* são palavras que indicam uma relação gramatical com outras palavras ou frases. Incluem verbos auxiliares, conjunções, advérbios, etc. A diferença entre a utilização de uma palavra-função ou outra pode parecer não discrimatória, porém vários trabalhos recentes [AL05] [ZZ05] demonstram a importância desta característica como criadora de um perfil

do autor [AC05].

Nesta classificação enquadram-se as características (palavras-função) utilizadas para este trabalho.

- Estruturais

  Características estruturais estão relacionadas com a organização do texto e com a disposição das informações. As características estruturais mais conhecidas e estudadas estão focadas na estruturas de palavras, como por exemplo, saudação inicial, encerramento, quantidade e tamanho dos parágrafos, etc. Embora estas características tenham sua importância, muitas outras características podem carregar importantes informações sobre o autor. Exemplo: fontes (tamanho, cor, tipo), formato de links, disposição de imagens, etc.

- Conteúdo Específico

  É similar ao grupo léxico baseado em palavras porém com um nível de refinamento ampliado. Estas características são palavras relacionadas ao contexto do documento. Por exemplo, em colunas de jornais sobre economia poderia utilizar-se palavras como LEI, CRISE, COMÉRCIO, INDÚSTRIA.

2.10.3 Vetores de Dissimilaridade

Para se trabalhar com identificação de autoria, comumente é considerado o modelo por autor (ou modelo pessoal). O modelo pessoal (ver Subseção 2.10.6.2) é baseado em duas diferentes classes geradas por documentos de mesmo autor e documentos de autores distintos. Estas classes são entrada para geração de um modelo de aprendizado para cada autor. A grande desvantagem deste processo é a necessidade da geração de novos modelos a cada inclusão de um novo autor assim como a importância de um grande número de documentos para geração de modelos confiáveis. Visando transpor estas limitações, esta pesquisa trabalha com a verificação das discrepâncias entre os documentos, representada através de vetores de dissimilaridade [PD02]. Uma vez as características extraídas e sendo V_i o vetor de características estilométricas dos documentos de referência e Q_i o vetor de características estilométricas dos documentos questionados, então, os vetores de dissimilaridade dados por $Z_i = ||V_i - Q_i||$ geram n instâncias diferentes utilizadas pelo classificador para obter suas decisões. Uma das principais vantagens na utilização de vetores de dissimilaridade, está na geração de um modelo global independente de autor e na geração de modelos mais robustos mesmo com poucas amostras para a abordagem através de modelo pessoal.

2.10.4 Dicotomia

Como já mencionado na subseção 2.10.2, a identificação da autoria, é um problema de classificação. Um cenário clássico para identificação da autoria é formado por m documentos de cada n autores e um documento questionado x. Quanto maior o número de autores n, maior a dificuldade em se determinar a autoria do documento x. Isto se deve ao fato de que o agrupamento no subconjunto de classes não provê a validade da individualidade da escrita [Cha01a]. Para se estabelecer um modelo de distinção de classes considerando toda a população n Cha [Cha01a] utilizou a técnica da dicotomia.

Para que seja considerada toda a população, a técnica da dicotomia modela o problema em duas classes. Autoria e não-autoria. A classe de autoria é formada pelos vetores de dissimilaridade (Subseção 2.10.3) de mesmo autor e classe de não-autoria formada pelos vetores de dissimilaridade de autores distintos. Exemplo de transformação policotômica para dicotômica pode ser visualizado na Figura 2.7.

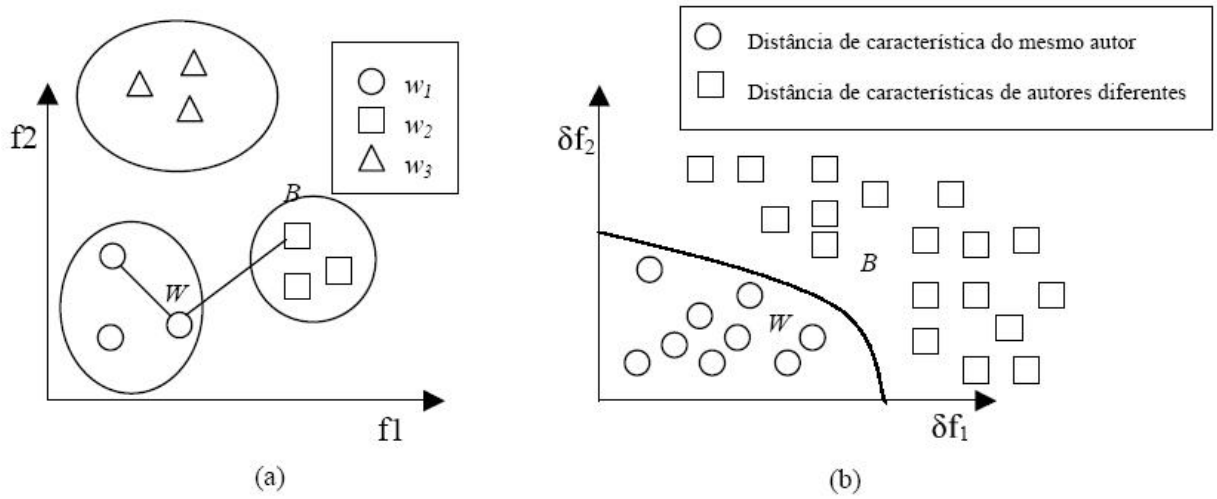

Figura 2.7: Exemplo de transformação: policotomia (a) → dicotomia (b)

As vantagens da dicotomia e razões para utilização da mesma nos métodos demonstrados neste trabalho são [Cha01a] [SJBS04]:

- Múltipla integração de tipos de características - Ao final, os vetores são compostos de valores escalares independente do tipo da característica;
- Solução para problemas de grandes número de classes;
- Determina a validade da individualidade do autor entre toda a população;
- Pode-se utilizar uma quantidade ilimitada de amostras;
- Nada impede a utilização de um dicotomizador em problemas de poucas classes;

- Modelo policotômico não é capaz de deduzir os resultados para toda a população, pelo fato de existirem classes invisíveis;
- Utilização em trabalhos de reconhecimento de padrões [Cha01a] [SJBS04].

2.10.5 Abordagens de Automatização

As três abordagens de automatização comumente utilizadas como ferramentas para atribuição da autoria são: (1) estatística, tem apresentado um alto índice de precisão através da testes como χ^2, t-test, etc. Tem a seu favor uma maior explanação, que pode ser útil para avaliar probabilidades e variâncias, principalmente em textos maiores; (2) Computacional, utiliza um conjunto de métodos (inclusive estatísticos, matemáticos), ferramentas e formulações direcionadas ao contexto das características extraídas; (3) Aprendizagem de Máquina, tem ultrapassado outras abordagens devido proporcionar grande escalabilidade m relação ao número de características que podem ser manipuladas bem como demonstra boa adaptação com textos pequenos [Cor03] [AC06].

2.10.6 Modelos de Classificação

As abordagens de automatização vistas na subseção anterior usualmente são utilizadas com dois modelos de classificação para identificação de autoria: global e pessoal [SJBS04]. O modelo pessoal utiliza um modelo por autor, enquanto que o modelo global faz uso de um modelo geral para todos os autores.

2.10.6.1 Modelo Global

No modelo global utiliza-se de dicotomia (Seção 2.10.4). A classe w_1 representa a classe de espécimes genuínos dos autores usados para treinamento e a classe w_2 representa o conjunto de espécimes pertencentes a autores distintos (representação na Figura 2.8). Na verificação, o modelo gerado é então utilizado para comparação com o documento desconhecido [Bar05]. O modelo global possui a desvantagem da generalização. No entanto, possui a vantagem de necessitar de um número reduzido de textos para cada autor e de não necessitar de um novo treinamento do modelo, na inclusão de novos autores.

2.10.6.2 Modelo Pessoal

O modelo pessoal, também conhecido como modelo por autor, baseia-se no conceito de policotomia (Subseção 2.10.4), ou seja, classificação do problema em n-modelos.

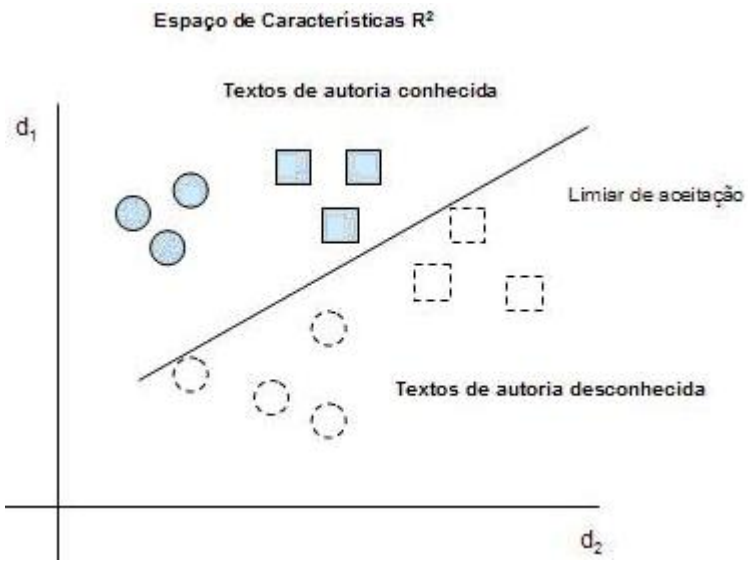

Figura 2.8: Modelo Global de Identificação da Autoria de Textos

Usualmente necessita de um grande número de documentos para a geração de modelos confiáveis. Tem como benefício descrever discrimantemente as variabilidades interpessoais, porém dependendo do número de classes gera um processo computacional lento e exige a geração de um novo modelo na inclusão de um novo autor [Bar05].

Comentários Finais

Neste capítulo foram abordados os principais pontos teóricos necessários para elaboração deste trabalho. Como principais seções apresentadas pode-se destacar a análise da língua portuguesa e sua individualidade, a contextualização do trabalho dentro do campo de pesquisa da lingüística, análise jurídica do atual tratamento de documentos digitais e estilometria para identificação da autoria e o conhecimento técnico dos métodos e processos para criação dos métodos e desenvolvimento dos experimentos e conclusões.

No capítulo seguinte, são descritas abordagens e trabalhos envolvendo estilometria para identificação de autoria.

Capítulo 3

Estado da Arte

O campo da análise da autoria é amplo e pode ser dividido conforme a Figura 3.1.

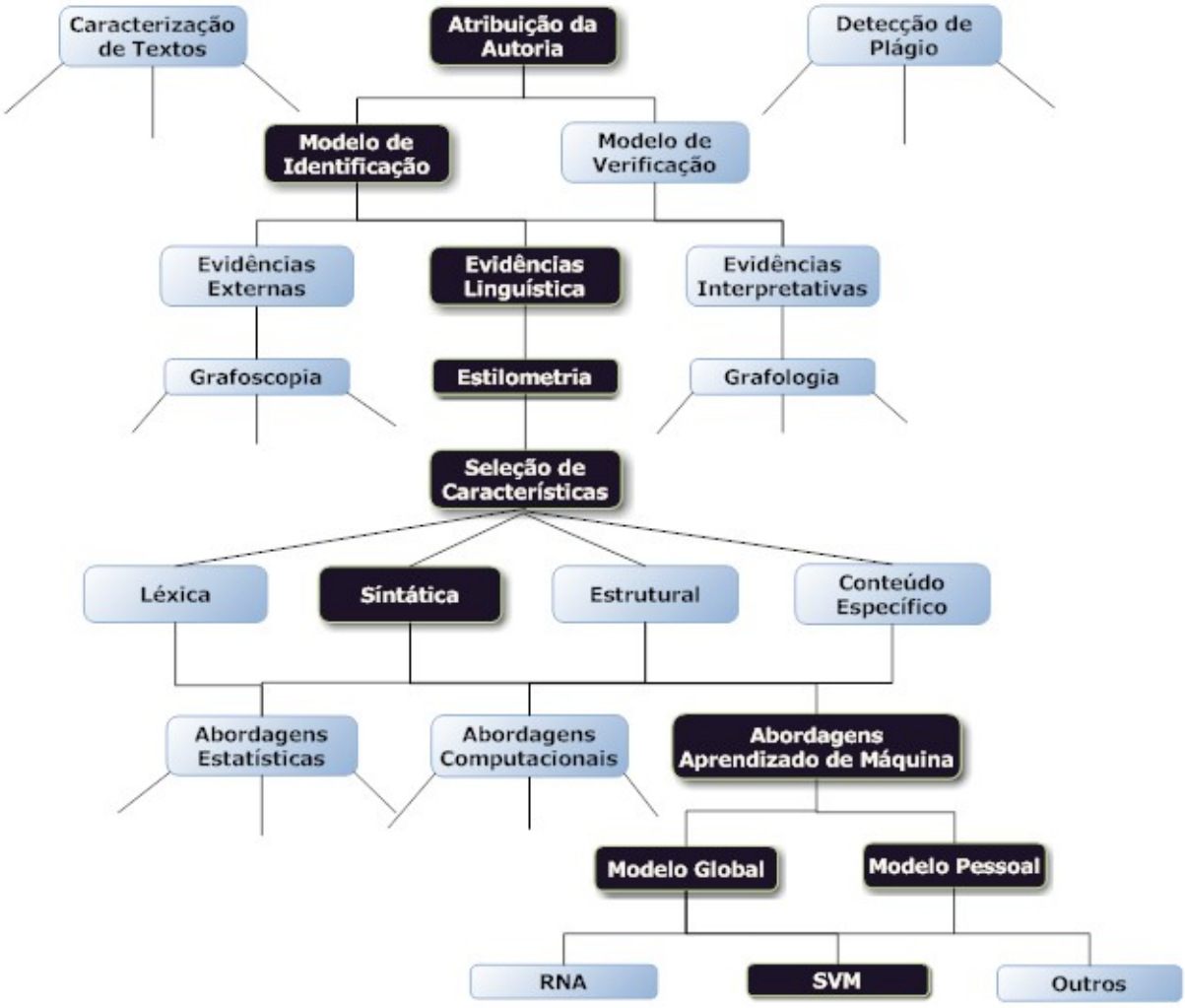

Figura 3.1: Exemplo de Divisão do campo da análise da autoria - Foco Estilometria. Os blocos pretos estão representando o foco deste trabalho.

De acordo com os blocos pretos (letra branca) da Figura 3.1, o foco deste trabalho se encontra em: Atribuição da Autoria → Modelo de Identificação (subseção 2.8.2) → Evidências Lingüísticas (subseção 2.10.2) → Estilometria (subseção 2.8.4) → Seleção de Características Estilométricas (subseção 2.10.2) → Abordagem com Aprendizagem de Máquina (subseção 2.10.5) → Abordagem Global e Pessoal (subseção 2.10.6) → SVM Binário e SVM Multi-classe (subseção 2.10.1). Cada um dessas subdivisões é analisada nas seções citadas.

Existem muitas pesquisas por diferentes caminhos no campo de pesquisa de estilometria [Hol85]. A atribuição automática da autoria de documentos digitais baseada na análise estilométrica apresenta diferentes abordagens, métodos e técnicas dependendo das características utilizadas e modelos de classificação. Neste capítulo, além de um breve histórico, são apresentados alguns importantes trabalhos que foram utilizados como referência para esta pesquisa. Ao final deste capítulo será apresentado um resumo separado por características utilizadas e abordagens de automatização.

3.1 Histórico

Cronologicamente os estudos em atribuição de autoria de documentos baseado na análise estilométrica incluem:

- 1901 - Mendenhall analisou a distribuição dos tamanhos de palavras de Bacon e Shakespeare analisando tamanho e a freqüencia relativa de palavras. Ele concluiu que a característica apresentava uma boa diferença interpessoal e que o método era bom para identificação da autoria (ver [Wil75]);

- 1932 - Zipf baseou-se em diferentes palavras utilizadas por autores. Zipf determinou que existe uma relação logarítmica baseada no número de ocorrências de uma palavra em um texto (Lei de Zipf) [Zip75];

- 1938 - Yule iniciou estudos baseados no tamanho de frases, porém, conclui que somente este caminho não era confiável. Mais tarde, Yule criou uma medida mais confiável, utilizando a lei de Zipf com uma distribuição *Poisson* [Yul38];

- 1940 - Williams analisou o tamanho das frases dos trabalhos escritos por Chesterton, Wells, and Shaw. Williams identificou que o log do número de palavras por frase ocorre de acordo com uma distribuição normal [Wil40];

- 1963 - Brinegar também utilizou distribuição do tamanho de palavras para determinar se Mark Twain tinha escrito as cartas Quintus Curtius Snodgrass (QCS (*Quintus*

Curtius Snodgrass)) [Bri63];

- 1963 - Mosteller e Wallace depois de usarem características como tamanho de palavras e tamanho de frase, utilizaram funções para contar palavras para identificar os trabalhos de Hamilton e Madison nos "Federalist Papers". Utilizaram teorema de Bayes para o problema ao invés das abordagens clássicas até então utilizadas. Identificaram 14 artigos como de James Madison [MW64];

- 1967 - Särndal trabalhou com distribuição quantitativa de palavras para determinar a probabilidade dos erros Tipo I (false rejeição) e Tipo II (falsa aceitação) para distinguir 2 autores. Criou várias hipóteses arbitrárias baseadas nesta característica [Sär67];

- 1975 - Willians analisou que Mendenhall cometeu erros em algumas de suas conclusões e que existem evidências que demonstram não ser totalmente confiável o método utilizado por ele [Wil75];

- 1985 - Holmes questionou os testes de Brinegar (1963), pois baseavam-se em uma hipótese duvidosa (amostras de um autor variavam a partir de uma freqüencia fixa de distribuição do tamanho das palavras) [Hol85];

- 1985 - Holmes fez uma revisão da análise do estilo literário listando possíveis fontes de características e técnicas para identificação de autoria. Dentre elas estão: freqüencia e distribuição baseada no tamanho de palavras, média de sílabas por palavra e distribuição das sílabas por palavra, tamanho de frase, freqüência de palavras, riqueza de vocabulário e léxica, distribuição do vocabulário, distribuição das freqüências de palavras [Hol85];

