PAULO JÚNIOR VARELA

UMA ABORDAGEM COMPUTACIONAL BASEADA EM ANÁLISE SINTÁTICA MULTILÍNGUE NA ATRIBUIÇÃO DA AUTORIA DE DOCUMENTOS DIGITAIS

Tese apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Doutor em Informática Aplicada.

CURITIBA

2017

PAULO JÚNIOR VARELA

UMA ABORDAGEM COMPUTACIONAL BASEADA EM ANÁLISE SINTÁTICA MULTILÍNGUE NA ATRIBUIÇÃO DA AUTORIA DE DOCUMENTOS DIGITAIS

Tese apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Doutor em Informática Aplicada.

Área de Concentração: *Ciência da Computação*

Orientador: Prof. Dr. Edson José Rodrigues Justino
Co-orientador: Prof. Dr. Flávio Bortolozzi

CURITIBA

2017

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central

V293a
2017

Varela, Paulo Júnior
Uma abordagem computacional baseada em análise sintática multilíngue na atribuição da autoria de documentos digitais / Paulo Júnior Varela ; orientador, Edson José Rodrigues Justino ; coorientador, Flávio Bortolozzi. – 2017.
xxi, 118 f. : il. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2017
Bibliografia: f. 109-118

1. Linguística – Processamento de dados. 2. Processamento de linguagem natural. 3. Semântica. 4. Informática. I. Justino, Edson José Rodrigues. II. Bortolozzi, Flávio. III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. IV. Título.

CDD 20. ed. – 004

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

PUCPR
GRUPO MARISTA

ATA DE SESSÃO PÚBLICA

DEFESA DE TESE DE DOUTORADO Nº 46/2017

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA – PPGIa
PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ - PUCPR

Em sessão pública realizada às 09h30 de 15 de Agosto de 2017, na Sala 203 – Escola de Negócios, ocorreu a defesa da tese de doutorado intitulada **"Uma Abordagem Computacional Baseada em Análise Sintática Multilíngue na Atribuição da Autoria de Documentos Digitais"** elaborada pelo aluno **Paulo Junior Varela**, como requisito parcial para a obtenção do título de **Doutor em Informática**, na área de concentração **Ciência da Computação,** perante a banca examinadora composta pelos seguintes membros:

Prof. Dr. Edson José Rodrigues Justino (orientador) - PPGIa/PUCPR

Prof. Dr. Flávio Bortolozzi (co-orientador) – UNICESUMAR

Prof. Dr. Julio César Nievola – PPGIa/PUCPR

Prof. Dr. Edson Emílio Scalabrin – PPGIa/PUCPR

Prof.ª Dr.ª Mauren Abreu de Souza – UTFPR

Após a apresentação da tese pelo aluno e correspondente arguição, a banca examinadora emitiu o seguinte parecer sobre a tese:

Membro	Parecer	
Prof. Dr. Edson José Rodrigues Justino	(X) Aprovada	( ) Reprovada
Prof. Dr. Flávio Bortolozzi	(X) Aprovada	( ) Reprovada
Prof. Dr. Julio César Nievola	(X) Aprovada	( ) Reprovada
Prof. Dr. Edson Emílio Scalabrin	(X) Aprovada	( ) Reprovada
Prof.ª Dr.ª Mauren Abreu de Souza	(X) Aprovada	( ) Reprovada

Portanto, conforme as normas regimentais do PPGIa e da PUCPR, a tese foi considerada:

(X) **APROVADO**
(aprovação condicionada ao atendimento integral das correções e melhorias recomendadas pela banca examinadora, conforme anexo, dentro do prazo regimental)

( ) **REPROVADO**

E, para constar, lavrou-se a presente ata que vai assinada por todos os membros da banca examinadora.
Curitiba, 15 de Agosto de 2017.

Prof. Dr. Edson José Rodrigues Justino

Prof. Dr. Flávio Bortolozzi

Prof. Dr. Julio César Nievola

Prof. Dr. Edson Emílio Scalabrin

Prof.ª Dr.ª Mauren Abreu de Souza

PPGIa PUCPR Programa de Pós-Graduação em Informática

Rua Imaculada Conceição, 1155 Prado Velho CEP 80215 901 Curitiba Paraná Brasil
Tel.: (41) 3271 1669 Fax.: (41) 3271 2121 www.ppgia.pucpr.br

À minha esposa Denise, ao meu filho Gabriel.

Aos meus pais Loury e Nerci Varela.

Agradecimentos

À Pontifícia Universidade Católica do Paraná, por tornar este trabalho possível pelo apoio financeiro e estrutural.

Ao meu orientador Prof. Dr. Edson J. R. Justino pelo direcionamento, incentivo, auxilio e por me conceder a honra de ser seu aluno.

Ao Prof. Dr. Flávio Bortolozzi, pelas sugestões e contribuições que delinearam este trabalho.

Aos professores Julio Nievola, Edson Scalabrin e Mauren de Abreu pelas importantes contribuições para melhoria do trabalho.

À minha esposa Denise, que sempre me incentivou e esteve ao meu lado nas horas mais difíceis deste caminho. Além de tudo, por gerar uma das melhores sensações que temos na vida, a de ser pai.

À ti meu filho Gabriel, que me despertou a coragem e a confiança de prosseguir.

Aos meus pais Loury e Nerci Varela, que sempre me incentivaram a continuar estudando!

E como a vida não ser doce sem paçoca, mel e chokito.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

θ	Número de amostras de textos por autor para treinamento
α	Número de autores
ξ	Número de amostras de textos de um determinado autor para testes
φ	Número de amostras de referências usadas nos testes
μ	Número de frases por amostra de texto
D_t	Subconjunto de vetores de características de cada autor
Fd_a	Cálculo da Decisão Final
F_k	Frases de um texto
Ft_i	Conjunto de vetores normalizados
N_k	Número de palavras da frase
P_r	Resultados Parciais
Q_a	Subconjunto de testes
R_p	Subconjunto de amostras de referência
S_c	Autor conhecido
S_d	Autor desconhecido
T	Amostra de texto
T_s	Conjunto de treinamento
Vt_i	Conjunto de Vetores de Características
Z_-	Amostras de textos de autores diferentes
Z_+	Amostras de textos de um mesmo autor

Lista de Abreviaturas

@>N	Adjunto Adnominal
CLEF	Base de Dados da *Conference and Labs of the Evaluation Forum*
CNG	*Common N-Grams*
CTT	Árvores de Copenhagen
DET	Pronome Determinante
HTML	*HyperText Markup Language*
J4.8	Algoritmo de árvore de decisão
KLD	*Kullback-Leibler Divergence*
LDA	*Latent Dirichlet Allocation*
M	Gênero Masculino
MDA	Análise Discriminante Múltipla
PAN/CLEF	Série de eventos e tarefas cientificas sobre textos digitais forenses
PDF	*Portable Document Format*
S	Singular (Tempo)
SVM	*Support Vector Machines*
VISL	*Visual Interactive Syntax Learning*
XVIII	Século 18

Resumo

Neste trabalho apresenta-se uma abordagem multilíngue baseada na linguística computacional utilizando características sintáticas da língua para aplicação em casos que envolvam a atribuição de autoria em documentos digitais. O principal problema da atribuição de autoria é saber se um determinado texto foi elaborado por um autor em específico, ou então, identificar o autor do documento entre uma lista de autores. Para solucionar este problema, aplicou-se um conjunto de características sintáticas, considerando a estrutura gramatical interna das frases, onde a ideia principal é extrair funções de cada palavra que são necessárias para compor uma frase, tais como: sujeito, predicado e complementos. Estes elementos denotam um padrão de escrita, traçando um perfil de estilo para cada autor. Para validar a proposta organizou-se uma base de dados de textos literários (textos longos) pertencentes a autores consagrados da literatura em cinco diferentes idiomas: português, espanhol, francês, alemão e inglês, e uma base de dados de textos jornalísticos (textos curtos) redigidos em língua portuguesa e inglesa. Para extrair as características estilométricas de cada texto foi feito uso do processamento de linguagem natural para rotulagem e classificação de cada palavra. Com isso, formaram-se cinco vetores de características estilométricas para cada amostra de texto. A formação dos vetores de características de cada autor é dada pela distância euclidiana entre cada uma das amostras, gerando assim parâmetros comparativos por meio da similaridade dos textos. Para a geração dos modelos de treinamento foram utilizadas duas abordagens (dependente e independente do autor), e duas abordagens no processo de testes (verificação e identificação de autoria). Avaliaram-se também o comportamento do modelo quanto a variação da quantidade de informação de cada amostra (10, 50, 100, 200, 300, 400 ou 500 frases por amostra), e o impacto da quantidade de amostras de referências usadas para cada autor (3, 5, 7 ou 9 amostras). Para realização dos experimentos utilizou-se o classificador SVM (*Support Vector Machines*) duas classes. Ao final, observou-se que a abordagem se mostrou estável e consistente nas taxas de acerto, atingindo 95-98% na verificação de autoria e 86-93% na identificação de autoria, produzindo assim, resultados similares nas diferentes línguas testadas, sendo elas latinas ou anglo-saxônicas. Então, a abordagem proposta se mostra um meio promissor em casos de atribuição de autoria.

Palavras-Chave: atribuição de autoria; características sintáticas; estilo; multilíngue.

Abstract

In this work, we show a multilingual approach based on computational linguistics using syntactic features of the language for application in cases that involve the authorship attribution in digital documents. The main problem of authorship attribution is whether a text was written by a particular author, or identify the author of the document between a list of authors. To solve this problem, we applied a set of syntactic features, considering the internal structure of sentences, where the main idea is to extract functions of each word that necessary to compose a sentence, such as: subject, predicate and complements. These elements denote a writing pattern, tracing a style profile for each author. In order, to validate the approach, a database of literary texts (long texts) belonging to consagrated authors of the literature was organized in five different languages: Portuguese, Spanish, French, German and English, and a database of newspapers texts (short texts) in Portuguese and English. To extract the stylometric attributes of each text was made use of natural language processing for labeling and classification of each word. Thus, five vectors os stylometric attributes were formed for each text sample. The formation of the vectors of attributes of each author is given by the Euclidean distance between each of the samples, thus generating comparative parameters through the similarity of texts. For the generation of the training models, two approaches are used (dependent and independent model), and two approaches in the testing process (verification and identification of the author). The model's behavior regarding the variation of the information quantity of each sample (10, 50, 100, 200, 300, 400 or 500 sentences per sample) was also evaluated, and the impact of the number of references samples used for each author (3, 5, 7 or 9 samples). To perform the experiments, the SVM (*Support Vector Machines*) classifier was used. At the end, it was observed that the approach was shown to stable and consistent in the hit rates, reaching 95-98% in the authorship verification and 86-93% in the authorship identification, thus producing similar results in the different languages tested, being they Latin or Anglo-Saxon. Then, the proposed approach is a promising in cases of authorship attribution.

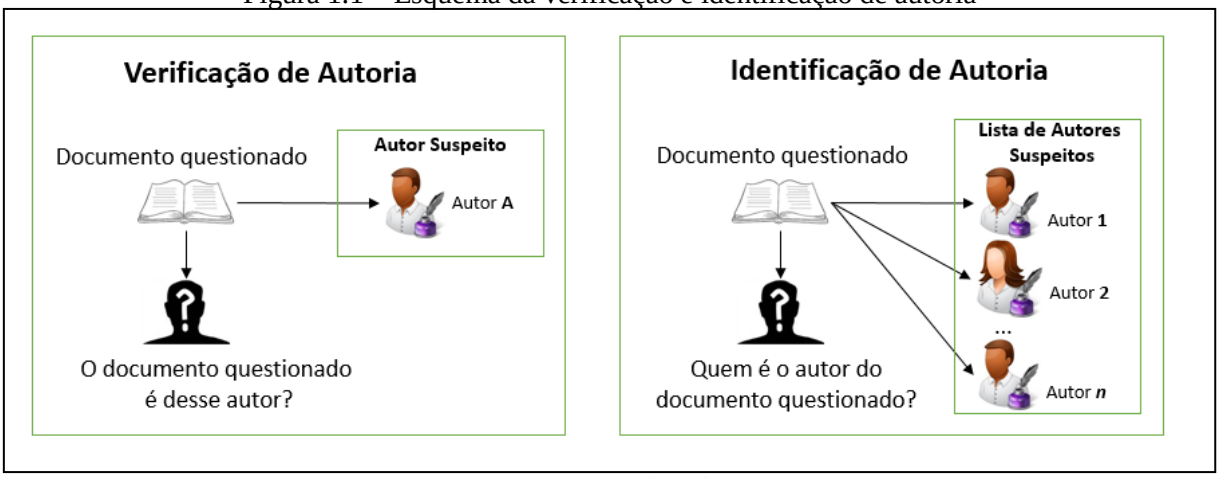
Keywords: authorship attribution; syntactic attributes; stylometry; multilingual.

Capítulo 1

Introdução

Com os meios de comunicação se tornando mais acessíveis e disponíveis, problemas relacionados à atribuição da autoria em documentos digitais se tornaram mais frequentes. Isso fez com que a quantidade de processos no meio jurídico que estão relacionados com a atribuição de autoria aumentasse, gerando assim, uma demanda por parte dos juristas e peritos. Um documento é dito questionado, quando não se sabe exatamente o autor que o escreveu. Então, o estudo de padrões da escrita, permite realizar uma análise computacional dos documentos questionados para descobrir se um texto foi elaborado por um autor em específico (verificação de autoria), ou então, identificar o autor do documento entre uma lista de autores (identificação de autoria), conforme pode ser observado na Figura 1.1.

Figura 1.1 – Esquema da verificação e identificação de autoria



Fonte: autor (2017)

Para solucionar este tipo de problema, onde muitas vezes uma análise manual não é possível, pela grande quantidade de informação a ser processada, o uso da aprendizagem de máquina é uma das mais usuais entre os pesquisadores [ARG03] [JUO06] [STA09]. A aprendizagem de máquina consiste em um método de análise de dados que automatiza o desenvolvimento de modelos analíticos por meio de algoritmos que aprendem por uma série de exemplos, com a finalidade de obter a melhor tomada de decisão [MIC94].

Entretanto, além das técnicas de aprendizagem de máquina é necessário o trato das questões associadas à determinação da autoria, quer seja na língua falada ou escrita. Para isso, pode-se fazer uso da linguística computacional, que tem por função usar meios computacionais para manipulação da linguagem humana. A linguística computacional permite que a atribuição da autoria de um documento possa ser feita independentemente da base de registro utilizada (papel ou formato digital), pois o processo consiste em rotular/classificar cada palavra das amostras de textos de acordo com o seu nível estrutural, morfológico ou sintático, por exemplo. Com isso, pode-se aplicar os conceitos da linguística no contexto judicial, ou seja, a linguística forense. Neste caso, a verificação e a identificação da autoria de um documento questionado podem ser executadas por intermédio da observação de atributos linguísticos, tais como os estilísticos, apresentados pelo autor do documento.

Todavia, com o decorrer dos anos, a aplicação da análise estilística em atribuição de autoria se tornou essencial [HOL98] [MCM02] [BEL08]. Entre as principais aplicações da estilística para resolução de problemas, estão [MCM02]: (*i*) a determinação se um autor escreveu os documentos cuja autoria é questionada, como por exemplo, obras literárias e discursos políticos; (*ii*) a comparação de amostras de textos com outros autores, quando não se possui autores óbvios, tal como a análise de mensagens de e-mail e postagens efetuadas de forma anônima; e, (*iii*) avaliação da semelhança de um documento questionado com documentos de autores suspeitos, por exemplo, bilhetes e mensagens de ameaça ou difamação.

Então, a estilística por sua vez, tem a função de avaliar e identificar as características que tornam um texto, bom ou ruim para discriminar o estilo. Já o estilo de cada autor ou grupo de autores é definido como sendo uma maneira particular de escrever, ou seja, um conjunto de características de uma obra ou de um autor. Sendo assim, o estilo é considerado um elemento variável do comportamento humano por diversos fatores, tais como: grupos sociais, conhecimento técnico, idade, nível de escolaridade, regionalismo e época. O estilo de

um autor é demonstrado pelo seu conjunto único de padrões gramaticais aplicados em sua escrita, que são conhecidos como marcadores de estilo ou características estilométricas [MCM02] [STA09]. Com a estilometria, que visa a aplicação do estilo linguístico aprendido no texto é possível realizar a parametrização das características de cada autor ou de grupos de autores, e assim, conseguir identificar padrões na escrita.

Por conseguinte, a especificação das classes de características linguísticas serve para limitar o campo de observação, ou seja, pode-se eliminar autores que não estão associados a um determinado tipo de escrita em especifico (grupo) [MCM02]. A autoria primeiramente pode ser identificada, ou seja, agrupada com as características pertencentes a sua classe. Posteriormente, a análise estilística tem por função verificar as características intrínsecas de cada autor, onde é evidenciado o conjunto combinações de características utilizadas por cada um dos autores analisados. Neste caso, traços individuais de escritas são denotados e são utilizados para distinguir e identificar os autores.

Diversas pesquisas têm sido apresentadas sob atribuição de autoria em textos nas últimas décadas, mostrando o aumento das taxas de precisão e da qualidade dos atributos [BRY62] [BRI63] [LEM94] [LOW95] [BAA96] [TWE98] [KUK01] [STA01] [BAA02] [DIE03] [ARG03] [GAM04] [CHA05] [UZU05] [ZHA05] [FIN06] [JUO06] [KOP06] [HIR07] [ZHA07] [ABB08] [PAV08] [STA09] [SAV11] [VAR11] [EBR13] [SAV13] [NEM15] [HAL16]. Entretanto, o teor das pesquisas é aplicado na sua grande maioria em um único idioma. E, um fator importante na avaliação de uma abordagem de atribuição de autoria é a sua capacidade de trabalhar com mais de uma língua, ou seja, de ser multilíngue. Neste caso, vale ressaltar que muitas características estilométricas são dependentes do idioma, pois são propriedades únicas e pertencentes a somente um grupo. Comumente, o problema principal é ligado a definição de quais características estilométricas são realmente relevantes para definir o estilo de cada autor. Em geral, as abordagens que utilizam um conjunto de atributos discriminantes em uma língua, podem transferir este mesmo conjunto para outras linguagens sem muitas adaptações.

Neste trabalho, propõe-se uma abordagem computacional para atribuição de autoria de textos em um ambiente multilíngue. Apresenta-se um conjunto de características sintáticas da língua, considerando as estruturas gramaticais internas das frases, por meio de seus níveis estruturais. A ideia principal é extrair funções sintáticas de cada palavra, necessárias para a

formação de uma frase, tais como: sujeito, predicado e complementos. Estes elementos estilométricos denotam um padrão da escrita, delineando um perfil para cada autor.

Assim sendo, aplica-se a abordagem em diferentes cenários, tais como: (*i*) em textos jornalísticos nos idiomas português e inglês; (*ii*) em textos literários nos idiomas, português, espanhol, francês, alemão e inglês. Para tanto, realizaram-se experimentos utilizando abordagens distintas no treinamento (dependente e independente do autor), e nos testes (verificação e identificação de autoria). Para averiguação do processo foi usado o classificador SVM (*support vector machines*).

1.1 Objetivo Geral

O objetivo principal deste trabalho é propor uma abordagem multilíngue baseada na linguística computacional que utilize características sintáticas de estilo para aplicação em casos que envolvam a atribuição de autoria em documento digitais.

1.2 Objetivos Específicos

Como objetivos específicos estão:

- Construir e disponibilizar uma base de textos multilíngue (nos idiomas português, espanhol, francês, alemão e inglês) para realização dos experimentos, já que são ínfimas as bases disponibilizadas para a comunidade que deseja testar suas abordagens;
- Trabalhar com textos heterogêneos quanto à: tamanho (curtos – textos jornalísticos, e longos – textos literários); assunto (diferentes tipos de temas tratados pelos textos); e, linguagem (textos escritos em língua distintas, tais como: a língua portuguesa, espanhola, francesa, alemã e inglesa);
- Identificar um grupo de características sintáticas discriminantes, que proporcione a abordagem proposta ser robusta e confiável quando aplicada em línguas divergentes, para que possam auxiliar em casos que envolvam a atribuição de autoria;
- Realizar testes em cenários de verificação e identificação de autoria para averiguar o comportamento da abordagem em diferentes situações;

- Avaliar a robustez do modelo quanto ao número de exemplares de referência usada na comparação, e quanto a quantidade de informação que compõe cada amostra de texto.

1.3 Justificativa/Motivação

Entre os fatores motivacionais, cita-se o contexto jurídico, onde muitos processos estão inter-relacionados com o questionamento da autoria de documentos impressos e digitais. Além do mais, com o avanço dos meios de comunicação, aumentaram também os casos em que é necessária a intervenção jurídica e pericial para realizar a análise de autoria de textos digitais. Isso, preconiza um problema ainda em aberto e nos motiva com a possibilidade de contribuição por meio da linguística computacional.

Um outro motivo é o fato de existirem grupos de características estilísticas que ainda não foram testadas pela literatura, abrindo mais uma lacuna para inserção da abordagem proposta. Neste caso, grupos de características da classe sintática, que tem por função analisar a estrutura gramatical interna das frases, ainda não foram devidamente experimentadas. Tais características podem denotar o estilo de cada autor, pelo uso de certas estruturas e funções que o diferenciam dos demais.

Relata-se também, que existem poucos trabalhos na literatura que versam sob a aplicação de uma abordagem em um ambiente multilíngue, ou seja, capaz de trabalhar com um mesmo grupo de características em diversos idiomas. Então, a concepção de uma base de textos multilíngue e o conjunto de atributos estilométricos, faz deste tópico, mais um elemento motivacional.

1.4 Originalidade/Contribuições

A principal contribuição desse trabalho é apresentar uma nova proposta de abordagem para auxilio no processo de atribuição de autoria. Neste caso, utilizando elementos estilísticos de classe sintática baseada em termos essenciais, integrantes e acessórios da frase. Tais elementos, apresentaram bons resultados, sendo superiores a outras abordagens já desenvolvidas, tornando assim, a abordagem proposta promissora em casos que envolvam a verificação e a identificação da autoria.

Em consonância, lembra-se que são raras as bases de dados disponibilizadas para a realização de experimentos. Sendo assim, isso levou a criação das bases de dados de textos digitais e disponibilizar de forma gratuita para que comunidade científica interessada possa fazer uso em seus estudos. Estas bases de textos, pertencem a autores consagrados da literatura mundial, escritos nos idiomas português, espanhol, francês, alemão e inglês, que tiveram suas obras literárias disponibilizadas em domínio público.

Cita-se também, a contribuição que a abordagem pode dar no meio jurídico, já que em muitos casos o papel exercido pelo perito linguista é de fundamental importância, na busca de esclarecimentos relativos ao material probante. No entanto, não são raros os casos em que peritos distintos acabem por apresentar laudos ou pareceres discordantes, em relação ao material periciado. Isso leva a concluir que existe, no decorrer do processo de análise pericial, um elevado grau de subjetividade na definição dos atributos relevantes e de como estimá-los, ou seja, não existe uma padronização de qual abordagem (características, métricas, estratégias) utilizar em casos que envolvam a atribuição de autoria, levando assim, que cada perito utilize a sua própria abordagem. Neste caso, a abordagem proposta evidencia um conjunto de atributos, métodos e ferramentas que podem auxiliar peritos, linguistas e juristas no processo de análise pericial de documentos digitais.

Além disso, alguns cenários são passíveis de aplicação, em que a abordagem proposta pode ser conveniente, tais como: a detecção de plágio, a categorização de gêneros textuais e a identificação de mensagens de ameaça e terrorismo.

1.5 Organização do Trabalho

Este trabalho está organizado em capítulos, sendo este o primeiro capítulo, e que se refere à introdução, composta pelos seus objetivos, justificativa e as contribuições. O Capítulo 2, apresenta a revisão da literatura, que engloba as premissas conceituais e teóricas da linguagem e o estado da arte em aplicações multilíngue. No Capítulo 3, é apresentada a metodologia, envolvendo desde a concepção das bases de dados até a rotulagem das palavras. Já no Capítulo 5 é demonstrada a proposta, que compreende uma passo-a-passo do desenvolvimento da abordagem, e por fim no capítulo 6, os resultados e análise dos experimentos.

Capítulo 2

Revisão da Literatura

Este capítulo tem por objetivo apresentar as premissas conceituais e teóricas, bem como estado da arte dos trabalhos já desenvolvidos em atribuição de autoria multilíngue.

2.1 Premissas Conceituais e Teóricas

Esta seção tem por objetivo apresentar a fundamentação teórica necessária para a compreensão da proposta. São discutidos temas como a linguagem e suas estruturas, linguística e estilística, bem como, os principais elementos da análise sintática.

2.1.1 Famílias Linguísticas

Existem no mundo cerca de 6000 línguas, que estão agrupadas em cerca de 20 grandes famílias. Entre estas famílias, se encontra a indo-europeia, que é responsável pela formação dos dois principais ramos das línguas oficiais dos países da Europa ocidental e das Américas, que são [CUL87] [WAL94] [EUR02]:

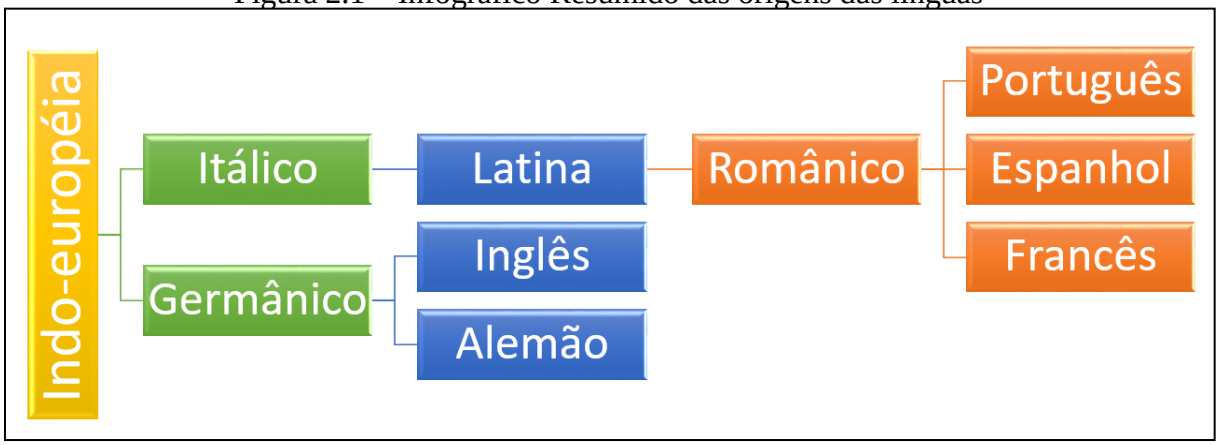
- Latim (português, espanhol, italiano e francês);
- Germânico (inglês e alemão);

Pesquisadores e linguistas utilizaram critérios linguísticos, históricos, sociais, culturais e geográficos para agrupar as línguas em suas respectivas famílias. As características

linguísticas utilizadas para estabelecer relações familiares entre as línguas são os seus sons, formação da palavras e estrutura das frases [EUR02] [MCM02].

Verifica-se na Figura 2.1, que os idiomas português, espanhol, italiano e francês surgiram de uma mesma vertente, ou seja, são descendentes das línguas itálicas, que posteriormente gerou o latim, que por sua vez derivou a língua românica, que em suas peculiaridades geraram cada uma das línguas subjacentes. Já a língua inglesa e alemã, possuem a vertente germânica.

Figura 2.1 – Infográfico Resumido das origens das línguas



Fonte: Adaptado de [EUR02]

Os sistemas de escrita de tantas línguas diferentes, também representa uma diversidade considerável. Para tanto, neste trabalho somente irá se fazer uso do sistema de escrita alfabético, ou seja, dos símbolos praticados nas línguas portuguesa e inglesa, por exemplo.

Um exemplo da herança que as linguagens herdaram ao longo dos tempos, é que algumas possuem a mesma família de origem (indo-europeia), sendo assim, muitos aspectos são semelhantes e podem denotar uma similaridade entre as características. Por exemplo, denota-se que em virtude das línguas portuguesa e espanhola serem oriundas de uma mesma língua mãe, é possível que as características usadas em ambos os idiomas sejam mais similares, do que quando comparadas com a língua chinesa. Entretanto, é essencial entender as formas de linguagem do sistema de escrita conforme descrito na seção 2.1.2.

2.1.2 Formas de Linguagem

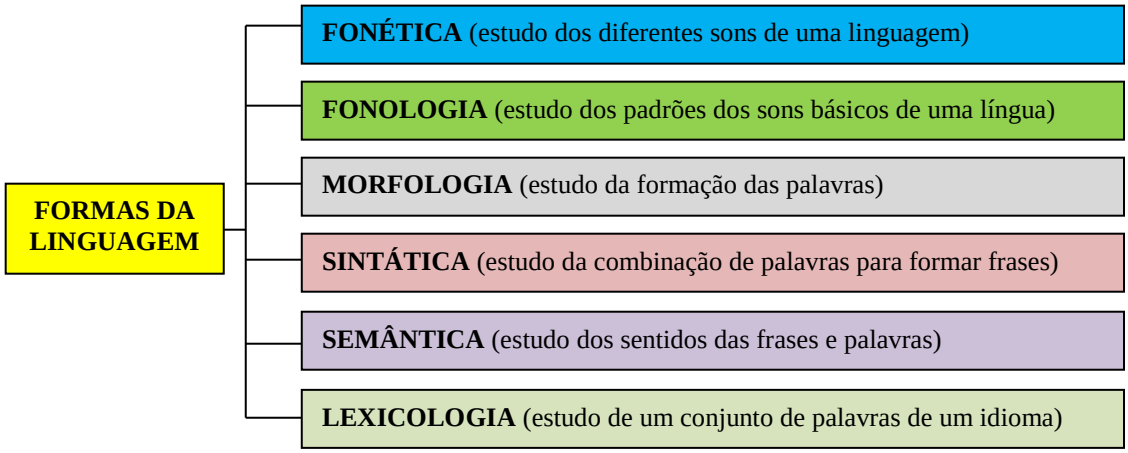
A linguagem é um sistema de comunicação que se expressa por meio de códigos que possuem um determinado significado. Combina sons com seus significados para produzir a

linguagem natural. O indivíduo que fala uma linguagem adquire a capacidade de combinar um ou mais sons em palavras, e palavras em estruturas maiores (combinação de palavras – que chamamos de frases). Sendo assim, tanto emissor como receptor fazem a associação mutua dos significados específicos em uma linguagem particular e dentro de um contexto [MCM02].

A linguagem pode ser estudada em planos complementares e inseparáveis, que são: forma e função. A forma corresponde à estrutura da linguagem, definindo o sistema linguístico. Função tem seu enfoque no uso da linguagem, sendo esta, definida como parte integrante da interação social humana, tal como a análise do discurso e a pragmática. As funções da linguagem relacionam a forma como a linguagem é usada nos contextos da fala e do fazer por meio da comunicação. Os linguistas rotineiramente quebram a linguagem em diversos componentes, conforme pode ser visto na Figura 2.2.

Esta divisão torna a linguagem mais fácil de estudar e descrever, e muitas vezes é útil para separar os componentes de estudo do idioma para melhor entender a linguagem. Por outro lado, a própria linguagem incorpora simultaneamente todos estes componentes juntos na fala e na escrita [MCM02] [WAL94].

Figura 2.2 – Subdivisão das Formas da Linguagem (Adaptado de [MCM02])



Fonte: Adaptado de [MCM02]

Neste trabalho, o aprofundamento é na forma de linguagem sintática, onde analisam-se os padrões de combinação de palavras de cada escritor para formação das frases, que é mais detalhado na seção 2.1.3.

2.1.3 Análise Sintática

A análise sintática é o estudo de como as palavras são combinadas em sequencias longas, tais como: frases (independente de verbo) e orações (dependente de verbo). Para padronizar e melhor entender adotou-se a palavra frase, como sendo frase e oração. O foco da sintaxe é analisar a estrutura gramatical interna das frases [MCM02]. A sintaxe estuda a disposição das palavras na frase e a das frases no discurso, bem como a relação lógica das frases entre si. Ao emitir uma mensagem verbal, o emissor procura transmitir um significado completo e compreensível. Para tanto, as palavras são relacionadas e combinadas. Sendo assim, a sintaxe é essencial para o manuseio das múltiplas possibilidades que existem para combinar palavras em frases [PLP13].

A frase é uma sequência linear de palavras, que podem ser descritas de forma isolada, uma a uma, da esquerda para a direita, nos idiomas português, espanhol, francês, alemão e inglês, por exemplo. No entanto, o receptor ou leitor tem que fazer mais do que interpretar cada palavra de uma frase/oração como uma unidade separada. Além de dimensionar a frase como uma sequência linear, o receptor ou leitor inconscientemente faz a decodificação por intermédio da identificação de um aglomerado de palavras que naturalmente se agrupam como frases, formando assim subpartes aninhadas de toda a frase [MCM02]. Por exemplo, unidades como sujeito e o verbo são universais. Um exemplo pode ser visto na frase “João beijou o sapo”, onde são identificados o sujeito (João) e o verbo (beijou/beijar).

É importante salientar que as pessoas adquirem a capacidade de construir estruturas sintáticas gramaticais aceitáveis rapidamente, e logo após, são capazes de produzir frases mais elaboradas. A tarefa dos linguistas é observar e encontrar maneiras de entender a linguagem, analisar o que os oradores e escritores fazem, e após isso explicar o como e o porquê.

A melhor forma para compreender e descrever a estrutura gramatical é pela análise dos níveis da frase e suas estruturas. Retomando o exemplo, “João beijou o sapo” é o mais alto nível da frase. Quando se realiza a separação da frase, encontrando o sujeito está se criando mais um nível da frase, por exemplo: “João” (sujeito) + “beijou o sapo” (predicado). E, subdividindo ainda mais a frase, pode-se chegar a mais níveis estruturais, tal como: “João” (sujeito) + “beijou” (verbo) + “o” (artigo) + “sapo” (substantivo). Evidentemente, o comprimento e a complexidade subjacente de uma frase pode ser um desafio para estabelecer

a sequência hierárquica de todos os elementos da frase, pois muitas vezes as frases possuem significados diversos.

2.1.3.1 Estrutura da Frase

A maioria das abordagens linguísticas possuem dois níveis, para descrever a estrutura da frase. A primeira foca na estrutura básica da frase de acordo com a linguagem, e a segunda estuda a variabilidade que os nativos de uma determinada língua manipulam as estruturas básicas para criar uma diversificada variedade de estruturas, para formação de frases mais complexas. Geralmente o emissor ou escritor efetua a manipulação das estruturas básicas, e de acordo com tais estruturas primárias o mesmo pode efetuar diversas transformações sobre elas, tais como: incluir, excluir e movimentar elementos de uma frase. Linguistas geralmente modelam esse processo, primeiramente especificando as estruturas básicas das frases, e posteriormente definindo as transformações dos enunciados que resultam em frases.

As frases são formadas por palavras e frases, que são alocadas em diversas categorias gramaticais. Tais categorias gramaticais têm por função se relacionar com a língua para indicar, por exemplo, tempo, modo e lugar relacionados nas frases. As principais categorias gramaticais são: pessoa, número, gênero, caso e tempo.

Para casos que envolvam a atribuição de autoria, as categorias gramaticais mais conhecidas e utilizadas são as partes do discurso. Neste caso, se fazem presentes, algumas características essenciais em uma linguagem, tais como: substantivos, verbos, adjetivos e advérbios. Uma outra categoria gramatical, que é utilizada para unir as palavras e formar frases também é utilizada, são exemplos: preposições, conjunções, pronomes e interjeições. E, por fim, outra categoria gramatical importante são os elementos que se relacionam com a função na frase, por exemplo: sujeito, predicado, complementos (adverbial, nominal e verbal) e transições de voz (passiva e ativa) [MCM02].

Em pesquisa na literatura, não foram encontrados relatos do desenvolvimento de trabalhos em atribuição de autoria que envolvam as características gramaticais desta última categoria, e, sendo assim, a abordagem proposta faz uso destas características para atribuição de autoria. Neste caso, destaca-se a subseção 2.1.3.2 que exemplifica a função sintática.

2.1.3.2 Função Sintática

A função sintática é definida como o papel que os elementos de uma frase exercem em seu interior, levando em consideração as relações existentes entre elas [WAL94]. Em outras palavras, é a análise de uma frase de uma forma que não fique igual a classe gramatical. Por exemplo, os substantivos ora exercem a função de sujeito, de adjunto adnominal, de objeto direto, indireto, etc. A função de cada termo da frase é determinada pela análise sintática. Nesse tipo de análise, cada termo da frase é estudado de acordo com o sentido e posição que ocupa na frase, estabelecendo relação com o restante dos termos.

Os termos de uma frase são categorizados em: essenciais, integrantes e acessórios. Os termos essenciais de uma oração dão o significado e o sentido a um determinado texto. São constituídos por sujeito e predicado. O sujeito é definido como sendo o termo sobre o qual o restante da oração diz algo. Já o predicado é o termo que contém o verbo e informa algo sobre o sujeito.

Os termos integrantes de uma oração são um complemento de significados e sentidos à determinados verbos ou nomes que não possuem sentido se não possuírem um complemento. Estes complementos podem ser verbais (direto e indireto), nominais e agente da passiva.

Os termos acessórios de uma oração, apesar de serem dispensáveis, são importantes para a compreensão do enunciado. Geralmente estes termos denotam e caracterizam o ser, determinam os substantivos e exprimem circunstâncias. Os termos acessórios mais importantes de uma oração são: adjunto adverbial, adjunto adnominal e aposto.

A principal diferença entre função sintática e classe gramatical, está no fato de que a função sintática desempenhada pela palavra é definida pela relação que esta estabelece com os outros termos da frase, sendo definida pela análise sintática. Já a classe gramatical é definida isoladamente por meio da análise morfológica da palavra, que é feita independentemente da posição da palavra na frase. Entre as classes gramaticais mais conhecidas estão: substantivo, adjetivo, verbo, artigo, pronome, advérbio, preposição, conjunção, numeral e interjeição.

Exemplifica-se esta diferença substancial pela análise das frases: “O Paulo leu o livro” e “O aluno aplicado estudava a lição”, que estão detalhadas nos exemplos das frases 1 e 2, apresentadas nas Figuras 2.3 e 2.4, respectivamente.

Figura 2.3 – Exemplo de classificação das palavras na frase 1.

Frase 1:	O	Paulo	leu	o	livro.
Classe Gramatical	Artigo	substantivo	Verbo	artigo	substantivo
Função Sintática	Adjunto adnominal	Núcleo do sujeito	Verbo principal	Adjunto adverbial	Núcleo do predicado

Fonte: autor (2017)

Figura 2.4 – Exemplo de classificação das palavras na frase 2.

Frase 2:	O	aluno	aplicado	estudava	a	lição.
Classe Gramatical	Artigo	substantivo	adjetivo	verbo	artigo	substantivo
Função Sintática	Adjunto adnominal	Núcleo do sujeito	Adjunto adnominal	Núcleo do predicado verbal	Adjunto adnominal	Núcleo do objeto direto.

Fonte: autor (2017)

Percebe-se nos exemplos que as classes gramaticais não possuem uma variação substancial, já que sua classificação é efetuada por meio da morfologia de cada palavra. No entanto, observando a função sintática, verifica-se que de acordo com a posição na frase, cada palavra pode mudar de função, indicando elementos variantes e também não variantes. E isso, leva a considerar que as funções sintáticas possuem um poder de discriminação quando aplicada em casos que envolvam o reconhecimento de padrões da escrita em documentos questionados.

Com o conhecimento adquirido sobre a análise sintática, pode-se realizar a aplicação da linguística por meios computacionais. Para tanto, se faz importante compreender como funciona a linguística que está detalhada na seção 2.1.4.

2.1.4 Linguística

Define-se linguística como o estudo cientifico da linguagem, ou seja, uma investigação da linguagem por meio de observações empíricas e controladas, que possam ser verificadas com base na estrutura geral da linguagem [LYO68].

A linguística aborda a compreensão dos sistemas de linguagem humana, por intermédio de características, regras e padrões presentes nos mais diversos idiomas, que se define como estudo teórico. A linguística também pode ser prática, na medida em que os conhecimentos teóricos possam ser utilizados para aplicações forenses, por exemplo. Vale ressaltar que a linguística é descritiva e não prescritiva, ou seja, o seu objetivo é compreender e descrever a linguagem e não estabelecer regras para a correta e adequada utilização da língua [MCM02].

Sabe-se que as diversas formas de linguagem são descritas por uma gramática (conjunto de regras individuais de uma língua), e esta gramática pode ser: descritiva e prescritiva. A gramática descritiva provê uma imagem objetiva e geral sobre todo o sistema de linguagem e não é avaliativa. Já a gramática prescritiva prevê regras de como a linguagem deve ou não ser usada, sendo assim, avaliativa, pois baseado no comportamento linguístico das pessoas pode se dizer se as mesmas estão fazendo uso correto e aceitável da língua [MCM93].

A linguística pode ser aplicada em diversas áreas do conhecimento humano para compreensão de diversas questões. É interdisciplinar, pois pode ser combinada com outras áreas, tal como, a computação [BAA13]. A linguística aplicada faz uso da coleta e análise de dados linguísticos, sintetiza resultados e obtém conclusões, caso que é característico da linguística aplicada em casos forenses.

2.1.4.1 Linguística Forense

A linguística forense é o estudo cientifico da linguagem aplicada em contextos judiciais. É uma área recente da linguística em relação aos seus mais de 2400 anos de história, e se faz uma crescente área da linguística moderna e aplicada [HON79] [MCM02] [GAM04].

Entre as principais aplicações da linguística forense no ramo cientifico, encontra-se a atribuição de autoria, cujo objetivo é identificar/verificar o autor por meio de traços linguísticos, sendo estes expressados pela fala e/ou da escrita.

Em casos forenses, a tarefa principal da ciência forense é descrever e medir as evidências que podem ser utilizadas como prova em um determinado delito. Neste caso, tais evidências são correlacionadas e comparadas com amostras dos suspeitos, onde é criada uma associação entre a evidência e a amostra do provável autor. O valor real de tais provas

encontra-se na comparação das amostras, onde pode-se identificar e individualizar traços entre dois objetos, tornando mais forte uma ligação entre as duas amostras [MCM93].

A classificação das áreas em que a linguística forense atua, evolui conforme o campo se desenvolve. Em geral, todas as classificações existentes na estrutura e na função da linguagem são usadas em aplicações forenses, tais como [MCM02]:

- **Semântica**: É o estudo dos significados expresso por palavras, frases e textos. O principal objetivo no contexto forense, está em sua aplicabilidade na compreensão e interpretação das linguagens que são difíceis de entender. Um exemplo de aplicação forense, está na interpretação de significados de expressões praticadas por criminosos.
- **Análise do Discursos e Pragmática**: Estudam os significados reais das frases e também o contexto social no qual foi escrito. Em aplicações forenses podem ser utilizadas em casos de estudos de contextos específicos, tais como: a análise do discurso falado e escrito de conversas e audiências.
- **Estilística**: Tem como objetivo forense prover a identificação do autor de um documento questionado. Todos os níveis de linguagem são possíveis características de estilo, tais como: classes de palavras, significados, gramática e uso da linguagem). Este é o foco central deste trabalho, e na seção 2.1.5.1 são abordados mais detalhes sobre a estilística.

Para melhor compreender as aplicações da linguística forense se faz necessário compreender as suas variações, que são apresentadas na seção 2.1.4.2.

2.1.4.2 Variações Linguísticas

A análise da variação da língua é importante nos estudos de atribuição de autoria, principalmente porque a variação pode deixar evidências que podem ser associadas com as características individuais ou de um grupo de autores. A variação linguística pode acontecer com somente uma ou algumas das pessoas do grupo ou até mesmo por todas.

Entre as principais variações de uma língua, pode-se citar a variação da ortografia, do vocabulário e da gramática. Geralmente as variações da língua ocorrem no decorrer dos anos pelo contato com outros grupos e outras línguas. Percebe-se então, que as línguas estão em constante evolução e receptivas a novas variações [BUR87] [BAY91] [KES03].

Em estudos de atribuição de autoria, a análise das variações da língua pode auxiliar no processo de distinção de grupos (exemplo: dialetos regionais, grupos sociais e técnicos), período (época) em que um determinado documento foi escrito (por exemplo: distinguir períodos literários diferentes: barroco e pós-modernismo), idade (exemplo: identificar a idade do autor por meio de traços linguísticos de sua geração), e origem da língua (saber se o texto original foi escrito em que língua ou se é uma tradução).

Uma vez que exista uma boa compreensão das variações linguísticas existentes em uma determinada língua, está se faz importante em casos de atribuição de autoria. As variações linguísticas podem ser determinadas pelas características de estilo existentes nas amostras dos textos conhecidos e questionados.

2.1.5 Estilo

O estilo é definido como sendo uma maneira particular de escrever, de exprimir o pensamento, ou seja, um conjunto de características de uma obra, de um autor ou de uma época [DIC14]. Sendo assim, o estilo é considerado um elemento variável do comportamento humano. Pode-se dizer que as ações humanas mais comuns, possuem parte invariante e parte variante. E, analisando as partes (principalmente as variantes), é possível verificar diferenças entre os indivíduos, podendo assim, realizar a verificação e a identificação de autoria por intermédio do estilo de escrita. O estilo de um autor é demonstrado pelo seu conjunto único de padrões gramaticais aplicados em sua escrita [MCM02] [GAM04] [BEL08] [STA09]. Estes padrões, geralmente são o resultado da utilização habitual e sistemática de algumas formas de escrita, e são conhecidos como marcadores de estilo ou características estilométricas [MCM02] [STA09].

O estilo uma vez desenvolvido pelo indivíduo, permanece com ele pelo restante de sua vida. Estilo é a parte mais profunda de nosso ser, que muitas vezes é desenvolvido e aplicado de forma inconsciente, ou seja, é um estilo comportamental que a medida que é praticado se torna parte da pessoa [KJE94] [SCH01]. O estilo da escrita é o resultado das escolhas que autor faz, e reflete hábitos subconscientes do autor, que faz a diferenciação com outros autores.

Quando aplicado em análise linguística escrita, o estilo se refere a variação linguística em que as formas variáveis são submetidas, tais como: gênero, períodos e situações. Por mais

que a relação entre estilo de escrita e estilo de fala compartilhem de um contexto social, o estilo da escrita é geralmente melhor definido por meio das características estruturais dos textos [BES88] [MEM94] [HEF95].

Para realizar a análise mais profunda do estilo, recorre-se a estilística, que tem por objetivo estudar cientificamente os processos do estilo, bem como aplicar métodos de análise para identificar o estilo de um autor [DIC14].

2.1.5.1 Estilística

A aplicação da estilística em casos de atribuição de autoria teve seus primeiros estudos desenvolvidos em meados do século 19. Os primeiros métodos desenvolvidos tinham como objeto de estudo a identificação de textos bíblicos e literários, cuja autoria era desconhecida ou contestada.

A estilística utiliza duas abordagens para a atribuição de autoria, que são [MCM02] [STA09]:

- **Qualitativa**: quando as características de escrita são identificadas, e posteriormente são descritas como características de um autor;
- **Quantitativa**: quando certas características são identificadas, e logo em seguida medidas, como por exemplo, a frequência relativa de um determinado conjunto de palavras. Os métodos quantitativos, também são conhecidos pelo termo estilometria (ver seção 2.1.5.2).

Os métodos qualitativos e quantitativos se complementam e podem ser utilizados em conjunto para identificar, descrever e mensurar a presença ou a ausência de características de estilo nos documentos questionados e conhecidos. Neste trabalho a abordagem mais presente e usual será a quantitativa.

A estilística tem como objetivo avaliar e identificar as características que tornam um texto ou parte deste, boa ou ruim para discriminar o estilo. Sendo assim, pode-se classificá-la também em prescritiva e descritiva. Prescritiva na medida em que especifica o que é correto, podendo ser útil em estudos de atribuição de autoria, pois pode ser usada para descrever a variação do estilo. Quanto ao método descritivo, reflete a aplicação de métodos analíticos, tal como identificar as características estruturais e funcionais, realizar a categorização e posteriormente analisar as regularidades [GAM04] [ARG07] [STA09]. Por conseguinte, as

frequências de ocorrências de características linguísticas são qualitativas, e muitas vezes podem ser quantificadas, fornecendo então uma base fundamental para a medição estatística do estilo, que é dada pela estilometria.

2.1.5.2 Estilometria

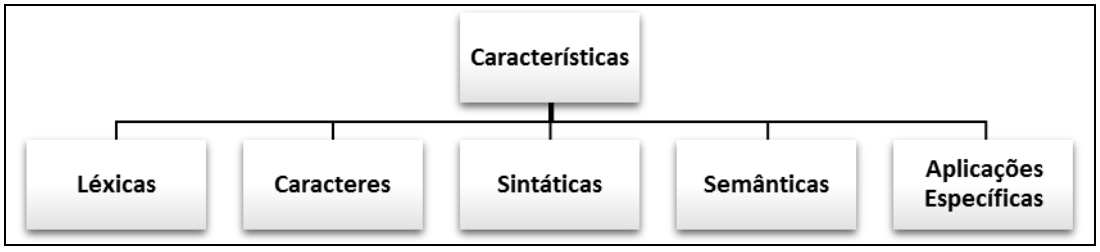
O termo estilometria tem origem do idioma grego, da junção de estilo + medição (*stylos* + *metron*) [BEL08] que significa o conjunto de qualidades características de um determinado autor quanto a sua maneira particular de escrever, ou seja, uma análise estatística do estilo da escrita [DIC14] [ZHE06]. A estilometria se transformou em um método cientifico para estudos do estilo linguístico estudado por diversos pesquisadores em diversas áreas, inclusive da identificação de autoria [BEL08].

A origem da estilometria remonta ao ano de 1851, quando o pesquisador e lógico inglês Augustus de Morgan sugeriu a um amigo que a autoria de um texto poderia ser resolvida por meio da comparação do tamanho das palavras usadas pelo autor, pois diferentes indivíduos certamente escrevem de formas diferentes [HOL98]. Este fato ocasionou o início dos estudos em estilometria e suas características, onde é possível fazer uso da técnica para verificar a real identidade de um determinado autor.

2.1.5.3 Características Estilométricas

As características estilométricas são atributos ou marcadores de estilo da escrita que são os mais eficazes discriminadores para atribuição da autoria [HOL98] [STA01] [KHM03] [ZHE06]. Existe uma grande variedade de recursos estilísticos que são utilizados no processo de atribuição de autoria pela computação (Figura 2.5), entre eles citam-se características: léxicas, baseada em caracteres, sintáticas, semânticas e de aplicações específicas [MEM93] [MEA95] [MAL06] [STA09].

Figura 2.5 – Variedade de Características Estilísticas



Fonte: autor (2017)

Características Léxicas: consideram um texto como uma simples sequência de símbolos ou caracteres. Foi com esta classe de características que os primeiros estudos estilométricos sobre identificação de autoria foram impulsionados. Conforme [STA09] as características léxicas podem ser subdivididas em algumas categorias que foram investigadas ao longo da história, que são: baseada em *tokens* (tamanho da palavra, tamanho da frase); frequência de palavras; riqueza de vocabulário; palavras *n-grams*; e, baseada em erros de ortografia [ZIP32] [YUL38] [YUL44] [MEW64].

Baseada em Caracteres: considera-se que um texto é uma simples sequencia de caracteres. Sendo assim, muitas medidas em nível de caracteres podem ser definidas, tais como: contagem de caracteres alfabéticos, contagem de digitos, quantidade de letras maiúsculas e minúsculas, frequencia de letras [DEV01] [ZHE06]. A abordagem por tipo de caractere é facilmente disponível para qualquer tipo de linguagem, e vem sendo bastante útil para quantificar o estilo de escrita [MOR65] [MAN95].

Características Sintáticas: utilizam a representação de um texto para fazer uso de métodos mais elaborados e complexos para análise da informação. O principal objetivo é extrair informações que os autores tendem a usar inconscientemente, e que possuam padrões sintáticos semelhantes. Então, a informação sintática é considerada mais confiável em comparação com os outros grupos de características em estudos de atribuição de autoria. São exemplos de características sintáticas: palavras-função, análise de partes da frase, estrutura das frases e regras de escrita. Maiores detalhes à respeito das características sintáticas usadas em atribuição de autoria são relatadas na seção 2.2.

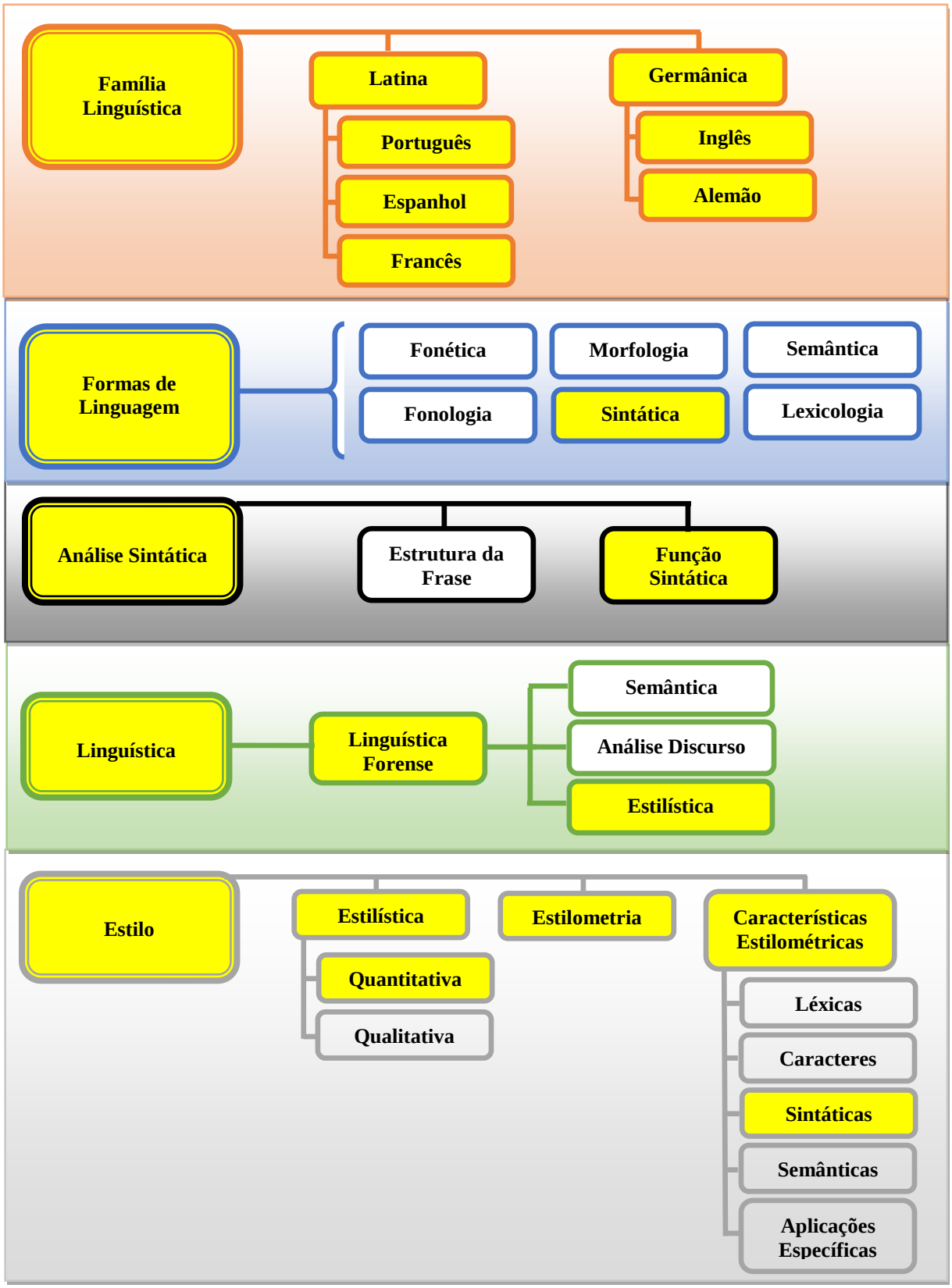
Características Semânticas: tem por função estudar o significado das palavras e frases. Os usos mais comuns da semântica são: dependências semânticas, sinônimos e hiperônimos, e associações funcionais. A literatura mostra que quando a informação

semântica é combinada com a informação léxica e sintática, melhora a precisão da classificação [GAM04].