- 1987 - Thisted e Efron utilizaram conceitos relacionados à riqueza do vocabulário para determinar a possibilidade de Shakespeare ser o autor de um novo poema questionado [TE87];

- 1995 - Lowe e Matthews, utilizaram redes neurais para comparar os perfis de Shakespeare a John Fletcher [LM95];

- 1996 - Merriam utilizou palavras que indicam um relacionamento gramatical (Exemplos: verbos auxiliares, preposições, conjunções) para comparar os estilos de Shakespeare e Christopher Marlowe [MM94];

- 1996 - Eliott e Valenza trabalharam na atribuição de poemas e peças a Shakespeare construindo um "estilo" Shakesperiano em comparação a outros autores baseado

principalmente nas incidências ou não de palavras (raras, novas, contrações, prefixos, sufixos, etc.) [EV91];

- 1996 - Foster estuda o poema "Uma Elegia Fúnebre", atribuindo-o a Shakespeare, utilizando como referência os trabalhos canônicos de Shakespeare, comparando o estilo, acidentes gramaticais, sintaxe e uso de palavras raras [Fos96a];

- 1996 - Foster analisa e atribui a autoria do texto "Primary Colors" - uma sátira ao presidente Clinton, publicado originalmente sem autoria [Fos96b];

- 1999 - Foster questiona os testes de Elliott e Valenza alegando "imperfeições tanto no projeto quanto na execução" [Fos99];

- 2000 - Foster contribui para identificação do "Unabomber" Ted Kaczynski comparando o "Manifesto Unabomb" com outros documentos fornecidos por seu irmão [Fos00];

- 2001 - Chaski não só realizou um bom trabalho científico, mas também preocupou-se com a aceitação e impacto dos processos perante a corte norte-americana [Cha97]. Chaski apresentou resultados empíricos divididos em 3 grupos de características: (1) Pontuação e estrutura de frase, (2) vocabulário, análise do conteúdo, complexidade frasal e (3) características relacionadas a erros, como por exemplo, erros gramaticais, erros de formação frasal, erros de pontuação. Para comparação das medidas de um autor com outros Chaski utilizou o método estatístico χ^2 [Cha01b];

- 2002 - Elliott e Valenza corrigiram algumas técnicas e afirmaram terem um teste afinado para identificação de autoria de Shakespeare [EV02];

- 2002 - Monsarrat afirmou que John Ford foi o autor de "Uma Elegia Fúnebre (W.S.)" atribuída a Shakespeare por Foster [Mon02];

- 2002 - Foster afirma que baseado nas evidências do trabalho de Monsarrat, a assinatura "W.S." de "Uma Elegia Fúnebre (W.S.)" está mais fortemente associada a Ford do que a Shakespeare (ver Apêndice A);

- 2002 - Smith e Kelly utilizaram características léxicas e de vocabulário para classificar trabalhos cronologicamente. Como resultado foi demonstrado que o estilo do autor varia durante o tempo (para estes testes os documentos tinham 10 anos de diferença) [SK02].

Devido ao grande desafio da área de pesquisa, muito se tem trabalhado e publicado em relação ao assunto de identificação de autoria nos últimos anos[1]. Dentro do objetivo desta pesquisa, os trabalhos recentes utilizados como referência são:

- 2003 - M. W. Corney - Utilizando várias características extraídas de textos com poucas informações (e-mails), sua dissertação de mestrado aborda uma metodologia robusta de estilometria e de aprendizado de máquina (SVM) [Cor03];
- 2003 - Diederich et al. utilizaram a vantagem do SVM trabalhar bem com grandes vetores e baseia-se em todas as palavras de textos de jornais alemães. Obteve resultados de 60% a 80% de acertos, além de descrever uma completa análise comparativa com outras abordagens [DKLP03];
- 2004 - Coutinho et al. - Utilizaram algoritmo de compressão PPM para identificar autoria com taxas de acerto de 78% [CMRJB04];
- 2007 - A. Garcia e J. Martin - Apresenta uma análise demonstrando que cada constante (Exemplo: Yule(K) e Zipf(Z)) deve ser utilizada para medir características específicas e como completar a outras [GMC07];

3.2 Identificação da Autoria - Estilística

O campo da análise da autoria pode ser subdividida em distintos focos, tais como, atribuição de autoria, caracterização da autoria e identificação de plágio [Cor03].

Na literatura referente a atribuição da autoria (foco desta pesquisa), três tipos de evidências podem ser utilizadas para atribuir a autoria: externas, lingüísticas e interpretativa [Cra98]. Evidências externas estão relacionadas a manuscritos, por exemplo. Evidências interpretativas estão relacionadas ao que o autor pretendia quando escreveu o documento e como podem ser comparados trabalhos de um mesmo autor. Evidências lingüísticas estão focadas nas palavras e padrões de palavras utilizadas em um documento. Este trabalho está focado nas evidências lingüísticas (palavras) descritas com detalhes na seção 2.10.2 [Cor03].

3.2.1 Probabilísticas e Estatísticas

Hänlein [Hän99], vê o trabalho de atribuição de autoria como uma combinação de métodos intuitivos e estatísticos. Em lingüística forense métodos estatísticos são uti-

[1] Aproximadamente 66 mil resultados para "Authorship Atributtion"no site de busca Google - 15/07/2007

lizados para medir probabilidades [Ols04]. Dentro do histórico visto na Subseção 3.1, estatísticos e matemáticos foram os primeiros pesquisadores da investigação da autoria de documentos. As características extraídas (tamanho de frases, palavras repetidas, quantidade de palavras, freqüência de palavras) eram comparadas entre autores através de métodos estatísticos. Estes testes tem como principal função avaliar a importância do relacionamento das variáveis (características extraídas) através de comparações. Um breve histórico dos testes estatísticos e probabilísticos e alguns trabalhos de referência:

- distribuição de freqüência: Descreve uma freqüência relativa das ocorrências de variáveis necessárias para aplicar todos os testes [McM02];
- t-test, t, erro padrão da diferença, análise da variância: Avaliam o potencial relacionamento das variáveis [Dav90];
- teste de proporção - z: Avalia o potencial relacionamento das variáveis, em porcentagem [Dav90];
- chi square - χ^2: Avalia o potencial relacionamento das variáveis, em freqüência [Dav90];
- coeficiente de correlação - r: Avalia a correlação das variáveis [Dav90];
- estimativa de freqüência - P - razão de verossimilhança - lambda: Estima a probabilidade comum da ocorrência das variáveis em corpora similares [McM02];

Várias dos testes estatísticos citados tem demonstrado um alto nível de precisão. Técnicas estatísticas têm com um de seus benefícios prover um maior potencial explicativo, que pode ser útil para avaliar tendências e discrepâncias, principalmente com grandes textos [AC06]. Por outro lado, segundo Olsson [Ols04], uma das desvantagens com abordagens estatísticas é que não trazem nenhum conhecimento lingüístico para o problema, ou seja, as variáveis são tratadas diretamente sem uma análise lingüística do contexto e relevância das mesmas.

3.2.2 Computacionais

Durante a década de 90 os líderes no campo da estilística eram Burrows, Baayen e co-autores e Holmes e co-autores. Todos trabalharam com pesquisas no sentido de definir melhores características e aplicar técnicas de classificação mais eficazes do que as atuais técnicas estatísticas [Cor03]. Dentre as técnicas computacionais utilizadas e trabalhos utilizados como referência podemos citar:

- Burrows em 1992 utilizou uma análise de padrões de palavras que apareciam mais freqüentemente em textos correlacionando cada palavra com todas usando método Pearson[2] [Bur92];

- Baayen et al. em 1996 executaram experimentos com uma base de dados "sintaticamente anotada". A idéia de Baayen et al. era que com uma anotação sintática pudesse reescrever regras geradas de um analisador de textos [Bea96];

- Em 2001, Holmes et al. utilizaram abordagens tradicionais e não tradicionais para tentar identificar 17 artigos de autoria desconhecida publicado no jornal *New York Tribune*. Através de método de discriminação, foram selecionadas 3000 palavras dos artigos e confrontadas com um corpus de 50 palavras comuns proposta por Burrows (1992) [Hea01];

- Benedetto et al. (2002) iniciaram a utilização de técnicas de compressão e medidas de entropia para caracterizar um estilo para o autor como também determinar a autoria. Em seu método documentos escritos pelo mesmo autor teriam níveis de compressão similar [Bea02].

- Em 2004, Coutinho et al. utilizaram algoritmo de compactação PPM-C para textos em língua portuguesa [CMRJB04].

Um dos objetivos deste trabalho, é comparar o modelo proposto por Coutinho et al. [CMRJB04], qual utiliza também uma base de documentos da língua portuguesa. Na seguinte subseção será explanado um pouco sobre essa abordagem.

3.2.2.1 PPM-C

Métodos de compressão têm sido aplicados para classificação de textos [Bea02] [Ter06]. O método de compressão de dados PPM (*Prediction by Partial Matching*) é um algoritmo utilizado em muito trabalhos de classificação de textos [Tea00] [Tha01].

A técnica de compactação do PPM consiste na modelagem estatística de contexto finito, ou seja, em estimar probabilidades para um caracter observando (n) caracteres anteriores, armazenando o contexto destes. Através de contadores de freqüência são calculadas as probabilidades em cada contexto, ou seja, segundo Coutinho et al. [CMRJB04] "a probabilidade condicional de x no contexto s, $P(x|s)$ é então estimada por $C(x|s)/C(s)$,

[2] o coeficiente de correlação de Pearson, também chamado de "coeficiente de correlação produto-momento" ou simplesmente de "r de Pearson" mede o grau da correlação (e a direção dessa correlação - se positiva ou negativa) entre duas variáveis de escala métrica.

onde $C(x|s)$ é o número de vezes que x segue s e $C(s)$ é o número de vezes que s é encontrada." [CMRJB04].

Para modelos baseados em palavras, os cálculos são similares (Figura 3.2 [CMRJB04]).

Ordem k = 2					Ordem k = 1					Ordem k = 0				Ordem k = -1			
Predição			c	p	Predição			c	p	Predição			c	p	Predição			c	p
ab	→	r	2	$\frac{2}{3}$	a	→	b	2	$\frac{2}{7}$	→	a	5	$\frac{5}{16}$	→	A	1	$\frac{1}{\|A\|}$
	→	*Esc*	1	$\frac{1}{3}$		→	c	1	$\frac{1}{7}$	→	b	2	$\frac{2}{16}$				
						→	d	1	$\frac{1}{7}$	→	c	1	$\frac{1}{16}$				
ac	→	a	1	$\frac{1}{2}$		→	*Esc*	3	$\frac{3}{7}$	→	d	1	$\frac{1}{16}$				
	→	*Esc*	1	$\frac{1}{2}$						→	r	2	$\frac{2}{16}$				
					b	→	r	2	$\frac{2}{3}$	→	*Esc*	5	$\frac{5}{16}$				
ad	→	a	1	$\frac{1}{2}$		→	*Esc*	1	$\frac{1}{3}$								
	→	*Esc*	1	$\frac{1}{2}$													
					c	→	a	1	$\frac{1}{2}$								
br	→	a	2	$\frac{2}{3}$		→	*Esc*	1	$\frac{1}{2}$								
	→	*Esc*	1	$\frac{1}{3}$													
					d	→	a	1	$\frac{1}{2}$								
ca	→	d	1	$\frac{1}{2}$		→	*Esc*	1	$\frac{1}{2}$								
	→	*Esc*	1	$\frac{1}{2}$													
					r	→	a	2	$\frac{1}{3}$								
da	→	b	1	$\frac{1}{2}$		→	*Esc*	1	$\frac{1}{3}$								
	→	*Esc*	1	$\frac{1}{2}$													
ra	→	c	1	$\frac{1}{2}$													
	→	*Esc*	1	$\frac{1}{2}$													

Figura 3.2: Modelo PPM-C depois do processamento da string abracadabra [CMRJB04]

Segundo Coutinho et al. [CMRJB04] algumas vantagens da classificação utilizando modelos de compressão de dados em relação a modelos clássicos de treinamento podem ser citadas: (1) não há um processo de extração e seleção de características, o algoritmo trabalha com o todo, (2) não é necessário fazer considerações simplificadoras a respeito das distribuições de probabilidades (3) a capacidade de construção adaptativa de modelos, por parte dos algoritmos de compressão, oferece um modo uniforme de classificar diferentes fontes de informação, (4) os métodos de classificação baseados em modelos de compressão são muito simples.

3.2.3 Aprendizado de Máquina

Abordagens com métodos de aprendizado de máquina no campo da estilometria iniciaram-se somente em meados da década de 90 e vem sendo utilizadas com sucesso nos últimos anos. Matthews e Merriam [MM94] utilizaram classificadores de redes neurais para comparar textos de Shakespeare, Marlowe e Fletcher. Em 1994, Kjell [Kje94b] utilizando como características freqüência relativa de 50 ou 100 mais significantes em

n-grams[3] efetuou comparações usando rede neural com classificadores Naïve e *Nearest Neighbour* [Kje94b] [KWF95]. Outros trabalhos com redes neurais, classificadores Naïve e k-NN continuaram com Hoorn em 1999, nos quais analisou textos em alemão obtendo classificação com precisão de até 86% indicando que tal classificação era aplicável a outros idiomas[HFKvdH99].

Lowe e Matthews em 1995 descreveram o uso de rede neural RBF (*Radial Basis Function*) para classificar peças de Shakespeare e Fletcher utilizando como características 5 palavras-função[4] (*function words*)[LM95].

De 1996 a 2000 os trabalhos avançaram com redes neurais (e variações) e algoritmos genéticos [HF95] [Hol98] [TSH96] [WAT00].

3.3 Análise Crítica do Estado da Arte

Um ponto verificado nos trabalhos de identificação da autoria através da estilometria, é que apesar de mais de um século de pesquisa, não existe um consenso entre os pesquisadores sobre quais características realmente são discriminatórias, quando usar, em quais tipos de documentos e principalmente qual a melhor forma de mensurá-las e com isso estabelecer a individualidade do estilo. Pesquisadores tem apresentado sua metodologia e resultados, defendendo-os como um novo método aceitável. Com isso, muito se tem publicado, abordando diversas características estilométricas, métodos estatísticos, probabilísticos, etc. Com essa avalanche de informações, muitos trabalhos recentes não mais propõem a criação de uma nova metodologia, mas sim fazendo *reviews* e estruturações sobre tudo que se tem publicado abrangendo uma determinada característica ou abordagem, etc. [Hol85] [Bea02] [Mal04] [ZQHC06] [GMC07]. A recente "compilação" de todas essas informações e uma estruturação, juntamente com os avanços tecnológicos estão estabelecendo novas premissas importantes para o notável avanço da linha de pesquisa.

[3] n-gram é a seqüência de letras de um pedaço de texto com n caracteres, também conhecido como POS (part of speech)

[4] palavras-função são grupo de palavras analisadas dentro de um contexto específico. Exemplos: conjunções, preposições, pronomes, contrações, advérbios, etc.

Capítulo 4

Método Proposto

Neste Capítulo serão apresentadas as etapas dos métodos de atribuição da autoria de documentos digitais utilizados, através da análise do estilo do autor. Um dos principais desafios proposto por essa metodologia está em trabalhar com características sintáticas da língua portuguesa brasileira. Através da utilização de características sintáticas ("palavras-função"), Seção 2.10.2, é demonstrado um modelo global de atribuição de autoria em comparação com um modelo por autor e método de compactação PPM-C proposto por Coutinho et al. [CMRJB04]. Como foram apresentados dois métodos neste trabalho, este capítulo não possui detalhadas as divisões da base de dados, combinações de documentos, etc. Estas informações estão apresentadas no Capítulo 5.

4.1 Abordagem

Como já visto no capítulo 3 (Estado da Arte) dentro do estudo da estilometria esta pesquisa está inserida em: Atribuição da Autoria → Modelo de Identificação (subseção 2.8.2) → Evidências Lingüísticas (subseção 2.10.2) → Estilometria (subseção 2.8.4) → Seleção de Características Estilométricas (subseção 2.10.2) → Abordagem com Aprendizagem de Máquina (subseção 2.10.5) → Abordagem Global e Pessoal (subseção 2.10.6) → SVM Binário e SVM Multi-classe (subseção 2.10.1).

4.2 Método de Identificação da Autoria

O método proposto de identificação de autoria será feito em 5 fases:

1. Coleta e Formação da Base de Dados;

2. Extração das Características;

3. Geração dos Vetores de Dissimilaridade;

4. Classificação - Produção de Modelos;

5. Processo de Decisão.

A Figura 4.1 apresenta de forma ampla o processo esquematizado para identificação da autoria.

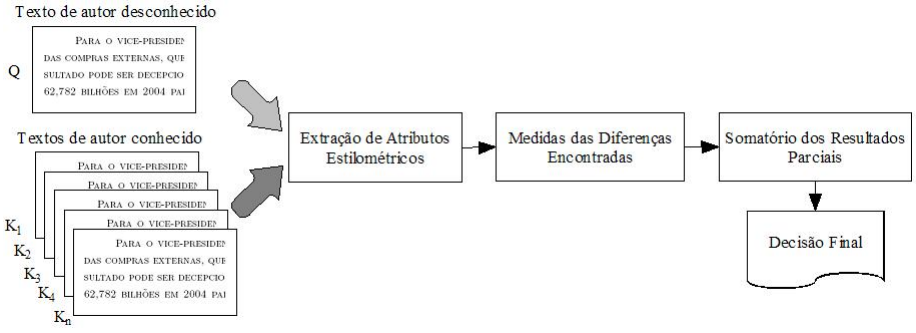

Figura 4.1: Diagrama esquemático das etapas - Estilometria

A seguir, cada etapa será detalhada demonstrando aspectos genéricos e as particularidades encontradas no método proposto.

4.3 Coleta e Formação da Base de Dados

Em casos forenses, textos cuja autoria está sendo questionada são freqüentemente pequenos e com isso, muitas vezes, são de difícil análise para identificação da autoria. Isto significa que, com documentos com poucas informações, lingüistas forenses apresentam seus relatórios com dados brutos, e não em um formato estatístico [McM02].

Para que a pesquisa seja mais próxima desta realidade a base de dados escolhida para provar os modelos foi colunas de 30 jornalistas brasileiros de diferentes jornais. Estas colunas tem um limite de informação desejado para esta pesquisa, em média 735 tokens[1].

Petições, e-mails, cartas demissionárias, cartas de seqüestro, cartas de ameaça, notas de suicídio, etc. possuem um contexto representado com pouca informação. Por isso, representa para esta pesquisa um grande desafio trabalhar com documentos com pouca informação e obter resultados satisfatórios e confiáveis para que sejam aceitos no meio jurídico.

Foram selecionadas colunas de jornais, obtidas pela internet, de 30 colunistas entre jornais do estado do Paraná e São Paulo (ver relação dos colunista no Apêndice B):

[1]palavras válidas para o extrator de características, ou seja, sem pontuação, algarismos, etc.

- Gazeta do Povo - http://www.gazetadopovo.com.br - Paraná;
- Tribuna do Paraná - http://www/parana-online.com.br - Paraná;
- Diário do Grande ABC - http://home.dgabc.com.br/ - São Paulo;
- Correio Popular - http://www.cpopular.com.br/ - São Paulo;
- Folha de São Paulo - http://www.folha.uol.com.br/ - São Paulo.

As colunas de jornais foram escolhidas por alguns motivos:

- Pequenas mas com tamanho suficiente para análise - Média de 3KB;
- Assuntos Diversos;
- Expressam a opinião pessoal do colunista.

A Tabela 4.1 demonstra um exemplo das colunas escolhidas com as informações de título da coluna, data, quantidade de tokens, tamanho em KB, quantidade de *hapax legomena* (quantidade de tokens que não se repetem).Todos os colunistas e seus textos (colunas) foram escolhidos aleatoriamente. As colunas selecionadas foram arquivadas no

Tabela 4.1: Colunas do autor Antônio Pietrobelli

Autor	Jornal	Assunto Principal
Antonio Pietrobelli	Tribuna do Paraná	Comércio Exterior

Colunas	Data	Palavras (Tokens)	Tamanho Kb	Hapax
01 AEB prevê superávit de US$ 42 bilhões, na balança	28/08/2005	800	5,34	403
02 Paraná busca negócios na França	21/08/2005	795	5,16	393
03 Exportações: projeção de US$ 114 bilhões	07/08/2005	847	5,49	313
04 Pacote Dificulta Importações	14/08/2005	871	5,56	412
05 Greve afetou a balança comercial	31/07/2005	732	4,79	356
06 Reflexos da queda do dólar	24/07/2005	892	5,59	424
07 Superávit de US$ 38 bilhões	17/07/2005	783	5,03	408
08 Argentina volta a atacar	10/07/2005	836	5,54	372
09 Exportando pela Infraero	03/07/2005	857	5,34	399
10 Salvaguardas contra a China	26/06/2005	784	5,21	405
11 Exportações crescem 10%	12/06/2005	762	5,02	337
12 Apoio da BM&F para empresários na China	05/06/2005	737	4,65	390
13 Milho transgênico é proibido	29/05/2005	673	4,37	369
14 Empresas perdem com queda do dólar	08/05/2005	774	4,95	371
15 Alca deixa de ser prioridade	24/04/2005	684	4,32	338

formato texto ASCII com acentuação e sem hifenização. Como pode ser visto na Tabela 4.1, os arquivos gerados pelas colunas de jornais são pequenos, variando em média de 3 a 6 KB. Os autores, identificados por suas iniciais, possuem perfil profissional variado. São economistas, empresários, políticos, ex-esportistas, etc. e cada um margeando assuntos específicos como: economia, política, comércio exterior, vinhos e cultura, esportes, humor e notícias gerais.

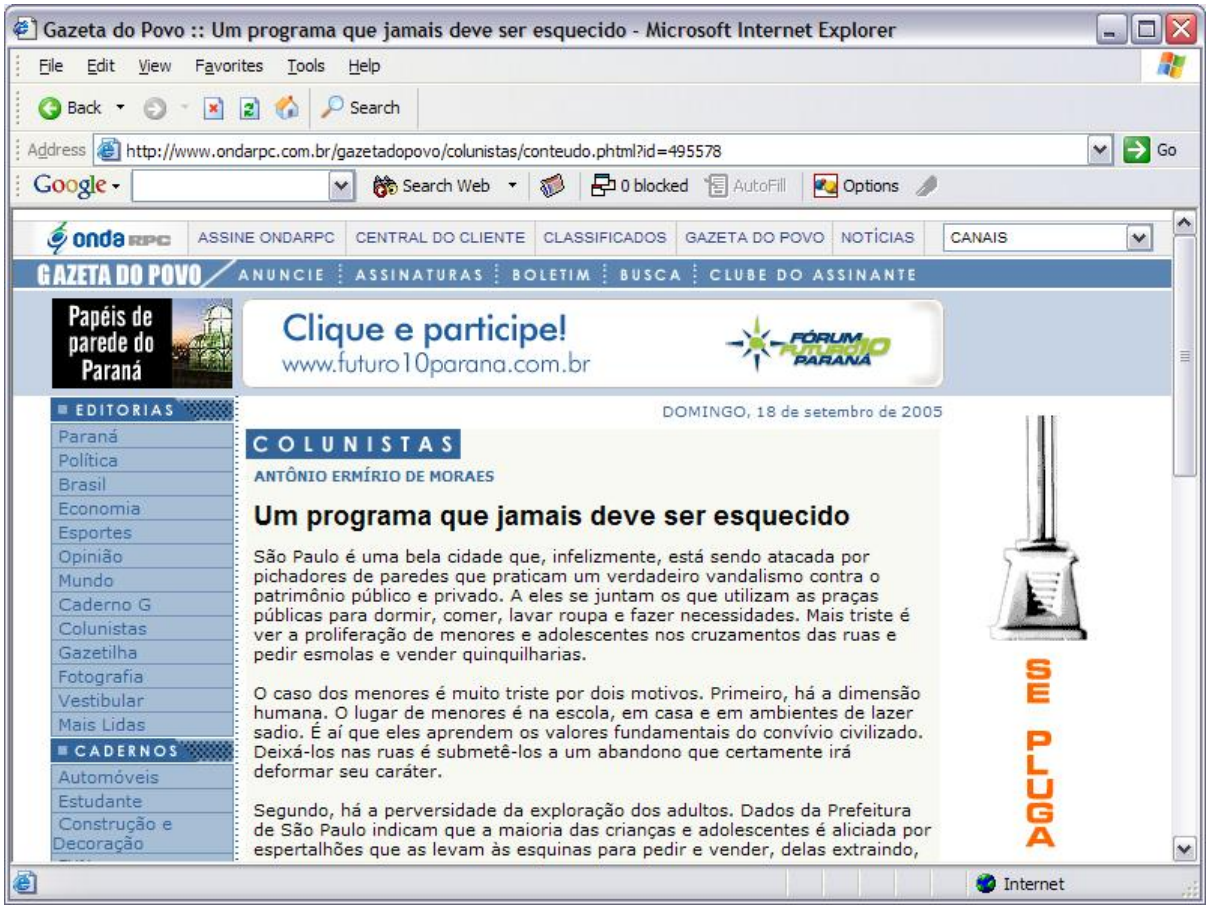

Figura 4.2: Apresentação eletrônica de uma coluna

A Figura 4.2 refere-se a apresentação eletrônica de uma coluna do jornal Gazeta do Povo.

A divisão da base de dados será detalhada por cada experimento nas seções Geração dos Modelos e Protocolos de Teste (Capítulo 5).

4.4 Extração das Características

Umas das etapas mais importante para obtenção de resultados satisfatórios está na escolha de atributos estilométricos que sejam discriminatórios dentro do contexto analisado. No contexto de identificação da autoria, estilo literário é um conjunto de elementos que personalizam o texto escrito de um autor. Texto é uma expressão verbal estruturada (em partes). As partes de um texto devem estar interligadas e suas orações agrupadas, dando uma idéia de continuidade e coesão. Para que orações e palavras sejam ligadas dentro dos textos, concluindo o pensamento, existe a necessidade da utilização de elementos de ligação conhecidos no estudo de uma língua como “conjunções”. Conjunção é uma palavra (ou mais) invariável que liga orações ou as palavras de uma mesma oração.

Advérbios estão relacionados à modificação que o autor dá a um verbo, adjetivo ou outro advérbio. É a "palavra" que indica as circunstâncias em que ocorre a ação verbal.

As conjunções e advérbios foram escolhidos por explorar tanto traços inconscientes do autor e também a opção de escolha do autor, ou seja, em alguns casos a utilização das conjunções e advérbios são automáticas e em outros casos cabe ao autor a escolha da palavra que expressaria da melhor forma sua intenção. Várias conjunções e advérbios podem ser trocados por outras sem modificar o sentido e a idéia do texto. Exemplos: "Ele é tal qual seu pai.", "Ele é tal e qual seu pai.", "Ele é tal como seu pai.", "Ele é que nem seu pai.", "Ele é assim como seu pai.". Ao digitar a frase citada como exemplo, a conjunção utilizada dependerá de cada autor, pois inconscientemente faz parte do estilo literário já formado no autor.

Como já visto na seção 2.10.2 para esta pesquisa estaremos utilizando características sintáticas, conhecidas como "palavras-função" ou *function words*. Vários estudos já foram feitos utilizando palavras-função de língua inglesa com resultados satisfatórios em relação a discriminatoriedade das características [GMC07] [ZZ05] [AL05]. Um ponto diferencial em relação a outros trabalhos com palavras-função é que no método proposto nesta pesquisa, as palavras-função não se restringem somente a uma palavra (Exemplo: embora) mas sim em um contexto de até 3 palavras (Exemplo: a menos que). Isto faz com que a característica analisada seja um teste novo, uma mistura de palavras-função com método de seqüência de palavras *n-gramas*[2], onde n=1..3.[BA04] [TWL02] [MPMyGR05].

No campo desta pesquisa as palavras-função utilizadas como características podem ser agrupadas em conjunções e advérbios. Foram extraídas de cada uma das colunas a quantidade de ocorrências de 171 palavras-função do tipo conjunções e advérbios. Estas palavras-função estão divididas em 77 conjunções e 94 advérbios listadas a seguir. Foram utilizadas conjunções de 12 grupos:

1. Coordenativas Aditivas: e, nem, mas também, mas ainda, senão também, bem como, como também.

2. Coordenativas Adversativas: porém, todavia, mas, entretanto, contudo, senão, no entanto, ao passo que, não obstante, apesar disso, em todo caso.

3. Coordenativas Conclusivas: logo, portanto, por conseguinte, por isso.

4. Coordenativas Explicativas: porquanto, que, porque.

5. Subordinativas Causais: como, visto que, visto como, já que, uma vez que, desde que.

[2] seqüência de palavras ou caracteres em um tamanho pré-definido

6. Subordinativas Comparativas: tal qual, tais quais, assim como, tal e qual, tal como, tão como, tais como, mais do que, tanto como, mais que, menos do que, menos que, que nem, tanto quanto, o mesmo que.
7. Subordinativas Conformativas: consoante, segundo, conforme.
8. Subordinativas Concessivas: embora, ainda que, mesmo que, ainda quando, posto que, por muito que, por mais que, se bem que, por menos que, nem que, dado que.
9. Subordinativas Condicionais: se, caso, contanto que, salvo que, não ser que, a menos que.
10. Subordinativas Consecutivas: de sorte que, de forma que, de maneira que, de modo que, sem que.
11. Subordinativas Finais: para que, fim de que.
12. Subordinativas Proporcionais: à proporção que, à medida que, quanto menos, quanto mais.

Os 94 advérbios utilizados foram:

1. Lugar: aqui, ali, aí, cá, lá, acolá, além, longe, perto, dentro, adiante, defronte, onde, acima, abaixo, atrás, em cima, de cima, ao lado, de fora, por fora.
2. Tempo: hoje, ontem, amanhã, atualmente, sempre, nunca, jamais, cedo, tarde, antes, depois, já, agora, então, de repente, hoje em dia.
3. Afirmação: certamente, com certeza, de certo, realmente, seguramente, sem dúvida, sim.
4. Dúvida: porventura, provavelmente, talvez.
5. Intensidade: ainda, apenas, de pouco, demais, mais, menos, muito, pouca, pouco, quase, tanta, tanto.
6. Negação: absolutamente, de jeito nenhum, de modo algum, não, tampouco.
7. Quantidade: todo, toda.
8. Modo: assim, depressa, bem, devagar, face a face, facilmente, frente a frente, lentamente, mal, rapidamente, algo, alguém, algum, alguma, bastante, cada, certa, certo, muita, nada, nenhum, nenhuma, ninguém, outra, outrem, outro, quaisquer, qualquer, tudo.

A seleção de palavras-função (conjunções e advérbios) como característica a ser analisada nesta trabalho para utilização na associação ou dissociação da autoria deve-se aos seguintes fatores:

- Explorar dados inconscientes do autor;
- Grande diversidade da língua portuguesa, o que proporciona muitas possibilidades;
- Características já utilizadas com sucesso em trabalhos com língua inglesa desde 1964 [GMC07] [ZZ05] [AL05] [MW64].
- Devido aos textos serem extraídos já eletronicamente, sofreram ações padronizadoras de editores de textos, referentes a layout, correções ortográficas, pontuação, etc. Isso elimina algumas características que poderiam ser utilizadas na criação do perfil do autor. Devem ser tratadas com outra abordagem pois são de outra subdivisão - Características Estruturais 2.10.2.

Devido as palavras-função utilizadas estarem em um contexto de *trigramas*, o processo de extração das características seguiu as seguintes diretrizes:

- Separador de Palavras: Espaço em branco, final de linha e caracteres não considerados Tokens. Isto quer dizer que em cada início de linha inicia-se uma nova palavra (sem hifenização);
- Palavras hifenizadas, mesóclises, próclises e ênclises foram consideradas palavras únicas. Exemplo: A frase "eu vou dar-te um pula-pula e também dar-te-ei um beijo, meu amor!" possui 12 tokens sendo 11 *hapax legomena* (palavra "um" ocorre duas vezes);
- Caracteres de pontuação não foram considerados tokens;
- Não houve diferenciação entre letras maiúsculas e minúsculas;
- Não foram considerados algarismos como Tokens;
- Caracteres especiais isolados não são considerados Tokens Ex: *,
- Mecanismo de busca da conjunção:
  - 3 tokens válidos do texto são concatenados;
  - Verifica-se a existência do padrão com os 3 tokens;
  - Caso não exista verifica-se a existência das duas últimas;

- Caso não exista verifica-se a existência da última;
- Para cada caso é feito o caminhamento de acordo com quantidade de tokens (1 a 3) que formam o padrão de busca.

Aclarando a função, o algoritmo de busca existe para que, por exemplo, a palavra-função "para que" seja utilizada como única e não duas palavras-função separadas, uma "para" e outra "que".

A melhor forma de trabalhar/comparar textos de tamanhos diferentes, também é objeto de estudos de várias pesquisas [SG06] [Cor03]. Neste método proposto, as quantidades de ocorrências das conjunções foram normalizadas em relação ao número total de tokens, ou seja, a análise direta dos dados obtidos não apresenta confiança ao utilizar textos de tamanhos diferentes. Algumas abordagens podem trabalhar com textos truncados [CMRJB04], contudo, na análise de palavras-função (conjunções e advérbios) os resultados poderiam estar distorcidos. Isto porque, como um texto é uma seqüência de idéias e em sua maioria tem um início, meio e fim, a probabilidade de encontrarmos algumas palavras-função em um texto truncado, seria menor do que em um texto completo.

4.5 Geração dos Vetores de Dissimilaridade

Na abordagem dicotômica, a métrica utilizada para medir as discrepâncias entre as características está diretamente relacionada com sucesso da classificação [Cha01a].

Com as características extraídas de todos os documentos são distribuídas em vetores, demonstrado na Equação 4.1:

$$fv_Q = (f_1, f_2, ..., f_L) \tag{4.1}$$

sendo f_v os conjuntos de características, Q o "*id* do documento" e L o número total de características.

O vetor de dissimilaridade com os valores escalares é gerado pelo módulo da diferença entre as características extraídas das colunas de acordo com os protocolos de treinamento e testes estabelecidos. Como exemplo, na Figura 4.3 tem-se duas colunas X e Y e o vector de dissimilaridade Z resultante de $Z_i = ||V_i - Q_i||$, sendo V_i e Q_i vetores de características estilométricas extraídas das colunas.

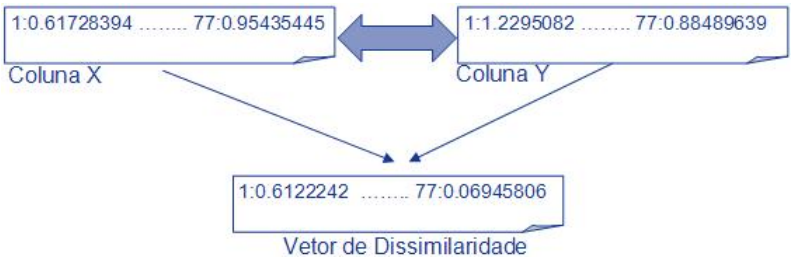

Figura 4.3: Exemplo de Vetor de Dissimilaridade

4.6 Geração dos Modelos

Existem dois estágios no que refere-se a modelos. O primeiro é o treinamento (criação), e o segundo no qual documentos questionados são confrontados ao modelo para que se tenha o resultado. Este último estágio chama-se classificação e será visto com mais detalhes na seção 4.8. Nesta seção será abordado o estágio de treinamento e criação dos modelos.

Este trabalho apresenta dois métodos. O primeiro utiliza um modelo global (2.10.6.1) utilizando SVM binário (2 classes) e o segundo utiliza um modelo pessoal (2.10.6.2) utilizando $SVM^{multiclass}$. Ambos estão descritos nas subseções a seguir.

4.6.1 Modelo Global

Como já visto na subseção 2.10.6.1 este modelo pode trabalhar com o conceito de dicotomia (ver subseção 2.10.4). Com as características extraídas são produzidos os vetores de dissimilaridade gerados de acordo com o protocolo de treinamento, entre documentos de mesmo autor (classe w_1) que representam a autoria e outros com os documentos de diferentes autores (classe w_2) representando a não-autoria ou dissociação.