Aplicações Específicas: características e medidas específicas de acordo com o conteúdo, a estrutura ou a linguagem para melhor representar nuances estilométricos de um determinado autor. Entre as características estruturais mais conhecidas, estão: a estrutura das palavras (que compreende algumas características, tais como: saudações iniciais e de encerramento, quantidade dos parágrafos, etc); e as características de ordem técnica (referentes a formatação – que podem evidenciar características importantes sobre o autor. Por exemplo, a formatação da fonte (tipos, tamanho, cor, alinhamento) [DEV01] [ABB05] [ZHE06]. O uso destes recursos só pode ser utilizado em determinados gêneros textuais, além do frequente uso em textos curtos, onde muitas vezes as outras abordagens não conseguem informações suficientes para uma boa classificação [ZHE06].

Em resumo, neste trabalho faz-se uso de textos escritos em línguas latinas e germânicas, tais como: portuguesa, espanhola, francesa, inglesa e alemã, que são pertencentes a família linguística indo-européia. Quanto as formas de linguagem, trabalhou-se com a forma sintática, por relatar uma maior riqueza de elementos textuais. Para realização da análise sintática utilizaram-se as funções sintáticas de cada palavra para denotar o padrão de escrita de cada autor. Perante os conceitos da linguística, fez-se uso da linguística forense e o uso da estilística. Já quanto ao estilo, trabalhou-se com a estilística de forma quantitativa, que é dada pela estilometria por meio das características sintáticas. Na Figura 2.6 é possível denotar o caminho aplicado (grifado em amarelo) para cada tópico dos elementos conceituais e teóricos.

Figura 2.6 – Caminhos dos Elementos Conceituais e Teóricos



Fonte: autor (2017)

Na próxima seção é apresentado o estado da arte sobre trabalhos que relatam a atribuição de autoria em ambiente multilíngue.

2.2 Estado da Arte

Um fator importante na avaliação de uma abordagem de atribuição de autoria é a sua capacidade de trabalhar com mais de uma língua, ou seja, de ser multilíngue. Vale ressaltar que muitas características estilométricas são dependentes do idioma, pois são propriedades únicas e pertencentes a somente um grupo. Em geral, as abordagens que utilizam características estilométricas podem ser transferidas para outras linguagens sem muitas adaptações. Diversos pesquisadores já desenvolveram seus estudos relatando os seus experimentos em diversas línguas. Na seção 2.2.1, são detalhados os estudos já realizados.

2.2.1 Estudos Relacionados

Em 1952, Fucks foi o primeiro pesquisador a realizar estudos da análise estilométrica em idiomas diferentes. Fez uso de 6 textos escritos por Shakespeare, Galsworthy e Huxley no idioma inglês; e 6 textos dos escritores Rilko, Carossa, Hesse, Mann e Jaspers no idioma alemão. As amostras dos textos variavam de 280 a 410 palavras no idioma inglês e de 460 à 890 no idioma alemão, não possuindo assim, uma padronização do número de *tokens* de cada amostra de texto. Fez uso de técnicas matemáticas e estatísticas para realizar as suas comparações entre os estilos adotados entre os autores, principalmente: frequência relativa, cálculo de distância e entropia. Realizou a contagem de letras, sílabas, palavras e frases em cada amostra de texto e descreveu a correlação e a distância média entre os autores. Mostrou a distribuição do tamanho e da média de sílabas por palavras de cada amostra, onde constatou que as frequências relativas das características estão diretamente associadas as características individuais de cada autor, o que pressupôs dizer que o número médio de sílabas por palavra é peculiar de um determinado autor. Percebeu também, que tal relação de ordem e coerência dos elementos textuais pode não ser discriminante em amostras de textos mais complexos, porém nos textos analisados uma certa sequência de números pode caracterizar certas nuances de propriedades individuais dos autores, podendo assim, definir a autoria ou o idioma de um determinado texto. Ao final, verificou que os valores de entropia eram maiores nas amostras

dos textos escritos em alemão do que os textos escritos em língua inglesa, o que poderia ser um dado discriminante [FUC52].

Em 2003, Peng *et al*. desenvolveram seu trabalho utilizando caracteres *n-grams* para identificação de autoria nas línguas inglesa, grega e chinesa. Propuseram uma abordagem simples para alcançar o melhor desempenho, descartando o uso demasiado de recursos e de pré-processamento. Fizeram uso da modelagem da linguagem por meio de recursos estatísticos (entropia) e incluindo a teoria da decisão bayesiana, onde representaram um modelo para cada autor, representando uma abordagem simples de atribuição de autoria independente de idioma. A base de dados em língua grega foi coletada em sites de jornais gregos, perfazendo dois conjuntos de dados anteriormente testados por [STA99] e [STA01]. Tais conjunto de dados foram categorizados em assuntos gerais (grupo A), e assuntos relacionados a ciência, história e cultura (grupo B). São 10 diferentes autores com 20 amostras de texto de cada autor em cada conjunto de dados. Fizeram uso de 10 amostras de cada autor para treinamento e as outras 10 amostras para testes. Os resultados apontaram que o uso de caracteres *3-gram* atingiu o melhor resultado (grupo A – 74%, e grupo B - 90%), inclusive com um ganho considerável na comparação com os trabalhos desenvolvidos por [STA99] e [STA01]. O conjunto de dados testados na língua inglesa foram textos de 8 autores da literatura de língua inglesa disponível em domínio público na internet. Para reduzir a dimensionalidade do modelo, utilizaram somente as 30 sequencias de caracteres mais frequentes, que representa mais de 99% das ocorrências de caracteres no corpus. Os resultados em língua inglesa apontaram que o modelo de *6-grams* se mostrou mais eficaz, obtendo uma precisão média de 98%. Já para o conjunto de dados em chinês (textos de 8 escritores chineses coletados na internet em domínio público), foram utilizados 2 amostras de textos para treinamento e 20 amostras para os testes (a diferença entre as amostras de treinamento deve-se ao fato de que a língua chinesa possui um vocabulário muito maior do que a inglesa e a grega). Também fizeram uso da redução da dimensionalidade, utilizando 2500 caracteres da língua chinesa, que corresponde a cerca de 99% das ocorrências identificados nas amostras. Os melhores resultados para a língua chinesa foram pelo uso de *3-grams*, que obteve o resultado de 94%. Entre as conclusões que os pesquisadores chegaram, estão: os caracteres *n-grams* se mostram simples e eficazes; *n-grams* podem ser aplicados em qualquer tipo de linguagem, incluindo sequencias musicais e cadeias genéticas; em linguagens asiáticas, onde a segmentação das palavras é maior, a aplicação ainda carece de mais estudos.

Perceberam que o método pode ser influenciado por alguns fatores, que pode afetar a precisão da abordagem, para tanto relatam a falta de experimentos e trabalhos em outras línguas e com outros aspectos tanto de técnicas, como de conjunto de dados [PEN03].

Abbasi e Chen (2005) utilizaram a técnica de aprendizagem de máquina, por meio dos classificadores J4.8 e SVM para realizar a classificação de textos extraídos de páginas de troca de mensagens na internet (fóruns) em língua inglesa e árabe, que relatavam suspeita de troca de mensagens entre prováveis terroristas e criminosos. O processo usado constou de coleta da base de dados, extração de características e realização de experimentos. O processo de coleta se deu pelo uso de softwares de busca na web, que identificava possíveis atos criminosos. Posteriormente, tais amostras de textos foram armazenadas em formato texto ou HTML. Sendo assim, a base de dados foi composta por 30 amostras de textos pertencentes à 5 autores distintos de cada idioma. No idioma inglês fizeram uso de 301 características, incluindo 87 léxicas, 158 sintáticas, 45 estruturais e 11 de conteúdo específico. No idioma árabe, realizaram os testes com 418 características, sendo 78 léxicas, 262 sintáticas, 62 estruturais e 15 de conteúdo específico. Realizaram a categorização das características para realização dos experimentos, ou seja, testaram primeiramente os grupos (léxico - contagem de palavras curtas, distribuição do tamanho das palavras e número de alongamentos (somente no idioma árabe); sintático - palavras função e origem da palavra (idioma árabe); estrutural - pontuação, tamanho e cor da fonte, imagens e links anexados; e de conteúdo específico), em separado e depois no conjunto. Perceberam que o classificador SVM obteve melhores resultados em ambos os idiomas, chegando a 97% de acurácia no idioma inglês e 94,8% no idioma árabe no teste conjunto de todas as características. Realizaram a otimização do processo por meio dos testes de Pairwise. Relatam que conforme as categorias são agrupadas o ganho do classificador é maior, sendo que as categorias léxicas e sintáticas refletem os melhores resultados. Ao final, realizaram comparação entre as duas línguas testadas, para verificar se o método é eficiente. Perceberam que o conjunto de características multilíngue aplicado teve resultados significativos para um grupo pequenos de autores, porém quando o número de autores aumenta para centenas, ainda é necessário o desenvolvimento de metodologias mais complexas. Um dos pontos negativos desta abordagem, é que as línguas inglesa e árabe não possuem a mesma origem, fazendo com que as características sejam diferentes para cada idioma e não unificadas em um único modelo. Um outro dado importante, que não é relatado na experiência é o número médio de palavras de cada amostra

de texto analisado, então, não se sabe ao certo se trabalharam com textos curtos ou longos [ABB05].

Em 2006, Zheng *et al.* realizaram estudos acerca de análise estilométrica em mensagens online captadas na internet nos idiomas inglês e chinês. Trabalharam com os quatro grandes grupos de características estilométricas, envolvendo um grupo de 270 características, que são: léxicas (87), sintáticas (158), estruturais (14) e de conteúdo específico (11). Para tanto, fizeram uso de técnicas de aprendizagem de máquina para realização das extrações de características por meio de algoritmos de aprendizagem indutivos, com o intuito de construir modelos de classificação eficientes para identificação de autoria. O processo consistiu em criação da base de mensagens, extração de características, geração do modelo e aplicação dos testes de identificação de autoria. Como base de dados, foram usadas amostras de 20 autores de língua inglesa que versavam em suas mensagens sobre *cybercrime*. O número de amostras variou entre 30 e 92, sendo o tamanho das amostras entre 84 e 346 palavras. Já para o idioma chinês, não foi possível criar um conjunto de dados comparável ao idioma inglês, neste caso, foram coletadas mensagens online sobre 7 temas diferentes (filmes, músicas, esportes, viagens, beleza, amor e romances). Sendo assim, foram coletadas mensagens escritas em chinês de 20 autores distintos, sendo que para cada autor foram usadas de 30 a 40 amostras de textos, com uma média de 807 palavras por texto. Para realização dos testes foram utilizadas 3 abordagens diferentes, que são: árvores de decisão (J4.8), redes neurais (*backpropagation*), e máquinas de vetor de suporte (SVM). Os resultados demonstraram que a abordagem proposta foi capaz de identificar autores de mensagens online com precisão entre 70-95%. Registraram que os quatro grupos de características estilométricas contribuíram de alguma forma para a discriminação dos autores, porém relatam que as características léxicas em separado não são eficazes para identificação de autoria no idioma chinês (52-57%), diferentes dos resultados proporcionados na língua inglesa (78-89%). Em uma análise dos resultados das características sintáticas, perceberam que quando adicionadas tais características aos testes em língua inglesa, os classificadores não mostraram resultados significativos, porém quando aplicados à base de dados em chinês, o resultado melhorou significativamente. Sendo assim, concluíram, que palavras-função não se saíram bem com textos curtos (língua inglesa), até mesmo porque, o número de palavras usada em uma mensagem online é pequeno, enquanto que o grupo de características usado como base era muito grande. Neste caso, aconselham realizar mais pesquisas para identificar um

conjunto adequado de palavras-função para cada tamanho de texto. Já quando o grupo sintático foi aplicado na base de textos em chinês, que continha textos relativamente maiores, tais características retornam resultado mais eficazes, mostrando que em textos mais longos, este grupo de características possui maior valor discriminante. Quando realizaram a inserção do grupo de características estruturais, o desempenho melhorou em ambas as linguagens, mostrando que o grupo estrutural parece ser um bom discriminante entre autores de mensagens online. Ao final realizaram a adição do grupo de características de conteúdo específico, e verificaram a melhoria do desempenho do conjunto de dados da língua inglesa, porém para o conjunto em língua chinesa a melhoria foi verificada em um nível pouco significativo. O destaque ficou para o SVM que superou as outras duas técnicas nos experimentos realizados. Um fator negativo, que pode prejudicar o modelo, é que o mesmo só pode ser aplicado para textos em língua inglesa e chinesa, pois as características de estilo das duas linguagens são completamente diferentes, até porque não possuem a mesma origem, podendo dificultar a adaptação do modelo para outras linguagens [ZHE06].

Em 2007 Zhang *et al*. aplicaram recursos sintáticos e semânticos em complementos verbais anteriores e posteriores a palavra função, em textos nos idiomas chinês, japonês e indiano (línguas orientais). Segundo os pesquisadores, conforme seus estudos em processamento de linguagem natural, não se pode utilizar somente regras sintáticas, e deve-se fazer uso de alguns recursos semânticos para complementar os resultados. Sendo assim, escolheram a palavra função “você” para realizar seus estudos, pois tal palavra possui a mesma pronuncia e escrita, porém significados diferentes conforme disposição em um texto nas línguas pesquisadas. O estudo se aprofundou principalmente nas diferentes características semânticas da palavra antecessora (verbos) que pode implicar que a palavra você possa assumir diferentes estruturas (polissemia), impactando no significado da palavra sucessora (substantivo). Para tanto, fizeram uso da abordagem de árvores de Copenhagen (CTT), que efetua o rastreamento da função das palavras em uma frase. Como base de dados, fizeram uso de corpus de pequenos textos da Universidade da China. Ao final, perceberam que a abordagem utilizada se mostra eficaz na língua chinesa, sendo assim, melhor aplicável do que nas outras línguas. Alguns pontos negativos do trabalho é que não são relatados a quantidade de autores, o número de amostras de cada autor, nem mesmo resultados claros nos comparativos de cada idioma testado. Também não são abordados os ganhos do uso desta abordagem em comparação com outras abordagens no que tange a atribuição de autoria.

Porém, relatam que ainda são necessários diversos estudos para aprimoramento da análise [ZHA07].

Em 2013, Savoy descreveu, avaliou e comparou o *Latent Dirichlet Allocation* – LDA, como uma abordagem para atribuição de autoria. Utilizou como base de dados para os experimentos, artigos extraídos de jornais escoceses e italianos. Foram extraídos 5408 artigos em língua inglesa escritos por 20 autores distintos, e 4326 artigos em língua italiana, também escritos por 20 autores diferentes. As amostras dos autores variaram de no mínimo 30 e no máximo 434 artigos por autor, sendo que o número mínimo de *tokens* é de 44 e o máximo de 4414, perfazendo uma média de 748 *tokens* por texto. Todos as amostras foram categorizadas por assunto, tais como: arte, politica, sociedade, negócios e esportes. Como pré-processamento, transformou todas as letras maiúsculas em minúsculas; nos artigos de língua inglesa retirou todos os acentos e símbolos diacríticos, já em língua italiana os acentos e diacríticos foram mantidos. Realizou comparações com diversas abordagens, entre eles: Delta, qui-quadrado, *Kullback-Leibler Divergence* - KLD e Naive Bayes. Na comparação com a regra Delta, extraiu as 400 palavras mais frequentes nos dois idiomas. Nos experimentos com qui-quadrado utilizou 653 características em inglês e 720 em língua italiana. Quando fez uso do KLD, utilizou 363 palavras em inglês e 399 em italiano. E, quando realizado os experimentos com o Naive Bayes, foram extraídas os 500, 1000, 2000 e 5000 termos mais frequentes. Em todos os experimentos realizados no LDA em comparação com as outras abordagens, foi feito uso de 20, 40, 50, 60 e 80 palavras mais comuns encontrados em cada língua. Como resultado da comparação efetuada, percebeu que a abordagem proposta pelo LDA resulta em melhores níveis de desempenho que a regra Delta e o qui-quadrado. Porém, quando comparado com o modelo KLD, a abordagem proposta mostrou desempenho mais baixo. Já quando comparado ao Naive Bayes, a abordagem do LDA pode demonstrar melhores resultados se bem ajustados os parâmetros em ambos os idiomas, porém em 62,5% dos experimentos o Naive Bayes se mostrou mais eficaz. Concluiu que o LDA superou a regra delta e a distância qui-quadrado com testes em um número restrito de termos. Já quando testado com um número elevado de termos a eficácia foi melhor que o método KLD [SAV13]. Ainda neste mesmo trabalho, Savoy efetuou uma análise a respeito do estilo de cada autor, mostrando o relacionamento de cada palavra com os assuntos relacionados nos artigos, porém faltaram detalhes mais aprofundados sobre tal relação.

Em 2013, Ebrahimpour *et. al.* aplicaram duas abordagens de atribuição de autoria em experimentos com textos em língua inglesa e grega. Fizeram uso de uma abordagem baseada na Análise Discriminante Múltipla – MDA e uma baseada no SVM, que foram testadas em ambas as bases de textos. Todos os experimentos são baseados nas frequências de palavras-função, pois segundo [STA09] [ZHO05] [EBR13], estas são algumas das melhores características para realizar a atribuição de autoria, pois a mesma se baseia em categoria de palavras. Como essas categorias possuem pouca ou nenhuma dependência em relação ao tema ou gênero dos textos, esta técnica pode ser aplicada em diferentes línguas e a diferentes tipos de textos. Realizaram um pré-processamento nos textos, retirando todas as informações que não fossem letras e espaços, justificando que esse processo é essencial para prover a portabilidade do método para diferentes tipos de textos. No processo de extração de características utilizaram um software para montar os vetores com as palavras-função escolhidas de forma normalizada. Para medir a precisão do método fizeram uso da validação cruzada. Realizaram os testes com uma base de 168 textos de 7 autores em língua inglesa, e posteriormente na base de 85 textos do *Federalist Papers.* Para complementar aplicaram o método em 32 textos gregos (cartas ao Hebreus). Na base de 168 textos, limitaram a frequência das primeiras 5000 palavras, e fizeram uso das palavras-função mais comuns, limitadas a 100 palavras-função, onde atingiram acurácia de 96,4% (100 palavras-função) de acerto pelo uso do MDA e 92,2% (95 palavras-função) com o SVM. Quando realizado os testes com a base do *Federalist Papers*, conseguiram uma classificação correta com o MDA de 97,1% e de 95,6% com o SVM, sendo que, em ambos os métodos o número de 75 palavras-função atingiu o mais alto nível de precisão. Na base de textos escritos em língua grega, os pesquisadores tiveram que concatenar todas as amostras de textos e dividi-la em 4 amostras, para que as amostras não ficassem tão diferentes quanto ao número de palavras, pois inicialmente tinham amostras com 5.000 palavras e amostras com 50.000 palavras. Em língua grega o método do MDA apresentou acurácia de 90,6% e o SVM de 87,5%. Concluem, evidenciando que as palavras-função são boas características para atribuição de autoria. Relatam que os passos essências para um bom método é realizar o pré-processamento dos textos, escolher e extrair as características e posteriormente realizar os experimentos utilizando algum método de classificação. Sobre as abordagens, perceberam que ambas as abordagens retornam resultados maiores que 90% de acurácia, sendo que o SVM se mostra um pouco limitante, pelo motivo de fornecer apenas decisões binárias. Já o MDA, permite

uma maior flexibilidade, permitindo o uso mais eficaz de técnicas estatísticas e de probabilidade, tanto é que pode ser utilizado em métodos para investigar o grau de colaboração entre autores [EBR13].

Frery *et. al.* (2014) apresentaram uma proposta baseada em aprendizagem de máquina para a tarefa de identificação de autoria. Fizeram uso de árvores de decisão para tomada de decisão. Usaram 8 conjuntos de atributos envolvendo caracteres e palavras *n-grams*, riqueza de vocabulário, sinais de pontuação e tamanho das frases. Trabalharam com a resolução de problemas em língua inglesa, espanhola, grega e holandesa. Sua proposta estimou resultados entre 62-71% em língua inglesa, 60-90% em língua holandesa, 77% em língua espanhola e 68% em língua grega. Segundo os autores, os resultados obtidos foram consistentes para o uso de uma árvore de decisão, já que é uma abordagem rápida e permite identificar como uso da predição, boas características [FRE14].

Potha e Stamatatos (2014) usaram o paradigma baseado no perfil do autor para verificação de autoria em inglês, espanhol e grego. Neste caso, todas as amostras de textos de um mesmo autor são tratadas de forma cumulativa, ou seja, são concatenadas em um grande documento e em seguida, uma única representação é extraída para formação do perfil do autor. Como atributos discriminantes usaram as mais frequentes *n-grams* (CNG – *Common N-Grams*). Usaram uma função de dissimilaridade como limiar de classificação. O corpus de textos da base foi dividido em 50% para treinamento e 50% para testes. Em língua inglesa atingiram resultados entre 79-88%, em grego a taxa de acerto média foi de 79%, e em espanhol cerca de 92% [POT14].

Ainda em 2014, Juola e Stamatatos reuniram em um único artigo os resultados de 18 equipes participantes do desafio de atribuição de autoria proposto pelo PAN/CLEF 2013. Apresentaram os métodos e as medidas de desempenho dos autores em um corpus de textos em língua inglesa, espanhola e grega. O melhor resultado para língua inglesa obteve taxa de precisão de 84%. Em língua grega 82% e em língua espanhola 93% [JUO14].

Em 2016 Halvani *et al.* apresentaram um método escalável para verificação de autoria em diferentes idiomas, gêneros e tópicos. Usaram corpus de testes de diferentes linguagens, incluindo: holandês, inglês, grego, espanhol e alemão. Propuseram um método fornecendo um limite universal por linguagem, sendo este usado para aceitar ou rejeitar a autoria de um documento questionado. O fator universal refere-se à capacidade de generalizar por meio de diferentes gêneros e tópicos dos textos. O modelo gerado é flexível podendo ser estendido de

forma incremental para lidar com novos idiomas, gêneros ou atributos. O método não envolve técnicas de processamento de linguagem natural e nem de aprendizagem de máquina, apresentando uma baixa complexidade computacional. Usaram 9 categorias de atributos envolvendo: pontuação *n-grams*, caracteres *n-grams*, palavras mais frequentes, prefixos, sufixos, prefixos e sufixos *n-grams*, prefixos-sufixos e sufixos-prefixos. Atingiram uma acurácia média de 75% [HAL16].

Na Tabela 2.1 é possível observar um resumo dos trabalhos desenvolvidos em atribuição de autoria com abordagem multilíngue.

Tabela 2.1 – Resumos de Trabalhos Multilíngue

Autor (es) /Ano	*Corpus*	*Número Autores*	*Amostras por autor*	*Tamanho (N° de palavras)*	*Método*	*Características utilizadas*	*Idioma*	*Resultados*
Fucks (1952)	Escritores em língua inglesa e alemã	3 5	2 1-2	280-410 460-890	Distância, entropia, frequência relativa	Léxicas (contagem de letras, palavras e frases).	Inglês Alemão	N/D N/D
Peng et. al (2003)	Textos literários e de colunas de jornais	20 8 8	20 N/D 22	N/D N/D N/D	Entropia, Bayes	Caracteres *n-grams.*	Grego Inglês Chinês	74-90% 98% 94%
Abbasi e Chen (2005)	Mensagens online	5 5	30 30	N/D N/D	SVM e J.48	Palavras-função, léxico, estruturais e de conteúdo.	Inglês Árabe	97% 94,8%
Zheng et al. (2006)	Mensagens online	20 20	30-92 30-40	84-346 807 (média)	SVM, J4.8 e Redes Neurais	Palavras-função, léxico, estruturais e de conteúdo.	Inglês Chinês	78-89% 52-57%
Zhang et al. (2007)	Frases curtas	N/D	N/D	N/D	Arvores de Copenhagen	Complementos verbais anteriores e posteriores à palavra-função (verbos)	Chinês Japonês Indonésio	N/D N/D N/D
Savoy (2012)	Artigos de jornais	20 20	30-433 52-434	44-4414 60-2935	Análise discriminante	Palavras-função	Inglês Italiano	82% 91%
Ebrahimpour et. al. (2013)	Livros em inglês *Federalist Papers* Cartas aos Hebreus	7 3 8	14-26 3-51 4	5000 N/D 1600-10000	Análise multivariada e vetores de máquina de suporte	Palavras-função	Inglês Inglês Grego	92-96% 95-97% 87-90%
Frery et. al (2014)	Diversos	N/D	N/D	N/D	Árvores de decisão	*n-grams*	Inglês Holandês Espanhol Grego	62-71% 60-90% 77% 68%
Potha e Stmatatos (2014)	PAN/CLEF	16	5-30	500-3000	Medidas de distância	*n-grams*	Inglês Espanhol Grego	79-88% 92% 79%

Juola e Stamatatos	PAN/CLEF	16	5-30	500-3000	Diversos	Diversos	Inglês	84%
							Espanhol	93%
							Grego	82%
Halvani et al. (2016)	Diversos	15-2000	N/D	N/D	Medidas de distâncias	Léxicas, caracteres *n-grams* e sintáticas	Holandês	73%
							Inglês	77%
							Grego	67%
							Espanhol	83%
							Alemão	78%

Fonte: autor (2017) N/D - Não Declarado.

2.3 Considerações do Capítulo

Neste capítulo, inicialmente foi apresentado as principais áreas de conhecimento necessárias para a compreensão da proposta deste trabalho em relação à atribuição de autoria. Sabe-se que existem muitos casos de atribuição de autoria que a linguística computacional poderá auxiliar a resolver, principalmente quando a autoria de um determinado texto ou documento é questionada. Diante disso, diversos conhecimentos devem ser postos em prática por parte de um perito ou linguista para solucionar o caso.

Uma das primeiras ações é tentar identificar características discriminantes que possam diferenciar um autor de outro. Para que isso seja possível, é necessário compreender traços do comportamento humano, principalmente por meio de seus sistemas de comunicação, neste caso, a linguagem. Em especial, o entendimento da comunicação escrita e sua relação com o estilo de cada autor ou de um grupo de autores. Geralmente o estilo de um autor está ligado a alguns fatores que podem auxiliar no processo de atribuição de autoria, tais como: classes sociais, regionalismo e escolaridade. Para mensurar o estilo de cada autor recorre-se ao uso da estilometria e suas diversas características, entre elas, as sintáticas. As características sintáticas por sua vez denotam o estilo de cada autor pelo uso inconsciente de certas palavras e estruturas que o diferenciam dos demais. As principais classes de características sintáticas são palavras-função, análise de parte da frase e regras de escrita. Entretanto, propõe-se o uso de características sintáticas baseado nas funções sintáticas e em seus níveis estruturais, mediante a identificação e extração dos termos essenciais, integrantes e acessórios que compõe cada frase de um texto.