Para que se estabeleça corretamente o modelo global, três pontos importantes necessitam ser considerados nesta etapa de treinamento:

- Os autores utilizados para a fase de treinamento devem ser exclusivos para tal, ou seja, estes autores não devem ser utilizados como referência ou testes;

- Deve-se evitar o sobre-treinamento, ou seja, o modelo de treinamento deve generalizar as discrepâncias criando um padrão global. Caso muitas amostras sejam utilizadas para esta etapa o modelo gerado estará especializado aos autores, impactando a classificação. Não existe uma fórmula específica para se evitar o sobre-treinamento. Existem métodos como *k-fold cross validation* e *holdout* [Cor03] para evitar o sobre-treinamento e sub-treinamento. Nos métodos apresentados foi utilizado *holdout*

simples, que pode ser visto com mais detalhes no capítulo de experimentos, protocolo de treinamento 5;

- A quantidade de vetores de autoria e não-autoria deve ser as mesmas [Jus02].

O modelo global dicotômico, possui vetores de classes de autoria (+1) e não-autoria (-1) gerados da seguintes forma:

- Geração dos vetores de dissimilaridade entre documentos de mesmo autor separados exclusivamente para testes, através da Equação 4.2;
- Geração dos vetores de dissimilaridade entre documentos de autores diferentes (escolhidos aleatoriamente).

Com os vetores de autoria (+1) e de não-autoria gerados (-1) o classificados SVM gerará o modelo único que será utilizado para os testes.

4.6.1.1 SVM^{light}

Foi utilizado o pacote freeware SVM^{light} [JOA02] para as etapas de treinamento e teste com o modelo global. Com relação a configuração do classificador, foram feitos testes com *kernel* linear e polinomial iniciado com 1 e parâmetros -r e -s em 0,01. A diferença apresentada foi muito pequena, chegando em um valor máximo de 0,5% na taxa de erro média. Considerando esta pequena diferença foi utilizado kernel linear neste trabalho. Maior detalhamento sobre as configurações e parâmetros do SVM pode ser encontrado em [JOA02].

4.6.2 Modelo Pessoal

A abordagem utilizada neste trabalho para criação de modelos pessoais foi através de um classificador multi-classe ($SVM^{multiclass}$ 2.10.1.2). Classificadores multi-classe (RNA (*Redes Neurais Artificiais*) com Backpropagation ou $SVM^{multiclass}$ visto na subseção 2.10.1.2), como no modelo global, geram um único modelo, porém cada vetor treinado possui a informação a qual classe ($w_{1..n}$) pertence, ou seja, somente vetores de autoria são treinados, cada qual referenciando a uma classe.

Mesmo tendo um único modelo, classificadores multi-classe enquadram-se no modelo pessoal, por necessitar criar um novo modelo quanto um novo autor é incluído.

Para geração do modelo multi-classe foram extraídas as características e gerados os vetores de dissimilaridade entre os documento de acordo com o protocolo de treinamento.

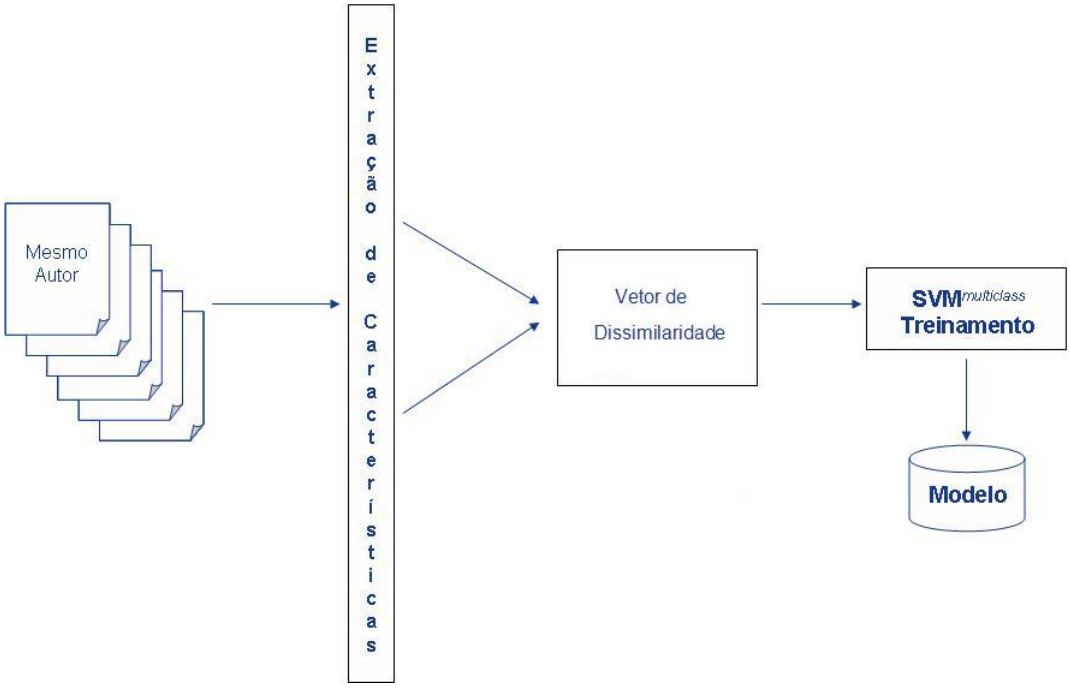

Figura 4.4: Fluxo com $SVM^{multiclass}$

Os vetores de autoria serão resultado das possibilidades dada pela Equação 4.2 onde A é o arranjo de d (neste caso 2) elementos em n (n = número de documentos separados para treinamento). Modelo de classificação com $SVM^{multiclass}$ está representado na Figura 4.4.

$$[b!]A_n^d = \frac{\frac{n!}{(n-d)!}}{2} \tag{4.2}$$

Com a técnica multi-classe (modelo pessoal) este trabalho se propõe também a comparar os resultados do método acima proposto com o método de compactação PPM-C apresentado por Coutinho et al. [CMRJB04].

4.7 Processo de Testes

Nesta fase os documentos questionados são classificados junto ao(s) modelo(s). O processo de teste ocorre especificamente para cada um dos modelos propostos respeitando o protocolo de testes. A Figura 4.5 apresenta o fluxo para associação ou dissociação da autoria.

4.7.1 Modelo Global

Para o modelo global a base de dados foi dividida em três: treinamento, referência e testes. Esta divisão está detalhada no Capítulo 5 - Experimentos Realizados. Lembrando

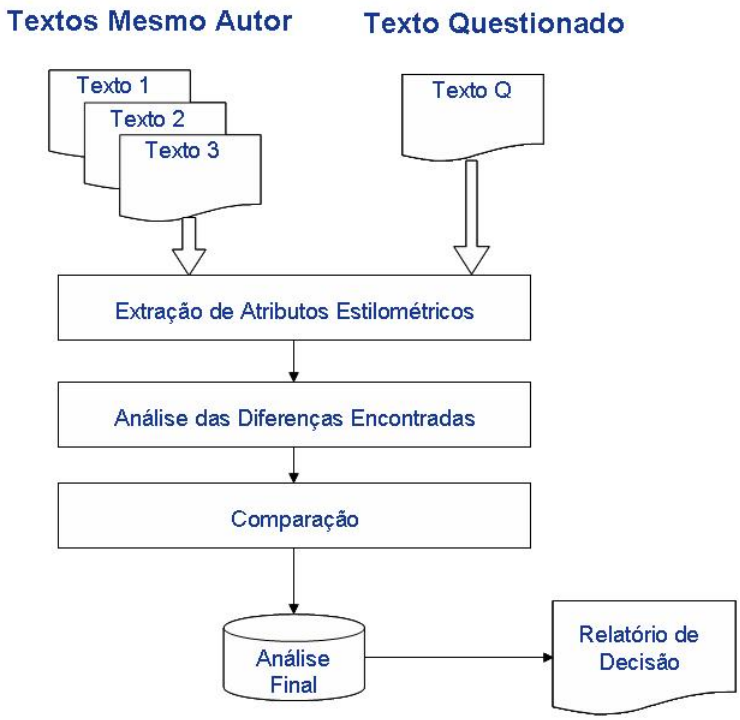

Figura 4.5: Fluxo do processo de testes

que os autores utilizados para treinamento são exclusivos de treinamento, o processo consiste em testar contra o modelo global gerado, vetores de autoria e de não-autoria, para que sejam obtidas as taxas de erros de falsa rejeição (Erro Tipo I) e falsa aceitação (Erro Tipo II).

Os vetores de autoria são gerados através das possibilidades dada pela Equação 4.2 onde A é o arranjo de d (neste caso 2) elementos em n (n = número de documentos separados para teste).

Para a geração dos vetores de não-autoria, são utilizados os documentos para treinamento como referência, ou seja, os vetores de dissimilaridade são formados por um documento de teste e um documento de treinamento de autor diferente (escolhido aleatoriamente).

Ao final são gerados a mesma quantidade de vetores para autoria (+1) e não-autoria (-1) os quais são classificados contra o modelo global.

4.7.2 Modelo por autor - Multi-Classe

No modelo por autor - multi-classe, da mesma forma que no modelo global, são utilizados documentos de treinamento como referência, ou seja, os vetores de dissimilaridade são formados por um documento de teste e um documento de treinamento porém de mesmo autor.

4.8 Processo de Decisão

Nesta fase se avaliam os resultados da classificação. O processo de decisão ocorre especificamente para cada um dos modelos propostos:

4.8.1 Modelo Global

Como resultado da classificação se obtém um arquivo com um número representando a classe atribuída (autoria (¿0) e não-autoria (¡0)) e o grau de confiança no resultado (magnitude), conforme exemplo apresentado na Figura 4.6. Com base no arquivo resul-

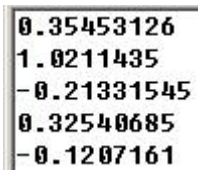

Figura 4.6: Exemplo de um arquivo resultado do SVM^{light}

tado gerado, é aplicado um processo de voto, analisando o grau de confiança de cada uma das comparações para o documento.

Exemplo: Se determinado protocolo, possui 5 documentos de referência, um documento questionado Q terá seus vetores de dissimilaridade gerados com cada um dos 5 documentos de referência gerando 5 vetores. Isto se repete entre mesmo autor, gerando vetores de autoria e autores distintos gerando vetores de não-autoria. A classificação final para o documento Q é então gerada pelo voto desses 10 vetores (5 de autoria e 5 de não-autoria). Com base neste voto é atribuída a autoria como verdadeira ou falsa podendo assim serem calculadas as taxas de erro de falsa rejeição (Erro Tipo I) e falsa aceitação (Erro Tipo II).

4.8.2 Modelo por autor - Multi-Classe

Após os vetores de testes de cada autor gerado, o documento é classificado, tendo como resultado uma classe associada, ou seja, o processo de decisão é gerado pela própria saída do classificador. Como arquivo resultado se gera uma matriz de confusão atribuindo o documento questionado a classe associada.

Comentários Finais

Neste capítulo foram apresentados os métodos propostos para identificação da autoria. As seções foram abordadas genericamente, sem detalhes da base de dados, quantidades, etc., uma vez que mais de um método está sendo proposto com alguns conceitos similares. No seguinte capítulo são apresentados detalhes de cada etapa já especificamente para cada método.

Capítulo 5

Experimentos Realizados e Análise dos Resultados

Este capítulo contém detalhamento de métodos, protocolos e divisão das bases em relação aos experimentos realizados.

5.1 Experimentos

Para testar os métodos neste trabalho foram feitos 2 experimentos e uma comparação com um modelo de identificação de autoria proposto por Coutinho et al. [CMRJB04]. Os experimentos abrangeram os modelos pessoal e global, com dicotomia.

5.1.1 Modelo Global Utilizando SVM

O modelo global possui a vantagem de não ser necessário gerar outro modelo com a inserção de documentos de um novo autor. Através das características geradas o classificador criará os modelos generalizando o processo de aprendizagem.

5.1.1.1 Divisão da Base de Dados

Como já apresentado na seção 4.3, para este experimento (Modelo Global) a divisão da base de dados ocorreu da seguinte forma:

- Modelo de aprendizado - 10 autores foram separados exclusivamente para gerar o modelo de treinamento com 5 documentos de cada.
- Modelo de testes - Outros 20 autores foram separados para testes. Foram utilizados 15 documentos de cada autor para os testes e como referência os 5 documentos de cada autor utilizados para treinamento escolhidos de forma aleatória e sem repetição.

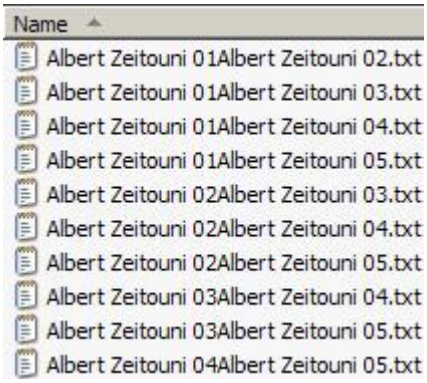

Figura 5.1: Exemplo da combinação para vetores de autoria - Treinamento

Importante comentar que em se tratando de um modelo global, nenhum dos autores participa do conjunto de testes através da divisão da base de dados, ou seja, o classificador busca autenticar autores nunca vistos.

5.1.1.2 Protocolo de Treinamento

Com os 5 documentos[1] separados para treinamento são gerados 2 conjuntos de vetores de dissimilaridade, autoria e não-autoria. O conjunto de vetores de dissimilaridade de autoria é gerado entre documentos de mesmo autor, e a quantidade de vetores é dada pela Equação 4.2. Neste caso o número de vetores será igual a 100, ou seja $A_5^2 = \frac{5!}{(5-2)!}{2}$ multiplicado por 10 autores. A Figura 5.1 apresenta um exemplo da combinação (ver seção 4.6.1) dos vetores de autoria (+1). Como é necessário que o modelo de treinamento seja balanceado, deve-se ter a mesma quantidade de vetores de dissimilaridade de autoria e não-autoria. Sendo assim para gerar os mesmos 100 vetores de não-autoria foram gerados para cada autor 10 vetores de não-autoria. Cada um dos 5 documentos gerou um vetor em comparação com um documento de outro autor (escolhidos tanto o autor quanto o documento aleatoriamente sem repetição, respeitando que os documentos necessitam estar entre os 5 utilizados para treinamento e os autores também exclusivos para treinamento). A Figura 5.2 apresenta um exemplo dos 10 vetores gerados para não-autoria para treinamento. Importante salientar que em se tratando de um modelo global, nenhum desses autores participa do conjunto de testes, ou seja, o classificador busca autenticar autores nunca vistos.

Com os vetores de dissimilaridade de autoria e não-autoria gerados, ambos são concatenados e submetidos ao treinamento (SVM, Subseção 4.6.1.1) gerando um modelo

[1] A utilização de somente 5 documentos dos 15 disponíveis, é importante para que as margens do SVM não estejam especializadas para este conjunto de autores. O que se busca no modelo global é a generalização do modelo.

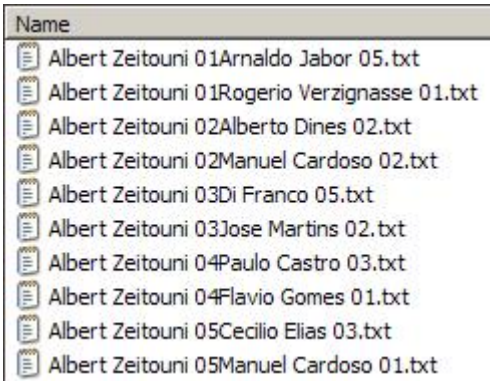

Figura 5.2: Exemplo das combinações para vetores de não-autoria - Treinamento

global.

5.1.1.3 Protocolo de Testes

Igualmente ao protocolo de treinamento no protocolo de testes também são necessários vetores de autoria e não autoria. Isto, como já comentado, para que sejam analisados os erros Tipo I e II (falsa rejeição e falsa aceitação). Para gerar os vetores de autoria são utilizados os 15 documentos separados para testes. Cada um desses 15 documentos gerará um vetor de autoria utilizando 5 amostras de seus próprios documentos como referencia (1500 vetores de autoria). A Figura 5.3 apresenta um exemplo dos vetores gerados pelos pares.

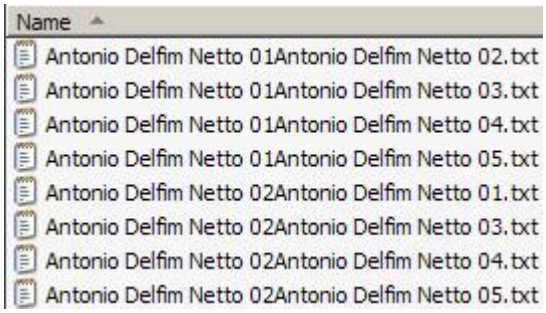

Figura 5.3: Exemplo das combinações para vetores de autoria - Testes

Para gerar os vetores de não-autoria cada um dos 15 documentos de testes utilizará como referência 5 documentos utilizado para treinamento escolhidos aleatoriamente, não repetido. Isto gerará *20 autores * 15 documentos * 5 de referência = 1500 vetores de não autoria*. Na Figura 5.4 está exemplificado a geração de vetores de não-autoria (-1) para etapa de testes. Como mencionado no protocolo, a base de treinamento é exclusiva. Para melhorar a confiabilidade dos resultados o processo de testes foi executados 3 vezes. Uma vez com uma base x de autores no treinamento e outra base yz como teste. Uma segunda

Figura 5.4: Exemplo das combinações para vetores de não-autoria - Testes

execução com a base y como treinamento e a base xz para teste. E uma última execução com a base z como treinamento e a base xy para teste.

Tabela 5.1: Protocolo de Testes - Modelo Global - SVM

Modelo Global - SVM				
Processo	Autores	Documentos	Vetores	
			Autoria	Não-autoria
Treinamento	1-10	1-5	100	100
Referência	1-10	1-5	1500	1500
Testes	11-30	1-15		
Voto Majoritário			1500	1500

5.1.1.4 Resultados

Devido a abordagem dicotômica e ao protocolo utilizado, na geração dos vetores de não-autoria, tanto para treinamento como testes é utilizado um técnica para selecionar aleatoriamente (sem repetições) o autor e o documento a ser utilizado como referência. Isto se deve para que haja balanço entre os vetores, uma vez que as possibilidades no vetores de características de autoria se esgotam e para os vetores de não-autoria sobram possibilidade de combinações. Para que não somente busque o melhor resultado, pela exaustão de execuções, o que impactaria na confiabilidade do modelo, foram verificadas em 15 execuções aleatórias qual seria a "semente[2]" de treinamento que apresentasse o menor erro médio entre falsa rejeição e falsa aceitação (ver Tabela 5.2). Com essa semente são verificadas novamente em 15 execuções (ver Tabela 5.3) e atribuída como resultado final a execução com "erro médio" médio (falsa rejeição e falsa aceitação), em outras palavras, uma "semente de referência" média.

Como resultado do protocolo, ver Tabela 5.1, é apresentado uma falsa rejeição (erro Tipo I) de 17% e uma falsa aceitação (erro Tipo II) de 38% gerando uma taxa de acerto média de 72,50%.

[2]parâmetro inicial utilizado no método de seleção do valor aleatório

Tabela 5.2: Execuções aleatórias para obtenção de uma boa "semente" de treinamento

Execuções SVM - Modelo Global			
Sementes	Falsa Rejeição	Falsa Aceitação	Erro Médio
Seed Train 1972446591 Seed Ref 1583725539	17.00%	38.67%	27.83%
Seed Train 1102014906 Seed Ref 560809877	13.00%	48.00%	30.50%
Seed Train 1849714140 Seed Ref 39233846	10.00%	55.00%	32.50%
Seed Train 1911441376 Seed Ref 173496445	12.33%	56.00%	34.17%
Seed Train 2131915993 Seed Ref 993494028	7.67%	61.00%	34.33%
Seed Train 1943423456 Seed Ref 1320678832	8.00%	63.33%	35.67%
Seed Train 1815875242 Seed Ref 655297178	9.67%	62.33%	36.00%
Seed Train 2071431856 Seed Ref 740729586	3.67%	72.33%	38.00%
Seed Train 512930230 Seed Ref 580310795	8.33%	68.67%	38.50%
Seed Train 1533459754 Seed Ref 1161872706	8.67%	71.00%	39.83%
Seed Train 1561018160 Seed Ref 995975093	5.67%	74.67%	40.17%
Seed Train 697493857 Seed Ref 1612175527	5.00%	77.67%	41.33%
Seed Train 249491504 Seed Ref 1981238045	4.33%	84.00%	44.17%
Seed Train 174675146 Seed Ref 1449259809	5.00%	83.33%	44.17%

Tabela 5.3: Execuções aleatórias para obtenção de uma boa "semente" de referência

Execuções SVM - Modelo Global			
Sementes	Falsa Rejeição	Falsa Aceitação	Erro Médio
Seed Train 1972446591 Seed Ref 1917470155	17.00%	34.67%	25.83%
Seed Train 1972446591 Seed Ref 934035782	17.00%	35.00%	26.00%
Seed Train 1972446591 Seed Ref 553512170	17.00%	36.00%	26.50%
Seed Train 1972446591 Seed Ref 1383049913	17.00%	36.33%	26.67%
Seed Train 1972446591 Seed Ref 1057290149	17.00%	36.67%	26.83%
Seed Train 1972446591 Seed Ref 1055116027	17.00%	37.00%	27.00%
Seed Train 1972446591 Seed Ref 1318205882	17.00%	38.00%	27.50%
Seed Train 1972446591 Seed Ref 1132687845	17.00%	38.00%	27.50%
Seed Train 1972446591 Seed Ref 1203009907	17.00%	38.33%	27.67%
Seed Train 1972446591 Seed Ref 1462030850	17.00%	38.67%	27.83%
Seed Train 1972446591 Seed Ref 4872942	17.00%	38.67%	27.83%
Seed Train 1972446591 Seed Ref 843961650	17.00%	40.67%	28.83%
Seed Train 1972446591 Seed Ref 1385269981	17.00%	41.00%	29.00%
Seed Train 1972446591 Seed Ref 1449105111	17.00%	41.33%	29.17%
Seed Train 1972446591 Seed Ref 2136437162	17.00%	41.67%	29.33%

5.1.2 Modelo por Autor Utilizando SVM Multiclasse

Através deste experimento foi possível verificar como se comportam os mesmos vetores em uma outra técnica de classificação com modelo por autor.

5.1.2.1 Divisão da Base de Dados

Para este experimento a divisão da base de dados ocorreu da seguinte forma:

- Modelo de aprendizado - 20 autores foram separados para gerar o modelo de treinamento com 5 documentos de cada.
- Modelo de testes - Os mesmos 20 autores são testados, porém com os documentos que não foram utilizados no treinamento. Cada um dos 10 documentos dos 20 autores utilizam como referência os 5 documentos utilizados para treinamento escolhidos de forma aleatória e sem repetição.

5.1.2.2 Protocolo de Treinamento

Com os 5 documentos separados para treinamento é gerado somente um conjuntos de vetores: autoria. O conjunto de vetores de dissimilaridade de autoria é gerado entre os mesmos 5 documentos de treinamento (de mesmo autor). A quantidade de vetores é dada pela Equação 4.2. Neste caso o número de vetores será igual a 200, ou seja $A_5^2 = \frac{5!}{(5-2)!}{2}$ multiplicado por 20 autores. A Figura 5.5 apresenta um exemplo da combinação (ver seção 2.10.1.2) dos vetores de autoria. No caso de um modelo multi-classe, cada vetor

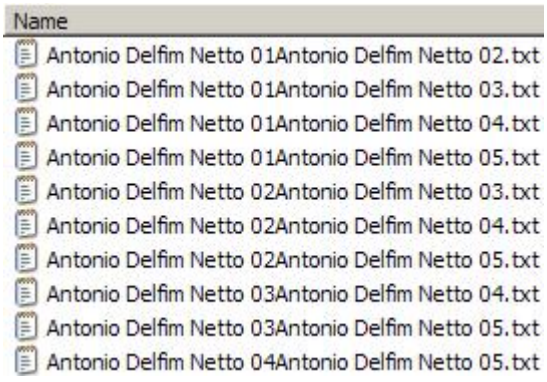

Figura 5.5: Exemplo dos vetores de autoria gerados para um modelo multi-classe

não representa autoria ou não-autoria como no método global aqui apresentado, mas sim a própria classe a qual pertence. Esta representação pode ser vista na Figura 5.6 onde a primeira coluna do arquivo indica a classe a qual pertence o vetor (as demais colunas são

o "id" da característica com seu valor de dissimilaridade. Exemplo 8:1.183069705963). Com os vetores de dissimilaridade de autoria gerados, os mesmos são submetidos ao treinamento ($SVM^{multiclass}$, ver Subseção 2.10.1.2) gerando um modelo único porém por autor.

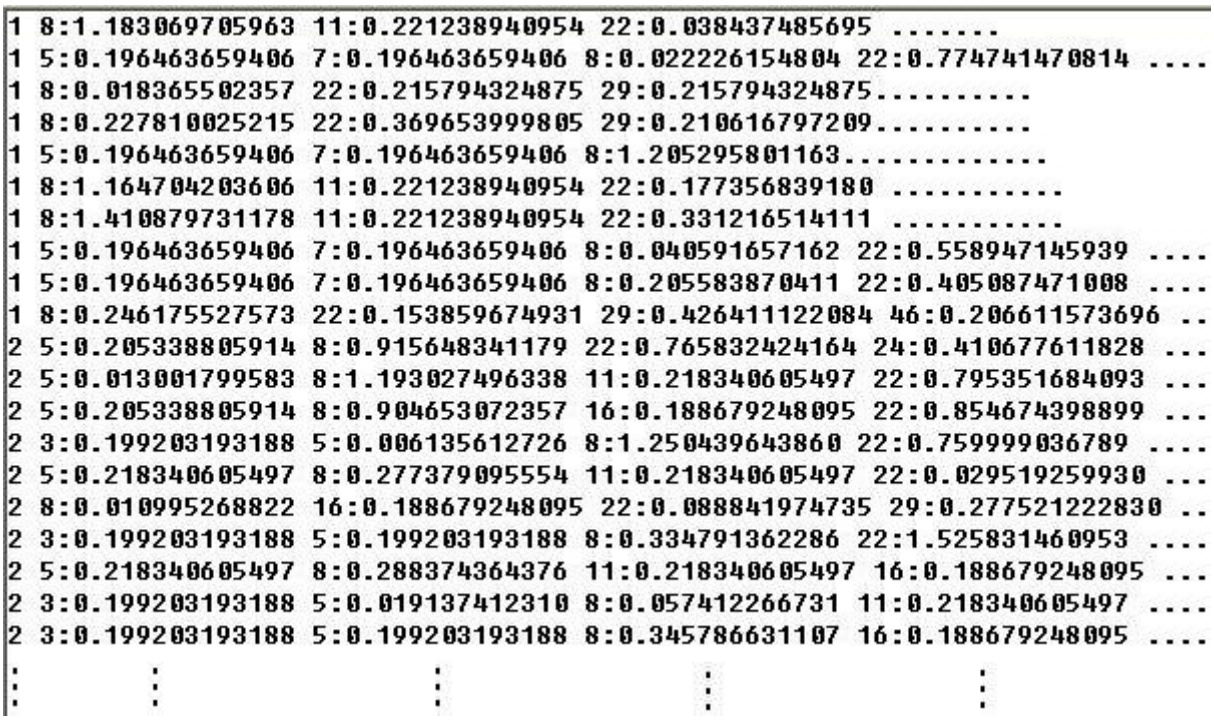

Figura 5.6: Exemplo dos vetores de autoria gerados para um autor para treinamento

5.1.2.3 Protocolo de Testes

Para teste no modelo multi-classe, cada um dos 10 documentos dos 20 autores não utilizados no treinamento geram 5 vetores, produto da comparação com 5 documentos do mesmo autor separado para treinamento, que neste caso são utilizados como referência. Ou seja, no caso de 20 autores serão 1000 vetores, sendo *20 autores * 10 documentos * 5 de referencia = 1000 vetores que serão testados.* A Figura 5.7 apresenta o fluxo de testes obtendo dados dos vetores de autoria gerados pelos pares.

Através da utilização de dicotomia, utilizando 5 documentos como referência, é necessário um voto entre as 5 respostas obtidas para o mesmo documento de teste. Neste trabalho foi utilizado o voto majoritário simples, ou seja, se como resultado dos cinco vetores referentes documento x se obteve atribuição as classes: 10,5,1,10,10 esse documento é atribuído a classe de autoria 10.

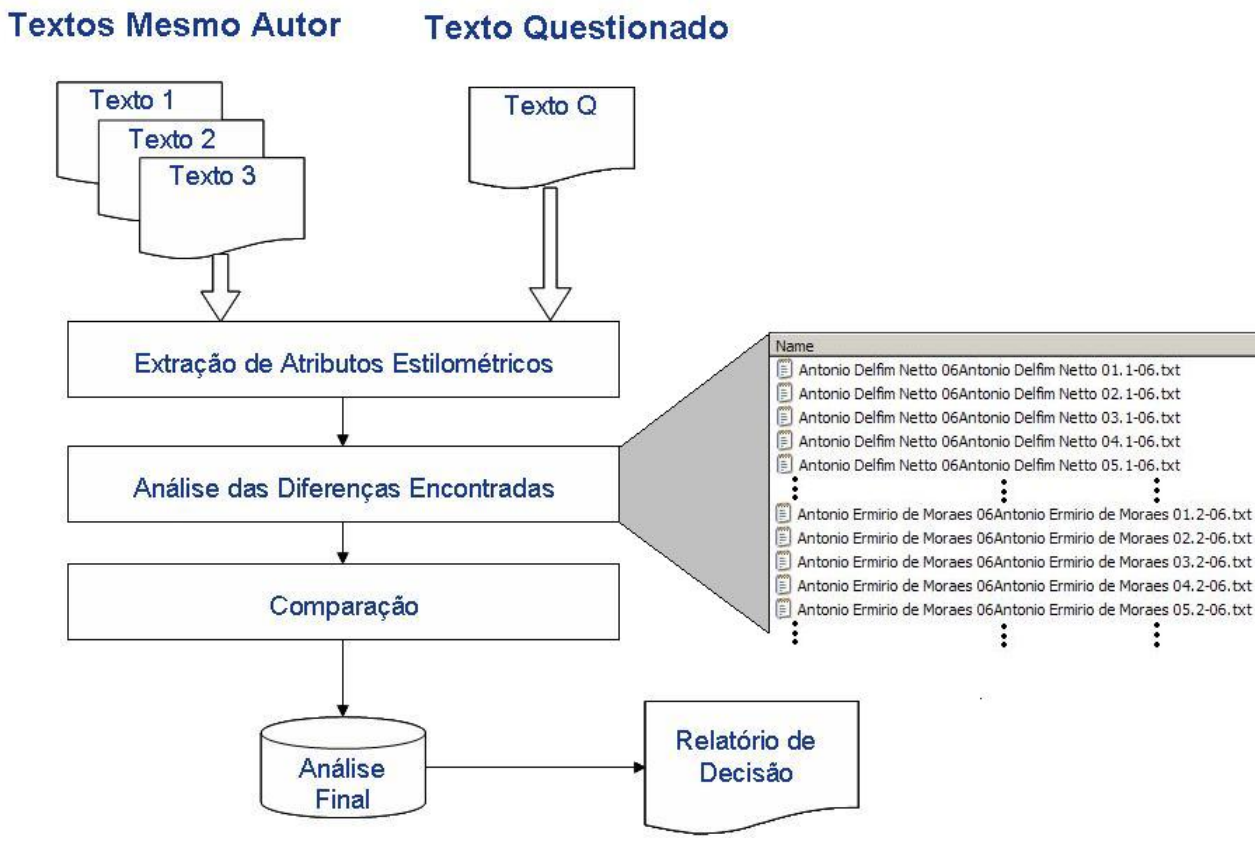

Figura 5.7: Exemplo Fluxo Testes - Modelo Multi-classe

A seguir estão listados os protocolos de testes utilizados para validação deste modelo, nas Tabelas 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11, 5.12.

Tabela 5.4: Protocolo Base Autores 1 a 10 - Documentos para treinamento 1 a 5

Modelo por Autor			
Processo	Autores	Documentos	Vetores Autoria
Treinamento	1-10	1-5	
Referência	1-10	1-5	500
Testes	1-10	6-15	
Voto Majoritário			500

Tabela 5.5: Protocolo Base Autores 1 a 10 - Documentos para treinamento 6 a 10

Modelo por Autor			
Processo	Autores	Documentos	Vetores Autoria
Treinamento	1-10	6-10	
Referência	1-10	6-10	500
Testes	1-10	1-5 e 11-15	
Voto Majoritário			500

Tabela 5.6: Protocolo Base Autores 1 a 10 - Documentos para treinamento 11 a 15

Modelo por Autor			
Processo	Autores	Documentos	Vetores Autoria
Treinamento	1-10	11-15	
Referência	1-10	11-15	500
Testes	1-10	1-10	
Voto Majoritário			500

Tabela 5.7: Protocolo Base Autores 11 a 20 - Documentos para treinamento 1 a 5

Modelo por Autor			
Processo	Autores	Documentos	Vetores Autoria
Treinamento	11-20	1-5	
Referência	11-20	1-5	500
Testes	11-20	6-15	
Voto Majoritário			500

Tabela 5.8: Protocolo Base Autores 11 a 20 - Documentos para treinamento 6 a 10

Modelo por Autor			
Processo	Autores	Documentos	Vetores Autoria
Treinamento	11-20	6-10	
Referência	11-20	6-10	500
Testes	11-20	1-5 e 11-15	
Voto Majoritário			500

Tabela 5.9: Protocolo Base Autores 11 a 20 - Documentos para treinamento 11 a 15

Modelo por Autor			
Processo	Autores	Documentos	Vetores Autoria
Treinamento	11-20	11-15	
Referência	11-20	11-15	500
Testes	11-20	1-10	
Voto Majoritário			500

Tabela 5.10: Protocolo Base Autores 1 a 20 - Documentos para treinamento 1 a 5

Modelo por Autor			
Processo	Autores	Documentos	Vetores Autoria
Treinamento	1-20	1-5	
Referência	1-20	1-5	1000
Testes	1-20	6-15	
Voto Majoritário			1000

Tabela 5.11: Protocolo Base Autores 1 a 20 - Documentos para treinamento 6 a 10

Modelo por Autor			
Processo	Autores	Documentos	Vetores Autoria
Treinamento	1-20	6-10	
Referência	1-20	6-10	1000
Testes	1-20	1-5 e 11-15	
Voto Majoritário			1000

Tabela 5.12: Protocolo Base Autores 1 a 20 - Documentos para treinamento 11 a 15

Modelo por Autor			
Processo	Autores	Documentos	Vetores Autoria
Treinamento	1-20	11-15	
Referência	1-20	11-15	1000
Testes	1-20	1-10	
Voto Majoritário			1000

5.1.2.4 Resultados

Os resultados de acordo com os protocolos de teste estão demonstrado na Tabela 5.13:

Tabela 5.13: Resultados - Modelo por Autor - SVM Multi-classe

Resultados - Modelo por Autor - SVM Multi-classe	
Protocolo	Taxa de Acerto
Base Autores 1 a 10 - Documentos para treinamento 1 a 5	80%
Base Autores 1 a 10 - Documentos para treinamento 6 a 10	80%
Base Autores 1 a 10 - Documentos para treinamento 11 a 15	72%
Base Autores 11 a 20 - Documentos para treinamento 1 a 5	87%
Base Autores 11 a 20 - Documentos para treinamento 6 a 10	88%
Base Autores 11 a 20 - Documentos para treinamento 11 a 15	91%
Base Autores 1 a 20 - Documentos para treinamento 1 a 5	83%
Base Autores 1 a 20 - Documentos para treinamento 6 a 10	83%
Base Autores 1 a 20 - Documentos para treinamento 11 a 15	85%

5.1.3 Modelo por Autor Utilizando PPM

Utilizando as mesmas bases de dados e o mesmo protocolo do modelo por autor multi-classe, foram executados testes através da técnica de compactação de dados PPM-C, proposta por Coutinho et al. [CMRJB04].