Por conseguinte, foi apresentado o estado da arte de trabalhos sobre atribuição de autoria em um ambiente multilíngue. Sabe-se que muitas abordagens já foram utilizadas em casos que envolvam a descoberta da autoria em documentos e textos questionados, porém, ainda muitos casos não foram solucionados por falta de métodos confiáveis e objetivos.

No desenvolvimento do estado da arte levou-se em considerações duas premissas básicas: primeiramente o uso de níveis estruturais da gramática da língua, ou seja, características sintáticas que possam ser utilizadas em separado ou em conjunto. E, consecutivamente a criação de uma abordagem mista que possa atender as línguas portuguesa, espanhola, francesa, alemã e inglesa, ou seja, multilíngue. Tais premissas justificam a criação da proposta deste trabalho (mais detalhes no Capítulo 4).

Diversos pesquisadores relatam a falta de trabalho em outras línguas que não seja a língua inglesa, bem como outras abordagens e outros conjuntos de dados [PEN03] [ABB05] [SAV12] [EBR13]. Neste caso, este trabalho se compromete a trabalhar com as características da língua com uma abordagem baseada em níveis sintáticos e com a formação de novas bases de dados composta por textos de obras literárias disponíveis em domínio público (para maiores detalhes consultar o capítulo 3).

Capítulo 3

Metodologia

Neste capítulo é apresentada a metodologia usada no desenvolvimento do trabalho. Inicialmente são apresentados as bases de textos e o conjunto de atributos utilizados para realização dos experimentos. São informadas as fontes de coleta, de disponibilização e de organização do conjunto de textos que compõe as bases. Quanto ao conjunto de atributos, detalham-se as características usadas no processo experimental. Em correlato, apresenta-se o software de rotulagem das palavras e o processo de transformação de informações textuais em informações numéricas.

3.1 Informações das bases de dados

Os documentos encontrados em casos que envolvem atribuição de autoria, geralmente são de diferentes tamanhos, variando de um texto curto (com poucas dezenas de palavras) a um texto mais longo (com centenas e milhares de palavras). Para tanto, não existem muitas bases de dados de acesso público e que permitam o desenvolvimento mais robusto de experimentos. Sendo assim, optou-se por duas frentes: a primeira consta de utilizar bases de dados já existentes e que estejam disponibilizadas para o acesso. Neste caso, utilizamos duas bases de dados de colunistas de jornais, sendo uma em língua portuguesa e uma em língua inglesa que são detalhadas na seção 3.1.1, as quais, chamamos de textos jornalísticos ou textos curtos. A segunda opção foi a criação da própria base de dados, que constam de textos literários em 5 idiomas diferentes (português, espanhol, francês, alemão e inglês), a qual chamamos de textos literários ou textos longos (ver detalhes na seção 3.1.2).

A base de dados em idiomas diferentes se dá pelo fato da realização de experimentos com as estruturas sintáticas de diferentes linguagens, a fim de verificar as semelhanças e as diferenças proporcionadas pelas mesmas estruturas em idiomas distintos. Todas as bases de dados utilizadas nos experimentos desenvolvidos neste trabalho, bem como a lista dos autores estão disponíveis para acesso em http://paginapessoal.utfpr.edu.br/paulovarela/databases.

3.1.1 Textos Jornalísticos

As bases de dados de textos jornalísticos são formadas por textos extraídos de colunas de jornais. São textos curtos, com média de 500 palavras por amostra. Para realização dos experimentos foram utilizadas duas bases de dados já existentes e disponibilizadas por [VAR10] e [PET04]. Para uma melhor organização das bases de textos, efetuou-se a divisão das bases pelo idioma, conforme pode ser visto na Tabela 3.1.

Tabela 3.1 – Resumo das Bases de Dados de Textos Jornalísticos

Idioma	Quantidade de Autores	Amostras por autor	Fonte
Português	100	30	Colunas de Jornais Brasileiros
Inglês	20	20	Colunas de Jornais Britânicos

Fonte: Autor (2017)

A base de dados em língua portuguesa é composta por textos extraídos de colunas de jornais entre os anos de 2008 e 2009 dos principais jornais brasileiros. É constituída de 100 autores diferentes com 30 amostras de textos por autor, e também separados em 10 classes de assuntos. Maiores informações podem ser consultadas em [VAR10]

A base de dados de textos em língua inglesa é a CLEF-2003 [PET04], e utilizada por exemplo, no trabalho de [SAV13]. A CLEF consta de textos curtos, extraídos de colunas de jornais em língua inglesa, pertencentes a 20 autores distintos, e informações adicionais sobre a base podem ser vistas em [PET04].

3.1.2 Textos Literários

As bases de dados de textos literários foram criadas especificamente por meio deste trabalho. São textos extraídos de obras literárias disponibilizadas em domínio público nos

idiomas português, espanhol, francês, alemão e inglês. Todas as amostras de textos possuem mais de 500 palavras e pertencem a autores da literatura mundial. As bases de dados de textos literários foram divididas por idioma, sendo que em todas, temos 100 autores distintos com 20 amostras de textos por autor como pode ser observado na Tabela 3.2.

Tabela 3.2 – Resumo das Bases de Dados de Textos Literários

Idioma	Quantidade de Autores	Amostras por autor	Fonte
Português	100	20	Domínio Público na internet (repositórios de obras literárias citadas em 4.3.1)
Espanhol	100	20	
Francês	100	20	
Alemão	100	20	
Inglês	100	20	

Fonte: Autor (2017)

Em língua portuguesa foram coletadas mais de 250 obras literárias em parte ou completas de 100 autores distintos, entre eles: Machado de Assis, Eça de Queiroz, Érico Veríssimo, Aluísio de Azevedo entre outros.

Para a base de textos em língua espanhola foram coletados textos de escritores de diversos países, principalmente da América Latina e da Europa. As obras literárias em espanhol foram coletadas principalmente no sítio do projeto Gutemberg (domínio público). Foram coletadas obras literárias de 100 autores diferentes, entre eles: Concha Espina, Garcilaso de La Vega, Roberto Bollanos, entre outros.

Em língua francesa são mais de 160 obras literárias originárias do idioma francês. Foram 100 autores com nacionalidade europeia, americana e africana. Entre estes autores pode-se destacar alguns, tais como: Molière, Júlio Verne e Alexandre Dumas.

Para a língua alemã, foram coletadas mais de 150 obras literárias de autores europeus que escreveram suas obras no idioma alemão. Entre estes autores, pode-se citar: Franz Kafka, Hermann Hesse e Johann Von Goethe.

A base de dados em idioma inglês consta de obras literárias de diversos autores da língua inglesa (principalmente de países como a Inglaterra e os Estados Unidos da América). Foram coletadas 192 obras literárias pertencentes a 100 autores distintos, entre eles: Agatha Christie, Isaac Asimov, Ian Fleming, entre outros.

O processo de garimpagem das obras literárias na internet para cada uma das línguas demorou cerca de 90 dias com dedicação média de 2 horas diárias.

Como pode ser denotado as obras coletadas para as bases de textos literários foram escritas e publicadas em diferentes épocas. Como a proposta é trabalhar com uma centena de autores por idioma, isso se fez necessário pelo fato de não ter disponível a quantidade de autores em um mesmo período literário, ainda mais, levando em consideração que as obras devem estar em domínio público. No entanto, selecionou-se somente os autores mais contemporâneos e que tiveram suas obras consagradas pela literatura.

Cita-se que um ponto a ser levado em consideração quanto as bases de textos, são as diversas reformas ortográficas e gramaticais ocorridas em cada uma das línguas no decorrer dos anos. Pode-se citar como exemplo, que um texto escrito no século XVIII possui diferenças ortográficas para um texto escrito nos dias atuais. No entanto, como a grande maioria das reformas linguísticas tratam da grafia e de aspectos gramaticais, não houveram alterações significativas nas funções sintáticas que pudessem inviabilizar o conjunto de atributos propostos, e, contudo, tais fatores não geraram interferências críticas que pudessem prejudicar o andamento dos experimentos.

Para coleta e tratamento dos textos que compõe as bases de dados, seguiu-se um protocolo para auxiliar no processo de captação e organização dos textos, o qual é descrito na subseção 3.1.3.

3.1.3 Pré-Processamento dos textos

A fonte de captação destes textos foi à rede mundial de computadores (internet). Foram visitados portais e sites de repositório de obras literárias, tais como [BND16] [BIB16] [DMP16] [EST16] [GUT16] [LIT16]:

- Biblioteca Digital Nacional;
- Biblio – Biblioteca Virtual de Literatura;
- Biblioteca de Literatura Brasileira;
- Portal Domínio Público; e
- Biblioteca Digital do Projeto Gutenberg.

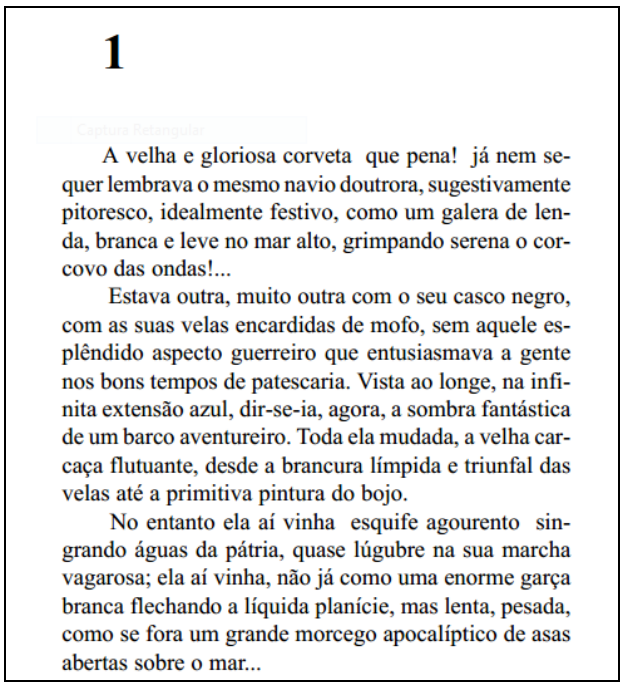
Para auxiliar no processo de garimpagem de textos e organização da base de dados foi elaborado um protocolo para a coleta dos textos, que incluem as seguintes orientações:

- Os textos obrigatoriamente serem pertencentes às obras literárias;
- Estar em domínio público e disponibilizado no formato digital;

- Possuir no mínimo 100 palavras, perfazendo o todo ou parte de capítulos das obras literárias;
- Versar em contos, histórias e novelas, ou seja, textos mais longos. Neste caso, foram descartados poemas e poesias pelo fato de diversas amostras possuírem um pequeno número de palavras;
- Realizar de pré-processamento dos textos, para eliminar as sujeiras textuais, tais como: cabeçalhos, numeração de páginas e outros elementos que não constam na obra original e que poderiam influenciar nos resultados de análise do estilo do autor;

Na maioria dos casos, os textos encontrados em domínio público estavam no formato PDF – *Portable Document Format* e HTML – *Hyper Text Markup Language*. Sendo assim, estes textos sofreram um pré-processamento, onde foram transformados para o formato de arquivo texto, de onde, posteriormente são extraídas as informações para gerar os vetores de características. Um exemplo destes textos encontrados de forma bruta está na Figura 3.1, onde é apresentada parte do texto da obra: O bom crioulo, de Adolfo Caminha.

Figura 3.1 – Amostra de texto disponibilizado em domínio público



1

A velha e gloriosa corveta que pena! já nem sequer lembrava o mesmo navio doutrora, sugestivamente pitoresco, idealmente festivo, como um galera de lenda, branca e leve no mar alto, grimpando serena o corcovo das ondas!...

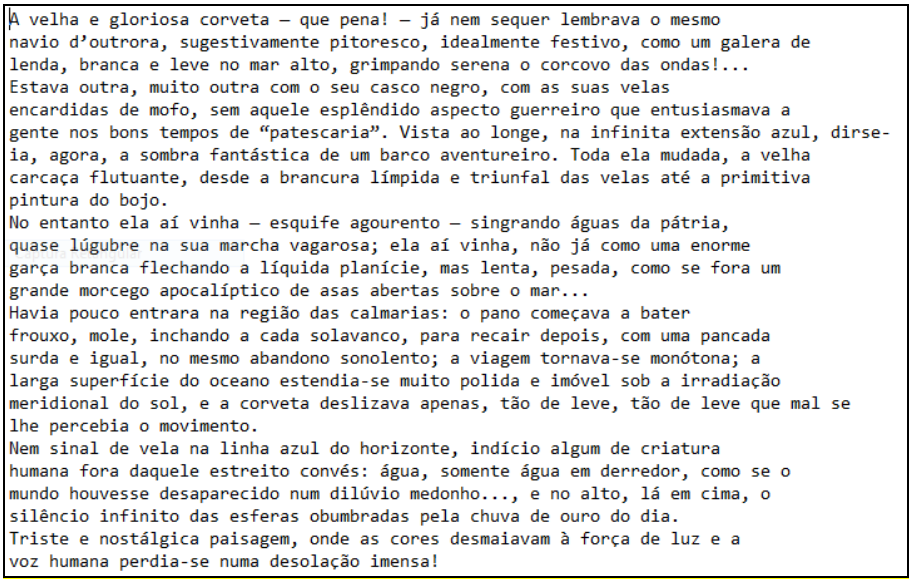
Estava outra, muito outra com o seu casco negro, com as suas velas encardidas de mofo, sem aquele esplêndido aspecto guerreiro que entusiasmava a gente nos bons tempos de patescaria. Vista ao longe, na infinita extensão azul, dir-se-ia, agora, a sombra fantástica de um barco aventureiro. Toda ela mudada, a velha carcaça flutuante, desde a brancura límpida e triunfal das velas até a primitiva pintura do bojo.

No entanto ela aí vinha esquife agourento singrando águas da pátria, quase lúgubre na sua marcha vagarosa; ela aí vinha, não já como uma enorme garça branca flechando a líquida planície, mas lenta, pesada, como se fora um grande morcego apocalíptico de asas abertas sobre o mar...

Fonte: Autor (2017)

Na Figura 3.2 é evidenciado o texto já tratado (em formato .txt) e pronto para uso no processo de extração de características.

Figura 3.2 – Amostra de texto tratado


A velha e gloriosa corveta – que pena! – já nem sequer lembrava o mesmo
navio d'outrora, sugestivamente pitoresco, idealmente festivo, como um galera de
lenda, branca e leve no mar alto, grimpando serena o corcovo das ondas!...
Estava outra, muito outra com o seu casco negro, com as suas velas
encardidas de mofo, sem aquele esplêndido aspecto guerreiro que entusiasmava a
gente nos bons tempos de "patescaria". Vista ao longe, na infinita extensão azul, dirse-
ia, agora, a sombra fantástica de um barco aventureiro. Toda ela mudada, a velha
carcaça flutuante, desde a brancura límpida e triunfal das velas até a primitiva
pintura do bojo.
No entanto ela aí vinha – esquife agourento – singrando águas da pátria,
quase lúgubre na sua marcha vagarosa; ela aí vinha, não já como uma enorme
garça branca flechando a líquida planície, mas lenta, pesada, como se fora um
grande morcego apocalíptico de asas abertas sobre o mar...
Havia pouco entrara na região das calmarias: o pano começava a bater
frouxo, mole, inchando a cada solavanco, para recair depois, com uma pancada
surda e igual, no mesmo abandono sonolento; a viagem tornava-se monótona; a
larga superfície do oceano estendia-se muito polida e imóvel sob a irradiação
meridional do sol, e a corveta deslizava apenas, tão de leve, tão de leve que mal se
lhe percebia o movimento.
Nem sinal de vela na linha azul do horizonte, indício algum de criatura
humana fora daquele estreito convés: água, somente água em derredor, como se o
mundo houvesse desaparecido num dilúvio medonho..., e no alto, lá em cima, o
silêncio infinito das esferas obumbradas pela chuva de ouro do dia.
Triste e nostálgica paisagem, onde as cores desmaiavam à força de luz e a
voz humana perdia-se numa desolação imensa!

Fonte: Autor (2017)

Posteriormente, cada base de textos foi organizada em ordem alfabética, disponibilizando em arquivo texto cada uma das amostras.

3.2 Conjunto de Atributos

Para a realização dos experimentos relatados neste trabalho foram usados um conjunto de características que denotam as funções sintáticas de cada palavra em uma frase. Ao todo, são 132 características extraídas e distribuídas nos vetores conforme as informações da Tabela 3.3. Tais características foram escolhidas pelo fato de estarem presentes nas cinco línguas em que efetuamos os experimentos, sendo assim, universais entre os idiomas português, espanhol, francês, alemão e inglês. O conjunto de atributos demonstrado na Tabela 3.3 é usado pelo fato de representarem o papel de determinada palavra dentro de uma frase, neste caso, cada termo da frase é estudado de acordo com o sentido e a posição que ocupa na frase, estabelecendo assim, uma relação com os restantes dos termos.

Cada vetor armazena um tipo de informação sintática que estão agrupadas por função sintática, que são: morfológicas, flexoras, sintáticas, sintáticas auxiliares e distâncias entre os

principais elementos sintáticos. Sendo assim, uma única palavra pode ter uma ou várias funções sintáticas, de acordo com o tipo da análise.

No vetor 1, agrupou-se o conjunto de atributos com funções morfológicas, ou seja, classes das palavras. Este grupo foi usado pois influencia diretamente nas regras de escrita de cada autor. As características flexoras foram agrupadas no vetor 2, e tem por função atuar com base na classe gramatical. Possuem influencias nas variações da voz e tempo verbal, por exemplo. As ditas características sintáticas principais são agrupadas no vetor 3. Tais atributos representam as funções essenciais, integrantes e acessórias que compõe uma frase, tais como: sujeito, predicado e complemento. No vetor 4, foram agrupadas as características sintáticas que atuam de forma secundária, porém com suma importância no processo de análise sintática. E, por fim no vetor 5 são agrupadas características que demonstram a distância euclidiana entre os principais elementos sintáticos dentro de uma mesma amostra de texto.

Tabela 3.3 Conjunto de Características Sintáticas

Vetores (Vt_i)	**Características**
1 - Morfológicas	Substantivo, nome próprio, especificadores, determinantes, pronome, adjetivo, advérbio, verbo, numeral, preposição conjunção coordenada, conjunção subordinada, interjeição.
2 – Flexoras	Número (singular, plural), gênero (masculino, feminino), pessoa (primeira, segunda, terceira), tempo (passado, presente, futuro), modo (imperativo, indicativo, subjuntivo), etc.
3 – Sintáticas	Sujeito, predicado, objeto direto, objeto indireto, verbo principal, verbo auxiliar, adjunto adverbial, adjunto adnominal, argumento adverbial, objeto proposicional, objeto complementar, agente da passiva, complementos (do sujeito, do predicado, do objeto), modificadores (adverbial, proposicional, cláusula), etc.
4 – Sintáticas Auxiliares	Artigo (definido, indefinido), pronomes (demonstrativo, quantificadores, possessivo, pessoal, reflexivo, coletivo, interrogativo), diferenciadores, pré e pós-posições (adjetivos, determinantes, adverbial), etc.
5 – Distância Euclidiana das funções Sintáticas	Distâncias entre: verbo principal, sujeito, predicado, objeto direto, objeto indireto, complemento, pronomes, advérbios, adjetivos, conjunções.

Fonte: Autor (2017)

Para realizar a rotulagem de cada função sintática, que é dada por cada palavra da frase usou-se o software de processamento de linguagem natural *Visual Interactive Syntax learning* – VISL implementado por [BIC16], que é destacado na seção 3.3.

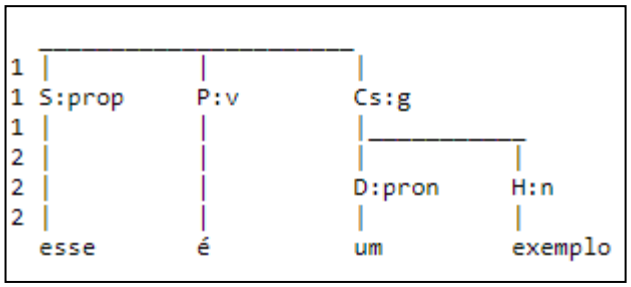
3.3 Ferramenta de Processamento de Linguagem Natural

O VISL é um software de processamento de linguagem natural que trabalha com rotulagens de palavras para mais de 25 linguagens, incluindo as línguas portuguesa, espanhola, francesa, inglesa e alemã. É um projeto de desenvolvimento e pesquisa do instituto de linguagem e comunicação da Universidade do Sul da Dinamarca [BIC16].

Tem por função rotular todas as palavras de uma frase de acordo com o seu nível de análise, podendo ser uma análise superficial, onde são anotadas somente as classes gramaticais de cada palavra ou uma análise mais profunda, chegando aos níveis estruturais e funcionais das palavras na frase.

A representação de cada rotulagem pode ser de forma horizontal ou no formato de árvore, como pode ser visto nos exemplos das Figuras 3.3 e 3.4.

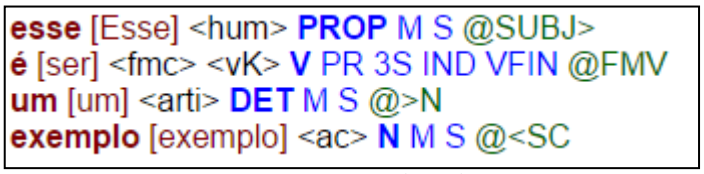
Figura 3.3 – Exemplo de Árvore Sintática do VISL



Fonte: Autor (2017)

No exemplo da Figura 3.3 é possível observar a representação de uma árvore sintática proposta pelo VISL. No entanto, neste exemplo somente são rotulados alguns elementos sintáticos, não demonstrando um nível mais detalhado de cada palavra na frase. Entretanto, na análise em um formato horizontal, que pode ser vista no exemplo da Figura 3.4, as rotulagens e as informações sintáticas são mais detalhadas, dando assim, possibilidades da realização de uma análise sintática mais aprofundada. Então, o formato usado para a extração de padrões de atributos sintáticos neste trabalho é baseado no modelo apresentado pela Figura 3.4.

Figura 3.4 – Exemplo de Rotulagem Sintática no VISL



Fonte: Autor (2017)

Pode-se denotar na Figura 3.4, que para cada palavra, podem haver diversas rotulagens, ou seja, uma palavra pode ter várias funções dentro de uma frase no aspecto sintático. Tomando o exemplo em língua portuguesa, se pode observar que a palavra "*é*" possui 5 rotulagens sintáticas. Neste caso, cada uma das rotulações com um significado perante o tipo de análise da frase. A rotulagem <*arti*> significa que a palavra é um artigo indefinido. A *tag* "DET" indica que a palavra perante a sua morfologia é um pronome determinante. Já sob o ponto de vista da flexão as rotulagens "M" e "S", significam que a palavra é do gênero masculino e está no singular. E por fim, a rotulagem "@>N", indica que é um adjunto adnominal, que tem por função acompanhar ou modificar o substantivo, ou seja, tem uma função adjetiva na frase.

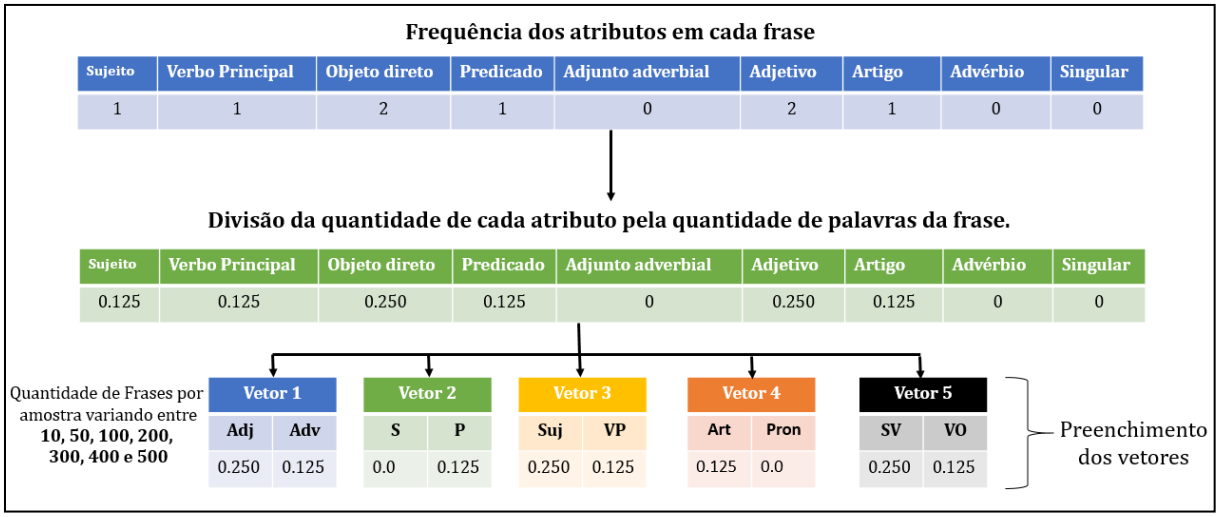
Após o processo de rotulagem das palavras, é necessário realizar a transformação das informações textuais em informações numéricas que são compreendidas pelas ferramentas de aprendizagem de máquina para geração dos modelos.

3.4 Transformação de Informação Textual em Informação Numérica

O processo de transformação das informações textuais em informações numéricas consiste inicialmente em realizar a contagem da frequência de cada característica de estilo existente em cada frase da amostra de texto. Após este processo, efetua-se a divisão da quantidade de cada atributo encontrado pela quantidade total de palavras na frase. Com isso, os vetores de 1 a 5 podem ser preenchidos pelas informações numéricas, que serão posteriormente processadas pelo classificador.

Na Figura 3.5 é possível observar um exemplo de como este processo de transformação ocorre.

Figura 3.5 – Exemplo de formação dos vetores de características



Fonte: Autor (2017)

Maiores informações a respeito do processo de extração de características, rotulagem de cada palavra e cálculos utilizados são apresentadas no Capítulo 4.

3.5 Considerações do Capítulo

As bases de textos são de suma importância para a realização dos experimentos em atribuição de autoria. Possuir acesso a bases com uma heterogeneidade quanto a sua linguagem, a quantidade de autores e a diversidades de textos se faz um ponto essencial para a realização de testes da abordagem. Neste caso, trabalhou-se com bases de dados com textos de diferentes tamanhos e assuntos, onde se conseguiu testar a abordagem em textos jornalísticos, bem como em textos literários. Além do mais, o uso de textos em cinco idiomas diferentes também proporciona verificar se a abordagem possui a robustez necessária para aplicação em idiomas diferentes. Considera-se a importância da coleção de bases de dados que foi utilizada neste trabalho. Destaca-se o legado que se pode proporcionar disponibilizando a outros pesquisadores que precisarem usar tais bases de dados.

Em outro ponto, foi apresentado o conjunto de atributos que foram utilizados como objetos discriminantes para o processo de atribuição de autoria. Usaram-se características de função sintática, que podem exercer diversas representações em uma frase.

Por conseguinte, foi apresentada o software de rotulagem das palavras e o processo de transformação das informações textuais contidas nas amostras de cada texto, bem como, a sua representação numérica perante o processo de geração de modelos de classificação.

Na continuidade do trabalho, a base de dados demonstrada neste capítulo é de fundamental importância para o desenvolvimento da abordagem e dos experimentos que serão evidenciados nos próximos capítulos.