5.1.3.1 Divisão da Base de Dados

Para este experimentos foram feitos testes com 2 bases de dados de 10 autores cada e depois com as duas concatenadas, ou seja uma base de 20 autores. A base de dados para este método foi utilizada da seguinte forma:

- Modelo de aprendizado - variando em número de autores de acordo com o protocolo, porém o número de documentos se manteve em 5 para que se mantenha um bom número de documentos para testes.
- Modelo de testes - mesmos autores de acordo com o protocolo porém com os 10 documentos não utilizados para treinamento

5.1.3.2 Pré-processamento

Alguns pré-processamentos são aplicados nos textos antes da criação dos modelos: eliminação de acentos, pontuação e números, conversão de maiúsculas em minúsculas e truncar as colunas para um tamanho comum. Este último é importante para o processo, pois modelos criados sobre textos maiores aumentam suas chances de comprimir melhor qualquer outro texto. Assim, autores com conjuntos de treinamento grandes tenderiam a "roubar" textos de autores com conjuntos de treinamento pequenos.

5.1.3.3 Protocolo de Treinamento

O processo de criação do modelo através do método de compressão consiste em agrupar os documentos selecionados para o treinamento de acordo com o protocolo, do mesmo autor, obtendo os modelos. É gerado um modelo para cada autor.

5.1.3.4 Classificação

Para classificação os documentos de testes serão comprimidos com cada modelo obtendo uma razão de compressão. O documento então é atribuído ao modelo que obtiver a maior razão de compressão, que é o tamanho do texto plano dividido pelo mesmo comprimido, para aquele documento.

5.1.3.5 Protocolo de Testes

Excluindo a parte da dicotomia também utilizada no processo multi-classe com SVM , os protocolos de testes para este método foram os mesmos utilizados no modelo multi-classe, para fins de comparação de acordo com as Tabelas 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11, 5.12.

5.1.3.6 Resultados

Os resultados de acordo com os protocolos de teste estão demonstrado na Tabela 5.14:

Tabela 5.14: Resultados - Modelo por Autor - PPM-C

Resultados - Modelo por Autor - PPM-C	
Protocolo	Taxa de Acerto
Base Autores 1 a 10 - Documentos para treinamento 1 a 5	77%
Base Autores 1 a 10 - Documentos para treinamento 6 a 10	80%
Base Autores 1 a 10 - Documentos para treinamento 11 a 15	79%
Base Autores 11 a 20 - Documentos para treinamento 1 a 5	89%
Base Autores 11 a 20 - Documentos para treinamento 6 a 10	91%
Base Autores 11 a 20 - Documentos para treinamento 11 a 15	93%
Base Autores 1 a 20 - Documentos para treinamento 1 a 5	84%
Base Autores 1 a 20 - Documentos para treinamento 6 a 10	83%
Base Autores 1 a 20 - Documentos para treinamento 11 a 15	86%

5.2 Comparativo entre os Experimentos

Analisar os resultados dos experimentos aqui apresentados envolve primeiramente dois fatores importantes: características e abordagens. Analisando primeiramente o experimento com o modelo global dicotômico, apresentando taxa de acerto médio de 72,50%, verifica-se que obteve uma taxa de acerto médio inferior ao dos outros dois métodos (taxa de acerto médio de 83,22% no modelo por autor multi-classe e 84,66% no modelo por autor com compressão PPM-C). Observa-se que as execuções do modelo global possuem uma alta variação dependendo das combinações formadas pelos vetores de não-autoria. Isto significa que testes com novos autores podem variar muito dependendo do autor ou documento utilizado para formar o dado escalar a ser classificado junto ao modelo pré-gerado. Devido a sua abordagem, o modelo global não pode ser comparado diretamente com os experimentos de modelos pessoais. O modelo global tem suas particularidades, como não necessitar da geração de um novo modelo a cada novo autor da base e conseguir trabalhar com um número reduzido de amostras obtendo o conhecimento sobre toda a população (o que torna os resultados mais confiáveis).

Passíveis de comparação, estão os modelos por autor através do método de compactação PPM-C e o modelo por autor multi-classe através do $SVM^{multiclass}$ baseado em características sintáticas (palavras) utilizando como características estilométricas a freqüencia de 171 palavras-função da língua portuguesa. Os dois métodos utilizaram-se da mesma base de dados e dos mesmos protocolos de testes apresentando os resultados demonstrados na Tabela 5.15. Os resultados apresentados demonstram alguns pontos importantes:

- A discriminatoriedade das palavras-função da língua portuguesa, (um dos objetivo deste trabalho);

- Importância das palavras-função com conceito de análise de trigramas (até 3 palavras consecutivas).
- Simplicidade e robustez do método PPM-C;
- Utilização da dicotomia com um modelo pessoal apresentou bons resultados.
- Resultados promissores através da nova ferramenta $SVM^{multiclass}$;

Tabela 5.15: Taxa de Acerto - Comparativo PPM-C e SVM Multi-classe

Taxa de Acerto - Modelo por Autor - PPM-C e SVM Multi-classe		
Protocolo	PPM-C	SVM
Base Autores 1 a 10 - Documentos para treinamento 1 a 5	77%	80%
Base Autores 1 a 10 - Documentos para treinamento 6 a 10	80%	80%
Base Autores 1 a 10 - Documentos para treinamento 11 a 15	79%	72%
Base Autores 11 a 20 - Documentos para treinamento 1 a 5	89%	87%
Base Autores 11 a 20 - Documentos para treinamento 6 a 10	91%	88%
Base Autores 11 a 20 - Documentos para treinamento 11 a 15	93%	91%
Base Autores 1 a 20 - Documentos para treinamento 1 a 5	84%	83%
Base Autores 1 a 20 - Documentos para treinamento 6 a 10	83%	83%
Base Autores 1 a 20 - Documentos para treinamento 11 a 15	86%	85%
Taxa de Acerto Médio	84,66%	83,22%

Analisando os resultados, a principal pergunta que claramente se identifica é por que os resultados do modelo por autor utilizando compressão PPM-C e o modelo por autor multi-classe foram tão semelhantes ? Dois pontos são observados: (1) A similaridade das características (mesmo que oculta) utilizadas: analisando o funcionamento do modelo de compressão PPM-C (subseção 3.2.2.1) verifica-se que ambos métodos trabalham com características sintáticas (palavras e caracteres) e com a freqüencia das mesmas; verifica-se que palavras-função (como as utilizadas no método multi-classe) analisando sua freqüencia em relação a outras palavras do textos, possuem o maior número de repetições, ou seja, se existem muitas ocorrências de palavras-função isso será benéfico também ao método PPM-C pois terá uma razão de compressão mais ajustada (strings iguais); (2) dicotomia e processo de voto majoritário corrigem as diferenças intrapessoais que possam existir, acertando a decisão final, mesmo em textos nos quais as palavras-função visivelmente não demonstravam discriminatoriedade.

Uma diferença importante entre os modelos por autor apresentados nesta pesquisa está relacionada a produção do modelo e treinamento. No método de compressão PPM-C geram-se n modelos, onde n é igual ao número de autores. O documento questionado é testado separadamente com cada modelo gerando uma razão de compressão. A autoria

do documento questionado é atribuída ao "dono" do modelo que gerou a maior razão de compressão. Já no modelo multi-classe, é gerado um único modelo com vetores de autoria de todas as classes, ou seja, o modelo treinado conhece o universo das características dos n autores, dificultando uma atribuição correta caso autores tenham uma pequena variabilidade interpessoal, porém apresentando uma maior confiabilidade nos resultados. No modelo global o universo contemplado pelo modelo é ainda maior, ou seja, além de basear-se nas características de autoria (de mesmo autor) utiliza-se características entre autores distintos (não-autoria) para geração do modelo, ou seja, adaptando o problema para o modelo por autor (em uma visão simplista), seria como criar autores (que não existem) baseados nos vetores de dissimilaridade de autores distintos e aplicar os testes em todos os modelos gerados. Em outras palavras, não só a taxa de acerto de um método deve ser considerada, mas também a confiabilidade da mesma.

Conclusão

O trabalho proposto apresenta duas abordagens para atribuição da autoria de documentos baseados na estilometria utilizando características da língua portuguesa. Como característica foram utilizadas freqüências de palavras-função ("palavras gramaticais"), ou seja, 171 palavras em ($n-gramas$, n=1..3) dentro das classes de conjunções e advérbios. Com as características extraídas utilizou-se de dicotomia para gerar vetores de autoria e não-autoria através de vetores de dissimilaridade. As abordagens apresentadas se diferenciam, uma com abordagem global (modelo global), e outra com abordagem pessoal (modelo por autor). Dentro da proposta deste trabalho, algumas importantes conclusões podem ser observadas, analisando os modelos, métodos, experimentos e comparativos aqui apresentados:

- Não é necessário um grande número de amostras para que se apresentem resultados promissores;
- Análise de características exclusivas da língua portuguesa podem, juntamente com outros métodos multi-idioma (Exemplo: características estruturais, etc.), melhorar resultados;
- Analisando que o método de compactação PPM-C faz uso de todo o texto e o método proposto utiliza somente da freqüência de 171 palavras-função, com base nos resultados de ambos, pode-se considerar como promissora a utilização das palavras-função aqui apresentadas como característica discrimante em textos de língua portuguesa.
- Comportamento dos métodos estável com número de amostras e tamanhos dos textos reduzidos;

O propósito deste trabalho não foi a comparação entre métodos já existentes, e sim, identificar métodos robustos juntamente com características estilométricas discriminantes da língua portuguesa brasileira para identificação da autoria e futura aplicação no contexto jurídico. Como trabalhos futuros encontram-se:

- Inclusão de novas características estilométricas da língua portuguesa;
- Criação de um software amigável para verificação de autoria com escolha de características da língua portuguesa;

- Criação de um corpus de palavras da língua portuguesa para análise automática de autoria baseada na estilometria.

Referências Bibliográficas

[AC05] A. Abbasi and H. Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005.

[AC06] Ahmed Abbasi and Hsinchun Chen. Visualizing authorship for identification. *Springer-Verlag Berlin Heidelberg - S. Mehrotra et al. (Eds.)*, ISI 2006, LNCS 3975:60–71, 2006. Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721, USA.

[AL05] S. Argamon and S. Levitan. Measuring the usefulness of function words for authorship attribution. *Association for Literary and Linguistic Computing*, 2005. University Of Victoria, Canada.

[BA04] R. Bekkerman and J. Allan. Using bigrams in text categorization. *CIIR Technical Report IR-408 Center for Intelligent Information Retrieval*, 2004.

[Bar05] Francis L. Baranoski. Identificação da autoria em documentos manuscritos usando svm. Master's thesis, Pontifícia Universidade Católica do Paraná, 2005. XXV Congresso da Sociedade Brasileira de Computação.

[Bea96] R. H. Baayen and et al. Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, (11(3)):121–131, 1996.

[Bea02] D. Benedetto and et al. Language trees and zipping. *Physical Review Letters*, 88(4):048702–1–4, 28 January 2002.

[BL03] M. Bulacu and Shomaker L. Writer identification using edge-based directional features. *IEEE - Computer Society*, 2:937–941, August 2003. Proc. of 7th Int. Conf. on Document Analysis and Recognition (ICDAR 2003).

[BNNH90] H. C. Black, J.R. Nolan, and J.M. Nolan-Haley. *Black's Law Dictionary*. West Publishing, 6 edition, 1990. St. Paul.

[Bri63] C. S. Brinegar. Mark twain and the quintus curtius snodgrass letters: A statistical test of authorship. *Journal of the American Statistical Association*, (58):85–96, 1963.

[Bue67] Francisco da Silveira Bueno Bueno. *Formação Histórica da Língua Portuguesa*. Editora Saraiva, São Paulo-SP, 3ª edition, 1967.

[Bur92] J. F. Burrows. Computers and the study of literature. *Computers and Written Text*, (Part 4):167–204, 1992. In C. Butler, editor, Applied Language Studies - Blackwell, Oxford.

[Cal99] Lélio Braga Calhau. O direito à prova as provas ilícitas e as novas tecnologias. *Jus Navigandi*, 4(36), Novembro 1999. Disponível em: ¡http://jus2.uol.com.br/doutrina/texto.asp?id=818¿. Acesso em: 23 mar. 2006.

[Cha86] W. L. Chafe. Writing in the perspective of speaking. 12, 1986. in Cooper, C.R. and Greenbaum, S.,Eds.,Studying Writing: Linguistic Approaches, Sage, Beverly Hills.

[Cha97] Carole E. Chaski. Who wrote it ? - steps toward a science of autorship identification. *National Institute of Justice - Journal*, (Issue No. 233):15–22, September 1997.

[Cha01a] Sung Hyuk Cha. *Use of the Distance Measures in Handwriting Analysis*. PhD thesis, University of New York at Buffalo, EUA, 2001. Doctor Thesis.

[Cha01b] Carole E. Chaski. Empirical evaluations of language-based author identification techniques. *The International Journal of Speech, Language and Law: Forensic Liguistics*, 8(1), 2001.

[Cha05] Carole E. Chaski. Who's at the keyboard? - autorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1), 2005. Spring 2005.

[CMRJB04] B. C. Coutinho, J. L. M. Macêdo, A. Rique Júnior, and L. V. Batista. Atribuição de autoria usando ppm. *XXV Congresso da Sociedade Brasileira de Computação - Unisinos - São Leopoldo/RS*, Julho 2004.

[Cor03] Malcolm Walter Corney. Analysing e-mail text authorship for forensic purposes, 2003. Thesis - Master of Information Technology.

[Cra98] C. Crain. The bard's fingerprints. *Lingua Franca*, (4):29–39, 1998.

[Cre95] Jean Pierre Cretez. A set of handwriting families: style recognition. *IEEE Computer Society Press.*, pages 489–494, August 1995. ProcIn Proc. of the 3th. International Conf. on Document Analysis and Recognition.

[CS01] K. Crammer and Y. Singer. On the algorithmic implementation of multi-class svms. *JMLR*, 2001.

[Dav90] L. M. Davis. Statistics in dialectology. *University of Alabama Press*, 1990. Tuscaloosa.

[DKLP03] J. Diederich, J. Kindermann, E. Leopold, and G. Paass. Authorship attribution with support vector machines. *Applied Intelligence*, (1), 2003.

[DOL81] G. Del Olmo Lete. *Mitos y Leyendas de Canaan según la Tradición de Ugarit*. Institución San Jerónimo & Ediciones Cristiandad, Madrid, 1981.

[dPdLP06] Comunidade dos Países de Língua Portuguesa. http://www.cplp.org, 2006. Data de acesso: 12/02/2006.

[EV91] W. E. Y. Elliott and R. J. Valenza. A touchstone for the bard. *Computers and the Humanities*, (25(4)):199–209, 1991.

[EV02] W. E. Y. Elliott and R. J. Valenza. So many hardballs, so few over the plate. *Computers and the Humanities*, (36(4)):455–460, 2002.

[Fab06] Nicholas Fabian. História do alfabeto e conecção com a escrita cuneiforme de ugarit. http://www.kfssystem.com.br/loubnan/fenicio.html, 2006. Data de Acesso: 01/02/2006.

[Fer04] Aurélio Buarque de Holanda Ferreira. *Novo Dicionário Eletrônico Aurélio versão 5.0*. Editora Positivo, 3 edition, 2004. revista e atualizada do Aurélio Século XXI, O Dicionário da Língua Portuguesa, contendo 435 mil verbetes, locuções e definições.

[Fos96a] D. Foster. A funeral elegy: W[illiam] s[hakespeare]'s "best-speaking witnesses". *Publications of the Modern Language Association of America*, (111(5)):1080, 1996.

[Fos96b] D. Foster. Primary culprit: An analysis of a novel of politics - who is anonymous? *New York, 26 February*, 1996.

[Fos99] D. Foster. The claremont shakespeare authorship clininc: How severe are the problems? *Computers and the Humanities*, (32(6)):491–510, 1999.

[Fos00] D. Foster. Author unknown: On the trail of anonymous. *Henry Holt and Company*, 2000. New York, NY.

[Fra06] Robert Fradkin. Course: History of the alphabets. http://www.wam.umd.edu/~rfradkin/alphapage.html, 2006. University of Maryland, Data de Acesso: 15/02/2006.

[GC98] A. Garton and Pratt C. *Learning to Be Literate: the Development of Spoken and Written Language*. Blackwell, Oxford, 2ª edition, 1998.

[GF03] Vicente Greco Filho. *Direito Processual Civil Brasileiro*, volume 2. Saraiva, São Paulo, 2003.

[GMC07] A. Garcia, M. Miranda, and J. Calle. Function words in authorship attribution studies. *Literary and Linguistic Computing*, 22(1):49–66, 2007.

[Gro06] The Handwriting Analysts Group. http://www.handwriting.org, 2006. Data de Acesso: 09/02/2006.

[Hea01] D. I. Holmes and et al. Stephen crane and the new-york tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, (35(3)):315–331, 2001.

[HF95] D. I. Holmes and R. Forsyth. The federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, (10(2)):111–127, 1995.

[HFKvdH99] J. Hoorn, S. Frank, W. Kowalczyk, and F. van der Ham. Neural network identification of poets using letter sequences. *Literary and Linguistic Computing*, (14(3)):311–338, 1999.

[Hän99] H. Hänlein. Studies in authorship recognition - a corpus-based approach. 1999. Peter Lang - Frankfurt.

[Hol85] D. I. Holmes. The analysis of literary style — a review. *J. R. Statist. Soc. A. 148*, (Part 4):328–341, 1985.

[Hol95] D. Holmes. Authorship attribution. *Computers and the Humanities*, (28):87–106, 1995. Kluwer Academic Publishers.

[Hol98] D. I Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, (13(3)):111–117, 1998.

[Inf01] Ulisses Infante. *Curso de Gramática Aplicada aos Textos*. Editora Scipione, São Paulo-SP, 6ª edition, 2001.

[JOA02] T. JOACHIMS. Optimizing search engines using clickthrough data. *ACM Conference on Knowledge Discovery and Mining (KDD)*, pages 1–10p, 2002.

[Joh00] B. Johnstone. *Qualitative Methods in Socioliguistics*. Oxford University Press, New York, 2000.