Capítulo 4

Proposta

Neste capítulo apresenta-se a proposta deste trabalho para os problemas de atribuição de autoria. Destaca-se o passo-a-passo para o desenvolvimento do tratamento das amostras de textos, construção dos modelos e o processo decisório.

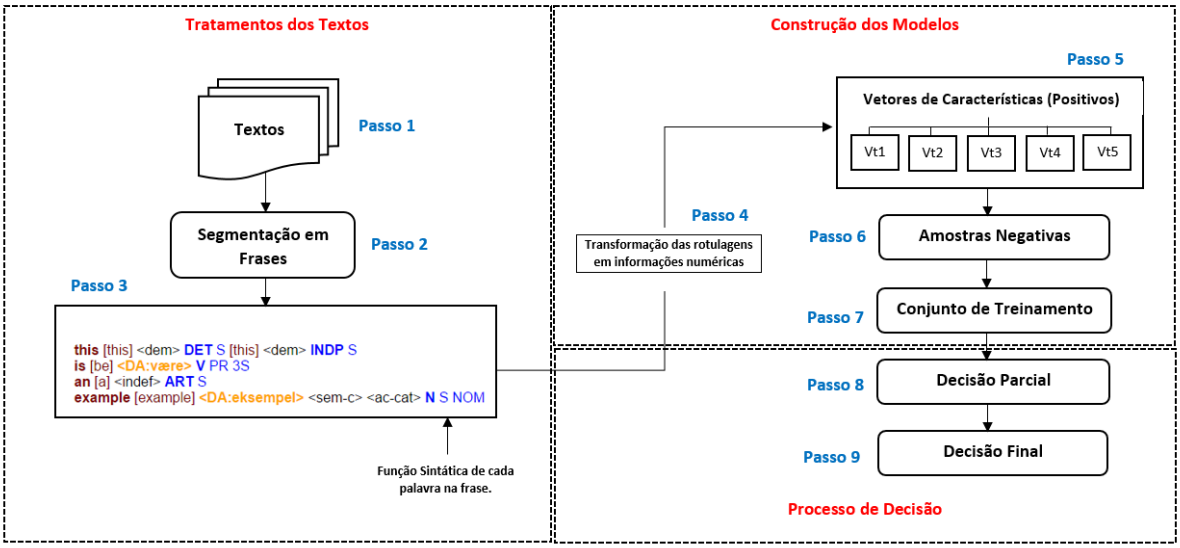
4.1 Visão Geral

No decorrer da história de atribuição de autoria, diversos métodos já foram aplicados. Entre os principais métodos, podem-se citar os baseados em funções numéricas, estatísticas e em aprendizagem de máquina. Neste contexto, a aprendizagem de máquina em conjunto com a categorização de textos provocou um importante marco nos estudos de atribuição de autoria. Tais métodos são de simples aplicação, pois consistem basicamente em transformar textos de autores conhecidos (treinamento) em vetores de representação numérica rotulados. Os métodos de aprendizagem são utilizados para encontrar os limites entre as classes, ou seja, as fronteiras que possam delimitar os estilos dos autores. O poder de separação das classes ou a natureza dos limites depende do método de aprendizado utilizado, podendo este auxiliar no processo de encontrar os limites de uma classe por meio de métodos que possam minimizar a distância entre as classes [KOP09]. Sendo assim, utilizou-se neste trabalho uma abordagem baseada em aprendizagem de máquina, que tem por função construir classificadores que são gerados a partir de amostras de textos de autores conhecidos e posteriormente aplicados à autores anônimos.

No entanto, a abordagem não consiste somente em métodos de classificação e aprendizagem de máquina, e sim, de um conjunto de processos desenvolvidos desde a

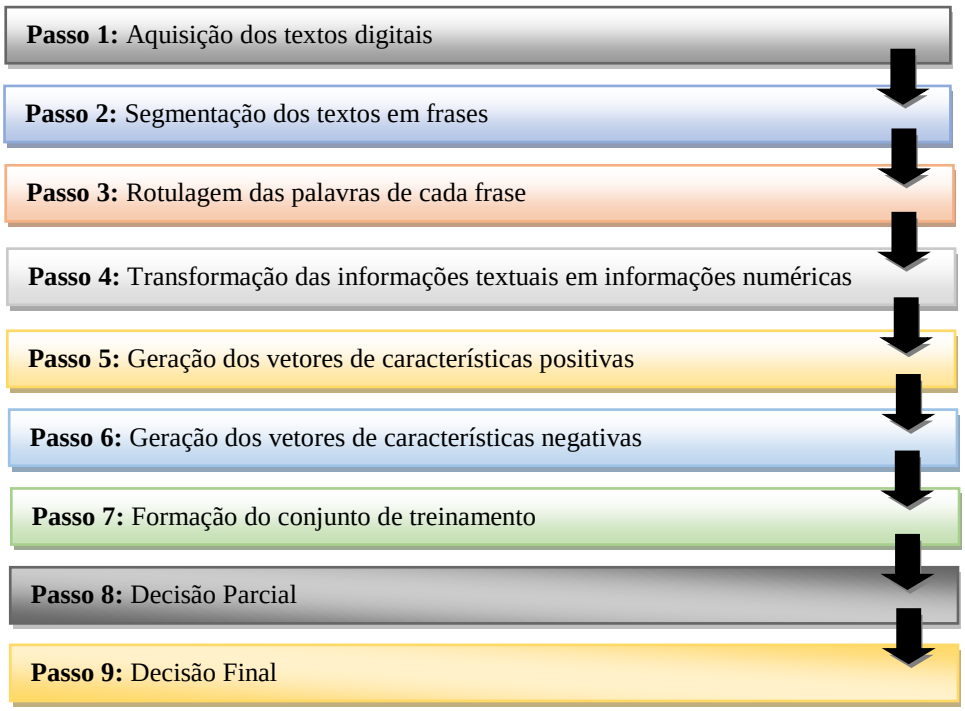
aquisição das bases de dados até o processo de tomada de decisão final. Na Figura 4.1, apresenta-se uma visão geral da abordagem. Na Figura 4.2, apresenta-se o fluxograma passo-a-passo, a qual é detalhada nas seções deste capítulo.

Figura 4.1 – Visão Geral da Abordagem



Fonte: autor (2017)

Figura 4.2 – Esquema Passo-a-Passo



Fonte: autor (2017)

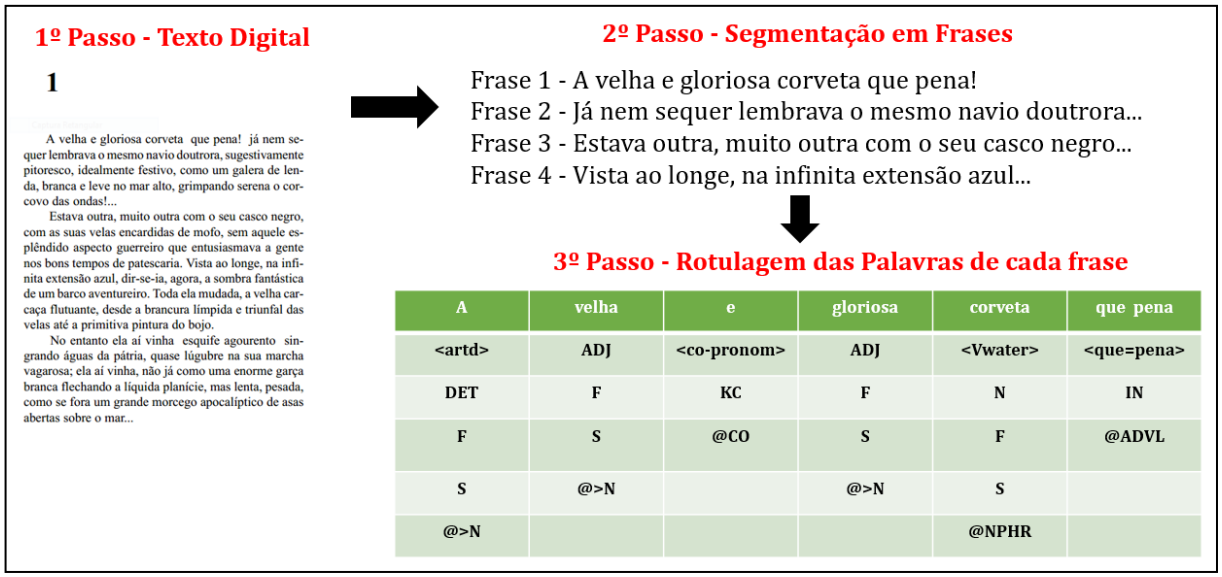
4.2 Tratamento dos textos

No processo de tratamento dos textos, evidencia-se a organização dos textos de cada autor, bem como a segmentação em frases e a rotulagem. Para melhor entender as fases deste processo, demonstra-se um passo-a-passo.

Passo 1: Inicialmente, todas as amostras de textos, de um autor conhecido (S_c) ou de um autor desconhecido (S_d) são organizadas para o processo de segmentação.

Passo 2: Cada amostra de texto é segmentada em frases, por intermédio de um processo automatizado. Neste caso, uma amostra de texto que é composta por diversas frases em um mesmo parágrafo são separadas, gerando um arquivo texto com uma frase por linha. Com isso, as frases podem ser processadas a fim de extrair as funções sintáticas de cada palavra, como pode ser visto na Figura 4.3.

Figura 4.3 – Exemplo do esquema de tratamento dos textos

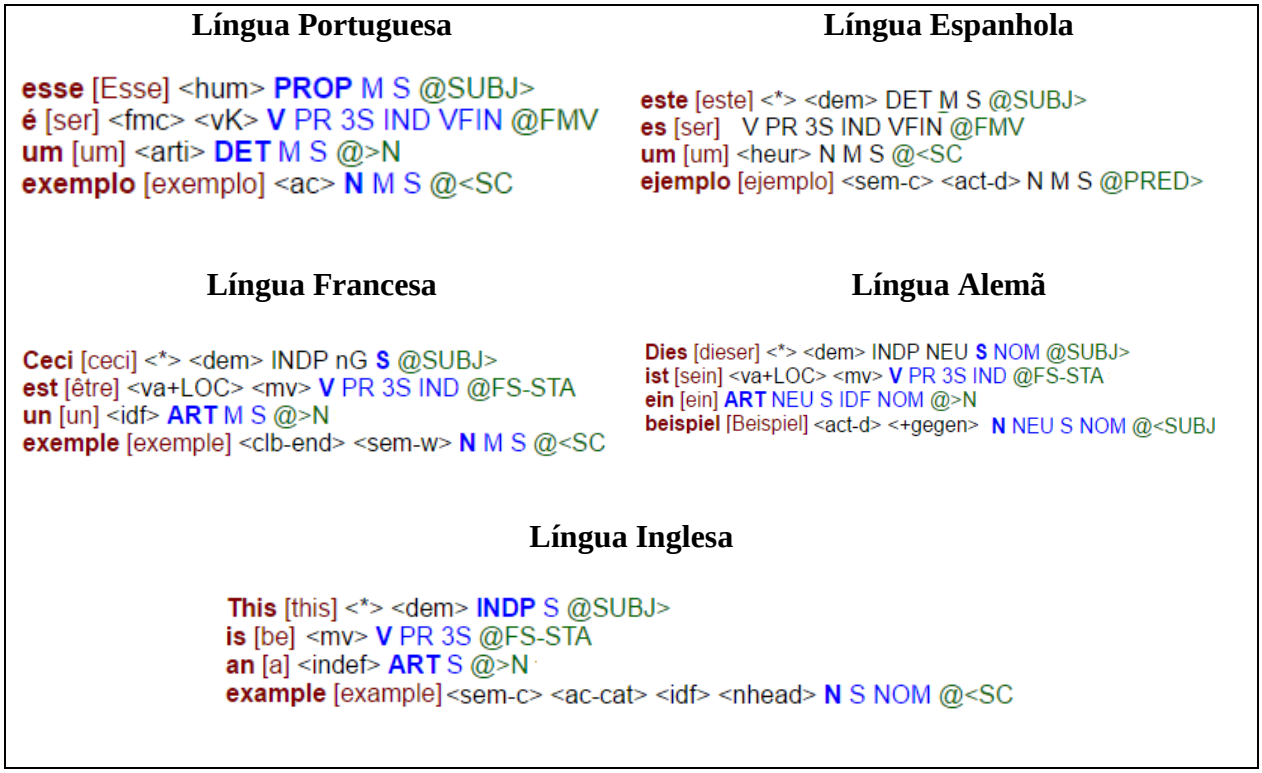


A	velha	e	gloriosa	corveta	que pena
<artd>	ADJ	<co-pronom>	ADJ	<Vwater>	<que=pena>
DET	F	KC	F	N	IN
F	S	@CO	S	F	@ADVL
S	@>N		@>N	S	
@>N				@NPHR	

Fonte: autor (2017)

Passo 3: As informações sintáticas de cada amostra e autor são extraídas por meio do processo de rotulagem de cada palavra, que é apresentado por [BIC16]. Na Figura 4.4, é apresentado um exemplo de rotulagem sintática da frase.

Figura 4.4 – Exemplo de rotulagem das frases em múltiplas linguagens



Fonte: autor (2017)

Passo 4: Após a rotulagem de cada palavra, realiza-se a transformação dos rótulos em informações numéricas que irão compor cada vetor de características. Usam-se 5 vetores de características separados por classes, para melhor entender o poder de discriminação das funções sintáticas. O vetor 1 (Vt_1) é composto por 12 atributos da classe morfológica das palavras, tais como: substantivo, verbo e adjetivo. O vetor 2 (Vt_2), possui 28 atributos que possuem funções de flexão nas palavras, tais como: gênero, pessoa e tempo. Já o vetor 3 (Vt_3) é composto por 45 atributos sintáticos, tais como: sujeito, predicado e complementos. No vetor 4 (Vt_4), são 35 atributos sintáticos que atuam de forma auxiliar, tais como: artigos e pronomes. E no vetor 5 (Vt_5), são 12 atributos que indicam as distâncias entre os principais elementos sintáticos em uma frase, tal como: distância do sujeito em relação ao objeto direto. Maiores informações a respeito do grupo de característica de estilo usadas neste trabalho podem ser consultadas em [BIC16]. O Algoritmo 1 apresenta o procedimento de cálculo para os vetores com as características de cada frase (ver Figura 4.5). Para cada texto T que contém F_k frases, onde $I = \{ \in N \mid 1 \leq i \leq 5\}$ e μ é o número de frases, são criados os vetores Vt_i, com base da Tabela 3.3. Para cada frase F_k é calculado o número N_k de palavras que a compõe. Em seguida, as quantidades de cada palavra rotulada, em cada classe, são computadas. Para cada

frase F_k são criados 5 vetores de características Vt_i, referindo-se às cinco classes da Tabela 4.3. Em seguida, divide-se os vetores Vt_i, pelo número de palavras de uma frase N_k, criando os vetores Ft_i.

Figura 4.5 – Computação dos Vetores de Características

Algoritmo 1

```
Entrada: Textos do autor com μ frases F[ ].
Rotulagem de cada frase F[ ] em classes sintáticas.
  for k ←1 to μ
  Computando N[k], número de palavras de F[k]
  for i ←1 to N[k] do
     for j ← to 12 /*classe 1 – tamanho do léxico */
       V1[j] ← ++ /*Quantidade de rotulagens das palavras da classe 1*/
     end for
     for j ← 1 to 28 do /*classe 2 – tamanho do léxico */
       V2[j] ← ++ /* Quantidade de rotulagens das palavras da classe 2*/
     end for
     for j ← 1 to 45 do /*classe 3 – tamanho do léxico */
       V3[j] ← ++ /* Quantidade de rotulagens das palavras da classe 3*/
     end for
     for j ← 1 to 35 do /*classe 4 – tamanho do léxico */
       V4[j] ← ++ /* Quantidade de rotulagens das palavras da classe 4*/
     end for
     for j ← 1 to 12 do /*classe 5 – tamanho do léxico*/
       V5[j] ← ++ /* /* Quantidade de rotulagens das palavras da classe 5*/
     end for
  end for /* Final das rotulagens de F[k]*/
    /* Ponderação das frases F[k] */
     for j ← 1 to 12 do
       F1[k] ← V1[j]/N[k]
     end for
     for j ← 1 to 28 do
       F2[k] ← V2[j]/ N[k]
     end for
     for j ← 1 to 45 do
       F3[k] ← V3[j]/ N[k]
     end for
     for j ← 1 to 35 do
       F4[k] ← V4[j]/ N[k]
     end for
     for j ← 1 to 12 do
       F5[k] ← V5[j]/ N[k]
     end for
  /* Final das ponderações*/
 end for k /* Final da Rotulagem das frases F[ ] */
```

Fonte: autor (2017)

Após a extração de características de cada amostra de texto, o processo de construção dos modelos é iniciado, e são descritos na seção 4.3.

4.3 Construção dos Modelos

A partir deste ponto, inicia-se o processo de construção de modelos, ou seja, de formação dos vetores de características, que em conjunto formam o modelo de cada autor para o processo de treinamento.

O treinamento tem por função gerar o modelo de classificação, ou seja, aplicar uma técnica de classificação que consiga diferenciar as classes e características dos autores. Neste trabalho o algoritmo de treinamento e classificação utilizado é o SVM duas classes com kernel linear. No processo de treinamento e geração dos modelos são utilizadas duas abordagens distintas: independente e dependente do autor.

No modelo independente é utilizado o conceito de dicotomia, ou seja, existem somente duas possibilidades: autoria ou não autoria. Nesta abordagem é gerado um modelo genérico de autoria e de não autoria por meio da combinação de amostras de um mesmo autor (amostras positivas – autoria) e a combinação de amostras de autores distintos (amostras negativas – não autoria). Neste caso, os autores que participam do treinamento não fazem parte dos testes, tendo assim um modelo que realizará a classificação de autores nunca vistos anteriormente.

Para construção deste modelo podem ser utilizadas um número reduzido de amostras de textos por autor, pois este modelo não necessita de um grande número de amostras por autor. A ênfase está na diversidade de autores e não na quantidade e variedade de amostras. Uma das vantagens do modelo independente é que para a inclusão de novos autores não é necessário a realização de um novo treinamento do modelo.

No modelo dependente do autor, para cada autor é gerado um modelo de atribuição baseado nas características de estilo sintático de escrita do autor. Este por sua vez, é baseado no conceito da policotomia, ou seja, na classificação do problema em diversos modelos. Neste modelo geralmente utiliza-se um grande número de amostras de texto por autor, pois a ênfase principal está nas características estilométricas de cada autor.

Na construção da base de treinamento do modelo dependente também são geradas amostras verdadeiras e amostras falsas. As amostras verdadeiras são constituídas pela combinação das amostras de um mesmo autor. Já as amostras falsas são constituídas pela

combinação de amostras de autores distintos. Ao final, o arquivo de treinamento possuirá o mesmo de número de amostras verdadeiras e falsas, ou seja, um vetor balanceado. Isso é necessário para que não haja desiquilíbrio ou tendência do modelo.

No modelo dependente do autor todos os autores da base participam do treino e também dos testes. Para tanto, as amostras de textos dos autores que fazem parte do treinamento não fazem parte dos testes, porém as amostras de textos separadas para testes são combinadas com amostras de referências que podem ou não fazer parte do treinamento.

Para calcular a similaridade entre os textos, faz-se uso da distância euclidiana entre cada uma das amostras. Neste caso, um vetor de similaridade é composto pelo módulo da diferença entre cada um dos elementos que compõe uma amostra de texto A de uma amostra de texto B, como pode ser visto na Figura 4.6 e representada matematicamente na equação 1.

Figura 4.6 – Exemplo de Cálculo da distância euclidiana entre amostras de textos

Amostra	Sujeito	Verbo Principal	Objeto direto	Predicado	Adjunto adverbial	Adjetivo	Artigo	Advérbio	Singular
A	0,123	0,233	0,032	0,079	0,021	0,100	0,153	0	0
B	0,123	0,123	0,059	0,103	0,028	0,050	0,033	0	0
Z	0	0,100	0,017	0,024	0,007	0,050	0,120	0	0

Fonte: autor (2017)

$$Z = | A - B| \qquad (1)$$

Dando continuidade ao passo-a-passo de detalhamento da abordagem proposta, apresenta-se:

Passo 5: Dado um conjunto de vetores de características $D_{p,}$ como visto na equação 2.

$$D_p = U_{i=1}^{5} F_i \qquad (2)$$

Onde $P = \{ \in N \mid 1 \leq p \leq \theta\}$ e θ o número de amostras por autor para o processo de treinamento. A partir do processo descrito no Algoritmo 1, é criado um conjunto de amostras genuínas, combinando amostras de um mesmo autor $Z_{(+)}$, representado na equação 3. Um subconjunto de amostras é usado para referências e treinamento e outro usado para os testes.

$$Z_{(+)} \leftarrow |R_p - R_k| \qquad (3)$$

Passo 6: O subconjunto falso $Z_{(-)}$ é gerado pela combinação de amostras de autores diferentes. Neste caso, um vetor de características do autor A é combinado com um vetor de características do autor B, conforme explicado na equação 4.

$$Z_{(-)} \leftarrow |A - B| \qquad (4)$$

Passo 7: A combinação de amostras positivas e negativas geram um conjunto de treinamento *Ts*, como na equação 5.

$$Ts \leftarrow \text{Conjunto de Treinamento } (Z_{(+)}, Z_{(-)}) \qquad (5)$$

Os processos das etapas 5, 6 e 7, que evidenciam os procedimentos para a geração do conjunto de treinamento (falso e positivo) podem ser vistos no Algoritmo 2 (Figura 4.7).

Figura 4.7 – Computação das amostras Positivas e Negativas

Algoritmo 2

```
/* Computação dos exemplos genuínos (positivos) - diferenças
     Entre vetores da mesma classe*/
Positivas ← 0
for p ← 1 to θ do
 /*Formação das amostras de Referência R*/
    R_p ← D_p
 /*Formação das amostras genuínos (positivas) Z(+)*/
    for k ← p + 1 to θ - 1 do
          Z(+) ← |R_p – R_k|
        Positivas++
   end for
 end for
 /* Computação dos exemplos falsos – diferença entre
      os vetores de classes diferentes*/
Falsas ← 0
while (Falsas != Positivas) do
  A ← R_Randon(p) /*obtidos aleatoriamente a partir dos vetores*/
 B ← R_Randon(p)≠A /*obtidos aleatoriamente, mas diferente de A*/
    Z(-) ← |A – B|
 Falsas++
end while
Ts ← Conjunto de Treinamento (Z(+), Z(-))
return Ts
```

Fonte: autor (2017)

4.4 Processo de Decisão

No processo decisório, o grupo de vetores de testes é usado para realizar a validação do modelo, ou seja, para verificar o poder de previsão do modelo de atribuição de autoria gerado pelo modelo de classificação. No processo de testes ou classificação são utilizadas duas abordagens quanto ao processo de atribuição de autoria, que são: verificação e identificação.

Na verificação de autoria o objetivo principal é verificar se o modelo criado no processo de treinamento é robusto o suficiente para conseguir classificar corretamente amostras de textos de um mesmo autor. A estratégia utilizada neste tipo de abordagem é um-contra-um. Neste caso, o autor que se deseja comparar é conhecido, sendo assim, realiza-se o processo de verificação com os modelos deste autor. O resultado é verificar se ele acerta ou erra.

No método de identificação de autoria o objetivo é confrontar toda a base de dados com todos os autores, em busca de identificar quem é o autor da amostra questionada. A estratégia utilizada nesta abordagem é um-contra-todos, ou seja, confrontar um texto questionado contra todos os modelos ou classes constantes do treinamento, afim de tentar identificar o provável autor. Neste caso, como o confrontamento é grande e de difícil decisão por parte do classificador, também realizamos a análise dos autores melhores classificados (Top 1, Top 3, Top 5 e Top10).

Todas as amostras de textos submetidos ao processo de treinamento ou testes foram retiradas da base de dados coletada (Capítulo 3). O número de amostras e de autores que selecionadas para o processo de treinamento e testes são definidos no protocolo de experimentos, definidos no capítulo 5.

No processo de decisão os passos 8 e 9 são apresentados para o melhor entendimento da abordagem.

Passo 8: Dado um conjunto de vetores de testes Q_a, onde A = { $\in$ N | $1 \leq a \leq \alpha$} e α é o número de autores, composto pelos subconjuntos dos vetores de características D_t onde T = { $\in$ N | $1 \leq t \leq \xi$} e ξ é o número de amostras de um determinado autor para o teste. O Algoritmo 3 apresenta o cálculo dos vetores de testes (Figura 4.8). O procedimento básico é computar o vetor de dissimilaridade entre uma instância de Q_a e um subconjunto de referências R_p de um autor. Neste caso, obtém-se um conjunto de resultados parciais Pr_{ap}.

Figura 4.8 – Processo de Testes - Saída

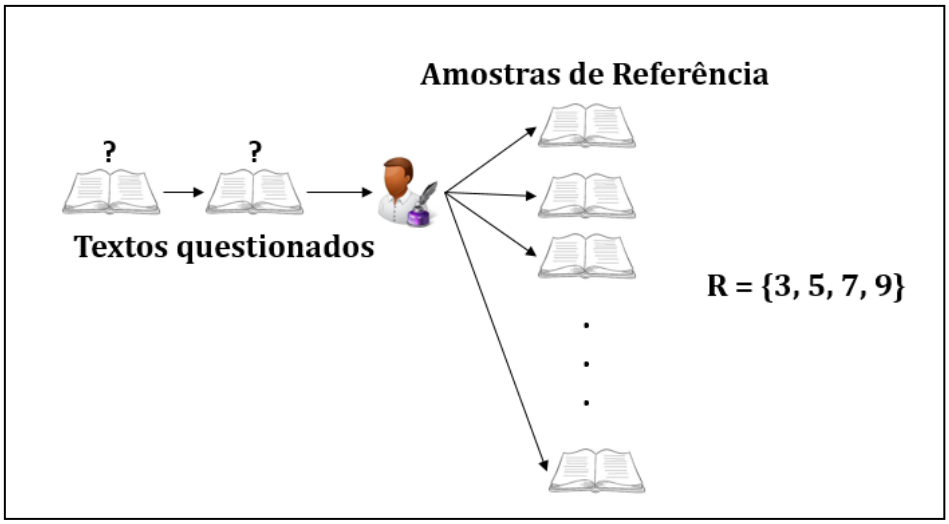
Algoritmo 3

Entrada: *um conjunto de vetores de testes Q_a e um subconjunto de referências R_p.*
Output*: um conjunto de resultados parciais Pr*
for *a* ← *1* ***to*** *α* ***do***
  for *p* ← *1* ***to*** *φ* ***do*** */* φ é o número de referências usado*/*
    $Pr_{ap} \leftarrow |Q_a - R_p|$
end for

Fonte: autor (2017)

Toda vez em que um texto questionado é submetido ao processo de testes, usam-se amostras de referência do autor que está se comparando o texto, conforme exemplo da Figura 4.9. As referências são usadas para que o processo de decisão tenha parâmetros comparativos entre o autor do texto questionado e o verdadeiro autor do documento. São usadas 3, 5, 7 e 9 amostras de referências por autor. Isso pelo fato de se trabalhar com no máximo 20 amostras por autor.

Figura 4.9 – Uso de amostras de Referência



Fonte: autor (2017)

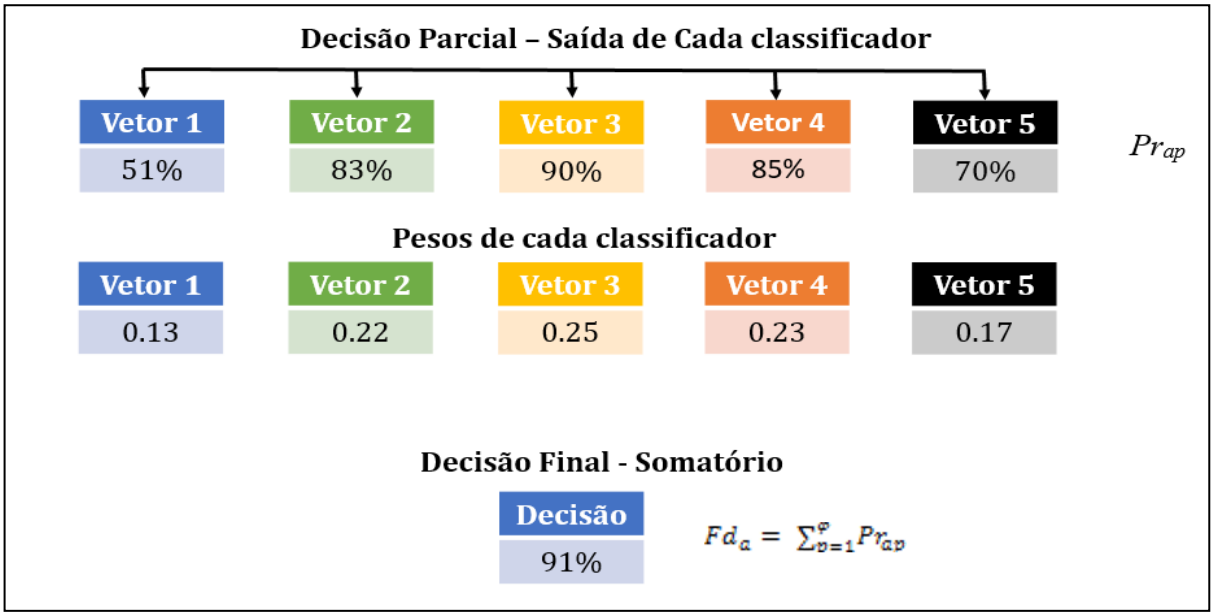
Passo 9: A decisão final Fd_a é calculada pelo somatório dos resultados parciais, conforme a equação 6.

$$Fd_a = \sum_{v=1}^{\varphi} Pr_{ap} \qquad (6)$$

Onde φ é o número de referências usadas.

Em resumo, para cada vetor de atributos existe um peso ponderado, que é dado pela saída de cada classificador. Na Figura 4.10 é possível observar um exemplo de como ocorre este processo. Ao final, o somatório dos pesos de cada classificador indica a taxa de acerto.

Figura 4.10 – Exemplo do Processo de Decisão



Fonte: autor (2017)

4.5 Considerações do Capítulo

Este capítulo apresenta a abordagem proposta para a resolução de casos que envolvam a atribuição de autoria em textos. Fica evidenciado que o uso de uma ferramenta de processamento de linguagem natural robusta e já testada se torna essencial quando se necessita realizar a rotulagem ou classificação de palavras.

A formação dos vetores de características de cada autor é dada pela distância euclidiana entre cada uma das amostras, gerando assim parâmetros comparativos por meio da similaridade dos textos. Para a geração dos modelos de treinamento foram usadas duas abordagens (independente e dependente do autor) e duas abordagens no processo de testes (verificação e identificação). Estes fatores são importantes, para que se possa melhor avaliar a estabilidade da abordagem proposta e também verificar as suas limitações.

No próximo capítulo são apresentados os resultados dos experimentos realizados neste trabalho.

Capítulo 5

Resultados Experimentais e Discussão

Neste capítulo apresentam-se os resultados dos experimentos em cada um dos cenários que foram montados para avaliação da proposta. Evidenciam-se os protocolos de treinamentos e testes, bem como as respectivas discussões de cada tópico. Em correlato, é mostrado um comparativo dos resultados proporcionado pela abordagem em relação a outras abordagens constantes na literatura.

5.1 Cenários

Para a realização dos experimentos foram utilizados diversos cenários de aplicação, que são relacionados a verificação e identificação de autoria. Entre eles, a variação da quantidade de informações em cada amostra (quantidade de frases por amostra) e a quantidade de amostras de referências.

Na verificação de autoria buscou-se testar se uma amostra questionada pertence a um determinado autor. Na identificação de autoria, o processo de testes consiste em colocar uma amostra de texto questionado contra todos os autores da lista, afim de procurar identificar o autor que contém as amostras mais semelhantes com o texto questionado.

Variou-se também a quantidade de informação de cada amostra (F_k), com o intuito de verificar se a abordagem proposta possui eficácia em ambientes com pouca e com grande quantidade de informação. Neste caso, realizaram-se experimentos em blocos de frases, ou seja, quantidade de frases por amostra. Então, considera-se uma baixa quantidade de informação quando se tem F_k = 10, ou seja, cada amostra de texto possui 10 frases. Uma média quantidade de informação sintática quando se tem F_k = 50 e 100 frases por amostra de

texto. E, uma grande quantidade de informação quando se tem F_k = 200, 300, 400 e 500 frases por amostra. Ao final se tem um conjunto F_k = {10, 50, 100, 200, 300, 400, 500}. Demais detalhes são evidenciados no decorrer deste capítulo.

Foram aplicadas diferentes quantidades de amostras de referência (R_p) para observar o comportamento da abordagem quando se tem a variação da disponibilidade de amostras de cada autor. Neste caso, como foi trabalhado com no máximo 20 amostras por autor, foram utilizadas R_p = {3, 5, 7, 9} amostras de referências por autor.

Então, foi montado um protocolo de experimentos que envolve a variação da quantidade de informações de cada amostra, do número de referências e das estratégias de treinamento e testes. No entanto, foram usados em todos os experimentos 20 amostras por autor, sendo 10 amostras para treinamento e 10 amostras para realização dos testes, ou seja, 50% para treino e 50% para testes.

Em todos os casos foi utilizado o classificador SVM duas classes com kernel linear. Inicialmente, foi trabalhado com computador pessoal core i5, com 4 Mb de memória RAM e capacidade de processamento de 2,5 Ghz. Entretanto, o processo de rotulagem e geração dos vetores de características de cada autor durava cerca de 24 horas, limitando nossos experimentos à mais de 500 dias de processamento. Então, optou-se por rodar os experimentos em uma *grid* composta por 50 computadores, que otimizou o tempo de processamento para ~1 hora por autor, processando toda a base em cerca de 3 semanas.

Os resultados apresentados neste capítulo possuem como base a precisão da abordagem, ou seja, é apresentada a taxa de acurácia, que possui por função representar a relação de proximidade entre o resultado experimental e o valor verdadeiro.

Nas seções 5.2 e 5.3 apresentam-se os resultados e a discussão da verificação e da identificação de autoria, respectivamente.

5.2 Verificação de Autoria

O objetivo principal da verificação de autoria é determinar se um texto foi escrito por um autor específico, ou seja, é um problema de duas classes: autoria (w_1) e não autoria (w_2).

Dado um vetor de características de um texto literário Q_a pertencente a um autor desconhecido S_u, a ideia é determinar se esse texto pertence ou não a um autor conhecido A. O objetivo é determinar se o texto questionado (Q_a) pertence à classe w_1 ou w_2. Se pertencer a

w_1, indica que a afirmação é verdadeira, ou seja, que o texto foi escrito pelo autor A, caso contrário, que pertença a w_2, a afirmação é falsa, ou seja, indica que o texto não é do autor A.

Primeiramente apresentam-se os resultados por linguagem, destacando os desempenhos em função dos modelos dependente e independente e também com base nos diferentes números de referência (R_p) e frases em cada amostra (F_k). Por conseguinte, são mostrados os resultados com base nos 5 grupos de características (vetores) que foram usados como atributos discriminantes, em separado e em conjunto para cada linguagem. O mesmo processo foi replicado usando textos em língua portuguesa, espanhola, francesa, alemã e inglesa.

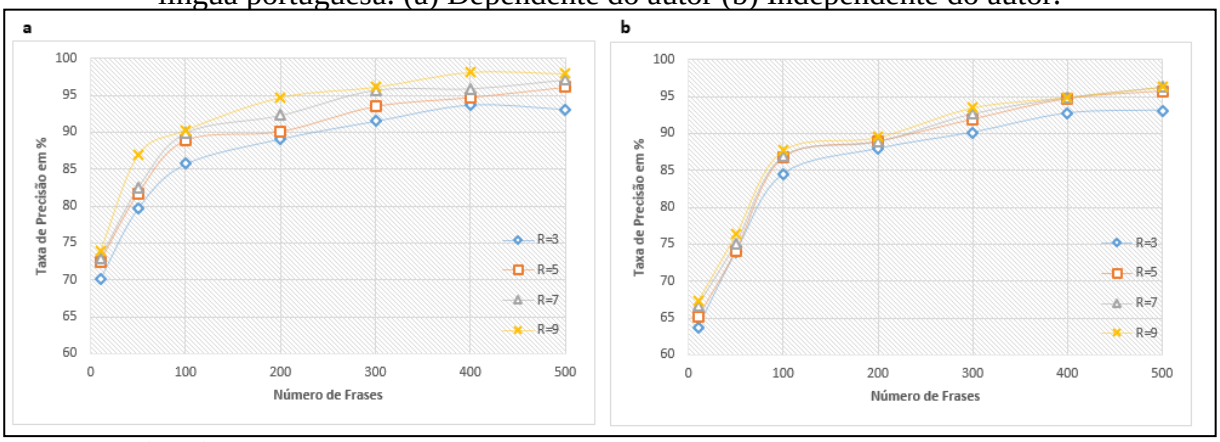
Inicialmente, apresentam-se os resultados para a língua portuguesa, que é considerada uma linguagem complexa gramaticalmente porque contém uma vasta variedade de elementos sintáticos. As Figuras 5.1 (a) e (b) ilustram que conforme é aumentada a quantidade de amostras de referências, aumenta-se também a taxa de precisão do modelo. Sendo assim, a decisão ótima indica o uso de R_p = 9, ou seja, o uso de 9 amostras de referências por autor. Entretanto, observa-se que o uso de R_p = 3, 5, 7 também relatam taxa de acerto acima de 70% em língua portuguesa.

Considerando a quantidade de texto em cada amostra, percebe-se que o desempenho melhora à medida que se aumenta a quantidade de informações, ou seja, um maior número de frases por amostra. A Figura 5.1 (a) mostra o resultado para o modelo dependente do autor. Neste caso, percebe-se que a precisão do modelo com baixa quantidade de informação sintática (F_k = 10) varia entre 70-73%. Quando se aumenta a quantidade de informação para (F_k = 50 ,100) a precisão tem uma significativa melhoria, passando a ter resultados entre 79-90%. E, finalmente com uma grande quantidade de informação (F_k >= 200) a precisão do modelo chega a 89-98%. Como é possível observar na Figura 5.1, o modelo independente do autor obteve uma menor taxa de precisão em relação ao modelo dependente do autor, como esperado. Entretanto, no modelo independente do autor, quando é usada uma baixa quantidade de informação sintática (F_k =10) a precisão possui uma variação entre 63-67%. Quando F_k = 50 ou 100, a precisão melhora para 73-87%, e, finalmente com uma grande quantidade de informação (F_k >= 200), a precisão atinge 87-96%.

Comparando ambos os modelos, observa-se que a medida que as quantidades de informações sintáticas nas amostras aumentam, a diferença na precisão dos modelos tende a diminuir (Tabelas 5.1 e 5.2), ou seja, que quanto mais frases são usadas por amostra mais as

taxas de precisão dos modelos dependente e independente se aproximam. No entanto, o modelo dependente do autor possui uma maior acurácia em comparação com o modelo independente do autor.

Figura 5.1 – Taxas de acurácia da abordagem conforme o número de frases e referência em língua portuguesa: (a) Dependente do autor (b) Independente do autor.



Fonte: autor (2017)

As Tabelas 5.1 e 5.2 mostram os resultados da verificação de autoria em língua portuguesa. Estratificam a taxa de precisão para cada conjunto de testes. Por exemplo, para um bloco de amostras de testes compostas por 10 frases, os testes foram efetuados com 3, 5, 7 e 9 amostras de referências.

Tabela 5.1. Resultados – Modelo Dependente do Autor - Língua Portuguesa

Número de Frases (F_k)	**Número de Amostras de Referência**			
	R_p = 3	***R_p = 5***	***R_p = 7***	***R_p = 9***
10	70.1	72.5	72.9	73.9
50	79.7	81.7	82.5	86.9
100	85.7	88.9	89.9	90.3
200	89.1	90.1	92.3	94.7
300	91.5	93.5	95.7	96.1
400	93.7	94.7	95.9	98.1
500	93.1	96.1	97.1	97.9

Fonte: autor (2017)

Tabela 5.2. Resultados – Modelo Independente do Autor – Língua Portuguesa

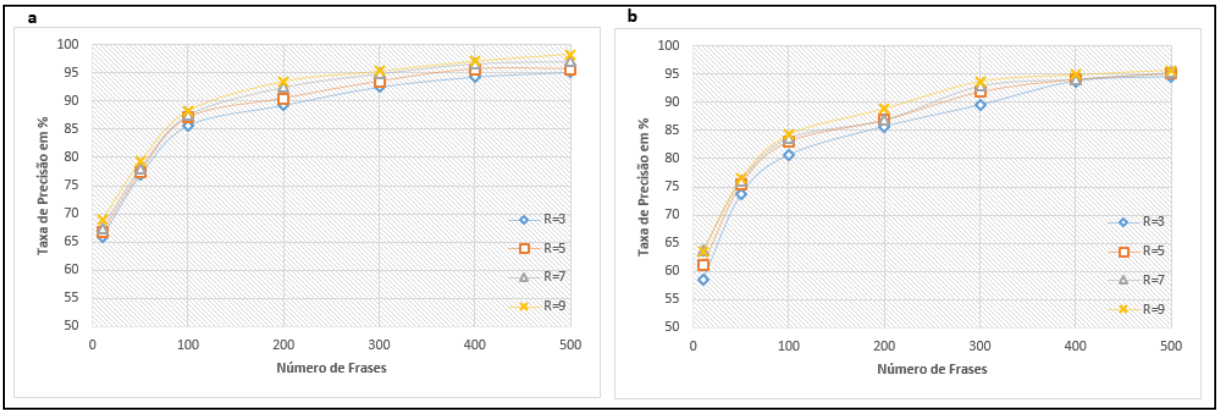
Número de Frases (F_k)	**Número de Amostras de Referência**			
	R_p = 3	***R_p = 5***	***R_p = 7***	***R_p = 9***
10	63.7	65.1	66.7	67.3
50	73.9	74.1	75.1	76.3
100	84.5	86.7	86.9	87.7
200	87.9	88.9	88.9	89.5
300	90.1	91.9	92.7	93.5
400	92.7	94.7	94.9	94.9
500	93.1	95.7	96.3	96.3

Fonte: autor (2017)

Em língua espanhola, observa-se que o comportamento da abordagem é estável, e apresenta resultados semelhantes ao da língua portuguesa. A semelhança dos resultados para as duas línguas pode estar ligada as suas estruturas sintáticas, uma vez que possuem uma gramática e uma sintaxe muito próximas, já que advém de uma mesma vertente linguística. Assim, não há diferenças substanciais nas estruturas das duas línguas. Basicamente o que muda de uma linguagem para outra é a escrita de elementos linguísticos tais como artigos, pronomes, preposições e etc.

Então, conforme pode ser observado nas Figuras 5.2 (a) e (b), os resultados evidenciam taxas de precisão semelhantes ao da língua portuguesa. Pode-se observar que o modelo dependente do autor fornece uma precisão de 65-68% quando F_k *=10*. Com quantidade de informação sintática média (F_k *= 50, 100*) a precisão é de 76-88%, enquanto que com alto teor de informações sintáticas em cada amostra a precisão varia de 85-98% de acordo com cada protocolo. No modelo independente do autor, obteve-se uma taxa de precisão entre 58-63% para F_k *= 10*. Já com F_k *= 50, 100* as taxas de acerto variam entre 73-84%. E por fim, com F_k *= 200, 300, 400, 500* a precisão alcança patamares entre 85-95%.

Figura 5.2 – Taxas de acurácia da abordagem conforme o número de frases e referência em língua espanhola: (a) Dependente do autor (b) Independente do autor.



Fonte: autor (2017)

Nas Tabelas 5.3 e 5.4, são apresentadas as taxas de precisão para a língua espanhola. Estes resultados são baseados pela quantidade de informação de cada amostra, dada por F_k e em relação a quantidade de amostras de referências, dada por R_p.

É possível observar na Tabela 5.3, que no modelo dependente do autor quando são usadas um número maior de amostras de referências as taxas de precisão são melhores. Por exemplo, em média tem-se uma precisão 3% maior quando se usa $R_p = 9$ comparado com o uso de $R_p = 3$.

No modelo independente do autor (Tabela 5.4), as taxas de precisão se mostram em média 5% menores em comparação com o modelo dependente do autor, indicando que quando se tem uma maior quantidade de informações do autor, obtem-se melhores resultados. Em relação ao número de amostras de referência, a mesma observação do modelo dependente é vista, indicando que um número maior de referências produz melhores taxas de precisão também em língua espanhola.

Tabela 5.3. Resultados – Modelo Dependente do Autor – Língua Espanhola

Número de Frases (F_k)	**Número de Amostras de Referência**			
	R_p = 3	***R_p = 5***	***R_p = 7***	***R_p = 9***
10	65.9	66.7	67.5	68.9
50	76.9	77.5	78.1	79.3
100	85.7	87.1	87.5	88.3
200	89.3	90.5	92.5	93.5
300	92.5	93.5	94.9	95.3
400	94.3	95.7	96.7	97.1
500	95.1	95.7	97.1	98.3

Fonte: autor (2017)

Tabela 5.4. Resultados – Modelo Independente do Autor – Língua Espanhola

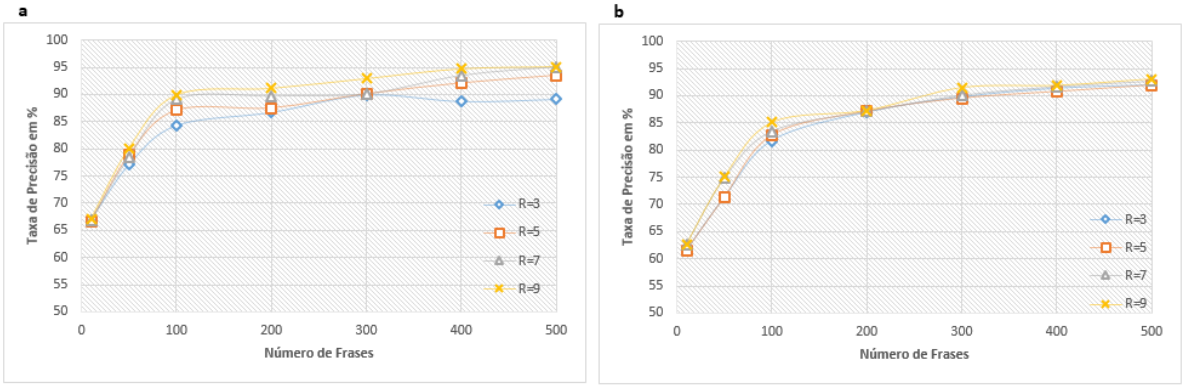
Número de Frases (F_k)	**Número de Amostras de Referência**			
	R_p = 3	***R_p = 5***	***R_p = 7***	***R_p = 9***
10	58.5	61.1	63.7	63.5
50	73.7	75.5	76.1	76.5
100	80.7	83.1	83.7	84.3
200	85.7	86.9	86.9	88.9
300	89.5	91.9	92.9	93.7
400	93.7	94.1	94.1	94.9
500	94.5	95.1	95.3	95.7

Fonte: autor (2017)

No idioma francês, os resultados podem ser observados nas Figuras 5.3 (a) e (b). Consegue-se ver que no modelo dependente do autor apresentado na Figura 5.3 (a) que quando usado $F_k = 10$ a taxa de acerto é de cerca de 67%. Já quando se tem $F_k = 50, 100$, a acurácia varia de 77-89%. E, quando se tem uma grande quantidade de informações, com $F_k >= 200$, obtêm-se resultados com precisão variando de 86-95%. Na Figura 5.3 (b), que apresenta modelo independente do autor percebe-se mais uma vez que os resultados apresentados evidenciam taxas de acerto menores que o modelo dependente do autor. Neste caso, quando se usa $F_k = 10$ a taxa de acerto média é de 62%. Com $F_k = 50, 100$ a precisão varia entre 71-85%. E, com $F_k >= 200$ a precisão é de 86-93%.

Nas Tabelas 5.5 e 5.6 pode-se observar que em consonância com as outras linguagens, as melhores taxas de precisão são proporcionadas quando se usa um número maior de amostras de referências. Com isso, pode-se perceber que o comportamento dos modelos em língua francesa se assemelha aos apresentados em língua portuguesa e espanhola. Isso pode estar ligado ao fato da língua francesa ser derivada do mesmo grupo linguístico que o português e o espanhol. Algumas expressões e regras gramaticais da língua francesa são semelhantes às do português e do espanhol, no entanto, elas diferem principalmente em gênero, preposições e formas verbais.

Figura 5.3 – Taxas de acurácia da abordagem conforme o número de frases e referência em língua francesa: (a) Dependente do autor (b) Independente do autor.



Fonte: autor (2017)

Tabela 5.5. Resultados – Modelo Dependente do Autor – Língua Francesa

Número de Frases (F_k)	**Número de Amostras de Referência**			
	R_p = 3	***R_p = 5***	***R_p = 7***	***R_p = 9***
10	66.9	66.7	66.9	67.1
50	77.1	78.9	78.5	80.1
100	84.3	87.1	89.1	89.9
200	86.7	87.5	89.7	91.1
300	89.9	90.1	90.1	92.9
400	88.7	92.1	93.5	94.7
500	89.1	93.5	95.1	95.1

Fonte: autor (2017)

Tabela 5.6. Resultados – Modelo Independente do Autor – Língua Francesa

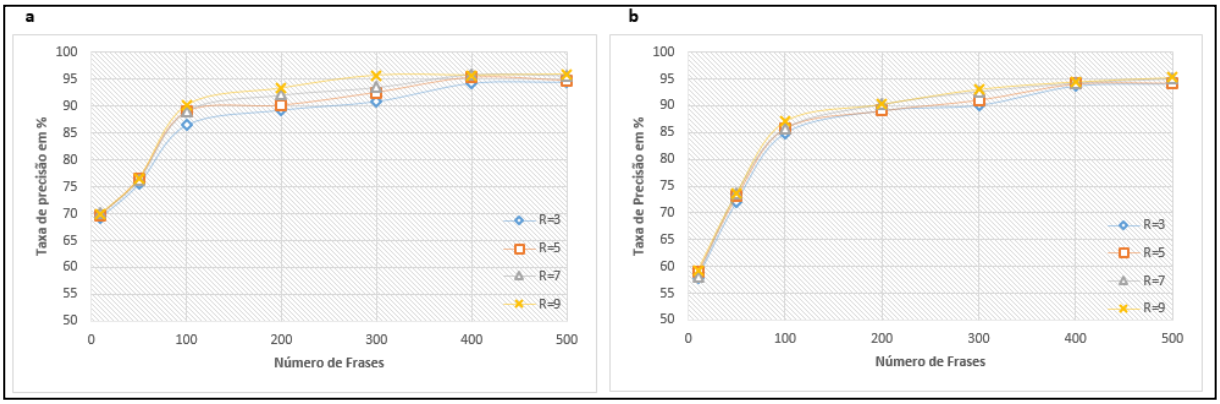
Número de Frases (F_k)	**Número de Amostras de Referência**			
	R_p = 3	***R_p = 5***	***R_p = 7***	***R_p = 9***
10	61.5	61.5	62.7	62.7
50	71.3	71.3	74.9	75.1
100	81.7	82.7	83.5	85.1
200	86.9	87.1	87.1	87.3
300	89.7	89.5	90.1	91.5
400	91.3	90.7	91.7	91.9
500	91.9	91.9	92.7	93.1

Fonte: autor (2017)

Para os testes realizados em língua alemã, as Figuras 5.4 (a) e (b) evidenciam a evolução dos resultados. Pode-se perceber com os resultados que em relação ao número de amostras de referências e a quantidade de informação por amostra, segue o mesmo padrão das outras línguas. Isso indica que o desempenho da abordagem proposta pode ser aplicado independentemente da origem da linguagem.

Na Figura 5.4 (a) é apresentado o resultado do modelo dependente do autor. A taxa de precisão varia de 58-59% para amostras com $F_k = 10$, enquanto que com $F_k = 50$ ou 100 a precisão é de 74-87%. Com $F_k >= 200$, a precisão atinge 86-95%. A Figura 5.4 (b) descreve os resultados para o modelo independente do autor, onde as taxas de precisão são ligeiramente menores do que o modelo dependente do autor. As diferenças nos resultados entre o modelo dependente e independente variam entre 2-6%, sugerindo que a abordagem é estável em relação ao volume de texto.

Figura 5.4 – Taxas de acurácia da abordagem conforme o número de frases e referência em língua alemã: (a) Dependente do autor (b) Independente do autor.



Fonte: autor (2017)

Nas Tabelas 5.7 e 5.8, são apresentados os resultados de forma estratificada para cada protocolo de teste. Percebe-se novamente que quanto maior o número de amostras de referências, melhor será o resultado do modelo.

Tabela 5.7. Resultados – Modelo Dependente do Autor – Língua Alemã

Número de Frases (F_k)	**Número de Amostras de Referência**			
	R_p = 3	***R_p = 5***	***R_p = 7***	***R_p = 9***
10	58.7	58.9	58.9	59.1
50	74.5	75.3	75.9	76.1
100	84.5	84.9	85.7	87.1
200	86.9	87.1	88.3	89.3
300	90.1	90.1	92.5	92.7
400	92.3	92.5	92.5	93.7
500	93.5	93.7	94.1	95.7

Fonte: autor (2017)

Tabela 5.8. Resultados – Modelo Independente do Autor – Língua Alemã

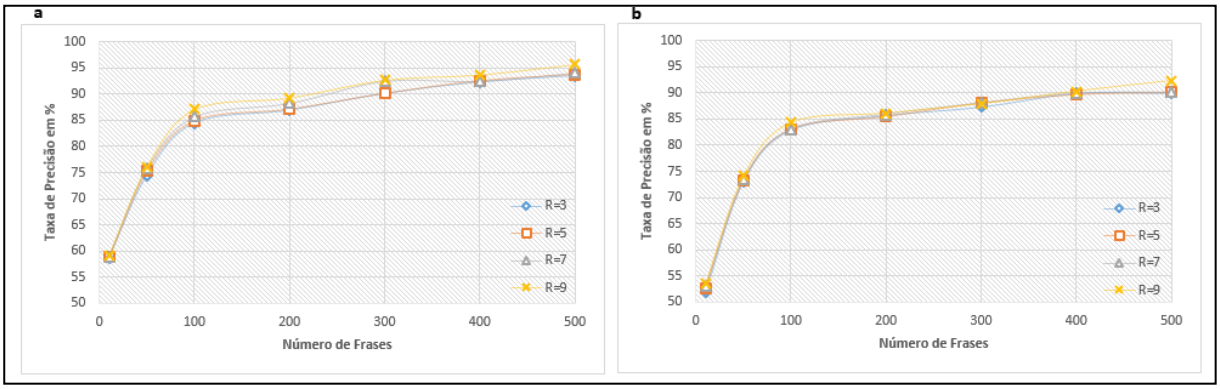
Número de Frases (F_k)	**Número de Amostras de Referência**			
	R_p = 3	***R_p = 5***	***R_p = 7***	***R_p = 9***
10	51.9	52.7	53.1	53.5
50	73.1	73.3	73.5	74.1
100	82.9	83.1	83.1	84.3
200	85.7	85.5	85.9	86.1
300	87.3	88.1	88.1	87.9
400	89.9	89.7	89.9	90.3
500	89.9	90.1	90.1	92.3

Fonte: autor (2017)

Finalmente, as Figuras 5.5 (a) e (b) descrevem os resultados para a língua inglesa, que tem origem germânica. Analisando o modelo dependente do autor, apresentado na Figura 5.5 (a), se tem uma precisão de 69-70% para amostras com F_k = *10*, enquanto que no modelo independente do autor, representado na Figura 5.5 (b) a precisão é de 57-59%. Com F_k = *50, 100* o modelo dependente do autor produz resultados com uma taxa de precisão entre 75-90%, enquanto que o modelo independente atinge 72-87%. Para amostras com *Fk >= 200*, ambos os modelos atingem uma precisão de 89-95%.

Em conformidade com o desempenho apresentado nas outras linguagens, as taxas de precisão em língua inglesa apresentam os melhores resultados com R_p = *9* e F_k *>= 200*. Isso é um indicador que existe estabilidade e consequentemente confiabilidade na abordagem proposta.

Figura 5.5 – Taxas de acurácia da abordagem conforme o número de frases e referência em língua Inglesa: (a) Dependente do autor (b) Independente do autor.



Fonte: autor (2017)

Tabela 5.9. Resultados – Modelo Dependente do Autor – Língua Inglesa

Número de Frases (F_k)	Número de Amostras de Referência			
	$R_p = 3$	$R_p = 5$	$R_p = 7$	$R_p = 9$
10	69.3	69.7	70.1	69.9
50	75.5	76.5	76.3	76.5
100	86.5	88.9	89.1	90.1
200	89.3	90.1	92.1	93.3
300	90.9	92.5	93.5	95.7
400	94.3	95.3	95.9	95.7
500	94.5	94.7	95.7	95.9

Fonte: autor (2017)

Tabela 5.10. Resultados – Modelo Independente do Autor – Língua Inglesa

Número de Frases (F_k)	Número de Amostras de Referência			
	$R_p = 3$	$R_p = 5$	$R_p = 7$	$R_p = 9$
10	57.9	58.9	58.1	59.3
50	72.1	73.1	73.7	73.7
100	84.9	85.8	85.8	87.1
200	89.1	89.1	90.3	90.3
300	90.1	91.1	92.7	93.1
400	93.7	94.1	94.3	94.5
500	94.1	94.3	95.1	95.3

Fonte: autor (2017)

Em correlato, nas Tabelas 5.11 e 5.12 são apresentados os resultados concatenados de todas as linguagens para os modelos dependente e independente do autor. Estes resultados são compostos pela média aritmética geradas tendo como base as amostras de referências (R_p).

Tabela 5.11. Precisão Média com base nas Referências – Dependente do autor

Número de Frases (F_k)	Linguagem				
	Portuguesa	**Espanhola**	**Francesa**	**Alemã**	**Inglesa**
10	72.4	67.3	66.9	58.9	69.8
50	82.7	78.0	78.7	75.5	76.2
100	88.7	87.2	87.6	85.6	88.7
200	91.6	91.5	88.8	87.9	91.2
300	94.2	94.1	90.8	91.4	93.2
400	95.6	96.0	92.3	92.8	95.3
500	96.1	96.6	93.2	94.3	95.2

Fonte: autor (2017)

Tabela 5.12. Precisão Média com base nas Referências – Independente do autor

Número de Frases (F_k)	Linguagem				
	Portuguesa	**Espanhola**	**Francesa**	**Alemã**	**Inglesa**
10	65.7	61.7	62.1	52.8	58.6
50	74.9	75.5	73.2	73.5	73.2
100	86.5	83.0	83.3	83.4	85.9
200	88.8	87.1	87.1	85.8	89.7
300	92.1	92.0	90.2	87.9	91.8
400	94.3	94.2	91.4	90.0	94.2
500	95.4	95.2	92.4	90.6	94.7

Fonte: autor (2017)

Analisando os dados das Tabelas 5.11 e 5.12, se vê que a amostras com $F_k = 10$ produzem resultados com ~63% de precisão. Quando se aumenta a quantidade de informação sintática para $F_k = 50$, a precisão média aumenta em ~13%, atingindo ~76%. Isso indica que um maior número de elementos sintáticos é um determinante em casos que envolvam a verificação de autoria. Para amostras com $F_k = (100, 200)$, a precisão média dos modelos é de 86% e 88%, respectivamente. Isto indica que, a partir de $F_k = 100$, os resultados tornam-se promissores. Com $F_k = (300, 400, 500)$ os resultados são satisfatórios, com uma precisão média de 91%, 93% e 94%, respectivamente. Na maioria dos casos, usando $R_p = 9$ e $F_k >= 300$, se obtêm resultados semelhantes, indicando que quando não se tem a disponibilidade de um conjunto de amostras com $F_k = (400, 500)$, o uso de amostras compostas por $F_k = 300$ apresentam uma precisão aceitável e superior a 90%.

Em geral, as linguagens testadas nos experimentos apresentam resultados promissores para a verificação de autoria. Destaca-se um desempenho levemente superior da língua portuguesa em relação as outras línguas. Isso se deve, provavelmente a riqueza gramatical e sintática da língua. Os resultados da língua espanhola expressam o segundo melhor desempenho, possivelmente devido ao fato de que ambas as línguas (português e espanhola) possuírem estruturas sintáticas semelhantes. A língua inglesa apresenta o terceiro melhor desempenho, com uma taxa de precisão superior as línguas francesa e alemã. No entanto, tais diferenças entre as línguas apresentam uma variação de ~5%. Pode-se concluir então, que o conjunto de atributos propostos é discriminante em um ambiente multilíngue.

Em todos os experimentos da verificação de autoria, o modelo dependente do autor obteve melhores resultados do que o modelo independente do autor. A diferença média entre os dois modelos é de 2-3%, indicando que ambos podem ser aplicados na verificação de autoria. Para todos os idiomas testados, os resultados com amostras que contém um número

maior de frases (quantidade de conteúdo) são melhores do que com amostras que possuem uma menor quantidade de informações. Isso se deve a reduzida variabilidade da estrutura linguística em textos menores - quanto maior a variabilidade melhor o desempenho.

Quanto ao número de amostras de referência, quanto maior o número de referências utilizadas no processo de decisão, maior o ganho. Isso ocorre porque há mais comparações com amostras positivas, permitindo uma maior taxa de discriminação pelo classificador.

Na próxima seção são apresentados os resultados para a identificação de autoria.

5.3 Identificação de Autoria

O processo de identificação de autoria tem como objetivo principal identificar o autor desconhecido de um determinado documento. Consiste em confrontar um autor A_c com uma base de possíveis autores. Então, se considera um texto S_u escrito por um autor desconhecido. Para identificar o autor A_c, onde $C = \{ \in N \mid 1 \leq c \leq \delta\}$ e δ é o número de autores da base de dados, realiza-se a maximização da relação $F_d = max\ \{D_i\ (x, R_c)\}$, onde D_i é o modelo treinado para os modelos dependente e independente do autor, que tem por função indicar a estimativa de probabilidade a posterior, indicando se S_u e a referência R_p pertencem a um mesmo autor. Neste caso, usa-se a mesma regra que na abordagem de verificação de autoria para determinar F_d.

O método proposto também fornece uma lista de amostras que se assemelham à amostra questionada. O tamanho dessa lista, que também é conhecida do *Top-list*, depende do número ψ dos resultados mais prováveis, onde $\psi = \{1, 3, 5, 10\}$. Os resultados são expressos pelas listas Top-1, Top-3, Top-5 e Top10. Isso significa que uma lista de eventos será considerada correta se pelo menos uma ocorrência estiver listada.

Portanto, uma taxa de precisão de 100% no Top-1 não é necessária em cada análise, por exemplo. Um especialista pode tomar uma decisão com base em uma lista de Top-5 ou Top-10. Contudo, é desejável um resultado próximo de 100% no Top-1.

Analisaram-se os resultados da identificação de autoria, levando em consideração a taxa de precisão em decorrência das quantidades de frases por amostra e do número de referências. Os resultados expressos nas Tabelas 5.13 a 5.17 são dados pela média aritmética das quatro diferentes quantidades de amostras de comparação (referências). Apresentam-se os

resultados para o Top-1, Top-3, Top-5 e Top-10 nos modelos dependente e independente do autor.

Na Tabela 5.13 são apresentados os resultados da identificação de autoria para a língua portuguesa. Para os textos em português os melhores resultados foram atingidos usando o modelo dependente do autor, pois a linguagem possui uma grande variabilidade de elementos gramaticais e sintáticos em relação as outras. Isso faz com que um modelo treinado para cada autor em específico se saia melhor que um modelo generalista, como é o caso do modelo independente. Quando usado $F_k = 10$ os resultados são relativamente baixos quanto a taxa de precisão, tendo no modelo dependente do autor 54,3% no Top-1, chegando a 70,7% no Top-10. Já no modelo independente, os resultados são ainda menores, tendo 46% de precisão no Top-1 e 58% no Top-10. Estes resultados demonstram que o uso de pouca informação sintática não é eficiente o suficiente para o uso no processo de identificação de autoria. Já quando se aumenta a quantidade de informação sintática em cada amostra, se percebe a melhoria dos resultados. Por exemplo, quando $F_k = 50$, a taxa de precisão tem um incremento de 24% no Top-1 e 15% no Top-10, para o modelo dependente. Assim, sucessivamente, há um incremento na taxa de precisão dos dois modelos a cada aumento da quantidade de frases por amostra, até chegar ao limite de 500 frases por amostra ($F_k = 500$). Neste caso, é onde os melhores resultados foram obtidos, com uma precisão de 89% para o Top-1 e de 99% no Top-10, no modelo dependente do autor.

Tabela 5.13. Identificação de Autoria – Resultados da Língua Portuguesa

Número de Frases (F_k)	**Dependente do Autor**				**Independente do Autor**			
	Top-1	**Top-3**	**Top-5**	**Top-10**	**Top-1**	**Top-3**	**Top-5**	**Top-10**
10	54.3	57.0	63.5	70.7	46.0	49.1	51.8	58.0
50	78.1	79.2	82.2	85.7	62.2	63.6	65.4	71.1
100	84.7	85.6	87.9	90.3	71.3	72.7	75.7	80.5
200	86.6	86.8	89.8	93.4	73.0	75.0	76.7	82.9
300	87.2	88.5	91.1	96.4	75.9	78.1	79.4	84.8
400	88.5	89.1	93.5	97.8	77.1	79.6	81.3	85.8
500	**89.6**	**90.2**	**94.4**	**98.8**	**79.0**	**79.7**	**82.1**	**86.5**

Fonte: autor (2017)

Em língua espanhola os resultados são detalhados na Tabela 5.14. Percebe-se que o desempenho para o espanhol é semelhante ao da língua portuguesa, uma vez que possuem similaridades em suas estruturas linguísticas. Em consonância com os outros resultados já apresentados, percebe-se que o uso de $F_k = 10$ não possui elementos suficientes para uma boa discriminação de autoria. Com uso de $F_k = 50$, ainda gera resultados com taxas de precisão

relativamente abaixo da média, porém, aplicáveis em alguns casos em que a precisão não precisa ser próxima a 100%. Entretanto, percebe-se que o conjunto de características propostas começa a obter resultados promissores quando se usa acima de 100 frases por amostra, indicando que quanto mais se saber sobre os padrões sintáticas de cada autor, melhor serão as chances de ter sucesso no processo de identificação de autoria.

Tabela 5.14. Identificação de Autoria – Resultados da Língua Espanhola

Número de Frases (F_k)	**Dependente do Autor**				**Independente do Autor**			
	Top-1	**Top-3**	**Top-5**	**Top-10**	**Top-1**	**Top-3**	**Top-5**	**Top-10**
10	56.4	62.9	65.9	70.8	46.7	49.2	53.6	59.8
50	73.9	76.0	77.4	81.2	62.8	65.6	70.3	77.6
100	83.3	85.9	86.2	89.2	71.4	73.8	75.5	81.7
200	85.8	87.1	87.3	90.5	71.9	74.2	76.5	83.0
300	88.3	88.1	89.1	92.0	73.9	76.7	78.7	86.8
400	88.8	88.3	90.2	94.2	75.8	77.5	80.6	88.0
500	**89.7**	**90.1**	**91.9**	**97.0**	**78.6**	**79.1**	**81.5**	**89.2**

Fonte: autor (2017)

Para os experimentos em língua francesa, obteve-se taxa de precisão aceitáveis, conforme pode ser observado na Tabela 5.15. Mais uma vez, o modelo dependente do autor é superior ao modelo independente, sendo 10% mais preciso na média dos resultados. Na lista do Top-1, observa-se que existe pouca variação quando $F_k >= 100$, com precisão média de 70-75% quando se usa o modelo independente do autor, e 80-83% para o modelo dependente do autor. Na lista do Top-10, a variação aumenta para em torno de 6% no modelo dependente e de 10% no modelo independente. No geral, constata-se que taxa de precisão em língua francesa se mostrou promissora e aplicável, chegando a taxas de acerto acima de 80% no Top-1, de 90% no Top-3, de 93% no Top-5 e 97% no Top-10. Portanto, os recursos utilizados na abordagem proposta também se mostram eficientes em língua francesa.

Tabela 5.15. Identificação de Autoria – Resultados da Língua Francesa

Número de Frases (F_k)	**Dependente do Autor**				**Independente do Autor**			
	Top-1	**Top-3**	**Top-5**	**Top-10**	**Top-1**	**Top-3**	**Top-5**	**Top-10**
10	59.4	62.4	65.1	75.5	51.3	53.2	56.9	66.6
50	69.5	73.5	76.4	82.5	61.8	63.6	65.5	72.9
100	80.3	82.5	85.2	91.0	70.2	71.6	73.0	80.8
200	81.1	83.9	85.8	91.1	73.1	75.7	78.2	82.7
300	81.1	86.2	88.5	93.1	74.1	77.6	80.4	85.6
400	**83.9**	89.5	90.8	93.8	75.0	77.8	81.2	86.8
500	83.8	**90.4**	**93.5**	**97.0**	**75.7**	**80.4**	**83.4**	**90.1**

Fonte: autor (2017)

Em língua alemã, os resultados são os mais baixos em termos de precisão, entre as línguas testadas. Na Tabela 5.16, pode-se observar que a diferença entre o melhor e o pior desempenho é de ~4% no modelo dependente do autor, e de ~2% no modelo independente do autor. Portanto, considerando os resultados de exatidão da lista do Top-1, que variam entre 45-86% e 65-96% de acordo com o número de frases e modelo de identificação, a abordagem permanece aceitável e promissora em língua alemã.

Tabela 5.16. Identificação de Autoria – Resultados da Língua Alemã

Número de Frases (F_k)	**Dependente do Autor**				**Independente do Autor**			
	Top-1	**Top-3**	**Top-5**	**Top-10**	**Top-1**	**Top-3**	**Top-5**	**Top-10**
10	53.2	58.1	62.6	69.9	47.0	48.4	52.8	65.7
50	68.1	71.1	74.1	78.3	64.6	67.3	70.5	79.4
100	80.2	82.5	85.3	90.1	71.8	74.1	75.9	82.5
200	83.4	83.8	86.8	91.4	73.0	75.9	76.4	84.0
300	84.7	84.5	88.4	92.5	76.5	79.5	81.7	86.0
400	85.5	85.4	90.0	94.7	77.3	80.1	82.5	88.7
500	**86.3**	**88.3**	**91.5**	**96.0**	**77.4**	**80.5**	**84.1**	**88.8**

Fonte: autor (2017)

A Tabela 5.17 resume os resultados para a língua inglesa. Neste caso, as taxas de precisão para o modelo dependente do autor com $F_k >= 50$ são superiores a 90%, enquanto que para o modelo independente do autor, as taxas de precisão estão acima de 79% no Top-10. Portanto, o modelo dependente do autor pode reduzir o número de autores na lista de investigação em mais de 90% dos casos.

Tabela 5.17. Identificação de Autoria – Resultados da Língua Inglesa

Número de Frases (F_k)	**Dependente do Autor**				**Independente do Autor**			
	Top-1	**Top-3**	**Top-5**	**Top-10**	**Top-1**	**Top-3**	**Top-5**	**Top-10**
10	57.8	62.1	65.6	74.9	49.8	53.5	56.8	62.9
50	79.3	82.1	85.6	90.2	69.3	71.7	74.2	80.3
100	81.1	83.8	87.4	90.9	73.3	75.8	76.9	82.3
200	84.1	85.6	87.8	92.1	73.7	76.0	77.2	84.6
300	85.2	85.4	89.2	95.5	74.1	77.0	79.9	86.4
400	85.8	86.8	90.0	96.4	75.0	78.8	81.4	88.0
500	**85.4**	**89.0**	**90.7**	**96.8**	**75.8**	**79.7**	**83.1**	**90.4**

Fonte: autor (2017)

Em geral, o modelo dependente do autor proporciona os melhores resultados. De fato, a abordagem dependente do autor gera um modelo para cada autor, que é um fator essencial no processo de identificação de autoria, pois o classificador procura um único autor entre todos os autores da lista.

Em resumo, para as 5 línguas testadas, amostras com $F_k = 10$, o desempenho varia entre 45-59% de precisão no Top-1. No entanto, a precisão aumenta conforme se aumenta o tamanho da lista. Para o modelo dependente do autor, amostras contendo $F_k = 50$ gerou taxas de precisão entre 68-78% em todas as linguagens do experimento, no Top-1. Já, quando foi analisada a lista do Top-10, as taxas de precisão têm um incremento entre 7-13%. Para $F_k = 100$, a precisão no Top-1 varia entre 80-84%, enquanto que no Top-10 os resultados são entre 89-91%. Para $F_k >= 200$, os resultados são promissores, pois as taxas de precisão são de 81-89% (Top-1), 83-90% (Top-3), 85-94% (Top-5), e 90-98% (Top-10). Estes resultados demonstram que a abordagem é eficaz para a identificação de autoria em múltiplas linguagens.

5.4 Experimentos Adicionais

Adicionalmente, foram realizados experimentos com 20 autores para cada língua, totalizando 100 autores. O objetivo foi verificar a confusão entre os autores de diferentes línguas. A Tabela 5.18 apresenta a matriz de confusão gerada nesta experiência. Foi utilizada a abordagem dependente do autor, pois obteve os melhores resultados nos experimentos anteriores.

Tabela 5.18. Matriz de Confusão

Língua	Portuguesa	Espanhola	Francesa	Alemã	Inglesa
Portuguesa	**90%**	10%	0%	0%	0%
Espanhola	10%	**85%**	5%	0%	0%
Francesa	5%	15%	**80%**	0%	0%
Alemã	0%	0%	5%	**75%**	20%
Inglesa	0%	0%	0%	10%	**90%**

Fonte: autor (2017)

Por exemplo, a Tabela 5.18 indica que os autores da língua portuguesa são classificados corretamente em 90% dos casos, ocorrendo 10% de confusão com autores da língua espanhola.

Em geral, a precisão media da identificação de autoria é de 84%, independentemente da língua. Estes resultados indicam que os autores das línguas latinas (português, espanhol e francês) só sofrem confusões entre si. A principal confusão com os autores das línguas advindas da vertente germânica é principalmente entre outros autores da língua germânica (alemão e inglês), com exceção de um erro de 5% ocorrido entre autores da língua alemã, que foram confundidos como textos escritos por autores de língua francesa. Portanto, afirma-se

que as linguagens similares e advindas de uma mesma corrente linguística, contem aspectos sintáticos que são herdados, e isto, indica que existem similaridades entre as línguas com a mesma raiz.

Por conseguinte, é apresentado o desempenho estratificado por vetor de atributos em separado e em conjunto para todas as linguagens dos experimentos. Neste caso, foi usado como parâmetro $R_p = 9$ que mostrou os melhores resultados. Quanto a quantidade de conteúdo de cada amostra, foram usadas $F_k = \{10, 100, 500\}$ frases por amostra, com o intuito de observar o comportamento quanto se tem uma variação de pouca a grande quantidade de informação em cada amostra. Como estratégias de treinamento e testes, foi feito uso da verificação de autoria por meio do modelo dependente do autor, que obteve no geral, os melhores resultados. Então, na Tabela 6.19 os resultados são apresentados.

Tabela 5.19. Resultados da Verificação de Autoria – Taxa de Acerto por Vetor

Língua	F_k	V_{t1}	$V_{t1} + V_{t2}$	$V_{t1} + V_{t2} + V_{t3}$	$V_{t1} + V_{t2} + V_{t3} + V_{t4}$	$V_{t1} + V_{t2} + V_{t3} + V_{t4} + V_{t5}$
	10	19.1	35.3	69.1	71.2	73.9
Portuguesa	100	23.1	39.7	82.3	87.5	90.3
	500	25.2	41.5	88.5	95.1	97.9
	10	9.9	29.3	61.3	65.7	68.9
Espanhola	100	17.9	30.7	81.1	85.5	88.3
	500	22.1	32.9	89.1	96.9	98.3
	10	17.1	24.1	60.9	64.3	67.1
Francesa	100	26.7	32.9	83.3	87.1	89.9
	500	29.9	35.1	89.3	94.7	95.1
	10	8.9	19.7	55.9	57.7	59.1
Alemã	100	18.9	29.9	79.9	84.9	87.1
	500	21.1	36.3	85.7	93.5	95.7
	10	17.5	30.1	61.1	66.7	69.9
Inglesa	100	21.1	34.3	84.3	88.1	90.1
	500	24.3	36.9	88.9	94.3	95.9

Fonte: autor (2017)

Como pode ser observado na Tabela 5.19, o grupo de características que compõe V_{t1} (morfológicas) possuem uma contribuição média de 14.5% na taxa de acerto quando usado $F_k = 10$, sendo a menor de 8.9% em língua alemã, e a maior contribuição em língua portuguesa, com 19.1% de acerto. Quando $F_k = 100$, a taxa de acerto média entre as línguas ganha cerca de 7% de incremento, chegando a 21.5%. E, com $F_k = 500$, a média tem um crescimento de 3% em relação à quando é usada 100 frases por amostra, o que equivale a uma taxa de acerto média de 24.5% em todas as línguas. Quando se adiciona o grupo de características flexoras (V_{t2}), a média de acerto passa a ser de 27.7%, 33.5% e 36.5% para $F_k = 10$, 100 e 500, respectivamente. Isso representa um incremento de aproximadamente 13% na taxa de acerto quando são incluídas as características de flexão em conjunto com as morfológicas.

Quando se incluem as características sintáticas, representadas por V_{t3}, obtem-se um incremento significativo nas taxas de acerto, chegando a 61.7% com $F_k = 10$, 82.2% com $F_k = 100$, e de 88.3 com $F_k = 500$. Isso significa que existe um incremento médio de 34%, 48.7% e 51.8%, conforme se aumenta a quantidade de frases por amostra. Isso é uma evidência que as características sintáticas são realmente discriminantes no processo de atribuição de autoria. Adicionalmente, quando é inserido o grupo de características sintáticas auxiliares (V_{t4}), obtêm-se um ganho médio de 3,5% quando $F_k = 10$, 4.5% quando $F_k = 100$, e 6.6% com $F_k = 500$. Isso pressupõe que o grupo de características apresentados em V_{t4}, contribuem de forma significativa para com os resultados. E, por fim, quando se usa V_{t5}, que é composto por características baseadas nas distâncias entre os principais elementos sintáticos de uma frase, há um incremento médio de 2.3% nos resultados. Isso infere, que embora a contribuição seja pequena ela classifica de forma correta um maior percentual de amostras.

Então, pode-se relatar que as características sintáticas de estilo são mais discriminantes quando atuam em conjunto do que em grupos separados. A colaboração dos grupos de características morfológicas e flexoras são de suma importância, uma vez que denotam para cada autor a construção das frases sob o ponto de vista estrutural. Já as características sintáticas apresentam uma análise mais complexa de cada frase, chegando a níveis estruturais e funcionais de cada frase, onde é possível segregar cada autor ou em pequenos grupos de autores. Sendo assim, tais características se mostram essenciais quando se necessita trabalhar com a atribuição de autoria de textos.

Complementarmente, aplicou-se a abordagem em bases de textos jornalísticos, ou seja, de textos curtos, com no máximo 1000 palavras por amostra em língua portuguesa e inglesa.

O conjunto de textos curtos, é formada por textos extraídos de colunas de jornais brasileiros e britânicos. Usaram-se duas bases de dados já existentes e disponibilizadas em [VAR10] e [PET04]. A base de dados em Português é formada por 100 autores com 30 amostras de textos por autor. Enquanto que a base de dados em Inglês (CLEF) é formada por 20 autores com 20 amostras por autor (ver Tabelas 5.20 e 5.21).

Tabela 5.20. Resultados da Verificação de Autoria em Textos Jornalísticos

Linguagem	**Abordagem**	
	Dependente	**Independente**
Portuguesa	90.6	95.1
Inglesa	86.8	90.0

Fonte: autor (2017)

Tabela 5.21. Resultados da Identificação de Autoria em Textos Jornalísticos

Linguagem	**Abordagem**
	Dependente
Portuguesa	85.0
Inglesa	75.5

Fonte: autor (2017)

A Tabela 5.20 apresenta os resultados da verificação de autoria, onde se pode observar que no idioma Português, a abordagem proposta apresenta taxas de acerto acima de 90%, no modelo dependente do autor, e de 95% no modelo independente, evidenciado resultados promissores para o método proposto.

Para os textos em língua inglesa foi usada a mesma abordagem, com o objetivo de verificar a eficácia em outro idioma cuja origem não fosse latina, como a língua portuguesa. Para tanto, foi usada a base de dados CLEF, onde foram constatados resultados considerados promissores, já que as estruturas usadas nos idiomas Português e Inglês possuem diferenças substanciais. No modelo dependente obteve-se, uma performance de 86.8%, enquanto que no independente obteve-se 90% de acurácia.

Nos testes de identificação de autoria, usou-se a abordagem dependente do autor, por ser a que apresentou os melhores resultados. Na Tabela 5.21, são apresentados os resultados para Top-1, onde é possível observar que, em textos curtos, a acurácia do modelo proposto apresentou um erro de 15% no idioma Português e de 24.5 % no Inglês. Com estes resultados

pode-se observar que a abordagem proposta obteve resultados considerados promissores em atribuição de autoria de textos jornalísticos.

5.5 Estudos Comparativos

Na Tabela 5.22, é apresentada uma comparação dos resultados proporcionados pela abordagem proposta com alguns estudos relacionados pela literatura. Esta comparação não é exata, porque os protocolos e bases de dados não são os mesmos. No entanto, é possível estimar as contribuições feitas pela abordagem proposta. Os resultados apresentados baseiam-se nos melhores resultados de cada abordagem em ambiente multilíngue.

Em língua portuguesa, compara-se a nossa abordagem com os trabalhos desenvolvidos por [PAV08] e [VAR10]. Neste caso, a abordagem proposta gera resultados entre 11-15% maiores nas taxas de precisão da verificação de autoria, e de 17-20% na identificação de autoria. Isso demonstra que a proposta apresentada se mostra eficaz e aceitável em língua portuguesa.

Em língua espanhola, a verificação de autoria atinge uma taxa de precisão de ~98%, cerca de 26% maior que a apresentada por [HAL16], embora utilizando uma base de dados diferente. Na identificação de autoria, a taxa de precisão é de 91%. Portanto, considera-se os resultados para língua espanhola em um nível aceitável de desempenho.

Para a língua francesa, a taxa de precisão para a verificação de autoria é de 95%. Na identificação de autoria, a taxa de precisão é de 86%, semelhante aos resultados relatados por [SAV11]. Assim, pode-se concluir que o conjunto de atributos propostos é aplicável em casos de atribuição de autoria em língua francesa.

Analisando os resultados em língua alemã, obteve-se uma taxa de precisão de 95% para a verificação de autoria, representando um incremento de 16% sobre o método proposto por [PET04]. Para a identificação de autoria, a taxa de exatidão é de 88%, o que significa ser ao menos 3% melhor que os resultados relatados por [HAL16]. Assim, se concluí que a abordagem também é eficaz em língua alemã, sugerindo a aplicabilidade às línguas anglo-saxônicas.

Na língua inglesa, a abordagem proposta obteve melhores resultados em termos de verificação de autoria do que os métodos apresentados em [ZHE06] [HAL16]. Para identificação de autoria, o método apresentado obteve melhores resultados do que os apresentados em [PEN03] [ZHE06] [SAV11] [EBR13]. No entanto, considera-se a

abordagem proposta apresenta resultados promissores em língua inglesa, porque se usa em média de 5 a 10 vezes mais autores que outras abordagens.

Com as comparações pode-se observar que quanto maior o número de frases por amostra, melhor o desempenho. Isso se aplica a todos os idiomas testados. Com um pequeno número de frase por amostra (F_k *=10*), os resultados tendem a ter baixa taxa de precisão. Usando F_k *= (50, 100, 200)* os resultados são satisfatórios e promissores. Em geral, com F_k *=>300*, as taxas de precisão indicam uma capacidade de discriminação maior e mais eficaz.

Pode-se observar que a abordagem proposta é robusta e aplicável em ambientes multilíngues que envolvam as línguas portuguesa, espanhola, francesa, alemã e inglesa. Descobriu-se que a abordagem é estável, produzindo praticamente o mesmo padrão de variação em diferentes línguas. Portanto, considera-se que a aplicação dos atributos definidos que descrevem as funções sintáticas que cada palavra exerce dentro de uma frase é discriminante e pode ser usada para a atribuição de autoria.

Tabela 5.22 Comparação dos Resultados com a Literatura

Literatura	Base de Dados	Características	Autores	Linguagem	Abordagens	
					Verificação	Identificação
Peng (2003)	Textos literários e colunas de jornais	Caracteres *n-grams*	20	Grega	-	90%
			8	Inglesa	-	98%
			8	Chinesa	-	94%
Abbasi (2005)	Mensagens *online*	diversas	25–100	Inglesa	-	69–88%
			25–100	Chinesa	-	49–100%
Zheng (2006)	Mensagens *online*	diversas	20	Inglesa	94%	89%
			20	Chinesa	82%	57%
Pavelec (2008)	Colunas de Jornais	Palavras-função	20	Portuguesa	83%	73%
Varela (2010)	Colunas de Jornais	Palavras-função	100	Portuguesa	87%	77%
Savoy (2011)	Livros de Romance	lemmas	9	Inglesa	-	92–100%
			15	Alemã	-	69–85%
			11	Francesa	-	70–100%
Savoy (2012)	Colunas de jornais	Palavras-função	20	Inglesa	-	82%
			20	Italiana	-	91%
Ebrahimpour (2013)	Textos literários e biblícos	Palavras-função	10	Inglesa	-	92–97%
			8	Grega	-	86–91%
Halvani (2016)	Variados	Variado	milhares	Inglesa	73%	-
				Espanhola	72%	-
				Alemã	79%	-
Nossa Abordagem	**Textos literários**	**Funções sintáticas**	**100 por linguagem**	**Portuguesa**	**98%**	**93%**
				Espanhola	**98%**	**91%**
				Francesa	**95%**	**86%**
				Alemã	**95%**	**88%**
				Inglesa	**95%**	**86%**
	Textos Jornalísticos		**20 por linguagem**	**Português**	**90-95%**	**85%**
				Inglês	**86-90%**	**75%**

Fonte: autor (2017)

5.6 Considerações do Capítulo

Neste capítulo foram apresentados os resultados dos experimentos deste trabalho. Aplicou-se a abordagem em cinco idiomas diferentes, caracterizando a aplicação do método em um ambiente multilíngue. Neste contexto, foram usadas duas estratégias nos experimentos, que são a verificação e a identificação de autoria.

Por conseguinte, no próximo capítulo são apresentadas as conclusões obtidas com a abordagem proposta, bem como, os futuros desdobramentos deste trabalho.

Conclusão

Neste trabalho foi proposta uma abordagem multilíngue baseada na linguística computacional utilizando características sintáticas de estilo para aplicação em casos que envolvam a atribuição de autoria em documento digitais. O principal propósito é resolver o problema da atribuição de autoria, que visa saber se um determinado texto foi elaborado por um autor em específico, ou então, identificar o autor do documento entre uma lista de autores. Então, para solucionar este tipo de problema e apresentar a originalidade deste trabalho, aplicou-se um conjunto de funções sintáticas exercidas por cada palavra, sendo que cada palavra pode mudar de função de acordo com sua posição na frase, indicando elementos variantes que denotam um perfil de estilo para cada autor.

Para averiguar a abordagem proposta, foi feito uso de textos literários escritos em cinco idiomas diferentes, sendo três línguas latinas (português, espanhol e francês) e duas línguas anglo-saxônicas (alemão e inglês). Como estratégias de treinamentos e testes foram utilizados os modelos dependente e independente do autor, bem como, a verificação e a identificação de autoria. Como classificador base foi usado o SVM duas classes com kernel linear, ao qual foi aplicado uma combinação de classificadores, de onde obteve-se um somatório de pesos ponderados de cada vetor de característica para obtenção do resultado final. Diante destes fatos, chegaram-se a uma série de conclusões sobre a nossa abordagem, as quais são relatadas nos próximos parágrafos.

Foi construída e disponibilizada uma base de dados de textos literários em cinco idiomas distintos, sendo eles: português, espanhol, francês, alemão e inglês. Para cada linguagem são 100 autores diferentes com amostras de obras literárias em domínio público. Neste caso, a base de dados literária é uma contribuição deste trabalho para com a comunidade cientifica que queira fazer uso de textos consagrados da literatura em seus experimentos.

Nos experimentos deste trabalho, foram efetuados testes com bases de dados com textos heterogêneos, sendo eles, curtos (textos jornalísticos), longos (textos literários), de diferentes assuntos e de diferentes linguagens. Percebemos que o modelo proposto gerou

resultados médios superiores à 90% de acerto em textos longos e com uma acurácia um pouco menor em textos curtos, chegando a uma média 85%. Isso se dá pelo fato, dos textos longos possuírem uma maior quantidade de elementos sintáticos, gerando assim, modelos mais específicos e detalhados para o processo de tomada de decisão.

Com base nos resultados, percebe-se que as funções sintáticas de cada palavra são bons elementos para uso em atribuição de autoria. Entre os 5 vetores de características que foram usados, se destaca o grupo composto por características sintáticas essenciais, evidenciados no vetor 3 (V_{t3}), tais como: sujeito, predicado, objeto direto, objeto indireto e adjunto adverbial, por exemplo. Neste caso, o ganho de acurácia deste grupo é responsável por cerca de 40% das taxas de acerto. É importante salientar, que todas os atributos sintáticos que foram usados nos experimentos são aplicáveis no conjunto de línguas utilizado nos testes.

Em relação as estratégias de treinamento e testes aplicados, percebe-se que o modelo dependente do autor, que busca extrair características únicas de cada autor, possui melhor desempenho em comparação ao modelo independente do autor, isso porque ele gera um modelo específico por autor. Entretanto, possui um maior custo computacional porque gera um modelo para cada autor da base de textos. Quanto as estratégias de verificação e identificação, em conformidade com a literatura constata-se que a verificação, que utiliza o método um-contra-um possui um melhor desempenho. No entanto, a identificação de autoria que trabalha com um processo mais complexo, já que confronta o texto questionado contra todos os autores da base, relatou taxas de acerto semelhantes com a literatura [PEG03] [ABB05] [ZHE06] [PAV08] [VAR10] [SAV12] [EBR13] [HAL16].

Na verificação de autoria obtiveram-se resultados com taxas de acurácia variando entre 63-97% em língua portuguesa. Em língua espanhola a performance variou entre 65-98%, conforme cada protocolo. O desempenho em língua francesa foi entre 66-95%. Já em língua alemã os resultados variaram entre 58-95% de acurácia. E, em língua inglesa obteve-se desempenho entre 69-95% de acerto. Então, pode-se observar que a abordagem produz uma variação similar nos resultados, indicando que o modelo é estável, podendo assim, ser aplicada em casos que envolvam a verificação de autoria.

Na identificação de autoria de língua portuguesa a abordagem proposta relatou resultados variando de 46-89% de acerto no Top-1 e de 58-98% no Top-10, conforme cada estratégia e protocolo aplicados. Em língua espanhola a variação das taxas de acerto do Top-1 foi entre 46-89%, e no Top-10 entre 59-97%. Nos experimentos com textos de língua francesa

a variação das taxas de acerto do Top-1 foi entre 51-83%, e no Top-10 houve uma variação entre 66-97%. Em língua alemã a performance variou entre 47-86% de acerto no Top-1 e de 65-96% no Top-10. E, nos textos de língua inglesa a abordagem aplicada relatou acertos entre 49-85% no Top1, e 62-96% no Top-10.

O protocolo experimental demonstrou a importância do uso de amostras de referência nos testes e seu impacto nos resultados. Quando se confrontam um texto questionado com 3, 5, 7 e 9 amostras de referências de cada autor da base de textos, consegue-se identificar que quanto mais amostras de referência se tem de um autor suspeito, melhores são os índices de taxa de acerto. No entanto, sabe-se que muitas vezes é difícil coletar um montante de amostras de textos que seja suficiente para análise, então, é possível perceber por meio dos resultados que uma quantidade mínima de 3 amostras de referências já se mostra viável para atribuição de autoria.

Paralelamente, constatou-se que os resultados tendem a ter melhores taxas de acerto quando há um incremento da quantidade de textos por amostra, ou seja, o melhor modelo está relacionado a um maior número de frases por amostra. Isso, evidencia que quanto mais material sintático se possui de cada autor, melhor é capacidade de discriminação da abordagem.

Com isso, pode-se relatar que a abordagem proposta se mostrou estável, produzindo praticamente o mesmo padrão de variação nas diferentes línguas testadas. Foi possível observar taxas de acerto semelhantes entre as línguas latinas e anglo-saxônicas, não havendo diferenças substancias entre as línguas. Isso indica, que os modelos gerados são consistentes.

Finalmente, conclui-se que a abordagem apresenta um meio promissor de atribuição de autoria para as cinco linguagens testadas. É notável que o conjunto de atributos que propomos é aplicável nos cinco diferentes grupos de linguagem, tais como as linguagens latinas e anglo-saxônicas.

Trabalhos Futuros

Outros trabalhos devem ser realizados com o objetivo avaliar o conjunto de atributos, utilizando diferentes gêneros e tipos de textos. Entre as sugestões, citam-se:

- Inserir novos conjuntos de atributos:
  - semânticos (sinônimos, antônimos, polissemia, homônimos e parônimos),

- figuras de linguagem (metáfora, hipérbole, metonímia e eufemismo, etc.),
- vícios de linguagem (barbarismo, neologismo, ambiguidade, cacófato, etc.).

- Aplicar a abordagem em diferentes gêneros e tipos de textos;
- Adaptar a abordagem para outras línguas, incluindo línguas indo-europeias não incluídas neste trabalho.
- Aplicar a abordagem em textos psicografados.

Referências Bibliográficas

[ABB05] ABBASI, A. CHEN, H. **Applying authorship analysis to extremist group web forum messages**. IEEE Intelligent Systems, Vol. 20, Nº 5. 67-75, 2005.

[ABB08] ABBASI, A. CHEN, H. **Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace**. ACM Transactions on Information Systems, Vol. 26, Nº 2. 7:1-29, 2008.

[ARG03] ARGAMON, S., KOPPEL, M., FINE, J. SHIMONI, A. **Gender, genre, and writing style in formal written texts**. Text, 23(3), 321–346, 2003.

[ARG07] ARGAMON, S., WHITELAW, C., CHASE, P., HOTA, S., GARG, N. LEVITAN, S. **Stylistic text classification using functional lexical features**. Journal of the American Society for Information Science and Technology, 58(6), 802–821, 2007.

[BAA96] BAAYEN, H., VAN HALTEREN, H. TWEEDIE, F.J. **Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution**. Literary and Linguistic Computing, 11, 121–131, 1996.

[BAA02] BAAYEN, H.VAN HALTERAN, H. NEIJT, A. TWEEDIE, F. **An experiment in authorship attribution**. Journees Internationales d'Analyse statistique des Donnees Textuelles, 2002.

[BAA13] BRITISH ASSOCIATION FOR APPLIED LINGUISTICS (BAAL). http://www.baal.org.uk Acessado em 10 de outubro de 2013.

[BAY91] BAILEY, G. **The apparent time construct, Language Variation Change**. 3:241–264, 1991.

[BEL08] BELAK, B. BELAK, S. PESA, A. R. **Stylometry – definition and development.** Annals of DAAAM & Procedings.85-94, 2008.

[BES88] BESNIER, N. **The linguistic relationships of spoken and written Nukulaelae registers**. Language, 64 (4), 707-736, 1988.

[BIB16] BIBLIOTECA VIRTUAL DE LITERATURA. http://www.biblio.com.br

[BIC16] BICK, E. **Visual Interactive Syntax Learning**. acessible at: http://beta.visl.sdu.dk/, 2016.

[BND16] BIBLIOTECA NACIONAL DIGITAL DO BRASIL. http://bndigital.bn.br/

[BRI63] BRINEGAR, C. S. **Mark Twain and the Quintus Curtius Snodgrass Letters: A statistical test of authorship**. Journal of the American Statistical Associations, 58, 85-96, 1963.

[BRY62] BRYANT, M. **English in the Law Courts: the part that articles, prepositions and conjunctions play in legal decisions**. Frederick Ungar: New York, 1962.

[BUR87] BURROWS, J. F. **Word patterns and story shapes: The statistical analysis of narrative style**. Literary and Linguistic Computing, 2, 61-70, 1987.

[CHA05] CHASKI, Carole E. **Who's at the keyboard? - authorship attribution in digital evidence investigations.** International Journal of Digital Evidence, 4(1), 2005. Spring 2005

[CUL87] CULBERT, S. S. **The principal languages of the World.** In: The World Almanac and Book of Facts - 1987, p. 216. Pharos Books, New York, EUA.

[DEV01] DE VEL, O., ANDERSON, A., CORNEY, M. MOHAY, G. **Mining e-mail content for author identification forensics**. ACM SIGMOD Rec. 30, 4, 55–64, 2001.

[DIC14] DICIONÁRIO ON LINE DE PORTUGUÊS. Disponível em: http://www.dicio.com.br. Acesso em: (24 de janeiro de 2014).

[DIE03] DIEDERICH, J. KINDERMANN, J. LEOPOLD, E. PAASS, G. **Authorship attribution with support vector machines**. Applied Intelligence, (1), 2003.

[DMP16] PORTAL DOMÍNIO PÚBLICO. http://www.dominiopublico.gov.br/

[EBR13] EBRAHIMPOUR, M. PUTNINS, T. J. BERRYMAN, M. J. ALLISON, A. NG, B. W. H. ABBOT, D. **Automated Authorship Attribution using advanced signal classification techniques**. PLOS One, Vol. 8. 2013.

[EST16] BIBLIOTECA VIRTUAL DE LITERATURA BRASILEIRA E PORTUGUESA. http://www.estudantes.com.br/

[EUR02] EURASIATIC GREENBERG. **Indo-European and Its Closest Relatives: The Eurasiatic Language Family**. Stanford University Press, 2002.

[FIN06] FINN, A. KUSHMERICK, N. **Learning to classify documents according to genre**. Journal of the American Society for Information Science and Technology. 57 (11), 1506-1518, 2006.

[FLE14] FLÉRY, J. LARGERON, C. JUGANARU-MATHIEU, M. **UJM at CLEF in author verification based optimized classification trees**. Notebook for PAN at CLEF 2014. 1042-1048, 2014.

[FUC52] FUCKS, W. **On the mathematical analysis of style**. Biometrika, 39,122-129, 1952.

[GAM04] GAMON, M. **Linguistic correlates of style: Authorship classification with deep linguistic analysis features**. In Proceedings of the 20th International Conference on Computational Linguistics (pp. 611–617). Morristown, NJ: Association for Computational Linguistics, 2004.

[GUT16] PROJECT GUTENBERG. http://www.gutenberg.org/

[GUY02] GUYON, I. WESTON, J. BARNHILL, S. VAPNIK, V. **Gene selection for cancer classification using support vector machines**. Machine Leraning, vol. 46, pp. 389–422, 2002.

[HAL16] HALVANI, O. WINTER, C. PLUG, A. **Authorship verification for different languages, genres and topics.** Digital Investigation, vol. 16, pp. 33–43, 2016.

[HEF95] HOLMES, D. FORSYTH, R. **The Federalist revisited: News directions in authorship attribution**. Literary and Linguistic Computing, 10, 111-127, 1995.

[HIR07] HIRST, G. FEIGUINA, O. **Bigrams of syntactic labels for authorship discrimination of short texts**. Literary and Linguistic Computing, 22(4), 405–417, 2007.

[HOL98] HOLMES, D. I. **The Evolution of Stylometry in Humanities Scholarship**. Literary and Linguistic Computing, Vol. 13, Nº 3. 111-117, 1998.

[HON79] HONORE, A. **Some simple measures of richness of vocabulary**. Association for Literary and Linguistic Computing Bulletin, 7(2), 172–177, 1979.

[JUO06] JUOLA, P. **Authorship attribution for electronic documents**. In M. Olivier & S. Shenoi (Eds.), Advances in digital forensics II (pp. 119–130). Boston: Springer, 2006.

[JUO14] JUOLA, P. STAMATATOS, E. **Overview of the author identification task at PAN 2013**. Conference and Labs of the Evaluation Forum, 2014.

[KES03] KESELJ, V., PENG, F., CERCONE, N., THOMAS, C. **N-gram-based author profiles for authorship attribution**. In Proceedings of the Pacific Association for Computational Linguistics (pp. 255–264), 2003.

[KHM03] KHMELEV, D.V. TEAHAN, W.J. **A repetition based measure for verification of text collections and for text categorization**. In Proceedings of the 26th ACMSIGIR (pp. 104–110). NewYork: ACMPress, 2003.

[KJE94] KJELL, B. **Authorship attribution of text samples using neural networks and Bayesian classifiers**. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 1660. San Antonio: IEEE, 1994.

[KOP06] KOPPEL, M., SCHLER, J., ARGAMON, S. MESSERI, E. **Authorship attribution with thousands of candidate authors**. In Proceedings of the 29th ACM SIGIR (pp. 659–660). NewYork: ACM Press, 2006.

[KOP09] KOPPEL, M., SCHLER, J., ARGAMON, S. **Computer Methods in Authorship Attribution**. Journal of the American Society for Information Science and Technology. 60(1), pp. 9-26, 2009.

[KUK01] KUKUSHKINA, O.V., POLIKARPOV, A.A., KHMELEV, D.V. **Using literal and grammatical statistics for authorship attribution**. Problems of Information Transmission, 37(2), 172–184, 2001.

[LEM94] LEDGER, G. MERRIAM, T. **Shakespeare, Flecther and the two noble kinsmen**. Literary and Linguistic Computing, 9, 235-248, 1994.

[LIT16] BIBLIOTECA DE LITERATURAS DE LÍNGUA PORTUGUESA – LITERATURA DIGITAL. http://www.literaturabrasileira.ufsc.br/

[LOW95] LOWE, D. MATTHEWS, R. **Shakespeare vs. Flecther: A stylometric analysis by Radial Basis Function**. Computer and the Humanities, 29, 449-461, 1995.

[LYO68] LYONS, J. **Introduction to Theorical Linguistics**. Cambridge University, Cambridge, 1968.

[MAL06] MALYUTOV, M.B. **Authorship attribution of texts: a review**. Information Transfer and Combinatorics, LNCS 4123, pp. 362–380, 2006.

[MAN95] MARTINDALE, C. MCKENZIE, D. **On the utility of content analysis in author attribution: The "Federalist"**. Computer and Humanities, 29, 259-270, 1995.

[MCM93] MCMENAMIM, G. R. **Forensic Stylistics**. Elsevier, Amsterdam, 1993.

[MCM02] MCMENAMIM, G. R. **Forensic Linguistics – Advances in Forensic Stylistics**. CRC Press, New York, 2002.

[MEA95] MEALAND, D. L. **Correspondence analysis of Luke.** Literary and Linguistic Computing, 10, 171-182, 1995.

[MEM93] MATTHEWS, R. MERRIAM, T. **Neural computation in stylometry: An application to the works of Shakespeare and Fletcher**. Literary and Linguistics Computing, 8, 203-209, 1993.

[MEM94] MERRIAM, T. MATTHEWS, R. **Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe**. Literary and Linguistics Computing, 9, 1-6, 1994.

[MEW64] MOSTELLER, F. WALLACE, D. L. **Inference and disputed authorship: The Federalist**. Reading, Addison-Wesley, 1964.

[MIC94] MICHIE, D. SPIEGELHALTER, D. J. TAYLOR, C.C. **Machine Learning, Neural and Statistical Classification**. 1994

[MOR65] MORTON, A. Q. **The authorship of Greek prose**. Journal of the Royal Statistical Society, 128, 169-233, 1965.

[NEM15] NEME, A. PULIDO, J. R. G. MUNOZ, A. HERNADES, S. DEY, T. **Stylistics analysis and authorship attribution algorithms based on self-organizing maps**. Neurocomputing, vol. 147, pp. 147–159, 2015.

[PAV08] PAVELEC, Daniel F; JUSTINO, E. J. R. ; BATISTA, Leonardo V.; OLIVEIRA, Luiz E. S. de . **Author Identification using Writer-Dependent and Writer-Independent Strategies.** In: 23th Annual ACM Symposium in Applied Computing (SAC2008), 2008, Fortaleza. Proceedings of the 23th Annual ACM Symposium in Applied Computing, 2008. v. 1. p. 414-418.

[PEK02] PEKALSKA, E. DUIN, R. P. W. **Dissimilarity representations allow for building good classifiers**, Pattern Recognition Letters, vol. 23(8), pp. 943–956, 2002.

[PEN03] PENG, F., SHUURMANS, D., KESELJ, V. WANG, S. **Language independent authorship attribution using character level language models**. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (pp. 267–274). Morristown, NJ: Association for Computational Linguistics, 2003.

[PET04] PETERS, C. BRASCHLER, M. GONZALO, J. KLUNCK, M. **Comparative evaluation of multilingual information access systems**. Berlin: Springer (LNCS, 3237), 2004.

[PLP13] PORTAL DA LÍNGUA PORTUGUESA. **Vocabulário Ortográfico da Língua Portuguesa**. Disponível em http://www.portaldalinguaportuguesa.org

[POT14] POTHA, N. STAMATATOS, E. **A Profile-based Method for Authorship Verification.** In *Proc. of the 8th Hellenic Conference on Artificial Intelligence* (SETN), LNCS, 8445, pp. 313-326, 2014.

[SAV11] SAVOY, J. **Who wrote this novel? Authorship attribution across three languages**. Travaux neuchâtelois de linguistique, vol. 55, pp. 59–75, 2011.

[SAV13] SAVOY, J. **Authorship attribution based on a probabilistic topic model**. Information Processing e Management, vol. 49(1), pp. 341–354, 2013.

[SCH01] SCHJELDAHL, P. **Ghosts: the dazzling mystery of the Kooning´s last paitings**. The New Yorker, 98-98, 2001.

[STA99] STAMATOS, E. KOKKINAKIS, G. FAKOTAKIS, N. **Automatic extraction of rules for sentence boundary disambiguation**. In Proceedings of the Workshop in Machine Learning in Human Language Technology, Advance Course on Artificial Intelligence, 88-92, 1999.

[STA01] STAMATOS, E. KOKKINAKIS, G. FAKOTAKIS, N. **Automatic Text Categorization in Terms of Genre and Author**. Computational Linguistics. Vol. 26 (4), 471-495, 2001.

[STA09] STAMATOS, E**. A survey of modern authorship attribution methods**. Journal of the American Society for Information Science and Technology. 60 (3), 538-556, 2009.

[TWE98] TWEEDIE, F. BAAYEN, R. **How variable may a constant be? Measures of lexical richness in perspective**. Computers and the Humanities, 32(5), 323–352, 1998.

[UZU05] UZUNER, O. KATZ, B. **A comparative study of language models for book and author recognition**. Springer Lecture Notes in Computer Science, 3651, 969–980, 2005.

[VAR10] VARELA, P. J. **O uso de atributos estilométricos na identificação da autoria de textos**. Dissertação de Mestrado. 89 p. Pontifícia Universidade Católica do Paraná, Curitiba, 2010.

[VAR11] VARELA, P. J. JUSTINO, E. J. R. OLIVEIRA, L. E. S. **Selecting syntactic attributes for authorship attribution**. Proceedings of the International Joint Conference on Neural Networks, pp. 167–172, 2011.

[VAR14] VARELA, P. J. JUSTINO, E. J. R. OLIVEIRA, L. E. S. **Uso de níveis estruturais e algoritmos genéticos na atribuição de autoria em língua Portuguesa**. In: 9ª Conferencía Ibérica de Sistemas y Tecnologías da Informácion - CISTI 2014, 2014, Barcelona. Sistemas y Tecnologías da Informácion - CISTI 2014 - Actas de la 9ª Conferencia Ibérica de Sistemas y Tecnologías da Informácion. Braga, Portugal: APPACDM, 2014. v. 2. p. 296-298.

[VAR16a] VARELA, P. J.; JUSTINO, E. J. R. ; BORTOLOZZI, F. ; BRITTO JUNIOR, A. **A Computational Approach for Authorship Attribution of Literary Texts using Syntatic Features.** In: International Joint Conference on Neural Networks - IJCNN, 2016, Vancouver, CA. The 2016 International Joint Conference on Neural Networks. Bandera, TX, USA: IJCNN, 2016. v. 1. p. 1-8.

[VAR16b] VARELA, P. J.; JUSTINO, E. J. R. ; BORTOLOZZI, F. ; OLIVEIRA, L. E. S. **A Computational Approach Based on Syntactic Levels of Language in Authorship Attribution**. Revista IEEE América Latina, v. 14, p. 259-266, 2016.

[WAL94] WALTER, H. **A Aventura das Línguas do Ocidente - A sua Origem, a sua História, a sua Geografia**. Terramar, Lisboa, Portugal, 1994.

[YUL38] YULE, G. U. **On sentence-length as a Statistical Characteristic of Style in Prose, with Application to Two Cases of Disputed Authorship**. Biometrika, 30: 363-90. (1944) The Statistical Study of Literary Vocabulary Cambridge University Press, Cambridge, 1938.

[YUL44] YULE, G.U. **The statistical study of literary vocabulary**. Cambridge University Press, 1944.

[ZIP32] ZIPF, G. K. **Selected Studies of the Principle of Relative Frequency in Language**. Harvard University Press, Cambridge, MA, 1932.

[ZHA05] ZHAO, Y. ZOBEL, J. **Effective and scalable authorship attribution using function words**. In: Proc. Second AIRS – Asian Information Retrieval Symposium. Springer, 174-189. 2005.

[ZHA07] ZHAO, Y. ZOBEL, J. **Searching with style: Authorship attribution in classic literature**. In Proceedings of the 30th Australasian Conference on Computer Science (Vol. 62, 59–68), Ballarat, Australia, 2007.

[ZHE06] ZHENG, R., LI, J., HUANG, Z. CHEN, H. **A framework for authorship analysis of online messages: Writing-style features and techniques**. J. Amer. Soc. Inf. Sci. Technol. 57, 3, 378–393, 2006.