[Jur05] Revista Consultor Jurídico. O dna da ação - arquivo de petição de luiz francisco foi gerado em empresa. http://conjur.estadao.com.br/static/text/28345,1, 2005. Notícia Publicada em 05/09/05 - Data de Acesso: 07/03/2006.

[Jus02] Edson J. Justino. *Análise de Documentos Questionados*. PhD thesis, Pontifícia Universidade Católica do Paraná, 2002. Tese de Doutorado.

[Kje94a] B. Kjell. Authorship attribution of text samples using neural networks and Bayesian classifiers, 1994. San Antonio - TX Man and Cybernetics.

[Kje94b] B. Kjell. Authorship determination using letter pair frequencies with neural network classifiers. *Literary and Linguistic Computing*, (9(2)):119–124, 1994.

[KSW04] A. Kaster, S. Siersdorfer, and G. Weikum. Combining text and linguistic document representations for authorship attribution. *Max-PlanckInstitute for Computer Science*, 2004. Germany.

[KWF95] B. Kjell, W. A. Woods, and O. Frieder. Information retrieval using letter tuples with neural network and nearest neighbor classifiers. *In IEEE International Conference on Systems, Man and Cybernetics*, (2):1222–1225, 1995. Vancouver - BC.

[LM95] D. Lowe and R. Matthews. Shakespeare vs. fletcher: A stylometric analysis by radial basis functions. *Computers and the Humanities*, (29):449–461, 1995.

[Mal04] M. B. Malyutov. Authorship attribution of texts: a review. *Proceedings of the program "Information transfer"held in ZIF*, page 17 pages, 2004. University of Bielefeld, Germany.

[McM02] Gerald R. McMenamin. *Forensic Linguistics - Advances in Forensic Stylistics*. CRC Press, Florida-USA, 1ª edition, 2002.

[Mig04] Luis Felipe Miguel. *Mídia e vínculo eleitoral: a literatura internacional e o caso brasileiro.*, volume 10. Opinião Pública, 2004.

[MM94] T. Merrian and R. Matthews. Neural compuation in stylometry ii: An application to the works of shakespeare and marlowe. *Literary and Linguistic Computing*, (9):1–6, 1994.

[Mon02] G. D. Monsarrat. A funeral elegy: Ford, w.s., and shakespeare. *Review of English Studies - Oxford University Press*, 18(210):186–203, 2002.

[Mor88] José Carlos Barbosa Moreira. *Temas de Direito Processual*. Saraiva, São Paulo-SP, 5ª edition, 1988.

[Mor00] Ron N. Morris. *Forensic Handwriting Identification Fundamental Concepts and Principles*. Academic Press, London-UK, 2000.

[MPMyGR05] R. M. C. Morales, L. V. Pineda, M. Montes-y Gómez, and P. Rosso. Authorship attribution using word sequences. *Laboratorio de Tecnologías del Lenguaje*, 2005. Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, España.

[MW64] F. Mosteller and D. L. Wallace. Inference and disputed authorship: The federalist. *Addison-Wesley, Reading, Massachusetts*, 1964.

[Ols97] D. R. Olson. On the relation between speech and writing. 4, 1997. in PonteCorvo, C., Ed., Writing Development: an Interdisciplinary View, John Benjamins, Amsterdan.

[Ols04] John Olsson. *Forensic Linguistics - An Introducion to Language, Crime and Law*. Continuum, New York-NY, 1ª edition, 2004.

[PD02] E. Pekalska and R. P. W. Duin. Dissimilarity representations allow for building good classifiers. *Pattern Recognition*, (23):943–956, 2002.

[Pin86] Edith Pimentel Pinto. *A Língua Escrita no Brasil*. Editora Ática, São Paulo-SP, 1ª edition, 1986.

[Rud98] J. Rudman. The state of authorship attribution studies: Some problems and solutions. *Computers and Humanities*, (31):351–365, 1998. Germany.

[SFK01] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and Humanities*, (35):193–214, 2001. Kluwer Academic Publishers.

[SG06] C. Sanderson and S. Guenter. Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. *Proceedings of International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 482–491, 2006.

[Sil91] César Antônio da Silva. *Ônus e Qualidade da Prova Cível*. Aide, Rio de Janeiro-RJ, 1991.

[Sil01a] Maria Cecília Perez de Souza e Silva. *Linguística Aplicada ao Portugues: Morfologia*. Editora Cortez, São Paulo-SP, 12ª edition, 2001.

[Sil01b] Ovídio A. Baptista da Silva. Curso de processo civil. *Revista dos Tribunais*, 2001. São Paulo-SP.

[SJBS04] C. R. Santos, E. J. R. Justino, F. Bortolozzi, and R. Sabourin. An off-line signature verification method based on the questioned document expert's approach and a neural network classifier. *The Ninth International Workshop on Frontiers in Handwriting Recognition*, pages 10–14p, 2004. Tokyo.

[SK01] Maria Cecília Perez de Souza e Silva and Ingedore Villaça Koch. *Linguística Aplicada ao Portugues: Sintaxe*. Editora Cortez, São Paulo-SP, 10ª edition, 2001.

[SK02] J. A. Smith and C. Kelly. Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works. *Computers and the Humanities*, (36):411–430, 2002.

[SNSSHCHASL02] Ph.D. Sargur N. Srihari, Ph.D. Sung-Hyuk Cha, M.E. Hina Arora, and M.S. Sangjik Lee. Individuality of handwriting. *Journal Forensic Science*, 47(2), 2002. Available online at: www.astm.org Paper ID JFS2001227_474.

[Soa06] Elvira Lobato Pedro Soares Soares. Juíza se afasta de casos com o opportunity. http://www1.folha.uol.com.br/folha/dinheiro/ult91u105686.shtml, 2006. Folha Online - Notícia Publicada em 04/03/06 - 09:45hs - Data de Acesso: 07/03/2006.

[Sär67] C. E. Särndal. On deciding cases of disputed authorship. *Applied Statistics*, (16):251–268, 1967.

[SSFC04] De Plácido e Silva, Nagib Slaibi Filho, and Gláucia Carvalho. *Vocabulário Jurídico*. Forense, Rio de Janeiro, 24 edition, 2004.

[TE87] R. Thisted and B. Efron. Did shakespeare write a newly-discovered poem? *Biometrika*, (74(3)):445–455, 1987.

[Tea00] W. J. Teahan. Proceeding of RIAO'00, 6th International Conference "Recherche d'Information Assistee par Ordinateur", 2000. Paris, France.

[Ter06] A. Teru. Authorship attribution by data compression program. *Libr Inf Sci*, 2006.

[Tha01] N. Thaper. Using compression for source based classification of text. *MIT*, 2001.

[THJA04] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector learning for interdependent and structured output spaces. *ICML*, 2004.

[TSH96] F. J. Tweedie, S. Singh, and D. I. Holmes. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities*, (30(1)):1–10, 1996.

[TWL02] C. M. Tan, Y. F. Wang, and C. D Lee. The use of bigrams to enhance text categorization. *Information Processing and Management*, 30(4):529–546, 2002.

[Vap98] V. Vapnik. Statistical learning theory. *Wiley, N. Y.*, page pp. 768, 1998.

[vHBT+05] H. van Halteren, R. H. Baayen, F. Tweedie, M. Haverkort, and A. Neijt. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 2005.

[WAT00] S. Waugh, A. Adams, and F. J. Tweedie. Computational stylistics using artificial neural networks. *Literary and Linguistic Computing*, (15(2)):187–198, 2000.

[Wil40] C. B. Williams. A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika*, 3/4(31):356–361, 1940.

[Wil75] C. B. Williams. Mendenhall's studies of word-length distribution in the works of shakespeare and bacon. *Biometrika*, 1(62):207–212, 1975.

[WN88] N. Wolf-Nelson. The nature of literacy. pages 11–28, 1988. in Nippold, M.A., Ed., Later Language Development: Ages Nine through Nineteen, College-Hill, Boston.

[Yul38] G. U. Yule. On sentence-length as a statistical characteristic of style in prose, with applications to two cases of disputed authorship. *Biometrika*, (30):363–390, 1938.

[Zip75] G. K. Zipf. Selected studies of the principle of relative frequency in language. *Harvard University Press*, 1975. Cambridge, MA.

[ZQHC06] R. Zheng, Y. Qin, Z. Huang, and H. Chen. A framework for authorship analysis of online messages: Writing-style features and techniques. *Journal of the American Society for Information Science and Technology*, (57(3)):378–393, 2006.

[ZZ05] Y. Zhao and J. Zobel. Effective and scalable authorship attribution using function words. *Lecture Notes in Computer Science*, (3689):174–189, 2005. Springer Verlag.

Apêndice A

Troca de E-mails sobre a autoria da “Uma Elegia Fúnebre”

SHAKSPER 2002: Abrams and Foster on ”A Funeral Elegy”

From: Hardy M. Cook (editor@shaksper.net)

Date: 06/13/02

Next message: Hardy M. Cook: ”Call for Proposals: 2004 Meeting of the Shakespeare”Previous message: Hardy M. Cook: ”Adaptations and Pop Shakespeare”Messages sorted by: [date] [thread] [subject] ———————————————————————— The Shakespeare Conference: SHK 13.1514 Thursday, 13 June 2002

[1] From: Rick Abrams ¡rabrams1@maine.rr.com¿ Date: Wednesday, 12 Jun 2002 15:54:47 -0400 Subj: WS’s Elegy

[2] From: Don Foster ¡foster@VASSAR.EDU¿ Date: Wednesday, 12 Jun 2002 16:20:55 -0400 Subj: WS’s Elegy [1]————————————————————————-

From: Rick Abrams ¡rabrams1@maine.rr.com¿

Date: Wednesday, 12 Jun 2002 15:54:47 -0400

Subject: WS’s Elegy

As a scholar who urged the attribution to Shakespeare of W.S.’s Funeral Elegy for William Peter, I wish to concede the force of the philological case for John Ford presented by Gilles Monsarrat in the latest issue of the *Review of English Studies*. Though I was aware of the high lexical correlation of Ford’s texts with the Elegy, until reading Monsarrat I underestimated the importance of the overlaps. Professor Monsarrat demonstrates that the overlaps are too pervasive to be waved away as ”influence.”I am now satisfied that the linguistic evidence for Ford is stronger than for Shakespeare. Cambridge University Press has recently announced the publication of Brian Vickers’s 628-page book on the Elegy, which argues Ford’s authorship on the basis of ”linguistic and statistical”evidence, according to the press advertisement, and Monsarrat hails Vickers’s book as ”definitive

and comprehensive."I'm not sure I need to see much more evidence to be convinced.

Less persuasive is Professor Monsarrat's speculation as to how the initials "W.S."came to be attached to Ford's poem and its dedication. Monsarrat posits a second hand (in the dedicatory epistle) and conjectures that the Elegy's autobiographical passages describe not Ford but an unknown W.S., whom Ford served as ghostwriter. I find this explanation strained and out of keeping with the period (Monsarrat's ghostwriting Ford sounds more like Gertrude Stein ventriloquizing Alice B. Toklas than, in Monsarrat's offered analogy, a scrivener wording an illiterate client's sentiments). Perhaps Vickers will clear up this point along with others, but until such time, I can only echo Leah Marcus's plea for literary history, voiced at an early stage of the discussion. To my mind, despite the overlaps of phrasing in WS's Elegy (1612) and Ford's *Christ's Bloody Sweat* (1613), the Elegy reads as a work of greater maturity, free from the adolescent high-jinks of Ford's youthful devotional writing. At the risk of extending Monsarrat's disintegrationist argument, I wonder whether not only a second voice but a second hand is present in Ford's poem. And on the basis of historical information I have turned up in the past several years, I find myself asking whether that editorial hand (or confessional voice) could be Shakespeare's.

Before such questions are engaged, I wish to offer public congratulations to Professor Monsarrat for throwing open many mysteries of a poem I once thought I knew.

[2]————————————————————————-

From: Don Foster ¡foster@VASSAR.EDU¿

Date: Wednesday, 12 Jun 2002 16:20:55 -0400

Subject: WS's Elegy

In 1996, having ventured an attribution of W.S.'s "A Funeral Elegy"to Shakespeare, I was blasted in the pages of TLS. But Shakespeare's authorship was not as easily disproved as some skeptics anticipated. Though several alternative attributions were advanced, they failed for a good reason. They were mistakes. Recently, though, the French scholar, G. D. Monsarrat, may have succeeded where English and American scholars have failed, demonstrating in an article in the *Review of English Studies* that the elegy looks like the work of the Jacobean dramatist, John Ford. I know good evidence when I see it and I predict that Monsarrat will carry the day. If I may quote the elegy, "what he spake / Seem'd rather answers which the wise embrace / Than busy questions such as talkers make."No one who cannot rejoice in the discovery of his own mistakes deserves to be called a scholar. Monsarrat's fine essay has compelled me, largely against my will, to return to an attribution and a text I have not considered in years. Years ago when Ford was first mentioned as a possible author, I scoffed at the attribution. Ford's rate of enjambment was too low. His use of "Shakespearean who"was largely confined to *Chris-

tes Bloudy Sweat* (1613). His distinctive vocabulary was not (and this was a downright mistake, as I have since discovered upon indexing the Ford canon more fully) as richly represented in the elegy as Shakespeare's. Ford would not, in a first-person funeral poem, attempt to deceive anyone about the author's identity. Etc. But I ought to have attended more closely to the internal evidence-something that, in an irony that I can only now fully appreciate, I myself insisted on in arguing the case for Shakespeare.

The 1612 quarto may have invited its first readers to take "W.S."for William Shakespeare, but that external evidence, I think, must now be viewed in a new light. I do not know how or why incorrect initials were tagged to the title page and author's dedication, nor how the text came to be published by Thomas Thorpe, nor why the elegist borrowed so heavily from Shakespeare and from texts known to Shakespeare. Monsarrat's hypothesis that Ford was employed as a ghost-writer for W.S. seems, to me, implausible for several reasons but I have no better solution to offer.

Since 1997 I have had a second career in criminology and forensic linguistics that has taken time from an unfinished project that remains, for me, a source of frustration. The Shaxicon database-which contributed to my own conviction, in 1996, that Shakespeare wrote the elegy-is still unpublished. Nor have I yet determined where I went wrong with the statistical evidence. Still, my experience in recent years with police detectives, FBI agents, lawyers, and juries has, I hope, made me a better scholar. Our courts have long exacted higher standards for the admissibility of evidence than literary journals. If authorship of "A Funeral Elegy" were a crime, no court in America would have allowed "expert witnesses"on the stand to opine that the offender was Sclater or Slayter or Strode or Simon Wastell. Nor, if Shakespeare were charged with the offense, would the courts have allowed a defense "expert"to opine that Shakespeare was simply not a man to write that sort of thing. My experience with the anonymous documents in criminal investigations indicates that competent and trusted people-math professors, parents, biowarfare experts-often commit acts or write texts that you wouldn't expect of them. Personal opinions cannot stand for evidence, nor can personal rhetoric. But in light of the evidence marshaled by Monsarrat, and possibly augmented by Brian Vickers' forthcoming book, the jury need not hold forth much longer on Shakespeare's authorship of "A Funeral Elegy."The kinds of linguistic and intertextual evidence I myself most trust-and that informs Monsarrat's essay-associate "W.S."more strongly with Ford than with Shakespeare.

S H A K S P E R: The Global Shakespeare Discussion List Hardy M. Cook, editor@shaksper.net The S H A K S P E R Web Site ¡http://www.shaksper.net¿

Apêndice B

Tabela de Autores - Base de Dados

Tabela B.1: Lista de Autores - Base de Dados

Autores	**Fonte**	**Temas/Detalhes**
Antonio Pietrobelli	Tribuna do Paraná	Temas relacionados com comércio exterior
Miguel Sanches Neto	Gazeta do Povo	Temas Gerais
Wilson Martins	Gazeta do Povo	Temas Gerais
Bernt Entschev	Gazeta do Povo	Temas Gerais
Paulo Coelho	Tribuna do Paraná	Colunas para Reflexão
Luiz Carlos Zanoni	Tribuna do Paraná	Vinhos e Cultura
Dante Mendonça	Tribuna do Paraná	Política com Humor
Tostão	Gazeta do Povo	Esportes
Antônio Ermírio de Moraes	Gazeta do Povo	Brasil (Economia, Política, etc.)
Carlos Nasser	Gazeta do Povo	Brasil (Economia, Política, etc.)
Antônio Delfim Netto	Gazeta do Povo	Economia e Política
Dom Moacyr Vitti	Gazeta do Povo	Arcebispo de Curitiba (2007) Temas sobre religião e ética
Reinaldo Bessa	Gazeta do Povo	Cultura e Cotidiano
Celso Nascimento	Gazeta do Povo	Economia
Carneiro Neto	Gazeta do Povo	Esportes (Futebol)
Carlos Brickmann	Diário do ABC	Economia e Política
Leone Farias	Grande Diário do ABC	Economia e Análise de Mercado Financeiro

continua na próxima página...

Tabela B.1: Lista de Autores - Base de Dados (cont.)

Autores	**Fonte**	**Temas/Detalhes**
Francisco Giovanni D. Vieira	Jornal Eletrônico www.wnet.com.br	Marketing
Reginaldo Aparecido Carneiro	Jornal Eletrônico www.wnet.com.br	Administração
Gilberto Dimenstein	Folha de São Paulo	Jornalismo Comunitário
Albert Zeutoni	Correio Popular www.cpopular.com.br	Assuntos Gerais e Cotidiano
Arnaldo Jabor	Correio Popular www.cpopular.com.br	Cotidiano, Economia e Política com humor
Cecílio Elias Netto	Correio Popular www.cpopular.com.br	Reflexão e Cotidiano
Alberto Dines	Correio Popular www.cpopular.com.br	Política
Carlos Alberto Di Franco	Correio Popular www.cpopular.com.br	Consultoria em Estratégia de Mídia
Flávio Gomes	Correio Popular www.cpopular.com.br	Esportes (Automobilismo)
Paulo R. Castro	Correio Popular www.cpopular.com.br	Psicanálise e psiquiatria
Rogério Verzignasse	Correio Popular www.cpopular.com.br	Temas Gerais
Manuel Carlos Cardoso	Correio Popular www.cpopular.com.br	Direito
Jose Pedro Martins	Correio Popular www.cpopular.com.br	Política

Apêndice C

Tabela de Atributos Estilométricos

C.1 Atributos da Língua Portuguesa

Tabela C.1: Características da Língua Portuguesa

Características da Língua Portuguesa	**Exemplos**
Aumentativos	casarão, balaço
Diminutivos	sopinha, pobrezinho
Superlativos	agradabilíssimo, amicíssimo
Conjunções	todavia, por conseguinte
Coletivos	matilha, vara
Pronomes	próclise, mesóclise e ênclise, pronomes de tratamento, etc.
Concordâncias	
Advérbios	absolutamente
Hifenização	super-homem
Porquês	Por que, porque, porquê e por quê
Plural de substantivos simples e compostos	troféus, canis
Gênero de substantivos	a sentinela, o cônjuge
Figuras de estilo	Eufemismos, Metonímia
Interjeição	ai! nossa! basta!
Vícios de Linguagem	Cacografias, Arcaísmo, Neologismos, Jargões

continua na próxima página...

Tabela C.1: Língua Portuguesa (continuação)

Características da Língua Portuguesa	**Exemplos**
Trema	agüentar, lingüística
Crase	Ele foi à casa de Pedro
Acentuação Gráfica	grátis, más, pêra, hífen
Numerais	cinqüenta, cincoenta, Hum, um
Iniciais maiúsculas	Brasil, Jesus
Sintaxe	
Ortografia	

Apêndice D

Recentes Casos de Uso (Brasil)

D.1 Juíza se afasta de casos com o Opportunity

Autora da decisão que levou ao afastamento do Opportunity do controle da Brasil Telecom, a juíza Márcia Cunha, da 2ª Vara Empresarial da Justiça do Rio, declarou-se na semana passada sob suspeição[1] o para julgar processos do banco e de companhias controladas pela instituição. Cunha protagonizou uma disputa pública com o Opportunity, do empresário Daniel Dantas, ao denunciar uma tentativa de suborno por supostos intermediários do banco.

O Opportunity e os fundos de pensão travaram o maior conflito societário em empresa privada da história recente do país, pelo controle da Brasil Telecom, que acabou ficando com os fundos. No ano passado, Dantas pediu anulação de uma decisão da juíza em favor dos fundos e pediu seu afastamento do caso, alegando parcialidade da magistrada.

A decisão em favor dos fundos foi mantida em segunda instância pelo Tribunal de Justiça do Rio. Mas isso não interrompeu a guerra paralela entre o Opportunity e Cunha, que culminou com a decisão dela de se declarar sob suspeição para julgar ações envolvendo o banco, afastando-se do processo da Brasil Telecom.

Em entrevista à Folha ontem, a juíza não quis falar sobre os motivos que a levaram ao seu ato. Mas, ao comunicar sua decisão no processo,

[1]Situação, expressa em lei, que impede os juízes, representantes do Ministério Público, advogados, serventuários ou qualquer outro auxiliar da Justiça de, em certos casos, funcionarem no processo em que ela ocorra, em face da dúvida de que não possam exercer suas funções com a imparcialidade ou independência que lhes competem.[Fer04]

afirma que o fez "por não ter força para enfrentar o poder econômico" do Opportunity.

No documento de duas páginas, declara que, desde que afastou o Opportunity da Brasil Telecom, tem "sofrido toda a sorte de infortúnios". Entre os quais, citou os boatos contra ela de que teria sido corrompida e comprado um apartamento em Ipanema (do qual a juíza diz ser locatária) e de que a decisão contra o Opportunity teria sido redigida por advogados dos fundos de pensão.

A juíza queixou-se ainda que teria sofrido intimidação por parte de estranhos, que seu gabinete teria sido invadido e ainda de que ela e o filho foram ameaçados na rua onde moram. Além disso, aponta as representações do Opportunity que a acusam de improbidade administrativa e falsidade.

O estopim de sua decisão, pelo que escreveu no processo, foi a apresentação por parte do Opportunity de quatro laudos periciais ao Conselho de Magistratura, onde corre processo disciplinar contra a juíza. Os laudos contratados pelo banco atestam, segundo documento obtido pela Folha e apensado no processo, não ser de autoria dela a sentença a favor dos fundos.

O laudo que mais incomodou a juíza foi o do membro da ABL (Academia Brasileira de Letras) Antonio Olintho. Sob o título "Perícia Estilística e de Identificação Autoral", o acadêmico compara várias sentenças da juíza e conclui que o texto da decisão "não está filiado ao estilo da juíza tanto na parte vocabular como na formação de frases". Para ele, a juíza "não é a verdadeira autora".

No processo contra a juíza, o Opportunity anexou ainda manifestações de duas professoras de português da UFRJ. Uma delas destaca como algo "inexplicável" o fato de a juíza ter grafado "Nova Iorque" na sentença contra o Opportunity, quando em outras escrevia "Nova York".

A primeira argumentação do Opportunity para afastar a juíza era que sua filha havia estagiado no escritório que advoga para os fundos nesse caso. Depois disso, começaram a circular boatos de que a juíza não foi autora da decisão e que teria recebido dinheiro para favorecer os fundos.[Soa06]

D.2 Caso Procurador Luiz Francisco de Souza

O segredo de um bom procurador da República está em suas fontes e na rapidez com que produz suas denúncias. O procurador Luiz Francisco de Souza reúne essas qualidades.

Mas na última quinta-feira (2/9), surgiu uma dúvida a respeito da alta produtividade do mais famoso integrante do Ministério Público no país. Uma Ação de Improbidade Administrativa combinada com Ação Civil Pública apresentada por ele um dia antes, apresentou uma esquisitice.

O arquivo em que foi digitada a ação não tem origem na Procuradoria, onde Luiz Francisco trabalha, mas no computador de um empresário que é parte interessada na causa em questão. O autor do arquivo seria o advogado do empresário, Marcelo Ellias.

O procurador rechaça com veemência que tenha apresentado uma ação que não seja de sua autoria. Mas não explicou porque ao se checar a origem do arquivo, verificando suas propriedades, o computador registrado é da Nexxy Capital Ltda., empresa de propriedade de Luiz Roberto Demarco.

A ação é contra 18 pessoas e empresas, mas o alvo principal é o administrador de fundos de investimentos Daniel Dantas. Demarco é seu desafeto, adversário e inimigo.

"Eu e só eu sou o autor intelectual desta ação em que trabalho há mais de três anos", garante Luiz Francisco. "Tenho aqui todos os documentos, todas as minutas que comprovam que o autor da representação sou eu".

O arquivo da petição foi enviado pela secretária do procurador ao site Consultor Jurídico. O nome do arquivo chamou a atenção por conter a expressão "Ufa Ufa Ufa". O inusitado provocou a curiosidade. Todo arquivo do editor de textos Word contém os dados básicos de sua criação, como a empresa em que está registrado o computador, o usuário da máquina, a data de criação do arquivo e até mesmo quando se deu a última impressão do arquivo. Uma rápida checagem mostrou que a data de criação do arquivo ou o dia em que fora gravado no computador da Procuradoria foi a última terça-feira (31/8). A petição tem data de 1º de setembro.

A primeira hipótese apresentada pelo site a Luiz Francisco foi a de

que ele poderia ter recebido um arquivo da Nexxy, apagado o conteúdo anterior e redigido nela sua petição. O procurador repeliu a possibilidade. Mais adiante, suscitou o fato de os computadores da Procuradoria serem máquinas apreendidas pela Receita, como a sugerir que o equipamento de seu uso pudesse ter sido antes da empresa. Luiz Francisco fez outras considerações. "Parte de meu trabalho é digitada no computador de minha secretaria e costumo usar o computador de minha casa também". Mas ele mesmo descartou a hipótese de um desses computadores pertencer ou ter pertencido a outrem.

Em outros telefonemas feitos para a redação da Consultor Jurídico, o procurador cogitaria de outras possibilidades, como a de ter usado um disquete que lhe foi emprestado há tempos por Marcelo Ellias, quando este advogava para a Caixa de Previdência do Banco do Brasil (Previ). Essa possibilidade, contudo, não parece combinar com a data de criação do arquivo, 31 de agosto último.

Em pelo menos três vezes, Luiz Francisco invocou como prova da sua absoluta correção, o fato de ser socialista e de ser sua tarefa "destruir o capital, como escrevi em meu livro". O Opportunity seria a incorporação do que há de mais nocivo na humanidade. E revelou que para livrar o país desse problema tem lançado mão de todos os recursos. "Já fui à CVM, à CPI do Banestado, à Advocacia-Geral da União, ao Senado, à Controladoria-Geral da União e vou onde puder ir para cumprir a minha missão".

Depois de mandar a ação, Luiz Francisco afirmou que a publicação da mesma não estava autorizada e que o envio serviu apenas para que se produzisse uma notícia a respeito. "Vou tirar até o último tostão de vocês se o site publicar essa história", avisou ele. "Não pra mim, que não quero dinheiro, mas para um asilo de cegos", acrescentou, completando que a "a partir de agora as portas do Ministério Público estarão fechadas para vocês".

A reportagem procurou o empresário Luiz Roberto Demarco e seu advogado, Marcelo Ellias, mas nenhum dos dois respondeu aos pedidos e recados deixados pela revista. Ao primeiro por meio de sua secretária, Magna. Ao segundo, pelo celular.

Cenas insólitas

Em um primeiro momento, Luiz Francisco convidou a reportagem

para verificar se, em seu computador, haveria algum vestígio de arquivo produzido fora da Procuradoria.

Ao ser procurado, em Brasília, pelo correspondente da revista Consultor Jurídico, Vicente Dianezi, o procurador adotou uma atitude incomum e inédita em sua história. Não permitiu a entrada em sua sala.

Pela primeira vez, recusou-se a receber um jornalista em seu gabinete. Afinal, os anais da imprensa registram atos do procurador como o de ter pegado emprestado o gravador do jornalista Andrei Meirelles para gravar, através da divisória do gabinete contíguo sua rumorosa conversa com o senador Antonio Carlos Magalhães, quatro anos atrás.

"Só converso com jornalistas investigativos e não sei quem são vocês", disse pelo telefone ao ramal da portaria. Acrescentou que poderia nos receber na quarta-feira, dia 8, quando apresentaria toda a documentação de três anos para cá do caso Previ/Opportunitty. Foi-lhe solicitado então que apenas enviasse para o saguão a cópia de qualquer outra petição gravada na fonte estrangelo edessa, a mesma da ação civil. Essa fonte, pouco usada, costuma ser encontrada em empresas de informática, como a Nexxy.

Como não tinha a fonte em seu computador, mandou por intermédio de sua assistente, a cópia de um ofício, datado de 1999, mas a fonte era arial. Dirigiu-se então ao saguão. Recusou-se a estender a mão ao jornalista. Muito alterado, foi dizendo que sua luta era pelos direitos humanos, pelos pobres e contra o capital. O procurador falava alto, borrifava saliva e envolvendo toda essa emoção cuspiu a obturação que passou procurar no chão, do alto de seus cerca de 1m80 de altura.

"Ninguém vai macular a minha imagem.Se tentarem isso, vou à Justiça buscar tostão por tostão e darei para o hospital da hanseníase". De pouco adiantou explicar que a única intenção era a de encontrar uma explicação para aquela fonte de texto incomum nas petições oficiais. "Talvez seja coisa do Marcelo Ellias que sempre vem aqui. Ele pode ter trazido um disquete". Indagado sobre a freqüencia com que se avista com Marcelo Ellias, recusou-se a responder. Ele esteve aqui este ano? "Sim, esteve". Quando foi a última vez? Nada respondeu. Disse apenas que Ellias era advogado da Previ.

Mais raivoso ainda disse que recebeu a informação "na semana passada"de que o Consultor Jurídico era "patrocinado pelo Grupo Opportu-

nity". "Estou investigando", acrescentou."Não são vocês que vão macular a minha imagem... Eu vivo do minguado salário que recebo aqui...". E dirigiu-se ao elevador, retirando-se, falando alto: "Não tenho nada com a Nexxy Capital. Eu sou contra o capital".[Jur05]

D.3 Exemplo de uma Coluna Inteira - Base de dados

AEB prevê superávit de US$ 42 bilhões, na balança - 28/08/05

O ritmo crescente das exportações na balança comercial brasileira levou a Associação de Comércio Exterior do Brasil (AEB) a ampliarem cerca de US$ 10 bilhões sua estimativa de superávit para 2005, de US$ 32,113 bilhões para US$ 42,133 bilhões. Segundo a entidade, as vendas externas devem chegar a US$ 114,555 bilhões, mantendo o registro anual de recordes históricos que começou em 2000. A AEB projeta, ainda, US$ 72,422 bilhões em importações.

Para o vice-presidente da AEB, José Augusto de Castro, no caso das compras externas, que também deverão bater um novo recorde, o resultado pode ser decepcionante. Isto porque o aumento previsto, de US$ 62,782 bilhões em 2004 para US$ 72,422 bilhões em 2005, deve-se, principalmente, à valorização cambial, que permite um cenário altamente favorável às importações, e não ao aquecimento da demanda interna.

Ilustrando a decepção com as compras no exterior, o montante estimado de US$ 14,884 bilhões para a aquisição de bens de capital é inferior aos valores recordes de US$ 16 bilhões em 1997 e US$ 16,057 bilhões em 1998, destacou Castro.

"A valorização cambial, que poderia ajudar as importações, sofre influência negativa do baixo nível de demanda doméstica", disse o vice-presidente da AEB.

Ele afirmou que, por outro lado, o mercado mundial está aquecido. Isso permite a manutenção da elevada demanda externa e das altas cotações internacionais das commodities exportadas pelo Brasil.

Conflitos com a China

As negociações entre Brasil e China, relacionadas ao comércio bilateral, estão prontas para começar. O ministro do Comércio chinês, Bo Xilai, respondeu a carta do ministro do Desenvolvimento, Indústria e Comércio Exterior, Luiz Fernando Furlan, encaminhada em julho. No

documento, Bo Xilai se diz disposto a negociar com o governo brasileiro, bem como "promover a resolução adequada dos problemas, que ambas as partes estão atentas". O encontro entre os dois ministros deverá acontecer na segunda quinzena de setembro, em Pequim, na China.

O secretário interino de Comércio Exterior, Armando Meziat, informou que uma missão técnica, coordenada por ele e por membros do Ministério do Desenvolvimento, viajará na primeira quinzena de setembro para iniciar as negociações com os chineses. "O grande objetivo é estreitar as relações entre Brasil e China", informou o secretário.

Segundo Meziat, as conversas com os setores que se sentem prejudicados, como têxtil e calçados, já começaram e que, até a próxima segunda-feira (29/08), esses segmentos tem que encaminhar dados ao MDIC que comprovem o dano que as importações chinesas estão ocasionando para sua indústria. Ele explicou que a idéia é levar esses números para que os chineses auto-limitem suas exportações para o Brasil, nos setores que houver necessidade. "A carta demonstra uma boa vontade do governo chinês em receber as queixas brasileiras, buscando solucioná-las", comentou.

Meziat lembrou ainda que a publicação do Decreto que regulamenta a salvaguarda específica para a China continuará seguindo seu caminho normal. "As reuniões com os chineses acontecem independente da publicação do Decreto", afirmou Meziat.

Brasil em Havana

Será realizada entre os dias 31 de outubro e 5 de novembro de 2005, no Expocuba, em Havana, Cuba, a 23.ª Feira Internacional de Havana. A Feira é realizada há 22 anos no mercado caribenho e centro-americano e representa um importante meio de prospecção e introdução de produtos brasileiros naquela região. A Fihav é um importante instrumento de ligação de promoção de produtos de exportação e também tem contribuído para estreitar as relações comerciais entre países, além de oferecer as empresas brasileiras participantes oportunidade de inseri-las e consolidá-las no mercado cubano.

Durante a realização da última edição da FIHAV, as transações comerciais alcançaram a cifra de US$ 247 milhões de dólares. A feira contempla os seguintes segmentos: alimentos in natura e processados, artigos de vestuário, calçados e acessórios, bebidas em geral, carnes e derivados, cerâmica decorativa e utilitária e pisos, equipamentos e instru-

mentos para laboratórios, equipamentos médicos e dentários, equipamentos, partes e peças para indústria açucareira, equipamentos, periféricos e insumos para informática, ferramentas em geral, frutas in natura, em massas ou caldas, máquinas e equipamentos para industria, comércio e agricultura, materiais de construção, materiais elétricos de alta e baixa tensão, móveis e artigos para decoração, papel para gráficas, produtos para higiene e limpeza, tintas e vernizes, utilidades domésticas e veículos e equipamentos para transporte.

A Agência de Promoção de Exportações e Investimentos - Apex-Brasil organizará e apoiará financeiramente a feira. O benefício concedido chegará a 70% do valor total do evento. A empresa expositora já se beneficia do apoio financeiro ao pagar sua participação à França Eventos e Negócios Internacionais Ltda., empresa representante oficial da Fihav e contratada pela APEX-Brasil para a realização da arregimentação de empresas e organização da feira. O pagamento desta participação é a contrapartida da empresa para o evento. Para participar, entre em contato com a organizadora oficial por meio dos contatos: Nilton França (27) 3324-3606 ou (27) 8111-8065 ou francaonline@terra.com.br; Tatiana: (11) 3259-6466 ou freedom- franca@yahoo.com.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Administração
Baixar livros de Agronomia
Baixar livros de Arquitetura
Baixar livros de Artes
Baixar livros de Astronomia
Baixar livros de Biologia Geral
Baixar livros de Ciência da Computação
Baixar livros de Ciência da Informação
Baixar livros de Ciência Política
Baixar livros de Ciências da Saúde
Baixar livros de Comunicação
Baixar livros do Conselho Nacional de Educação - CNE
Baixar livros de Defesa civil
Baixar livros de Direito
Baixar livros de Direitos humanos
Baixar livros de Economia
Baixar livros de Economia Doméstica
Baixar livros de Educação
Baixar livros de Educação - Trânsito
Baixar livros de Educação Física
Baixar livros de Engenharia Aeroespacial
Baixar livros de Farmácia
Baixar livros de Filosofia
Baixar livros de Física
Baixar livros de Geociências
Baixar livros de Geografia
Baixar livros de História
Baixar livros de Línguas

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo