UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Development of new models for authorship recognition using complex networks

Vanessa Queiroz Marinho

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Vanessa Queiroz Marinho

Development of new models for authorship recognition using complex networks

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Diego Raphael Amancio

USP – São Carlos
September 2017

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

M337d Marinho, Vanessa Queiroz
Development of new models for authorship
recognition using complex networks / Vanessa
Queiroz Marinho; orientador Diego Raphael Amancio. -
- São Carlos, 2017.
103 p.

Dissertação (Mestrado - Programa de Pós-Graduação
em Ciências de Computação e Matemática
Computacional) -- Instituto de Ciências Matemáticas
e de Computação, Universidade de São Paulo, 2017.

1. Authorship attribution. 2. natural language
processing. 3. complex networks. I. Amancio, Diego
Raphael, orient. II. Título.

Vanessa Queiroz Marinho

**Desenvolvimento de novos modelos para reconhecimento
de autoria com a utilização de redes complexas**

Dissertação apresentada ao Instituto de Ciências
Matemáticas e de Computação – ICMC-USP,
como parte dos requisitos para obtenção do título
de Mestra em Ciências – Ciências de Computação e
Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e
Matemática Computacional

Orientador: Prof. Dr. Diego Raphael Amancio

USP – São Carlos
Setembro de 2017

ACKNOWLEDGEMENTS

Agradeço primeiramente a Deus pela vida cheia de alegrias, saúde e muitas oportunidades.

Agradeço também à toda minha família. Principalmente aos meus pais, Sirlei e Dener, e minha irmã Isabela pelo amor, por serem meu porto seguro, e por sempre apoiarem os meus sonhos, principalmente o de estudar na USP.

Ao meu orientador Dr. Diego Raphael Amancio por todos os ensinamentos, pela amizade, pela paciência e a disponibilidade em ajudar. Agradeço também pela liberdade que sempre me deu para realizar esse trabalho.

I would like to express my sincere gratitude to Dr. Graeme Hirst, who accepted me in his research group and treated me as one of his students. Thank you so much!

Ao meu namorado, amigo e confidente Fábio por todo o amor, carinho, apoio nas minhas decisões e pela ajuda em incontáveis momentos.

Aos amigos e professores do NILC, em especial ao Edilson, Fernando, Leandro e Nathan. Aos colegas do IFSC, Henrique, Filipi e Prof. Luciano da F. Costa. Gostaria de agradecer também ao Prof. Francisco A. Rodrigues. Aos amigos que a USP me deu em 2009 e que a vida manteve, Débora e Leandro. Agradeço também às minhas amigas Mayara, Lívia, Vanessa L., Janine e Sheena por entenderem a minha ausência (e atrasos) em vários momentos.

I'm also grateful to all the amazing friends I've made at the University of Toronto. Thank you so much Katie, Nona, and Patricia for making me feel welcome and for all the nice moments we've shared. Hope to see you all again someday.

À FAPESP e a CAPES pelo apoio financeiro e pelo interesse nesse estudo. Ao ICMC e à USP por serem uma segunda casa nesses quase 9 anos.

This research was supported by FAPESP (grant numbers 2015/05676-8 and 2015/23803-7) and CAPES. The opinions, assumptions, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of FAPESP and CAPES.

ABSTRACT

MARINHO, V. Q. **Development of new models for authorship recognition using complex networks**. 2017. 103 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Complex networks have been successfully applied to different fields, being the subject of study in different areas that include, for example, physics and computer science. The finding that methods of complex networks can be used to analyze texts in their different complexity levels has implied in advances in natural language processing (NLP) tasks. Examples of applications analyzed with the methods of complex networks are keyword identification, development of automatic summarizers, and authorship attribution systems. The latter task has been studied with some success through the representation of co-occurrence (or adjacency) networks that connect only the closest words in the text. Despite this success, only a few works have attempted to extend this representation or employ different ones. Moreover, many approaches use a similar set of measurements to characterize the networks and do not combine their techniques with the ones traditionally used for the authorship attribution task. This Master's research proposes some extensions to the traditional co-occurrence model and investigates new attributes and other representations (such as mesoscopic and named entity networks) for the task. The connectivity information of function words is used to complement the characterization of authors' writing styles, as these words are relevant for the task. Finally, the main contribution of this research is the development of hybrid classifiers, called *labelled motifs*, that combine traditional factors with properties obtained with the topological analysis of complex networks. The relevance of these classifiers is verified in the context of authorship attribution and translationese identification. With this hybrid approach, we show that it is possible to improve the performance of network-based techniques when they are combined with traditional ones usually employed in NLP. By adapting, combining and improving the model, not only the performance of authorship attribution systems was improved, but also it was possible to better understand what are the textual quantitative factors (measured through networks) that can be used in stylometry studies. The advances obtained during this project may be useful to study related applications, such as the analysis of stylistic inconsistencies and plagiarism, and the analysis of text complexity. Furthermore, most of the methods proposed in this work can be easily applied to many natural languages.

Keywords: Authorship attribution, natural language processing, complex networks.

RESUMO

MARINHO, V. Q. **Desenvolvimento de novos modelos para reconhecimento de autoria com a utilização de redes complexas**. 2017. 103 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Redes complexas vem sendo aplicadas com sucesso em diferentes domínios, sendo o tema de estudo de distintas áreas que incluem, por exemplo, a física e a computação. A descoberta de que métodos de redes complexas podem ser utilizados para analisar textos em seus distintos níveis de complexidade proporcionou avanços em tarefas de processamento de línguas naturais (PLN). Exemplos de aplicações analisadas com os métodos de redes complexas são a detecção de palavras-chave, a criação de sumarizadores automáticos e o reconhecimento de autoria. Esta última tarefa tem sido estudada com certo sucesso através da representação de redes de co-ocorrência (ou adjacência) de palavras que conectam apenas as palavras mais próximas no texto. Apesar deste sucesso, poucos trabalhos tentaram estender essas redes ou utilizar diferentes representações. Além disso, muitas das abordagens utilizam um conjunto semelhante de medidas de redes complexas e não combinam suas técnicas com as utilizadas tradicionalmente na tarefa de reconhecimento de autoria. Esta pesquisa de mestrado propõe extensões à modelagem tradicional de co-ocorrência e investiga a adequabilidade de novos atributos e de outras modelagens (como as redes mesoscópicas e de entidades nomeadas) para a tarefa. A informação de conectividade de palavras funcionais é utilizada para complementar a caracterização da escrita dos autores, uma vez que essas palavras são relevantes para a tarefa. Finalmente, a maior contribuição deste trabalho consiste no desenvolvimento de classificadores híbridos, denominados *labelled motifs*, que combinam fatores tradicionais com as propriedades fornecidas pela análise topológica de redes complexas. A relevância desses classificadores é verificada no contexto de reconhecimento de autoria e identificação de *translationese*. Com esta abordagem híbrida, mostra-se que é possível melhorar o desempenho de técnicas baseadas em rede ao combiná-las com técnicas tradicionais em PLN. Através da adaptação, combinação e aperfeiçoamento da modelagem, não apenas o desempenho dos sistemas de reconhecimento de autoria foi melhorado, mas também foi possível entender melhor quais são os fatores quantitativos textuais (medidos via redes) que podem ser utilizados na área de estilometria. Os avanços obtidos durante este projeto podem ser utilizados para estudar aplicações relacionadas, como é o caso da análise de inconsistências estilísticas e plágios, e análise da complexidade textual. Além disso, muitos dos métodos propostos neste trabalho podem ser facilmente aplicados em diversas línguas naturais.

Palavras-chave: Reconhecimento de autoria, processamento de línguas naturais, redes complexas.

LIST OF FIGURES

Figure 1 – All possible directed motifs involving three nodes.	35
Figure 2 – All possible undirected motifs with three and four nodes.	35
Figure 3 – Transition probabilities to the nodes of two networks.	36
Figure 4 – Backbone and merged symmetries	37
Figure 5 – Co-occurrence networks	57
Figure 6 – Extended co-occurrence networks	58
Figure 7 – Mesoscopic approach proposed by Arruda <i>et al.</i> (2017)	59
Figure 8 – PCA of the texts in two scenarios: original and without stopwords.	64
Figure 9 – Co-occurrence network and some motifs extracted from it	66
Figure 10 – Partitions from books of the Brontë sisters compared in terms of the frequency of words and labelled motifs	67
Figure 11 – Accuracy rates in assigning the authorship of books from Dataset 1 for several values of $ W $	69
Figure 12 – Accuracy rates in assigning the authorship of books from Dataset 3 for several values of $ W $	69
Figure 13 – Named entity networks	74
Figure 14 – Accuracy rates in the pairwise classification using mesoscopic networks. . .	76
Figure 15 – PCA of the books written by Darwin, Hardy, Poe and Twain	77
Figure 16 – Mesoscopic networks of the 20 books written by four selected authors. . . .	78

LIST OF TABLES

Table 1 – Summary of the related work presented in Section 3.1	53
Table 2 – Pre-processing: removal of stopwords	56
Table 3 – Pre-processing: lemmatization process	56
Table 4 – Accuracy rates when <i>Further Neighborhood</i> networks were extracted from Dataset 1	62
Table 5 – Accuracy rates when <i>Further Neighborhood</i> networks were extracted from Dataset 2	62
Table 6 – Accuracy rates when the frequency of motifs was extracted from Dataset 1	63
Table 7 – Accuracy rates when network features were extracted from Dataset 1	64
Table 8 – Accuracy rates when labelled motifs were extracted from Dataset 1 and 3	68
Table 9 – Accuracy rates when labelled motifs and the frequency of motifs were extracted from Dataset 4	71
Table 10 – Accuracy rates in discriminating the debates from the Canadian Hansard	72
Table 11 – Accuracy rates in discriminating the debates from the European Parliament	72
Table 12 – Accuracy rates when named entity networks were created from Dataset 1 and 4	75
Table 13 – Accuracy rates when mesoscopic networks were created from Dataset 2	76
Table 14 – Accuracy rates when simplified function word networks were created from Dataset 3	79
Table A.1 – Dataset 1	95
Table A.2 – Dataset 2	96
Table A.3 – Dataset 3	97
Table A.4 – Dataset 4	98

LIST OF ABBREVIATIONS AND ACRONYMS

BA	Barabási-Albert
CN	Complex Network
ER	Erdős-Rényi
kNN	<i>k</i> -Nearest Neighbors
NER	Named Entity Recognizer
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
PCA	Principal Component Analysis
POS	part-of-speech
SVM	Support Vector Machines
WS	Watts-Strogatz
XML	Extensible Markup Language

CONTENTS

1	INTRODUCTION	23
1.1	Motivation and Goals	24
1.2	Structure of the document	25
1.3	Final Remarks	26
2	BACKGROUND	27
2.1	Complex Networks	27
2.1.1	Network Models	28
2.1.1.1	<i>Erdős-Rényi Model</i>	28
2.1.1.2	<i>Watts-Strogatz Model</i>	29
2.1.1.3	<i>Barabási-Albert Model</i>	29
2.1.2	Networks applied to Language Studies	30
2.1.2.1	<i>Dorogovtsev-Mendes Model</i>	30
2.1.3	Complex Network Measurements	31
2.1.3.1	<i>Degree</i>	31
2.1.3.2	<i>Assortativity</i>	31
2.1.3.3	<i>Average Degree of Neighbors</i>	32
2.1.3.4	<i>Clustering Coefficient</i>	32
2.1.3.5	<i>Matching Index</i>	33
2.1.3.6	<i>Average Geodesic Distance or Average of Shortest Paths</i>	33
2.1.3.7	<i>Betweenness Centrality</i>	34
2.1.3.8	<i>Motifs</i>	34
2.1.3.9	<i>Accessibility</i>	35
2.1.3.10	<i>Symmetry</i>	36
2.2	Authorship Attribution	37
2.2.1	Stylometric attributes	38
2.2.1.1	<i>Lexical features</i>	38
2.2.1.2	<i>Character features</i>	39
2.2.1.3	<i>Syntactic and Semantic features</i>	39
2.2.2	Attribution Methods	41
2.3	Final Remarks	41
3	RELATED WORK	43

3.1	Related Work	43
3.1.1	<i>Antiqueira et al. (2006)</i>	44
3.1.2	<i>Amancio et al. (2011)</i>	45
3.1.3	<i>Mehri, Darooneh and Shariati (2012)</i>	45
3.1.4	<i>Amancio, Oliveira Jr and Costa (2012b)</i>	46
3.1.5	<i>Lahiri and Mihalcea (2013)</i>	47
3.1.6	<i>Segarra, Eisen and Ribeiro (2013), Segarra, Eisen and Ribeiro (2015)</i>	47
3.1.7	<i>Amancio (2015a)</i>	48
3.1.8	<i>Amancio (2015b)</i>	49
3.1.9	<i>Amancio (2015c)</i>	50
3.1.10	<i>Amancio, Silva and Costa (2015)</i>	50
3.1.11	<i>Akimushkin, Amancio and Oliveira Jr. (2017)</i>	51
3.2	Final Remarks	52
4	RESULTS	55
4.1	Materials and Methods	55
4.1.1	<i>Datasets</i>	55
4.1.2	<i>Pre-processing</i>	56
4.1.3	<i>Network models</i>	57
4.1.3.1	<i>Co-occurrence Networks</i>	57
4.1.3.2	<i>Mesosopic Networks</i>	58
4.1.4	<i>Features</i>	58
4.1.5	<i>Machine Learning Methods</i>	59
4.2	Results	61
4.2.1	<i>Extensions of co-occurrence networks</i>	61
4.2.2	<i>Motifs</i>	62
4.2.3	<i>Labelled Motifs</i>	65
4.2.3.1	<i>Labelled Motifs for Authorship Attribution</i>	66
4.2.3.2	<i>Labelled Motifs for Translationese Identification</i>	70
4.2.4	<i>Other network representations applied to authorship attribution</i>	73
4.2.4.1	<i>Named Entity Networks</i>	73
4.2.4.2	<i>Mesosopic Networks</i>	75
4.2.4.3	<i>Function word networks</i>	77
4.3	Final Remarks	80
5	CONCLUSION	81
5.1	Contributions	82
5.2	Limitations and Future Work	83
5.3	Publications	84

BIBLIOGRAPHY 85

**APPENDIX A DATASETS USED FOR AUTHORSHIP ATTRIBU-
TION 95**

APPENDIX B CANADIAN HANSARD AND EUROPARL 99

B.1 Canadian Hansard 99

B.2 Europarl 100

APPENDIX C LIST OF STOPWORDS 103

INTRODUCTION

Complex networks have been employed to model a great variety of systems found in the real-world ([ALBERT; BARABÁSI, 2002](#)). Some examples include food webs, which can be described as networks of species connected according to their respective predator-prey relationships, and the Web, a network of billions of Web pages connected by the hyperlinks among them ([MIHALCEA; RADEV, 2011](#)). Given their multifaceted nature, most studies in complex networks have benefited from ideas of several areas, such as mathematics, physics, biology, computer and social sciences ([NEWMAN, 2010](#)).

The ever increasing data availability and computational capacity have fostered the development of efficient algorithms in several areas. This has allowed the analysis of huge networks with millions of nodes and even billions of edges. Of particular interest to the goals of this Master's project, textual networks may represent the syntactic ([CANCHO; SOLÉ; KÖHLER, 2004](#); [AMANCIO *et al.*, 2012](#)), semantic ([LIU, 2009](#)) or empirical ([LUDUEÑA; BEHZAD; GROS, 2014](#)) relationships between words. Despite their differences, many nodes of these networks can be reached with a few steps and their degree distributions follow a power law, which are the universal properties known as *small-world* and *scale-free*, respectively.

Word co-occurrence (or adjacency) networks ([CANCHO; SOLÉ, 2001](#)) are a widely used representation. In this model, nodes represent words while edges connect adjacent words. In particular, co-occurrence networks can be understood as a simplification of syntactic networks, as most of the syntactic relationships occur in very short scales ([CANCHO; SOLÉ; KÖHLER, 2004](#)). The representation of texts as co-occurrence networks has proven successful in many tasks, such as to identify literary movements ([AMANCIO; OLIVEIRA JR; COSTA, 2012a](#)), to distinguish prose from poetry ([ROXAS; TAPANG, 2010](#)), automatic summarization ([AMANCIO *et al.*, 2012](#)), and to discriminate informative and imaginative documents ([ARRUDA; COSTA; AMANCIO, 2015](#)). Several studies on the properties of co-occurrence networks have proved that most topological measurements extracted from such networks capture characteristics related to the syntax and style ([ANTIQUERA *et al.*, 2006](#); [AMANCIO *et al.*, 2011](#); [MEHRI;](#)

DAROONEH; SHARIATI, 2012; MIHALCEA; RADEV, 2011). As a result, these networks are more adequate to handle stylistic tasks. In this context, we intend to use word co-occurrence networks to characterize different writing styles in the authorship attribution task.

Authorship attribution methods have attracted widespread interest due to their numerous applications, such as to classify literary works (MATTHEWS; MERRIAM, 1993; BURROWS, 1987), solve copyright disputes (GRANT, 2007), and identify patterns of terrorist communication (ABBASI; CHEN, 2005). One of the first statistic-based authorship attribution approaches was conducted by Mosteller and Wallace (1964). In that work, the authorship of several political essays, known as *The Federalist Papers*, were investigated. Since their seminal work, researchers have proposed several attributes to quantify writing styles. A general assumption in many works is that authors have their own signatures (known as *authorial fingerprints*), which can be used to differentiate their writings (JUOLA, 2006). The attributes traditionally used in the task are lexical, character, syntactic, and semantic features (STAMATATOS, 2009). The first two categories, which are based on statistical properties of words and characters, include some of the most popular traditional features, as reported by Grieve (2007), Koppel, Schler and Argamon (2009), and Stamatatos (2009).

The usefulness of Complex Network (CN)-based techniques to the authorship attribution task has been observed in many works (ANTIQUEIRA *et al.*, 2006; AMANCIO *et al.*, 2011; MEHRI; DAROONEH; SHARIATI, 2012; LAHIRI; MIHALCEA, 2013), in which the topological features extracted from the networks were able to distinguish several authors. This Master's project aimed to extend the methodology employed in those works with the introduction of new representations and features. Traditionally, the topological analysis of texts usually disregards the textual context after the networks are devised. However, this information might be useful to characterize the networks. In this work, this contextual information was included in hybrid classifiers that combine networked with traditional techniques usually employed in the authorship attribution studies.

1.1 Motivation and Goals

Since the advent of the Internet, studies in authorship attribution have experienced considerable changes. Its rapid popularization, with almost half of the world's population with Internet connection¹, has allowed immediate access to a vast amount of electronic texts (such as emails and messages from blogs and forums). This has fostered the development of efficient methods to automatically handle this content. Moreover, there has been a paradigm shift in the attribution methods due to the increase of computational capacity, from computer-assisted to computer-based, which has led to the development of fully-automated systems (STAMATATOS, 2009).

¹ <http://www.internetlivestats.com/internet-users/>

The enormous amount of texts has unveiled the potential of authorship attribution analysis in different contexts. This task is quite relevant inside the Natural Language Processing (NLP) area, and has implied in many new developments in several disciplines, ranging from literature to computer forensics (MATTHEWS; MERRIAM, 1993; TWEEDIE; SINGH; HOLMES, 1996; ABBASI; CHEN, 2005; GRANT, 2007; FRANTZESKOU *et al.*, 2006; STEIN; LIPKA; PRETTENHOFER, 2011). The relevance of such a task is even more evident when we want to estimate the similarity among texts (AMANCIO *et al.*, 2013a).

Even though the task of authorship attribution with complex networks has already been addressed (AMANCIO *et al.*, 2011; MEHRI; DAROONEH; SHARIATI, 2012; LAHIRI; MIHALCEA, 2013), it is still possible to obtain better results. One of the goals of this Master's project was to improve the current network-based representations with the extension of the co-occurrence model, as well as the identification of other representations for the task. Moreover, this project also aimed to develop hybrid classifiers, in which two components were analyzed. The first, the topological component, is obtained with measurements extracted from textual networks, while the second is achieved with the application of traditional methods, such as the frequency of specific words. One of the main hypothesis of this Master's project was an improvement of performance in the authorship attribution task, mostly because the disadvantages of each technique should be overcome with the hybrid classification.

The methods proposed in this Master's research might be useful not only to authorship attribution studies, but also to other NLP tasks in which texts with similar structures have to be distinguished, such as plagiarism detection and the identification of translationese (KOPPEL; ORDAN, 2011). In summary, the research goals of this Master's project are described formally in the following paragraph:

“This Master’s research aimed to develop new models for authorship attribution based on complex networks. In particular, one of the goals was to extend the co-occurrence model in a twofold manner: with the addition of function words and with the connection of words in larger scales (not only immediate neighbors as in the traditional co-occurrence models). Furthermore, we also aimed to improve the current authorship attribution methods combining traditional features with the ones obtained with topological measurements in a hybrid classifier.”

1.2 Structure of the document

The remainder of this document is organized as follows:

- Chapter 2 presents fundamental concepts on the two main areas related to this project, complex networks and authorship attribution.

- Chapter 3 addresses the main related works and the state-of-the-art on authorship attribution with complex networks.
- Chapter 4 reports the materials, the methods and the main results achieved with this Master's project.
- Chapter 5 discusses the contributions and limitations of this research, along with some remarks for future work.

1.3 Final Remarks

Throughout this manuscript, we use the terms *function words* and *stopwords* interchangeably. The same happens with the terms *authorship attribution* and *authorship recognition*.

BACKGROUND

In this Chapter, we present the fundamental concepts from the main areas in which this work is included, complex networks and authorship attribution. Section 2.1 presents some definitions, the main network models and measurements used in this work. In Section 2.2, we present some definitions, the first works in authorship attribution and how they have been enhanced to incorporate other attributes. The main stylometric features and the two attribution methods are also explained in Section 2.2.

2.1 Complex Networks

Complex networks have been used to represent a great variety of complex systems (ALBERT; BARABÁSI, 2002; COSTA *et al.*, 2011). According to Barabási (2014), the properties emerging from a complex system cannot be easily inferred only by the analysis of its components. There are several examples of complex systems, such as social and biological systems, the Internet, and the human society. In this context, several researchers in complex networks have dedicated to study and explain how specific behaviors and patterns emerge from these systems (NEWMAN, 2010; COSTA *et al.*, 2011), such as a giant component comprising most of the nodes and remarkably short distances between nodes. Traditionally, the study of networks was related to the analysis of random graphs. The mathematician Leonard Euler was one of the pioneers in graph theory. In 1736, he solved the well-known problem of the Königsberg Bridges (BARABASI, 2003). Since his seminal work, graph theory has been the focus of several studies (BOCCALETTI *et al.*, 2006).

In order to model interactions among elements, networks are formed by a set $V = \{v_1, v_2, \dots, v_n\}$ of nodes, which represents the elements, and a set $E = \{e_1, e_2, \dots, e_m\}$ of edges representing relationships among elements from V . An adjacency matrix A can be used to represent the network connectivity, where the element A_{ij} is equal to 1 iff node i is connected to node j . There are directed and undirected networks. In undirected networks, the matrix A is

symmetric, i.e. $A_{ij} = A_{ji}$ for all pairs i and j . The connections in a network can be differentiated by assigning a weight to each edge (COSTA *et al.*, 2007). As a consequence, a weighted network is defined by three sets, the set V and E presented above, and a set $W = \{w_1, w_2, \dots, w_m\}$ containing the weights.

In recent years, the field of complex networks has drawn the attention of the scientific community, mainly due to the finding that several real-world systems could be represented by networks whose characteristics cannot be explained by random network models (COSTA *et al.*, 2007). Instead, these networks present non-trivial patterns (NEWMAN, 2010). Other factors have contributed to the growing interest in complex networks (ALBERT; BARABÁSI, 2002). The expanding boundaries of several disciplines, such as biology and sociology, has allowed the access to many specific databases from their domains. Such online availability fostered the development of huge datasets of many real networks, such as power-grid networks (WATTS; STROGATZ, 1998), US airport networks (COLIZZA; PASTOR-SATORRAS; VESPIGNANI, 2007), and metabolic networks (DUCH; ARENAS, 2005). Moreover, the increase of computational capacity has allowed the analysis of huge networks comprising millions of nodes.

According to Newman (2003), the most popular and studied networks can be divided into four categories: social, information, technological, and biological networks. These categories are not mutually exclusive, i.e. a network could be included in one or more categories. Motivated by such diversity of real networks, researchers have compared them in terms of their properties. As a consequence, several common properties of these networks have been unveiled. Real-world networks share several characteristics, such as community structure (GIRVAN; NEWMAN, 2002), motifs (MILO *et al.*, 2002; KASHTAN *et al.*, 2004b), and the degree distribution following a power law (CLAUSET; SHALIZI; NEWMAN, 2009).

2.1.1 Network Models

Mathematical models are one of the most effective ways to understand the impact of several network properties (NEWMAN, 2010). Network models have been increasingly used for the investigation of several phenomena. These models replicate the connection patterns and dynamic behaviors found in real-world networks in an attempt to understand their implications. The three main network models are described below.

2.1.1.1 Erdős-Rényi Model

The model proposed by Erdős and Rényi (1959) is considered as one of the simplest network models. In this model, there are n nodes connected by m edges, which are randomly chosen from $\frac{n(n-1)}{2}$ possible edges. In an equivalent definition given by a binomial model, m is not fixed and there is a probability p of connecting each pair of nodes. The Erdős-Rényi (ER) networks are also called Poisson Random Graphs (NEWMAN, 2003) because their degree distribution follows a Poisson distribution. Due to its simplistic approach, the ER model is not

appropriate to model real-world networks, because it does not reflect some properties of these networks (COSTA *et al.*, 2007), such as the presence of many loops of size three (triangles) and the degree distribution following a power law (CLAUSET; SHALIZI; NEWMAN, 2009).

2.1.1.2 Watts-Strogatz Model

Watts and Strogatz (1998) published a model known as *Small-World Networks*. The *small-world* property is found in many real-world networks, in which many nodes can be reached through a few steps in the network. To create *small-world* networks, construct a ring lattice with N nodes, where each node is connected to its m nearest neighbors on the left side and m nearest neighbors on the right side. Then, for each node i , each edge that connects i to j in a clockwise sense is rewired with probability p . The end of the edge is a node j' chosen randomly, avoiding self-connections. This process is repeated until all nodes are considered. As a consequence of the rewiring procedure, the *small-world* networks are placed between regular and random networks. When $p = 0$, no edge is reconnected and the regular graph presents many loops involving three nodes and long average path lengths. Conversely, for $p = 1$, all edges are rewired and the obtained network is similar to a random network, which presents few cycles and small shortest paths. They probed that there is a region between those two boundaries in which the *small-world* network presents small shortest paths and many cycles. Even though Watts-Strogatz (WS) networks present considerably more cycles than ER networks, the degree distribution still follows a Poisson distribution, not a power law.

2.1.1.3 Barabási-Albert Model

The degree distribution following a power law is a property found in many real-world networks, such as the Web and citation networks (NEWMAN, 2003). This is probably a consequence of the process known as preferential attachment, which is one of the mechanisms of the network model proposed by Barabási and Albert (1999). The Barabási-Albert (BA) networks are called *scale-free* networks. Two mechanisms play an important role in order to obtain *scale-free* networks:

- Growth: At each time step, a new node i with m edges is added to the network.
- Preferential attachment: The probability of the edge connecting the new node i to an existing node j depends on the degree of j .

As a consequence of the preferential attachment, the most connected nodes are more likely to receive new edges. This phenomenon is known as *the rich get richer*. In *scale-free* networks, there are some nodes – called *hubs* – that present a considerable fraction of the total number of connections (COSTA *et al.*, 2007). Finally, this model generates networks whose degree distribution follows a power law with an exponent $\gamma = 3$ (NEWMAN, 2003).

2.1.2 Networks applied to Language Studies

In recent years, a growing body of literature has been employing complex networks to model and analyze the human languages (CONG; LIU, 2014). According to findings from studies considering several languages, human languages are also a complex system (LARSEN-FREEMAN; LYNNE, 2008; BECKNER *et al.*, 2009), with some properties emerging from experience, and social interaction, for example. Therefore, the models and tools of complex networks constitute a relevant methodology to study the language. In this context, a network N is given by $N = (V, E)$, where the set of nodes V represents linguistic units and the set of edges E represents relationships among those units. Some examples of linguistic units are words, phonemes, and morphemes (CONG; LIU, 2014). The relationships among those units could be extracted from different linguistic levels, such as co-occurrence, syntactic, or semantic. The co-occurrence relationships represent the order of the words in a sentence. In this model, two words are connected if they are adjacent in at least one sentence. In the syntax-based models, each edge represents a syntactic dependency, which connects a head to a modifier word. Finally, the extraction of semantic relationships requires a deeper analysis (CONG; LIU, 2014). Despite the differences, complex networks modelling any of those relationships display some properties found in many real networks. For instance, Cancho and Solé (2001) showed that the co-occurrence networks of human language display the *small-world* and *scale-free* properties. Their networks presented more than 450,000 nodes and the relationships were obtained from an extract of the British National Corpus. In addition, Cancho, Solé and Köhler (2004) extracted syntactic dependency networks from three European languages, namely Czech, German, and Romanian. They found the presence of *small-world* structures and discovered that their degree distributions follow a power law. Taken together, these findings suggest that human language presents a structure that could be described by universal patterns. Dorogovtsev and Mendes (2001) proposed a network model to explain language evolution, in which the language is represented as an evolving network of interacting words. This model is briefly described in the following section.

2.1.2.1 Dorogovtsev-Mendes Model

The degree distribution of the words obtained by Cancho and Solé (2001) presents a peculiar characteristic, two regions following different power laws. To account for these properties, Dorogovtsev and Mendes (2001) proposed a theory about the evolution of the human language based on an evolving network of interacting words. At each time step, a new word i is added to the network and t represents the total of nodes. As in the model proposed by Barabási and Albert (1999), the new word i is connected to an existing word j with a probability proportional to its degree k_j . In addition, at each time step, $c * t$ new edges connecting existing words (i.e. all except from the new word i) are added, where c is a constant. The probability of each new edge connecting words j and j' is proportional to the product $k_j k_{j'}$.

With these two rules, [Dorogovtsev and Mendes \(2001\)](#) successfully represented the word network described by [Cancho and Solé \(2001\)](#). The two regions with different power law exponents are obtained by the two growth rates employed in this model: a constant one involving the edges of the new nodes and an increasing rate for the edges among the existing nodes ([BIEMANN, 2012](#)).

2.1.3 Complex Network Measurements

The most important network measurements for the development of this Master's project are described below.

2.1.3.1 Degree

The degree of a node is one of the simplest measurements that can be extracted. In undirected networks, it indicates the amount of edges connected to vertex i and can be obtained by

$$k_i = \sum_{j=1}^N A_{ij}, \quad (2.1)$$

where A_{ij} is an element from the adjacency matrix A and N is the number of nodes. In directed networks, the degree of a node is defined in a two-fold manner:

- k_i^{in} : It indicates the number of incoming edges at node i , also known as *in-degree*.
- k_i^{out} : It indicates the number of outgoing edges from node i , known as *out-degree*.

2.1.3.2 Assortativity

In some real-world networks, similar nodes tend to be connected ([NEWMAN, 2002](#)). For the case of the degree of two connected nodes i and j , three configurations for k_i and k_j are possible:

- $k_i \sim k_j$ - Hubs are connected to hubs.
- $k_i \neq k_j$ - Hubs are connected to nodes with low degree.
- There is no relation between the values of k_i e k_j .

The assortativity measurement determines the degree correlation, which can be obtained with the Pearson correlation coefficient. It can be calculated as ([COSTA et al., 2007](#))

$$r = \frac{\frac{1}{M} \sum_{j>i} k_i k_j A_{ij} - \left[\frac{1}{M} \sum_{j>i} \frac{1}{2} (k_i + k_j) A_{ij} \right]^2}{\frac{1}{M} \sum_{j>i} \frac{1}{2} (k_i^2 + k_j^2) A_{ij} - \left[\frac{1}{M} \sum_{j>i} \frac{1}{2} (k_i + k_j) A_{ij} \right]^2}, \quad (2.2)$$

where M represents the total number of edges, k_i and k_j are the degrees of nodes i and j , respectively, and A_{ij} is an element from the adjacency matrix A .

If $r > 0$, the most connected nodes tend to connect to others with similar degree and the network is classified as assortative (NEWMAN, 2002). If $r < 0$, the network is classified as disassortative. In disassortative networks, the most connected nodes tend to connect to others with fewer connections. Finally, if $r = 0$, there is no relation among the degree of the nodes, and the network is classified as non-assortative. In most of the cases, word adjacency networks are disassortative.

2.1.3.3 Average Degree of Neighbors

The measurement $k_{nn}(i)$ (PASTOR-SATORRAS; VÁZQUEZ; VESPIGNANI, 2001) indicates the average degree of the neighbors of a node. This can be obtained by

$$k_{nn}(i) = \frac{1}{k_i} \sum_{j=1}^N k_j A_{ij}, \quad (2.3)$$

where k_i and k_j are the degrees of node i and j , respectively, and A_{ij} is an element from the adjacency matrix A , and N is the number of nodes.

2.1.3.4 Clustering Coefficient

The clustering or transitivity is commonly found in friendship networks. In such networks, if A and B have a friend in common, it is likely that A and B are also friends. The transitivity is related to the presence of triangles in networks. Different from the random networks of Erdős-Rényi, real-world networks present high frequency of loops involving a few nodes (COSTA *et al.*, 2007).

The clustering coefficient of a given node i , $cc(i)$, indicates the probability of two neighbors of i being connected with each other. This measurement is obtained by

$$cc(i) = \frac{2e_i}{k_i(k_i - 1)}, \quad (2.4)$$

where e_i represents the number of edges among the neighbors of node i and k_i is the degree of node i . If $k_i = 1 \mapsto cc(i) = 0$.

To characterize the whole network, the $\langle CC \rangle$ measurement is given by the average of the clustering coefficient of all nodes:

$$\langle CC \rangle = \frac{1}{N} \sum_{i=1}^N cc(i), \quad (2.5)$$

where N is the number of nodes in the network. In this measurement, each node i has the same weight, regardless of its degree k_i .

An alternative way to characterize the whole network is the clustering coefficient given by the transitivity equation. This measurement is obtained as

$$C = \frac{3 \times N_{\Delta}}{N_3}, \quad (2.6)$$

where N_{Δ} is the number of triangles in the network and N_3 is the number of connected triples. The possible values for this measurement are $0 \leq C \leq 1$, in which $C = 0$ corresponds to a network without triangles. On the other hand, $C = 1$ presents a network with complete transitivity. Different from the previous measurement, C gives the same weight to each triangle in the network. As a consequence, $\langle CC \rangle$ and C result in different values for the same network, because nodes with high degree probably have more triangles when compared to low degree nodes (COSTA *et al.*, 2007). Regarding this measurement, Cancho and Solé (2001) observed that the clustering coefficient in textual networks is higher than the expected value for the corresponding random network.

2.1.3.5 Matching Index

The matching index (KAISER; HILGETAG, 2004) is a measurement that assigns a value to each edge of the network. Given an edge e connecting nodes i and j , the matching index quantifies the similarity between the two network regions connected by e , in terms of the number of neighbors shared by i and j . It can be computed as (COSTA *et al.*, 2007)

$$\mu_{i,j} = \frac{\sum_{k \neq i,j} A_{ik} A_{jk}}{\sum_{k \neq j} A_{ik} + \sum_{k \neq i} A_{jk}}, \quad (2.7)$$

where A_{ij} is an element of the adjacency matrix A . When the value of $\mu_{i,j}$ is low, the edge is connecting two dissimilar network regions. Conversely, high values of this measurement indicate that two similar regions are connected by an edge.

2.1.3.6 Average Geodesic Distance or Average of Shortest Paths

A path in a network is defined by a sequence of distinct edges connecting a sequence of distinct nodes. The length of a path between nodes i and j is equal to the number of edges between these two nodes. The geodesic path between nodes i and j is one of the paths connecting them with the shortest length (COSTA *et al.*, 2007). The geodesic distance between nodes i and j is represented by l_{ij} .

The average geodesic distance, L , is a measurement that characterizes the internal structure of a network. This measurement quantifies the average distance between two nodes and it can be computed as

$$L = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j=1}^N l_{ij}, \quad (2.8)$$

where N is the number of nodes and l_{ij} is the geodesic distance between nodes i and j . If the network has more than one component, some pairs i and j will not be connected by a path. In

this case, $l_{ij} = \infty$ and, $L = \infty$. To avoid this, one possible solution would consider only the giant component in the calculation of L . In textual networks, L quantifies the relevance of each word according to its distance to the most frequent words (AMANCIO *et al.*, 2011).

2.1.3.7 Betweenness Centrality

The betweenness centrality is used to measure the traffic that goes through a node or an edge. Considering that all messages exchanged among pairs of nodes travel by shortest paths, one can measure the node (or edge) relevance by the amount of shortest paths that run through that node (or edge) (BOCCALETTI *et al.*, 2006). The betweenness centrality can be computed as

$$B_i = \sum_{st} \frac{n_{st}^i}{g_{st}}, \quad (2.9)$$

where n_{st}^i is the number of shortest paths between nodes s and t that run through node i and g_{st} is the total number of shortest paths between nodes s and t , performed for all pairs of nodes s and t . Nodes with high values of this measurement may influence the network because they control the information that goes to the other nodes (NEWMAN, 2010).

In a textual network, frequent words usually present higher values of this measurement. However, some words may act as *articulation points* connecting concepts from different communities and, therefore, also present a high values of betweenness centrality. For this reason, this measurement quantifies the diversity of contexts in which a word can be employed (AMANCIO *et al.*, 2011).

2.1.3.8 Motifs

Complex networks can also be characterized by the extraction of motifs. Motifs are recurrent interconnection patterns found more often in real-world networks than in randomized ones (MILO *et al.*, 2002). These patterns are expressed as small subgraphs, usually with three or four nodes. The set of directed motifs with three nodes is presented in Figure 1. Figure 2 presents the undirected motifs involving three and four nodes. Milo *et al.* (2002) discovered the presence of motifs in several real-world networks. They found that transcription and neuronal connectivity networks presented the same set of motifs, which suggested a structural similarity in those networks. Moreover, the motifs extracted from a set of networks representing electric circuits were able to split these networks into two classes, without any external knowledge (MILO *et al.*, 2002; KASHTAN *et al.*, 2004b). Remarkably, each class represented a different functionality of the electric circuits.

Motifs have already been employed to characterize texts (MILO *et al.*, 2004; KRUMOV *et al.*, 2011; EL-FIQI; PETRAKI; ABBASS, 2011; BIEMANN; ROOS; WEIHE, 2012; AMANCIO *et al.*, 2013a; CABATBAT; MONSANTO; TAPANG, 2014; MESGAR; STRUBE,

2015; MARINHO; HIRST; AMANCIO, 2016). Because the motifs extraction is computationally expensive in very large networks, we mainly used the motifs presented in Figure 1 and 2.

Figure 1 – All possible directed motifs involving three nodes.

Source: [Marinho, Hirst and Amancio \(2017\)](#).

Figure 2 – All possible undirected motifs with three (a) and four nodes (b).

Source: Adapted from [Silva and Stumpf \(2005\)](#).

2.1.3.9 Accessibility

The accessibility quantifies the number of nodes effectively accessible from an initial node ([VIANA; BATISTA; COSTA, 2012](#)). This measurement is defined as

$$a^{(h)}(i) = \exp\left(-\sum P_h(i, j) \ln P_h(i, j)\right), \quad (2.10)$$

where $P_h(i, j)$ is the probability of reaching a node j at the h^{th} concentric level, centered at node i ([TRAVENTÇOLO; VIANA; COSTA, 2009](#)). Using that definition, $0 < a^{(h)}(i) \leq N_i(h)$, where $N_i(h)$ represents the number of nodes at the h -th concentric level. The maximum value is obtained when the transition probabilities to a given concentric level are the same. In that case, all nodes can be equally accessed. Figure 3 illustrates some transition probabilities to the nodes of two toy networks. Different from other measurements, the accessibility does not correlate with the word frequency. Interestingly, this measurement can be used to detect keywords ([AMANCIO;](#)

OLIVEIRA JR; COSTA, 2012a), which have been proven relevant for the authorship attribution task (AMANCIO *et al.*, 2011).

Figure 3 – In the graph on the left, the transition probabilities to the nodes in the second concentric level are the same, therefore $a^{(2)}(1) = 4$. On the right, the transition probabilities are different due to the addition of two edges. Visually, there is a tendency to access node 8, therefore $a^{(2)}(1) < 4$.

Source: Elaborated by the author.

2.1.3.10 Symmetry

Silva *et al.* (2016) proposed two measurements that quantify the local symmetry around a node, i.e. the heterogeneity of accessing neighbors of a given node. These measurements, called *backbone* and *merged* symmetry, can be understood as a normalization of the accessibility (VIANA; BATISTA; COSTA, 2012), where the number of reachable nodes is the normalization factor.

To obtain these measurements, concentric random walks are adopted to avoid transitions to previously visited concentric levels. To disregard edges between nodes from the same concentric level, some transformations need to be performed. In the backbone symmetry, all edges between nodes from the same concentric level h are removed. On the other hand, nodes that share edges in the same level h are combined in a super node to obtain the merged symmetry. After these topology transformations, the transition probabilities $P_h(i, j)$ are calculated for each pair of nodes i and j . Figure 4 illustrates the topological transformations needed for each symmetry measurement and their respective transition probabilities.

The backbone symmetry is calculated as

$$Sb_i^{(h)} = \frac{\exp(-\sum P_h(i, j) \ln P_h(i, j))}{|\xi_i^{(h)}|}, \quad (2.11)$$

where $P_h(i, j)$ is the probability of reaching a node j at the h^{th} concentric level, centered at node i , and $\xi_i^{(h)}$ represents the set of reachable nodes at a distance h from node i . The merged symmetry

Figure 4 – In the original graph, there are some edges between nodes from the same concentric level, such as the ones connecting nodes A and B, and B and C. These edges are removed in the backbone symmetry. In the merged symmetry, nodes A, B, and C are merged into a super node X. Note that these topological transformations modify the transition probabilities.

Source: Elaborated by the author.

is obtained in a similar fashion, but the edge weights are considered in order to calculate the transition probabilities.

2.2 Authorship Attribution

A typical authorship attribution method assigns a text whose authorship is unknown to the most likely author, from a set of candidate authors (STAMATATOS, 2009). The first activities supported by statistical techniques date back to the XIX century. One of the first approaches, and still widely used, involves probabilistic models. These methods attempt to maximize the probability of identifying the real author of a text. Mosteller and Wallace (1964) showed that the frequencies of common words, such as *and* and *to*, depend on the authorship of the text and are relevant features to distinguish different authors. In that study, they investigated the authorship of hundreds of political essays, *The Federalist Papers*.

The seminal work carried out by Mosteller and Wallace (1964) was one of the first statistic-based methods for authorship attribution. Since then, many works have been dedicated to define new attributes to characterize writing styles (HOLMES, 1994), a research line called *stylometry*. According to Stamatatos (2009), stylometric features can be divided into some categories, such as lexical, character, syntactic, and semantic features.

Since the advent of the Internet, there have been a few changes in the authorship attribu-

tion studies (STAMATATOS, 2009). The availability of an ever growing amount of texts on the Web (such as emails and blog posts) has increased the need for methods that analyze this content. Moreover, this has fostered improvements in several areas, such as information retrieval, machine learning and NLP. Such improvements have had a great impact on the authorship attribution field.

This publicly available content unveiled the applicability of authorship analysis in several areas. In addition to the traditional literary research (MATTHEWS; MERRIAM, 1993; BURROWS, 2002), authorship attribution methods have been used in history (TWEEDIE; SINGH; HOLMES, 1996), intelligence (ABBASI; CHEN, 2005), civil and criminal law (GRANT, 2007; CHASKI, 2005), computer forensics (JUOLA, 2006; FRANTZESKOU *et al.*, 2006), and others. Another relevant application of authorship attribution is in the context of plagiarism detection (WHITE; JOY, 2004; STEIN; LIPKA; PRETTENHOFER, 2011), because it may be possible to identify pieces of text that were copied or present stylistic inconsistencies using the attribution methods.

2.2.1 Stylometric attributes

Some attributes traditionally used in stylistic tasks, such as the authorship attribution, are described below.

2.2.1.1 Lexical features

Lexical features consider the text as a sequence of words. Examples of lexical features commonly used in authorship attribution tasks are word and sentence lengths, frequency of words, vocabulary richness and others (JUOLA, 2006; GRIEVE, 2007; STAMATATOS, 2009). Considered as one of the first authorship attribution approaches using lexical features, Mendenhall (1887) used the length of sentences and words to identify the authorship of documents. Currently, most of the attribution studies use, at least partially, lexical features to characterize writing styles (STAMATATOS, 2009).

An important finding related to lexical features is the fact that common words (such as articles, prepositions, and pronouns), referred to as *function words*, are one of the best characteristics to distinguish a set of authors (BURROWS, 1987; GARCÍA; MARTÍN, 2007; KOPPEL; SCHLER; ARGAMON, 2009). Once function words are topic-independent and are mostly used unconsciously, they may capture writing choices of each author (STAMATATOS, 2009; KOPPEL; SCHLER; ARGAMON, 2009). For example, Uzuner and Katz (2005) used a corpus comprising 50 books written by 8 authors. The highest accuracy, 87%, was achieved with the frequency of function words. Some of the most predictive function words were: *the*, *not*, *she*, and *'s*. In addition, frequent words account for the successful approach proposed by physics in (HAVLIN, 1995). In that work, the rank distances of the word frequencies were used to calculate the similarity between texts.

Another possible lexical-based approach is to extract the most frequent words in a text (BURROWS, 1987; STAMATATOS, 2009; KOPPEL; SCHLER; ARGAMON, 2009). Therefore, each text is represented as a vector of word frequencies. Then, suitable machine learning techniques can be applied to distinguish these vectors. There is no consensus regarding the number of frequent words that should be used as classification features, particularly now that classification algorithms can handle thousands of features. Remarkably, if this number is small, the most frequent words will be dominated by function words, and content-words will be rare in the feature set. Features based on the frequency of words do not consider the word order, which may also reveal stylistic information. To account for this information, the frequency of n subsequent words (word n -grams) can be employed as textual features (PENG; SCHUURMANS; WANG, 2004).

2.2.1.2 Character features

In this scenario, a text is analyzed as a simple sequence of characters. Some examples of character features are alphabetic and digit characters count, frequency of letters and punctuation marks, and others (STAMATATOS, 2009). Another approach is to extract frequencies of character n -grams (STAMATATOS, 2012), i.e. subsets of n adjacent characters. An advantage of this type of information is that character features can be easily extracted from any natural language with low computational cost (GRIEVE, 2007). Therefore, these are language-independent features.

In the context of authorship verification, it has been observed that the frequency of the most common n -grams of characters is a relevant feature for the problem (JANKOWSKA; MILIOS; KESELJ, 2014). Grieve (2007) conducted an in-depth study in which several features were extracted to characterize the authorship of texts. An accuracy of 72% was reached when lexical attributes and punctuation marks were combined using texts from 20 authors. Sapkota *et al.* (2015) reported that not all character 3-grams are relevant for the authorship attribution task. In particular, the most important ones are those that capture information about affixes and punctuation marks.

2.2.1.3 Syntactic and Semantic features

The extraction of syntactic and semantic characteristics requires a deeper textual analysis. According to Stamatatos (2009), syntactic information is more reliable than lexical information to characterize the authorship of texts, because syntactic patterns tend to be employed unconsciously by authors. In addition, word statistics patterns are prone to manipulation. For instance, an author trying to hide his/her identity can easily mimic such patterns from another (BRENNAN; GREENSTADT, 2009). The usefulness of syntactic features is related to the relevance of function words in characterizing writing styles, once these words are usually employed in some syntactic structures.

Some syntactic approaches include the extraction of rewrite rule frequencies. Rewrite

or production rules indicate how a sentence may be decomposed into its constituent parts. A constituent part is formed by one or more words behaving as a single unit (JURAFSKY; MARTIN, 2000). Each constituent part has a main word (called *head*) and the others, if any, are *dependent* words. A sentence is considered syntactically valid if it is possible to generate all its terminals (words) according to some rewrite rules (KUMAR, 2012). An example of a rewrite rule is

$$S \rightarrow NP VP, \quad (2.12)$$

where S means Sentence, NP means Noun Phrase and VP means Verb Phrase. This rewrite rule means that a sentence is formed by a noun phrase followed by a verb phrase. Using that rule, the following sentence "*The student conducted all the experiments*" would be divided into *The student* (noun phrase) and *conducted all the experiments* (verb phrase). Rewrite rules are important not only because they carry syntactic information, but also because they express ways to combine words into phrases.

In addition to the rewrite rules, part-of-speech (POS) tags and the analysis of sentences and chunks are also employed as syntactic features (STAMATATOS, 2009). POS tagging is the process of assigning a marker to a word to indicate its grammatical class (JURAFSKY; MARTIN, 2000), such as noun, adjective, and others. One of the first works that used syntactic features for the authorship attribution task was carried out by Baayen, Halteren and Tweedie (1996). They extracted rewrite rule frequencies and those measurements provided better results than the ones obtained with traditional word-based methods. Hirst and Feiguina (2007) combined the idea of bigram frequencies with syntactic analysis. Their syntactic label bigrams were found useful to distinguish different authors, even when the methods were applied in short texts.

Semantic dependency graphs can be used as a source of semantic features (GAMON, 2004). Another semantic approach may consider information about synonyms of the words (MC-CARTHY *et al.*, 2006). This information could be extracted from databases such as the Word-Net (MILLER, 1995). In addition, Argamon *et al.* (2007) defined a taxonomy to associate words or phrases with semantic functions. According to this taxonomy, the words *and* and *moreover* are classified as *Conjunctions* expanded to *Extension*, because both are used to add more information. The extracted features consisted of simple statistics, such as "the number of *Conjunctions* that were expanded to *Extension*".

For the cases where text analysis requires more detail, such as to extract syntactic and semantic features, the obtained results will be less accurate and more likely to include noisy information (STAMATATOS, 2009). As a result, these features usually play a complementary role in most approaches and are combined with lexical and character features.

2.2.2 Attribution Methods

The main components of an authorship attribution task are: a set of candidate authors, texts whose authorship is known (training set) and one or more texts whose authorship is unknown (test set). Each text from the test set has to be assigned to one of the authors. The authorship attribution approaches can be divided into two categories. In one of them, each text from the training set is used individually (*instance-based* approaches), while the other uses the training texts cumulatively (*profile-based* approaches). The two approaches are briefly described below:

- *Profile-based* approach: In this approach, the texts in the training set from an author a are combined in a single text t_a . Then, stylistic properties are extracted from text t_a . The obtained properties are used to characterize the *profile* of author a . Each test sample is compared with the t_a texts from all authors. This comparison can be done with distance measurements, so that the distances from the test sample to all profiles are calculated.
- *Instance-based* approach: Most of the authorship attribution approaches consider each text from the training set separately (STAMATATOS, 2009). By doing so, each training text is considered as an instance and it is represented by a set of attributes. Classification algorithms are trained on these instances in order to obtain an attribution model. Then, the obtained model will be used to classify texts whose authorship is unknown. In order to obtain a reliable authorship attribution model, the provided instances have to be representative of each class. In addition, the instances must present a similar length; otherwise, texts should be split into partitions with similar sizes (SANDERSON; GUENTER, 2006).

2.3 Final Remarks

In this chapter, we presented the main concepts from the two research areas related to this Master's project, complex networks and authorship attribution. It is estimated that there are more than 100 measurements that characterize the topology of complex networks (COSTA *et al.*, 2007). In Section 2.1.3, we described only the ones used in this project. Most of them are available at the Igraph software package (CSARDI; NEPUSZ, 2006).

In the next Chapter, we present many related works that tackle the authorship attribution problem with complex network methods.

RELATED WORK

In the field of complex networks, there are several approaches that have been dedicated to quantify and characterize authors' writing style. In this chapter, we present the main related work to this Master's research. In particular, we describe some works that used complex networks to tackle the authorship attribution problem.

3.1 Related Work

In recent years, there has been considerable interest in applying complex network methods to investigate linguistic properties. Networks are used to model objects and the connections among them. This representation can be used to model data from different types and it is the subject of study of distinct fields, such as mathematics, physics and computer science. In the latter, graph-based techniques have been applied in the analysis and construction of software architectures (MOURA; LAI; MOTTER, 2003), spam filters (KONG *et al.*, 2006), and NLP systems (MIHALCEA; RADEV, 2011; CONG; LIU, 2014).

In textual analysis, CN-based methods have improved the performance of some NLP tasks (AMANCIO, 2015b). These methods could be applied to study events from several linguistic subdisciplines, such as syntax (CANCHO; SOLÉ; KÖHLER, 2004; AMANCIO *et al.*, 2012) and semantics (SILVA; AMANCIO, 2013; MASUCCI *et al.*, 2011). The word adjacency (or co-occurrence) network is a well-known representation in which nodes represent distinct words while the edges are established between adjacent words. Because this representation captures syntax and stylistic features (AMANCIO *et al.*, 2013a), it has been employed to model texts for syntactic complexity analysis (AMANCIO *et al.*, 2013b), detection of literary movements (AMANCIO; OLIVEIRA JR; COSTA, 2012a), and stylometry studies (GRABSKA-GRADZINSKA A. KULIG; DROZDZ, 2012).

One of the most important findings related to co-occurrence networks was reported by

Cancho and Solé (2001). In that work, they proved that co-occurrence networks present two relevant properties usually found in real-word networks. These properties are *scale-free* and *small world*. The *scale-free* property could be understood as a consequence of an optimization process, so that the efforts to send and receive a message are reduced (CANCHO; SOLÉ, 2003; AMANCIO; OLIVEIRA JR; COSTA, 2012a). Such a process could be related to the emergence of Zipf's Law (ZIPF, 1949), which states that the relationship between the absolute frequency of words and their rank in the frequency table follows a power law. Moreover, these two universal properties were also found in syntactic dependency networks obtained from three European languages (CANCHO; SOLÉ; KÖHLER, 2004).

3.1.1 Antiqueira et al. (2006)

Antiqueira et al. (2006) represented several books as complex networks in order to characterize their authorship. They investigated the correlation between the books written by the same author and the measurements extracted from their respective networks. They selected books written by 8 distinct authors from different literary genres, such as fiction, poetry, and scientific works. Stopwords were removed and the remaining words were lemmatized in the pre-processing phase. The lemmatization is a process that transforms plural words into their singular forms, verbs with inflection into their infinitive forms and names to their masculine version. As a consequence, words that refer to the same concept will be represented by the same node, regardless of their inflections. After the pre-processing, each book was represented by a directed co-occurrence network. The direction of the edge is from the first to the following word. The weight of the edge connecting i and j represents the number of times i is followed by j in the text.

In order to extract features from the co-occurrence networks, they used the average out-degree, the clustering coefficient, and the assortativity, as well as a measurement proposed by them based on the connectivity of the network. The obtained values for some of these measurements varied among the authors. After combining some of these measurements, it was possible to cluster some authors with similar writing styles.

The largest difference in writing styles was between Charles Darwin and William Wordsworth. This is a consequence of their different types of work: Charles Darwin wrote about his *scientific theories*, observations and findings; while the works of William Wordsworth were poetry. Antiqueira et al. (2006) suggested that the networks of each author present specific characteristics and those are useful not only to capture authorial fingerprints but also to be used in the authorship attribution task.

3.1.2 *Amancio et al. (2011)*

The work proposed by *Amancio et al. (2011)* investigated the relationship between the writing style of different authors with word intermittency and complex network measurements. The selected dataset comprised 5 books for each one of the 8 authors. In the pre-processing phase, stopwords were removed and the remaining words were lemmatized. Even though function words have already been used in some authorship attribution approaches, they decided to remove them in order to consider only words with more semantical content. The texts were represented as co-occurrence networks.

Complex network measurements – such as clustering coefficient, the average of the shortest paths, and betweenness – were extracted from the networks. The frequency of words was also used to characterize each book. Another measurement, the word intermittency (burstiness), quantifies the uneven distribution of words along the text. This measurement is a good indicator for topic-related words, such as names of characters and places. Because most of these measurements apply to a single node, some statistics such as the average, standard deviation, and skewness (third moment), were extracted in order to obtain a global characterization.

Three machine learning algorithms were employed to evaluate the ability of those features in distinguishing different authors. The accuracy rate of one algorithm in discriminating the authorship of books reached 50%, when all features were used. When feature selection was attempted, the accuracy rate enhanced to 65%. In their work, the features more related to the authorship were the average of shortest paths and the skewness in the distribution of word intermittency. The average of the shortest paths quantifies the distance of the words to the network *hubs* (frequent words). The other feature can be understood as the fraction of all keywords in the text.

Amancio et al. (2011) conducted other experiments to illustrate how the selected network measurements could be complementary to more traditional methods. For example, they extracted the frequency and the word intermittency of the 5 most frequent words of each book. Interestingly, the accuracy rate increased even more and reached 80%. They concluded that complex network measurements and word intermittency are able to extract several characteristics related to the authorship of texts.

3.1.3 *Mehri, Darooneh and Shariati (2012)*

Mehri, Darooneh and Shariati (2012) used complex networks to assign the authorship of 36 books written by 5 famous Persian authors. The pre-processing consisted in the removal of punctuation marks and numbers. No lemmatization process was performed; and all words were lower-cased. An undirected co-occurrence network was created for each text.

Complex network measurements – such as degree, average degree of the neighbors, and clustering coefficient – were extracted from the co-occurrence networks. In addition, they defined

two measurements, the q -parameter and the α -exponent. The q -parameter is related to the degree distribution and it can be understood as a power law generalization. The α -exponent is given by $N_e \approx N_v^\alpha$, where N_e and N_v represent the number of edges and nodes, respectively.

Different from other complex network approaches devoted to authorship attribution, the methodology proposed by [Mehri, Darooneh and Shariati \(2012\)](#) did not use machine learning algorithms to induce classifiers. Instead, each author a was represented by a profile vector v_a , where each element $v_a[i]$ is the average of one measurement over all books written by author a . In a similar fashion, a text X of unknown authorship is also represented by a vector v . The probability of X being written by author a is calculated, for every author a in the set of candidate authors. In particular, this probability is related to the distance between the elements of v_a and v . The larger distances between v and v_a , the lower will be the probability of X being written by author a .

In order to validate their methodology, a single book b is tested at each step while the remaining 35 are used to create the profile vectors of each author. The authorship of b is assigned to the author with the maximum probability. They reported an accuracy rate of 91%, based on the number of false (and true) positives and negatives. However, the other works described in this subsection use the accuracy rate as the percentage of instances correctly classified. In this case, the obtained accuracy rate would be 77.7% because only 8 books were incorrectly assigned. They also investigated the relevance of the selected features and concluded that the parameter q and the exponent α yielded good results for the authorship attribution task. In addition, these attributes could be combined with others for linguistic-related tasks.

3.1.4 [Amancio, Oliveira Jr and Costa \(2012b\)](#)

In order to measure the similarity among texts, [Amancio, Oliveira Jr and Costa \(2012b\)](#) proposed methods that combine semantic characteristics with the topology of co-occurrence networks. This approach was illustrated in different contexts, such as authorship attribution, and the identification and quality evaluation of machine translated texts.

Several measurements, such as the degree, shortest paths, clustering coefficient, and betweenness, were extracted from the co-occurrence networks. In addition, the frequency of all motifs involving three nodes was extracted. [Amancio, Oliveira Jr and Costa \(2012b\)](#) defined three similarity indices to evaluate and classify the translated texts and the authorship of texts. The first and the second indices were based only on semantic and topological information, respectively. The third one combines both sources of information.

In the authorship attribution context, [Amancio, Oliveira Jr and Costa \(2012b\)](#) investigated the relevance of semantic and topological information. Two sets of texts were analyzed, one comprising several poems and the other comprising non-verse texts (i.e. prose). In their experiments, both features were relevant to characterize the authorship of texts. They observed

that some authors share a similar semantic content but present different writing styles. Similarly, there are some authors, such as Bram Stoker and Charles Darwin, that display similar writing styles (i.e. topology features are similar), even though they write about completely different topics (i.e. very distant semantically).

As suggested in this work, the relevance of each method depends on the purpose of the application under study. For example, authors with similar writing styles could be discriminated by the semantic content of their texts. On the other hand, authors whose texts are semantically similar, could be discriminated by their individual writing styles. Moreover, both sources of information can be used in an authorship attribution scenario with many authors.

3.1.5 *Lahiri and Mihalcea (2013)*

Lahiri and Mihalcea (2013) conducted an in-depth study in which they explored several network features for the task of authorship attribution. They used three different datasets. The first was a collection of more than 3,000 electronic books written by 142 authors extracted from the Project Gutenberg repository. The other two datasets were extracted from authorship attribution competitions. No pre-processing technique was applied to the texts and stopwords were kept. They used directed unweighted co-occurrence networks to model each text.

Their approach was conducted according to a twofold perspective: global and local. The former, called *summary features*, consisted of a vast set with 127 features extracted from each network. Some of these measurements include the number of nodes and edges, the clustering coefficient, and statistical properties of the degree distribution. In the latter, called *local features*, 10 local properties were extracted from selected nodes. The node selection was based on several representative lists of words, such as a list with the 571 most common words.

They tested several machine learning classifiers and their method achieved an accuracy rate of around 35% using summary features in the Project Gutenberg dataset. Considering that the choice baseline for the Project Gutenberg dataset is less than 1%, the obtained result with summary features is substantially better than that. Remarkably, local features performed even better and the classification results were enhanced to almost 80%. Such a good discrimination obtained with local features confirms the usefulness of stopwords (the majority of the words in the representative lists) for the task of authorship attribution.

3.1.6 *Segarra, Eisen and Ribeiro (2013), Segarra, Eisen and Ribeiro (2015)*

Research on authorship attribution via complex networks has tended to focus on the development of new measurements and attributes rather than a new representation. In general, texts are modelled as co-occurrence networks and function words are usually removed. However, a number of studies have found these words relevant for the task (*GARCÍA; MARTÍN, 2007*;

GRIEVE, 2007; STAMATATOS, 2009; KOPPEL; SCHLER; ARGAMON, 2009). So as to account for function words, Segarra, Eisen and Ribeiro (2013) proposed a network formed only by function words.

In the proposed network, nodes represent function words and two words i and j are connected by a directed edge if they appear in the same sentence by less than D positions apart from each other. The weight of that edge is the likelihood of finding j in the neighbourhood of i considering all sentences, i.e. the conditional probability of encountering j given that the word i was observed. These likelihoods can be interpreted as the transition probabilities between function words in a Markov chain. By doing so, these chains can be compared using relative entropies.

The function words adjacency networks were applied to the authorship attribution problem using a dataset with 130 books from 18 authors. Each author was then represented by a profile, which consisted of a single Markov chain extracted from its texts. Next, authors are compared in terms of the entropy between their respective Markov chains. When all 18 authors were considered, their results achieved an accuracy rate of 88%. The obtained results outperformed traditional techniques based on word frequencies. Moreover, the combination of both approaches have led to higher accuracies.

As an extension of the previous work, Segarra, Eisen and Ribeiro (2015) demonstrated that function words adjacency networks can also be used to identify authors in collaborative works, as well as to classify texts according to their time period, genre, and gender of the author.

3.1.7 Amancio (2015a)

Even though many works have been dedicated to the recognition of patterns in text, only a few have analyzed the stylistic fluctuations along the text. Useful information could be hidden in such fluctuations. For example, fluctuations around the average of the distribution of word frequencies could be used to detect the most relevant concepts. In this context, Amancio (2015a) studied the stylistic variability across texts using the burstiness of word occurrences and methods from complex networks.

Initially, Amancio investigated whether the stylistic variation in texts is a useful feature for the authorship attribution task. To analyze the stylistic evolution, each text was split in shorter chunks with L words, where $500 \leq L \leq 1,300$. Then, each chunk (subtext) was represented by a word co-occurrence network. Some measurements, such as the accessibility, the average of the shortest paths, the clustering coefficient, and the betweenness, were extracted from those networks. By doing so, each measurement X could be associated to a temporal series where each element x_i is the value of measurement X in the subtext i . The temporal series of each book were decomposed in terms of the Fourier transform and some of those components were used as features for the classification algorithms.

The obtained results are interesting in several ways. The lowest accuracy rate was obtained with $L = 500$ words while the highest one was obtained with $L = 1,300$ words, 35% and 45%, respectively. The most relevant feature to distinguish different authors was the variability of the average of the shortest paths. The results of this study prove that stylistic variations could be used to distinguish the authors' writing styles. Moreover, the proposed methodology could be used to extract complementary attributes to be combined with traditional features.

In another experiment, Amancio verified if the burstiness of some words was useful to characterize the authorship of books. In this context, texts without pre-processing were considered. To characterize each text, the burstiness of the 100 most frequent words was used as feature for the classifiers and the obtained accuracy rate was 65%. This value was enhanced to 90% when some authors – the ones leading to most of the errors – were removed from the dataset. Although many attributes were used, function words – such as *but*, *and*, *I* and *who*, and *as* – were the most relevant ones for the task. Finally, this study suggested that the uneven distribution of some words could be useful to the authorship attribution task.

3.1.8 Amancio (2015b)

Even though there is a substantial amount of works dedicated to tackle linguistic applications with complex networks, only a few have investigated the complementary role of network methods in NLP tasks to improve their performance and, therefore, the state-of-the-art at those tasks. In some cases, the best results are still obtained with NLP traditional methods. In this context, Amancio (2015b) defined methods in order to improve the performance of some text classification tasks, such as authorship attribution and style identification (informative or imaginative).

After pre-processing the texts, word co-occurrence networks were created. Some network measurements were extracted from those networks, such as the average degree, accessibility, and assortativity. Traditional features usually employed in NLP, such as the frequency of common words and character bigrams, were also obtained from the texts. Amancio proposed two different ways to combine the traditional and the network components. The first one, called *Hybrid*, linearly combines the traditional and networked features. On the other hand, the *Tiebreaker* method uses the network features only when the results obtained through traditional features are not reliable.

In the proposed experiments, several combinations of networked and traditional features were tested. In general, there was an improvement in the accuracy for authorship identification when both features were used. For many cases, the results obtained with the *Hybrid* method outperformed those obtained with traditional or networked features separately.

In another context, Amancio demonstrated that there are also performance gains when both features are combined to classify texts according to their styles, namely imaginative or

informative. Therefore, the measurements obtained from the network topology can complement the characterization of writing styles. The technique proposed in this work is quite relevant because it improved the performance of traditional approaches.

3.1.9 *Amancio (2015c)*

Unlike most of the previously mentioned works, which usually consider large pieces of texts or books, *Amancio (2015c)* studied the stability of topological measurements in short pieces of text. The main goal was to investigate whether the topology of networks modelling subtexts is relevant for text analysis. This methodology was verified in the authorship attribution context.

Each subtext was pre-processed and modelled as a co-occurrence network. Several measurements, such as the clustering coefficient, the intermittency, and the accessibility, were extracted from the networks. In the first experiment, the variability of the measurements was analyzed in a dataset with 50 books. Each book was split into samples with L words, where $300 \leq L \leq 2,100$. The obtained results revealed that most networked measurements are stable and display low variability in the shorter samples. According to *Amancio (2015c)*, authors tend to maintain their writing style across different parts of the same book.

In order to verify the applicability of the proposed method, another experiment was carried out. *Amancio (2015c)* investigated the influence of text length on the performance of the authorship attribution task. The dataset consisted of 5 books for each one of the 4 selected authors. The subtexts length varied from 500 to 21,400 words, where the latter corresponds to the size of entire books. Four supervised machine learning methods were employed to measure the effects of sampling on the authorship identification. The obtained accuracy rates proved that subtexts are able to distinguish the authorship of texts. The lowest accuracy rates were achieved with samples of $L = 500$ words.

In most of the cases, the highest accuracy rates did not occur when entire books (21,400 words) were used, but rather with shorter texts. Remarkably, high accuracies were achieved with subtexts comprising only 8% of the original size of the book. These results suggest that improvements could be achieved with a proper textual sampling. One important contribution of this work was that it proved that shorter texts can be analyzed with methods and concepts of complex networks. On the other hand, the proposed method can only be applied to large pieces of texts, because samples obtained from short documents usually display high variability.

3.1.10 *Amancio, Silva and Costa (2015)*

In this work, *Amancio, Silva and Costa (2015)* used symmetry measurements to investigate the connectivity patterns in word co-occurrence networks. Their method was applied to the authorship attribution task mainly because authors tend to make several stylistic choices

while writing. For example, one has to decide about using the same word or using a synonym. Therefore, it is assumed that these characteristics would reflect in the network topology.

One of the experiments in this paper consisted in investigating whether the symmetry measurements are able to characterize different writing styles. To identify the authorship of books, 40 books written by 8 different authors were analyzed. Instead of extracting the symmetry measurements for all nodes, these values were calculated for only the shared vocabulary among books, which consisted of 229 words. The relevance of the symmetry features was evaluated with four machine learning algorithms. The obtained accuracy rates varied from 20% to 82.5% when several combinations of algorithms, symmetry measurements and concentric levels h were tested.

[Amancio, Silva and Costa \(2015\)](#) concluded that symmetry measurements are able to characterize stylistic marks left by each author. This is because each author presents specific bias towards the usage of different patterns and this is quantified through the homogeneity of accessibility of nodes. They also proved that these measurements are not correlated with others, such as the clustering coefficient and the betweenness. Therefore, these features could play a complementary role in the characterization of networks for several tasks.

3.1.11 [Akimushkin, Amancio and Oliveira Jr. \(2017\)](#)

There is a considerable amount of work that uses complex networks to tackle the authorship attribution problem. However, most of them consider the texts as static structures. [Akimushkin, Amancio and Oliveira Jr. \(2017\)](#) proposed a new methodology for the identification of authorship based on the topology evolution of co-occurrence networks modelling short texts.

A collection of novels and tales comprising 80 texts (10 texts per author) with varied lengths was considered. Some pre-processing steps, such as the removal of stopwords, were applied. In this work, each text is not represented by a single network, but rather by a series of independent networks obtained from non-overlapping partitions with 200 tokens each. Twelve measurements, such as the network transitivity, betweenness centrality, and degree, are used to characterize each partition. Therefore, a sequence of networks leads to a sequence of extracted measurements.

Each book is thus represented by twelve time series, one for each measurement. The classification features correspond to the first four moments of each series. Supervised learning algorithms covering distinct paradigms were used, such as k -Nearest Neighbors (kNN), Naive Bayes, and J48. The obtained accuracy rates were relatively high and the highest one, 88.75%, was reached with kNN.

The methodology proposed by [Akimushkin, Amancio and Oliveira Jr. \(2017\)](#) clearly has an advantage over the previous works. In many approaches, texts are usually truncated to the length of the shortest one. By doing so, the obtained co-occurrence networks can be compared.

However, the discarded pieces of text might contain useful information about its authorship. Remarkably, the representation of textual content as time series allows texts of different sizes to be compared, regardless of the size of their series.

3.2 Final Remarks

In this Chapter, we presented some related work in authorship attribution using complex network methods. Most of these approaches extract topological measurements from complex networks and only a few have combined networked methods with traditional techniques usually employed in NLP. As we already mentioned in Chapter 1, one of the goals of this Master's research was to combine traditional techniques in authorship attribution with topological measurements extracted from the networks. A comparison among all works presented in this Chapter is given in Table 1. It is important to note that some factors – such as pre-processing steps, the number of candidate authors, and the number and length of texts – have influence over the accuracy rates obtained by each approach.

The methodology and the results achieved with this Master's research are presented in the next Chapter.

Table 1 – Summary of the related work presented in Section 3.1

Related Work	Number of Texts	Language	Number of Authors	Number of Features	Highest Accuracy Rate
Antiqueira et al. (2006)	44	English	8	4	Not applicable ^a
Amancio et al. (2011)	40	English	8	15	65.00%
Mehri, Darooneh and Shariati (2012)	36	Persian	5	12	77.70%
Amancio, Oliveira Jr and Costa (2012b)	20 Several poems	English English	4 4	Not available	Not applicable ^b Not applicable ^b
Lahiri and Mihalcea (2013)	3,036	English	142	127 (summary features) other values (local features)	78.85%
Segarra, Eisen and Ribeiro (2013)	130	English	18	Not applicable ^c	88.00%
Amancio (2015a)	40	English	8	16 (variation analysis) 100 (word intermittency)	65.00%
Amancio (2015b)	40	English	8	29 (networks), 340 (intermittency and stopwords), 640 (character bigrams)	Not applicable ^d
Amancio (2015c)	20	English	4	11	86.67%
Amancio, Silva and Costa (2015)	40	English	8	229	82.50%
Akimushkin, Amancio and Oliveira Jr. (2017)	80	English	8	48	88.75%

^a [Antiqueira et al. \(2006\)](#) did not apply their methods to authorship attribution, they just suggested that co-occurrence networks could distinguish different authors.^b [Amancio, Oliveira Jr and Costa \(2012b\)](#) did not provide accuracy rates. Instead, they illustrated their methods with several hierarchical clusterings of the texts.^c There are no classification features, [Segarra, Eisen and Ribeiro \(2013\)](#) assigned unknown texts to the author with the most similar Markov chain.^d The accuracy rates reported by [Amancio \(2015b\)](#) are not percentage values, rather they provided relative accuracies considering a baseline method.

RESULTS

In this Chapter, we report the methodology and the results achieved during this Master’s research. In Section 4.1, we describe the selected datasets, the pre-processing steps applied to the texts and the processes to create networks from texts. In addition, we also present the classification features and the machine learning algorithms. The main results obtained with this research are reported in Section 4.2.

4.1 Materials and Methods

In this Section, we present the materials and methods used during this Master’s research. First, the datasets employed in our experiments are described. Then, we present the pre-processing steps and the techniques used to model texts as complex networks. Finally, we describe the process to obtain the classification features and the selected machine learning methods.

4.1.1 *Datasets*

In this work, we used several datasets for authorship attribution. They are presented in Tables A.1, A.2, A.3, and A.4 in the Section A of the Appendix, henceforth referred to as Dataset 1, 2, 3, and 4, respectively. All the selected books are in English and those from Dataset 1, 2, and 3 were extracted from the Project Gutenberg¹. Dataset 4 has some books under copyright rules. They were available at the Department of Computer Science from the University of Toronto. We also used debates extracted from the Canadian and the European Parliaments for the identification of translationese. Details about these debates are presented in the Section B of the Appendix.

¹ <<https://www.gutenberg.org/>>

4.1.2 Pre-processing

Some pre-processing steps can be applied before modelling texts as complex networks. One of the basic steps is to remove all punctuation marks. In all experiments conducted during this Master's research, contractions (such as *I'm*) were not expanded. The decision of using *I'm* instead of its equivalent form *I am* is probably related to the writing style of each author. Therefore, the contractions are mapped differently in the network when they are not expanded.

Another widely used approach is the removal of stopwords (or function words), which are mainly adverbs, articles, and prepositions. These words are usually disregarded because they convey little semantic content. Table 2 illustrates an example in which the stopwords were removed. The list of stopwords considered in this work is presented in the Section C of the Appendix.

Table 2 – Example of the removal of stopwords applied to sentences from the book *The Adventures of Sherlock Holmes*, written by Arthur Conan Doyle.

Original	No stopwords
<i>There are three men waiting for him at the door, said Holmes.</i>	<i>three men waiting door said holmes</i>
<i>Oh, indeed! You seem to have done the thing very completely.</i>	<i>oh indeed seem done thing completely</i>
<i>I must compliment you.</i>	<i>must compliment</i>
<i>And I you, Holmes answered.</i>	<i>holmes answered</i>

Another pre-processing step usually employed in many studies is known as lemmatization. This process is applied using a POS tagger. The tagger used in this work was the Natural Language Toolkit (NLTK) (BIRD; KLEIN; LOPER, 2009). This process modifies the words so that plural words, verbs and names are changed to their singular, infinitive and masculine forms, respectively. By doing so, words related to the same concept are associated to the same node, despite their different inflections. Table 3 illustrates the lemmatization process in the same extract presented in Table 2.

Table 3 – Example of the lemmatization step applied to sentences from the book *The Adventures of Sherlock Holmes*, written by Arthur Conan Doyle.

Original	After lemmatization
<i>There are three men waiting for him at the door, said Holmes.</i>	<i>there be three man wait for him at the door say holmes</i>
<i>Oh, indeed! You seem to have done the thing very completely.</i>	<i>oh indeed you seem to have do the thing very completely</i>
<i>I must compliment you.</i>	<i>i must compliment you</i>
<i>And I you, Holmes answered.</i>	<i>and i you holmes answer</i>

It is important to highlight that these steps are optional and their usage is related to the purpose of the application under study.

4.1.3 Network models

Most experiments conducted during this Master’s project used one of following network representations: word co-occurrence (or adjacency) and mesoscopic networks. Both representations are described below. It is important to highlight that network-based representations and methods are generic. Therefore, this methodology can be applied to several NLP tasks, not only authorship attribution.

4.1.3.1 Co-occurrence Networks

As a consequence of being formed by linear chains of words, one of the simplest ways to represent the written language is to connect adjacent words. This type of representation, known as co-occurrence network, is widely used in the literature (CANCHO; SOLÉ, 2001; AMANCIO *et al.*, 2011; ROXAS; TAPANG, 2010). Because most of the syntactic relationships occur among nodes in the first neighborhood, co-occurrence networks can be understood as an approximation of syntactic networks (CANCHO; SOLÉ; KÖHLER, 2004). In a co-occurrence network modelling text, the nodes represent words while the edges are established between adjacent words. We usually disregard sentence and paragraph boundaries, so that the last word of a sentence or paragraph is connected to the first word of the next sentence or paragraph. Figure 5 illustrates toy co-occurrence networks for the sentence *To be or not to be*. On the left, we presented a directed network while its undirected version is presented on the right. Both networks were created connecting each word to its immediate neighbor.

Figure 5 – Co-occurrence networks for the sentence *To be or not to be* extracted from the well-known play *Hamlet* written by William Shakespeare. In this approach, each word is connected to its immediate neighbor. A directed network is presented on the left while its undirected version is presented on the right.

Source: Elaborated by the author.

A key aspect in the co-occurrence representation is related to the size of the words window J , which is an indicator of the context around a word. The simplistic approach that connects words to their immediate neighbors, as the networks illustrated in Figure 5, disregards some syntactic and semantic relationships that may occur among distant words (ALVAREZ-LACALLE *et al.*, 2006). To account for those relationships, the size of J can be increased. For example, Figure 6 illustrates toy co-occurrence networks in which $J = 2$. Therefore, each word is connected to its first and second closest neighbors.

Figure 6 – Extended co-occurrence networks for the sentence *To be or not to be* extracted from the well-known play *Hamlet* written by William Shakespeare. In this window-based representation, each word is connected to its first and second closest neighbors. A directed network is presented on the left while its undirected version is presented on the right.

Source: Elaborated by the author.

A number of studies have found that co-occurrence networks are useful to model texts in several contexts (AMANCIO *et al.*, 2011; AMANCIO *et al.*, 2012; LAHIRI; MIHALCEA, 2013; CONG; LIU, 2014; ARRUDA; COSTA; AMANCIO, 2015). In particular, extended co-occurrence networks have been employed to capture the topical structure of texts (ARRUDA; COSTA; AMANCIO, 2016) and to automatically identify cognitive impairments from transcripts (SANTOS *et al.*, 2017).

4.1.3.2 Mesoscopic Networks

Mesoscopic-based approaches are alternative ways to represent texts as complex networks. Such networks are able to portray the topical structure and the text unfolding along time (ARRUDA *et al.*, 2017). The methodology to create mesoscopic networks is illustrated in Figure 7. Initially, the text T is divided into paragraphs so that $T = (p_1, p_2, \dots)$, where p_i is the sequence of words forming paragraph i (see Figure 7-a). In mesoscopic networks, each node represents a sequence of Δ consecutive paragraphs, as shown in Figure 7-b. Then, the *tf-idf* (MANNING; SCHÜTZ, 1999) technique was used to quantify the importance of the words in each sequence of paragraphs, which is presented in Figure 7-c. Each pair of nodes i and j is connected by an edge whose weight is given by the cosine similarity between their respective *tf-idf* maps (see Figure 7-d). Finally, unweighted networks can be obtained following specific processes. For instance, the edges with the lowest weights can be removed in order to obtain a fixed percentage of edges or to reach a given average degree (ARRUDA *et al.*, 2017; MARINHO *et al.*, 2017). Figure 7-e shows the obtained unweighted network with average degree equals to 2.

4.1.4 Features

After modelling the texts as networks using one of the representations discussed above, the measurements described in Section 2.1.3 can be extracted from those networks. For the cases when the measurements are globally defined, i.e. there is a single value for the whole network, such as the assortativity and the frequency of motifs, the values of these measurements will be

Figure 7 – The mesoscopic approach proposed by [Arruda et al. \(2017\)](#) includes five main steps. In (a), the text T is divided in terms of its paragraphs. Then, overlapping windows with Δ subsequent paragraphs are extracted from the text, in this case $\Delta = 2$ (b). To account for the words in each window, its respective tf-idf map is computed (c). The weight of the edge connecting i and j is given by the cosine similarity between their respective tf-idf maps (d). The edges with the lowest weights are discarded following a specific process (e), such as to reach a given average degree. For example, the edges with the lowest weights in (e) were removed until the network reached an average degree of 2.

Source: [Marinho et al. \(2017\)](#).

directly used as features for the machine learning process. On the other hand, most measurements presented in Section 2.1.3 are locally defined, i.e. a value is assigned to each node. In such cases, we calculated the average $\langle X \rangle$, standard deviation $\sigma(X)$, and skewness (third moment) $\gamma(X)$ of these measurements over all nodes.

$$\text{Average:} \quad \langle X \rangle = \frac{1}{M} \sum_{i=1}^M X_i, \quad (4.1)$$

$$\text{Standard Deviation:} \quad \sigma(X) = \sqrt{\frac{\sum_{i=1}^M (X_i - \langle X \rangle)^2}{M - 1}}, \quad (4.2)$$

$$\text{Skewness:} \quad \gamma(X) = \left\langle \left(\frac{X - \langle X \rangle}{\sigma(X)} \right)^3 \right\rangle, \quad (4.3)$$

where X represents a locally defined network measurement and M is the number of nodes. The obtained statistics from each measurement were then used as features for the machine learning methods.

4.1.5 Machine Learning Methods

In order to quantify the ability of the selected features to characterize written texts, we used some supervised methods to induce classifiers from a training set. Before describing each method, consider the following definitions. The training set $X_{\text{training}} = \{(x_1, y_1), \dots, (x_l, y_l)\}$ contains l tuples, where the first component of the i -th tuple, $x_i = (f_1, \dots, f_d)$, represents the

classification features. The second component, y_i , is the class of the instance. The goal of a supervised method is to learn the mapping $x \mapsto y$ from the training set. A test set $X_{test} = \{x_{l+1}, \dots, x_{l+u}\}$ is used to verify the quality of the learned models. The selected techniques are described below.

- **Support Vector Machines (SVM):** In this technique, the training instances are divided into several regions according to their features. This is performed by specific functions that aim to maximize the separation margin (FACELI *et al.*, 2011). By doing so, new instances are classified according to their placement in one of those regions.
- **kNN:** This technique is based on a voting process over the k -closest instances from the training set (k_{nn}), in a normalized space involving all attributes (AHA *et al.*, 1991). If most of the instances from the set k_{nn} belong to the class y' , then this class will be assigned to the unknown instance.
- **Naive Bayes:** Based on the Bayes rule, this method states that the class y' of a test instance is the one that satisfies the following condition:

$$P(y'|f_1, \dots, f_d) > P(y_k|f_1, \dots, f_d), \quad (4.4)$$

where $P(y_k|f_1, \dots, f_d)$ is the probability of assigning the class y_k to a test instance given the features $F = \{f_1, \dots, f_d\}$, for each $y_k \neq y'$ (MITCHELL, 1997).

- **Decision Trees:** These methods, such as *C4.5* and *J48*, create a decision tree based on how each feature splits the training instances (MITCHELL, 1997). Different metrics can be used, such as the information gain and the Gini index (QUINLAN, 1993).
- **Random Forests:** Random forests construct various decision tree classifiers from several samples of the training data and combine their predictions in a majority rule voting (BREIMAN, 2001).

For all methods, well-known machine learning libraries, such as *Scikit-learn* (PEDREGOSA *et al.*, 2011) and *Weka* (HALL *et al.*, 2009), were used. In most cases, we employed the default configurations of these methods. These algorithms were applied to a training set independent of the test set using cross validation techniques. We usually employed *leave-one-out* or 10-fold cross validation. In the former, at each cycle one instance is used as test while all the others are used in the training process. In the latter, one tenth of the instances are used as test whereas the other nine tenths are used to train the classifier. For the cases when feature selection was attempted, we used *wrapper*-based methods, in which the subsets of attributes are evaluated by using a selected classification model (KOHAVI; JOHN, 1997).

4.2 Results

This Section describes the main results achieved during this Master’s research. First, in Section 4.2.1, we present the techniques used to extend the co-occurrence models. Then, in Section 4.2.2, we show that small interconnection patterns found on co-occurrence networks are relevant for the authorship attribution task. Section 4.2.3 describes our hybrid classifier (called *Labelled motifs*), which combines topological features from co-occurrence networks with the frequency of common words. The performance of labelled motifs is evaluated in the contexts of authorship attribution and translationese identification. Finally, we present in Section 4.2.4 some network representations (other than co-occurrence networks) applied to the authorship attribution task.

4.2.1 Extensions of co-occurrence networks

While several studies have been devoted to apply networked representations to NLP tasks, only a few works have tried to extend the word co-occurrence models (ARRUDA; COSTA; AMANCIO, 2016; SANTOS *et al.*, 2017). In this Master’s research, we propose some extensions to the well-known word co-occurrence model for the goal of grasping styles in a more adequate and accurate manner. These extensions include the connection of words based on further hierarchies and the addition of syntactic and relevant links. However, the accuracy rates obtained with syntactic and relevant links were not as high as the others. For simplicity’s sake, those results were not included in this manuscript.

The traditional co-occurrence model captures the stylistic properties of texts by connecting only immediate adjacent words. However, this model is not without its share of problems. For example, co-occurrence networks might fail at capturing the relationship between distant words. Constantoudis *et al.* (2015) reported that long-range correlations in written texts occur due to the multidimensional mapping of thoughts and ideas in chains of words. In order to account for the presence of relevant links between non-adjacent words, we proposed an extension called *Further Neighborhood*. In this method, we connect every pair of words that are separated by less than $J - 1$ intermediary words, where $J = \{1, 2, 3\}$. It is important to highlight that for larger values of J , the precision of the obtained links will be low. For $J = 1$, the traditional co-occurrence network is obtained. Figure 6 illustrated two networks (directed and undirected) created with the *Further Neighborhood* method for $J = 2$.

Initially, we selected the texts from Dataset 1. We removed stopwords and the remaining words were lemmatized. After the networks were created, the following measurements were extracted from them: assortativity, betweenness centrality, clustering coefficient, selectivity of words, the shortest paths, and the average degree of neighbors. We used four machine learning methods implemented in *Weka* with 10-fold cross validation. The accuracy rates are reported in Table 4. We compared our *Further Neighborhood* methods with a simplistic co-occurrence

extension, called *Sentence-based*. In this technique, all words from the same sentence are connected in a clique. This is based on the assumption that sentences represent substructures organized in a meaningful way to express an idea. A sentence was considered as a sequence of words separated by periods, quotation, exclamation or question marks, as well as colons and semi-colons. In addition, we also applied the *Further Neighborhood* method in the texts from Dataset 2. These results are then presented in Table 5.

Table 4 – Percentage (%) of books correctly classified from Dataset 1 with the *Further Neighborhood* and *Sentence-based* methods. Two sets of attributes were considered: (i) all features (AF), and (ii) the features obtained from the feature selection method (FS).

	J48		kNN		SVM		Naive Bayes	
	AF	FS	AF	FS	AF	FS	AF	FS
Further neighborhood $J = 1$	32.5	62.5	42.5	55.0	40.0	35.0	35.0	50.0
Further neighborhood $J = 2$	42.5	62.5	52.5	65.0	42.5	40.0	50.0	60.0
Further neighborhood $J = 3$	47.5	70.0	52.5	72.5	40.0	45.0	60.0	65.0
Sentence based	35.0	52.5	40.0	60.0	40.0	22.5	40.0	45.0

Table 5 – Percentage (%) of books correctly classified from Dataset 2 with the *Further Neighborhood* method. Two sets of attributes were considered: (i) all features (AF), and (ii) the features obtained from the feature selection method (FS).

	J48		kNN		SVM		Naive Bayes	
	AF	FS	AF	FS	AF	FS	AF	FS
Further neighborhood $J = 1$	25.0	40.0	35.0	46.0	12.0	7.0	37.0	40.0
Further neighborhood $J = 2$	25.0	39.0	38.0	44.0	17.0	7.0	38.0	41.0
Further neighborhood $J = 3$	23.0	41.0	40.0	45.0	29.0	2.0	33.0	39.0

The results summarized in Table 4 and 5 indicate that the *Further Neighborhood* extension usually characterizes better the texts for the authorship attribution task, when compared to the traditional co-occurrence representation. These results confirm our initial hypothesis which stated that the connection of words in a longer context could improve the performance of some authorship attribution scenarios. The best accuracy rate in Dataset 1 was enhanced to 72.5%. Even though the *Sentence-based* extension provided accuracy rates higher than the expected chance baseline for Dataset 1, 12.5%, those were not as high as the ones achieved with the *Further Neighborhood* extension.

4.2.2 Motifs

Current CN-based solutions to authorship attribution apply a myriad of network measurements, most of them were described in Section 2.1.3. However, there has been little discussion on the usage of motifs to characterize and distinguish writing styles for authorship attribution. The number of *motifs* found on a network can be used to characterize its topology. In this experiment, we tested the hypothesis that fingerprints left by different authors can be captured in a simplistic

manner via the frequency of network motifs. As such, the main goal was to apply the concept of motifs to characterize the authorship of texts and probe the relevance of such structures as features for the problem. To evaluate the significance of the occurrences of the motifs, their frequencies are typically compared with the expected ones in random networks (MILO *et al.*, 2002). For simplicity's sake, we disregarded the occurrences of those structures in random networks and the word motifs is henceforth used as a synonym for subgraphs.

For the experiments presented in this Section, four input scenarios were considered: (i) original text, (ii) without stopwords, (iii) after the lemmatization process and (iv) after the removal of stopwords and the lemmatization process. We selected those scenarios in order to investigate some pre-processing steps that can be applied before creating the networks. For each input scenario, one directed network was derived for each book from Dataset 1, shown in Table A.1. Then, those networks were characterized by the absolute frequency (i.e. raw count) of the thirteen directed motifs with three nodes, which were presented in Figure 1. We only extracted those motifs because we aimed to use only a few attributes. The frequencies were extracted with a script developed during this Master's project and the values were checked against the ones obtained with the software *mfinder* (KASHTAN *et al.*, 2004a). Those values were then used as features for the selected machine learning algorithms available in *Weka* with 10-fold cross validation. The accuracy rates are presented in Table 6.

In addition, a random verification step was conducted to probe whether the motifs are actually extracting a real pattern. In this step, each instance maintains the same features (i.e. the frequency of all 13 motifs); however, its class y is randomly selected from the set of candidate authors. It is important to highlight that the correct author is also included in this set. This process was run 10 times and used the instances from the input scenario (iii). The accuracy rates are presented in the last row of Table 6. These results are quite similar to the expected chance baseline for this problem, 12.5%, because each one of the 8 candidate authors has the same probability of being randomly selected.

Table 6 – Percentage (%) of books correctly classified when the absolute frequencies of directed motifs were used as the only classification features.

	J48	kNN	SVM	Naive Bayes
(i) Original text	40.0	55.0	45.0	45.0
(ii) Without stopwords	27.5	32.5	0.0	30.0
(iii) Lemmatization process	57.5	45.0	45.0	52.5
(iv) Without stopwords + lemmatization	22.5	27.5	2.5	30.0
Random verification step	12.0	9.0	13.0	10.8

Taken together, the results presented in Table 6 confirm that the frequency of motifs is able to significantly extract patterns related to the authorship of texts and discriminate different writing styles. The best accuracy rate achieved with the frequency of motifs was 57.5%, which is considerably higher than the expected chance baseline for the problem. Remarkably,

Figure 8 – PCA of the texts in two scenarios: original and without stopwords. The letters in the legend represent the following authors: Arthur Conan Doyle (A), Bram Stoker (S), Charles Dickens (D), Edgar Allan Poe (P), Hector Hugh Munro (M), Pelham Grenville Wodehouse (W), Thomas Hardy (H), William Makepeace Thackeray (T).

Source: Elaborated by the author.

function words played an important role in such structures. This is confirmed by the very low accuracy rates obtained when those words were disregarded. The Principal Component Analysis (PCA) (JOLLIFFE, 2002) of the texts in scenario (i) and (ii) are presented in Figures 8(a) and 8(b), respectively. A much better discrimination is achieved when stopwords are considered.

For comparison purposes, we extracted five traditional network measurements from the undirected co-occurrence networks. The selected measurements were the average degree of neighbors, shortest paths, betweenness centrality, clustering coefficient and the assortativity. The accuracy rates for the four scenarios are presented in Table 7.

Table 7 – Percentage (%) of books correctly classified when other network features were extracted from the books

	J48	kNN	SVM	Naive Bayes
(i) Original text	50.0	42.5	42.5	55.0
(ii) Without stopwords	37.5	45.0	27.5	37.5
(iii) Lemmatization process	47.5	50.0	37.5	45.0
(iv) Without stopwords + lemmatization	32.5	37.5	32.5	40.0

Comparing the results presented in Tables 6 and 7, the accuracy rates achieved for the input scenarios (i) and (iii) are higher than the ones retrieved with the selected traditional network measurements. However, some of these measurements, such as the betweenness centrality, are correlated with the word frequency. As a consequence, this correlation may improve the

performance of such measurements.

To the best of our knowledge, this was one of the first works in which network motifs were used to identify the authorship of texts. The goal of this experiment was not to provide state-of-the-art results for the task. Instead, the characterization based on the absolute frequency of motifs may be seen as a complementary feature. Moreover, this approach can be combined with traditional techniques employed in stylistic studies.

4.2.3 *Labelled Motifs*

Traditionally, CN-based approaches to text analysis usually disregard the textual context after the networks are devised. Moreover, the topological measurements do not even consider such information. The textual context – such as the words associated to each node, i.e. node labels – might be useful to characterize the networks. However, only a few works have investigated the benefits of combining both paradigms. So inspired, we proposed a hybrid classifier, henceforth referred to as *Labelled Motifs*, that combines the frequency of common words with small structures known as motifs. This was achieved by considering the node labels during the extraction of motifs.

In this hybrid classifier, instead of using the frequency of motifs, we calculated the frequency of a given word w in all the 13 directed motifs. In particular, the frequency of a labelled motif that combines word w with motif m , $f_{w,m}$, is calculated as

$$f_{w,m} = \frac{\tilde{f}_{w,m}}{f_m}, \quad (4.5)$$

where the numerator is the number of occurrences of word w in motif m , while the denominator represents the number of occurrences of motif m , irrespective of the node labels. The word w belongs to the set of words W . Several criteria could be used to select the set of words W . In this experiment, we considered W as the set of the most frequent words from the training dataset. This description corresponds to the first version of the devised method, referred to as *LMV1*. The second version considers the word position inside the motif in terms of the different configurations of the nodes, referred to as *LMV2*. Some frequencies of labelled motifs extracted from the toy network depicted in Figure 9 are described below:

- *LMV1*: The frequency of word *question* in Motif 2. The word *question* is one of the nodes in 3 occurrences of Motif type 2, i.e. $\tilde{f}_{\text{question},2} = 3$. Motif 2 occurs 7 times ($f_2 = 7$). Therefore, the frequency of word *question* in Motif 2 is $f_{\text{question},2} = 3/7$.
- *LMV2*: The frequency of word *question* as the central node in Motif 2. The word *question* appears only once in such configuration in Motif type 2, i.e. $\tilde{f}_{\text{question},\text{central},2} = 1$. Therefore, the frequency of word *question* in this node configuration in Motif 2 is $f_{\text{question},\text{central},2} = 1/7$.

Figure 9 – An example of a co-occurrence network is presented on the left. On the right, we show all motifs of type 2 extracted from this network, with a total of 7, i.e. $f_2 = 7$.

Source: [Marinho, Hirst and Amancio \(2017\)](#).

The ability of labelled motifs in characterizing texts was initially probed in a case study. We selected the following books: *Agnes Grey* and *The Tenant of Wildfell Hall*, written by Anne Brontë, *Jane Eyre* and *The Professor* from Charlotte Brontë, and *Wuthering Heights* from Emily Brontë. The Brontë sisters are very hard to distinguish ([KOPPEL; SCHLER; MUGHAZ, 2004](#)). Each book was split into non-overlapping partitions of 8,000 words, with a total of 76 instances. No pre-processing step was applied other than the removal of punctuation marks. The frequencies of the words *a* and *to* extracted from each partition are presented in Figure 10. We also extracted the frequency of both words in motif 2, using *LMV* 1.

A careful examination reveals that there is a large overlap between the partitions from Emily Brontë (orange squares) and Charlotte Brontë (blue triangles) in Figure 10-(a), while a much better discrimination was achieved with the labelled motifs in Figure 10-(b). These results suggest that labelled motifs are useful to discriminate texts. Moreover, for some cases, this approach achieves better discrimination results than the ones obtained with only the frequency of words.

4.2.3.1 Labelled Motifs for Authorship Attribution

To address the authorship attribution task, we considered Dataset 1, presented in Table A.1, and Dataset 3, presented in Table A.3. For this experiment, the books from Dataset 1 were truncated to the size of the shortest novel. On the other hand, the books from Dataset 3 were used to evaluate the performance of our hybrid classifier in characterizing shorter pieces of text. Therefore, the books from Dataset 3 were split into several non-overlapping partitions with 8,000 words each. To avoid issues with imbalanced classes, we selected the same number of partitions per author. We considered W as the set of the most frequent words. By doing so, the classification

Figure 10 – Two feature sets are extracted from 76 partitions of the books of Anne Brontë (AB), Charlotte Brontë (CB) and Emily Brontë (EB). In (a), the partitions of the books of Charlotte and Emily Brontë are characterized by the frequency of two words, *a* and *to*. In (b), the same data is visualized according to the frequency of word *a* and *to* in Motif type 2, i.e. $f_{a',2}$ and $f_{to',2}$. Finally, the partitions from Anne Brontë are added in (c).

(a) Frequency of word *a* and word *to* in partitions of Charlotte Brontë and Emily Brontë. (b) Frequency of word *a* and word *to* in Motif type 2 in partitions of Charlotte Brontë and Emily Brontë.

(c) Frequency of word *a* and word *to* in Motif type 2 for the three sisters.

Source: Elaborated by the author.

features consist of the frequencies of all combinations of words from W appearing in the motifs, considering the versions $LMV1$ and $LMV2$. We also run the experiment for the case when the features were based solely on the frequency of the W most frequent words, referred to as MFW . We employed four machine learning algorithms available in *Weka* using 10-fold cross-validation. These results are presented in Table 8, where $|W| = \{5, 10, 20\}$.

Table 8 – Percentage (%) of texts correctly classified when labelled motifs were extracted from books of Dataset 1 and 3. For Dataset 1, the best result of our technique surpasses by 15 percentage points the best one obtained with the frequency of common words. However, for Dataset 3, the gain in performance was less than 2 percentage points.

Dataset	Methods	$ W $	J48	kNN	SVM	Naive Bayes
Dataset 1	LMV1	5	45.0	65.0	62.5	30.0
	LMV1	10	37.5	60.0	67.5	27.5
	LMV1	20	60.0	65.0	75.0	25.0
	LMV2	5	55.0	50.0	62.5	22.5
	LMV2	10	47.5	65.0	77.5	15.0
	LMV2	20	45.0	60.0	80.0	25.0
	MFW	5	30.0	57.5	22.5	50.0
	MFW	10	45.0	52.5	27.5	42.5
	MFW	20	52.5	62.5	65.0	45.0
Dataset 3	LMV1	5	58.7	65.1	74.3	69.2
	LMV1	10	61.7	83.7	91.6	81.3
	LMV1	20	66.8	88.3	95.4	78.7
	LMV2	5	62.1	67.4	82.0	69.1
	LMV2	10	65.5	80.9	91.6	75.3
	LMV2	20	68.1	88.0	96.0	77.5
	MFW	5	58.3	67.8	57.8	73.5
	MFW	10	65.8	83.6	85.3	83.8
	MFW	20	70.3	91.1	94.4	91.5

The best results presented in Table 8 were achieved with the SVM. In addition, the second version of our hybrid classifier yielded better results than those from the $LMV1$ version. Such a pattern reinforces the relevance of function words in specific node configurations. The relevance of combining the node labels with the local structure of networks becomes evident when we compare our results with the ones obtained with the frequency of words. For Dataset 1, the best accuracy rate obtained with the frequency of common words was 65%, while our technique reached 80%. However, the gain in performance in Dataset 3 was only 1.6 percentage points. A possible explanation is that this classification scenario can be considered easier. Because each book was split into several parts, some partitions used during the test phase are likely to be very similar to others (from the same book) used during training, which made the task easier for the classifiers. In addition, Dataset 3 has fewer books per author, for instance Emily Brontë and Nathaniel Hawthorne had only one book each, which resulted in less variance of writing styles.

The results presented in Table 8 show that, in most cases, larger values of $|W|$ yield better accuracy rates. To investigate the dependency between the accuracy rates and the size of the set W , we run the authorship attribution task for $1 \leq |W| \leq 40$. The accuracy rates obtained for each value of $|W|$ in Dataset 1 and 3 are presented in Figures 11 and 12, respectively.

Figure 11 – Accuracy rates in assigning the authorship of books from Dataset 1 for several values of $|W|$, in the two versions of our hybrid classifier.

Source: Elaborated by the author.

Figure 12 – Accuracy rates in assigning the authorship of books from Dataset 3 for several values of $|W|$, in the two versions of our hybrid classifier.

Source: Elaborated by the author.

A careful examination of Figures 11 and 12 reveals that, in most cases, SVM outperforms the other classifiers. Moreover, this classifier does not require many words to provide excellent results. Therefore, we did not use more than 20 words in the other experiments, once there is an efficiency loss related to the inclusion of more words.

In the last experiment, we analyzed the texts from Dataset 4, which is presented in Table A.4. Initially, we verified how distinct was the group of authors (Agatha Christie, Iris Murdoch, P. D. James, and Ross Macdonald) with a simple pairwise classification. We used the full length of the books and we did not apply any pre-processing steps. The novels from all the three time periods of each author were used. We devised one co-occurrence network for each book and we extracted the labelled motifs as features for the classification, with $|W| = 20$. For comparison purposes, we also extracted the absolute frequency of the directed motifs with 3 and 4 nodes (total of 199 different motifs) and we compare those results with the majority class baseline. In order to simplify our table, we just presented the results for the classifier with the best accuracies, SVM. The results are presented in Table 9. Once again, these results confirm that motifs and labelled motifs are able to characterize different writing styles for the authorship attribution task.

4.2.3.2 Labelled Motifs for Translationese Identification

In order to probe the ability of our hybrid classifier to discriminate texts in a task other than authorship attribution, we employed labelled motifs in the task known as translationese identification. In this task, the main goal is to evaluate whether a text in a language L was originally produced in L or it was translated into L . One of the first studies on translationese was carried out by Gellerstam (1986). He analyzed texts produced in Swedish and texts produced in other languages and then translated into Swedish, and noticed that the main differences between them are not related to the quality of the translation. Instead, these differences can be understood as an influence of the source language on the target one. Since then, several works have been dedicated to propose methods that automatically distinguish original from human-translated texts (BARONI; BERNARDINI, 2006; HALTEREN, 2008; ILISEI *et al.*, 2010; POPESCU, 2011; AVNER; ORDAN; WINTNER, 2016; RABINOVICH; WINTNER, 2015). Most of these works apply their methods in a range of parallel resources, such as versions in several languages of literary works, news articles, and transcripts from parliamentary debates.

In this experiment, we selected the Canadian Hansard and the Europarl (KOEHN, 2005). These two datasets are described in the Section B of the Appendix. The high quality of their translations is a consequence of the good translation standards the public organizations have to follow. Therefore, the task of translationese identification using such corpora is challenging. Moreover, it provides another scenario to investigate the capabilities of the proposed hybrid methodology.

We did not employ any pre-processing step apart from removing punctuation marks from both datasets. The Canadian Hansard dataset was studied in terms of its two languages, English and French. For the English analysis, each one of the 463 sessions was divided into two files, one containing all sentences produced in English and the other with the sentences translated into English, representing the classes *Original* and *Translated*, respectively. One co-occurrence

Table 9 – Percentage (%) of books from Dataset 4 correctly classified from each pair of authors and from the complete set of authors.

Authors	Baseline	Methods	SVM
Christie-Murdoch	68.9	<i>LMV1</i>	92.6
		<i>LMV2</i>	93.5
		Motifs of Size 3	79.0
		Motifs of Size 4	83.2
Christie-James	56.6	<i>LMV1</i>	96.5
		<i>LMV2</i>	97.9
		Motifs of Size 3	90.2
		Motifs of Size 4	91.8
Christie-Macdonald	57.0	<i>LMV1</i>	97.1
		<i>LMV2</i>	97.2
		Motifs of Size 3	78.6
		Motifs of Size 4	87.6
Murdoch-James	62.9	<i>LMV1</i>	98.0
		<i>LMV2</i>	97.1
		Motifs of Size 3	81.2
		Motifs of Size 4	86.9
Murdoch-Macdonald	62.6	<i>LMV1</i>	98.2
		<i>LMV2</i>	99.0
		Motifs of Size 3	79.5
		Motifs of Size 4	93.2
James-Macdonald	50.3	<i>LMV1</i>	99.2
		<i>LMV2</i>	99.5
		Motifs of Size 3	83.4
		Motifs of Size 4	95.1
All four	37.9	<i>LMV1</i>	91.5
		<i>LMV2</i>	93.7
		Motifs of Size 3	63.1
		Motifs of Size 4	76.3

network was devised for each file and labelled motifs were extracted. For comparison purposes, we also extracted the frequency of common words. The French analysis was conducted in a similar way. We employed four machine learning algorithms available in *Weka* using 10-fold cross validation. The accuracy rates obtained with this dataset are presented in Table 10. These results are relatively high, which suggests that labelled motifs are able to capture information about French to English and English to French translations. Remarkably, in almost all cases, these features achieved results higher than the ones based solely on the frequency.

In the Europarl, we studied translationese using four target languages (English, French, Italian, and Spanish) and six source languages (English, Finnish, French, German, Italian, and Spanish). Those source languages were chosen because they were employed in a previous study conducted by [Koppel and Ordan \(2011\)](#). For the English analysis (i.e. English as the

Table 10 – Accuracy rates (%) in discriminating the debates from the Canadian Hansard into two classes (*Original* and *Translated*).

Target Language	Methods	W	J48	kNN	SVM	Naive Bayes
English	LMV1	20	90.6	89.6	97.1	90.7
	LMV2	20	90.7	94.0	98.3	88.4
	MFW	5	72.2	75.3	57.8	53.1
	MFW	10	74.8	76.0	60.6	53.7
	MFW	20	78.5	80.2	64.4	54.3
French	LMV1	20	94.6	87.0	98.3	89.8
	LMV2	20	95.5	89.4	98.7	89.3
	MFW	5	70.7	70.0	59.3	53.5
	MFW	10	72.0	72.5	56.4	53.8
	MFW	20	87.3	87.2	63.0	55.3

target language), all the sentences in English were split into six files, according to their source languages. In this case, the file whose source language is English belongs to the class *Original*, and the files whose source languages are Finnish, French, German, Italian, and Spanish are examples of the class *Translated*. These 6 files were divided into non-overlapping partitions with 8,000 words each. In order to avoid issues with imbalanced data, we selected approximately $5n$ partitions from English and n partitions from the other 5 source languages, where $n = 180$ partitions. We devised one co-occurrence network from each partition. The remaining steps are similar to those applied to the Canadian Hansard. The analyses with French, Italian, and Spanish – the other three target languages – were conducted in a similar way, where $n = 128$, $n = 55$, and $n = 69$ partitions, respectively. The accuracy rates achieved with the Europarl are presented in Table 11. For simplicity’s sake, we just reported the results for the classifier with the best accuracies.

Table 11 – Accuracy rates (%) in discriminating the debates from the Europarl into two classes (*Original* and *Translated*).

Target Language	Methods	W	SVM	Target Language	Methods	W	SVM
English	LMV1	20	90.6	Italian	LMV1	20	93.5
	LMV2	20	92.8		LMV2	20	95.9
	MFW	5	68.6		MFW	5	85.6
	MFW	10	68.8		MFW	10	90.7
	MFW	20	78.8		MFW	20	92.7
French	LMV1	20	87.3	Spanish	LMV1	20	91.0
	LMV2	20	88.1		LMV2	20	93.4
	MFW	5	62.2		MFW	5	88.4
	MFW	10	78.1		MFW	10	90.8
	MFW	20	82.7		MFW	20	92.0

Taken together, the results presented in Tables 10 and 11 confirm the relevance of frequent words as features for translationese identification, as reported by Koppel and Ordan (2011). Once again, the characterization provided by labelled motifs captures information about translationese, even when more than one source language is present in the class *Translated*. However, the gain in performance depends on the language being studied. For instance, our method surpassed the accuracy obtained with word frequencies by a margin of 14 percentage points for English, while the gain was only 1.47 percentage points for Spanish. In the latter case, excellent results were achieved using only the frequency of the 5 most frequent words.

In the literature, some works have already used the Canadian Hansard and the Europarl to investigate the translationese. For example, Kurokawa, Goutte and Isabelle (2009) used the 35th to 39th Parliaments from the Canadian Hansard. Their method achieved accuracies as high as 90% using word bigram frequencies. The sentences from the dataset were analyzed individually and as blocks with varying sizes. Koppel and Ordan (2011) analyzed 2,000 English chunks from the Europarl. Their best accuracy rate, 96.7%, was achieved with the frequency of 300 function words. However, they only detected translationese with English as the target language.

4.2.4 *Other network representations applied to authorship attribution*

In this Section, we describe some approaches to authorship attribution based on representations other than the co-occurrence model. In Section 4.2.4.1, we report the results obtained with networks created from named entities. In addition, mesoscopic networks are used to model the texts in Section 4.2.4.2. Finally, we present a simplified version of the function word networks proposed by Segarra, Eisen and Ribeiro (2013) in Section 4.2.4.3.

4.2.4.1 *Named Entity Networks*

Named entities are words that name people, places or organizations. Elson, Dames and McKeown (2010) extracted social networks from several nineteenth-century British novels and discovered that the majority of novels do not fit some characterizations provided by literacy scholars. For example, some theorists have suggested that the novel's setting (urban or rural) would have an effect on the structure of its social network. However, the authors found that the number of characters and speakers in the urban novels were not significantly greater than those found in the rural novels.

Assuming that authors may connect the named entities in their texts in very different ways, the focus of this activity was to evaluate whether the network of named entities could extract information about the authorship of an unknown text. To identify these entities, we employed a technique called named entity recognition, which identifies people, places and organizations in texts. We used the Stanford Named Entity Recognizer (NER) (FINKEL; MANNING, 2009).

This approach was based on the finding that concepts appearing in the same context are likely to be semantically related (MATHIESEN; YDE; JENSEN, 2012). Therefore, we connected all entities placed in the same context in order to capture the existent semantic relationships. We connected two entities whenever they are distant to each other by less than X words, where $X = \{250, 500, 750, 1000\}$. To illustrate, if we used $X = 3$ in the following sentence *Romeo and Juliet lived in Verona*, the entities *Romeo* and *Juliet* (and *Juliet* and *Verona*) are connected by edges. On the other hand, the entities *Romeo* and *Verona* are distant to each other by more than 3 words and, therefore, are not connected.

Figure 13 – Example of a named entity network for the sentence *Romeo and Juliet lived in Verona* using $X = 3$.

Source: Elaborated by the author.

The named entities from Dataset 1, which is presented in Table A.1, and Dataset 4, shown in Table A.4, were extracted. We used the original texts because the NER does not work properly when the input text is pre-processed. In addition to some of the measurements described in Section 2.1.3, we extracted two more measurements, the number of nodes (i.e. the number of entities in the text) and the number of edges (i.e. the number of connections among entities). We used four machine learning methods implemented in *Weka* with 10-fold cross validation (for Dataset 4) and leave-one-out (for Dataset 1). We did not extract motifs or labelled motifs in this experiment. The accuracies are presented in Table 12. The expected chance baseline for Dataset 1 is 12.5% while the one for Dataset 4 is around 30%.

The accuracies obtained in this experiment are lower than the ones we have already reported; however, this is still a relevant finding. This experiment shows that not only stylometric features extract information and characterize the authorship of books, but also the named entity networks retrieved from them. Remarkably, the way that characters, locations and organizations are connected along the books carries information about their authorship. Therefore, this information can be used as a complementary feature in other authorship attribution classifiers.

Table 12 – Percentage (%) of books correctly classified when named entity networks were extracted from Dataset 1 and 4.

Dataset	Context size	J48	kNN	SVM	Naive Bayes
Dataset 1	X = 250	20.0	20.0	27.5	25.0
	X = 500	32.5	15.0	17.5	42.5
	X = 750	35.0	25.0	20.0	40.0
	X = 1000	27.5	32.5	25.0	47.5
Dataset 4	X = 250	64.1	64.1	70.1	77.6
	X = 500	61.2	68.6	68.6	73.1
	X = 750	64.1	70.1	70.1	68.6
	X = 1000	62.6	67.1	68.6	64.2

4.2.4.2 Mesoscopic Networks

The mesoscopic networks proposed by [Arruda et al. \(2017\)](#) have proven successful at capturing the story flow. For instance, they described the topology of the mesoscopic network derived from the book *Alice’s Adventures in Wonderland* according to the events that occurred along the book. Moreover, their method was used to discriminate real from shuffled texts, where no story exists. The goal of this experiment was to investigate if there is a dependency between authors’ writing styles and the story flow of their books. In particular, we tested the hypothesis that fingerprints left by each author are present at a mesoscopic scale.

As explained in Section 4.1.3.2, some criterion has to be selected in order to obtain unweighted mesoscopic networks. For this experiment, we decided to remove the edges with the lowest weights until each network reached a fixed average degree $\langle k \rangle$. Instead of selecting a single value for $\langle k \rangle$, we used average degrees ranging from 5 to 50, by steps of 5. The network measurements extracted from the networks were: degree, average degree of neighbors, assortativity, clustering coefficient, accessibility and symmetry (for $h = \{2, 3\}$). We tested several classifiers and the SVM and Random Forest were chosen. They are available in *Scikit-learn* and we used the *leave-one-out* cross validation technique. The texts from Dataset 2, shown in Table A.2, were pre-processed and a mesoscopic network was devised for each book. The accuracy rates are presented in Table 13.

Considering that the expected chance baseline for Dataset 2 is only 5%, our best scenario achieved results as high as 35%. A pairwise classification with the complete set of candidate authors is illustrated in Figure 14. For that classification, we combined the features obtained for all the average degrees listed in Table 13 and used the SVM classifier.

The accuracy rates presented in Figure 14 are considerably good, except the ones in the lighter colors. From the set of 20 authors, we selected four – Charles Darwin, Thomas Hardy, Edgar Allan Poe, and Mark Twain – due to the diversity of their works: such as *novels*, *short stories* and *scientific theories*. The obtained accuracy rates in classifying the authorship into

Table 13 – Percentage (%) of books correctly classified when mesoscopic networks were extracted from the books of Dataset 2.

Average Degree	Random Forest	SVM
$\langle k \rangle = 5$	10	12
$\langle k \rangle = 10$	18	14
$\langle k \rangle = 15$	22	25
$\langle k \rangle = 20$	25	24
$\langle k \rangle = 25$	21	17
$\langle k \rangle = 30$	21	23
$\langle k \rangle = 35$	16	17
$\langle k \rangle = 40$	16	23
$\langle k \rangle = 45$	18	25
$\langle k \rangle = 50$	16	20
All combined	26	35

Figure 14 – Fraction of books correctly assigned in the pairwise classification when mesoscopic networks were extracted from Dataset 2.

Source: Adapted from [Marinho et al. \(2017\)](#).

those 4 authors increased to 65% and 50% with Random Forests and SVM, respectively. In Figure 15, we illustrate the PCA of the books written by the selected authors.

Figure 15 – PCA considering the books of the four selected authors: Charles Darwin, Thomas Hardy, Edgar Allan Poe, and Mark Twain.

Source: [Marinho et al. \(2017\)](#).

Such a good partitioning illustrated in Figure 15 is a direct consequence of the peculiarities present in the mesoscopic networks of those books, shown in Figure 16. For example, books that contain tales or short stories present a similar chain-like topology, this is the case of the books written by Edgar Allan Poe and the book *A Changed Man and Other Tales*, from Thomas Hardy. Interestingly, that book resulted next to the ones of Edgar Allan Poe and Charles Darwin in the PCA results, rather than the other books of Thomas Hardy. Remarkably, the books written by Charles Darwin are next to the ones of Edgar Allan Poe in the PCA results. Such similarity is visually confirmed in Figure 16, in which the mesoscopic networks from his books also present a chain-like topology. This is probably a consequence of the scientific nature of Charles Darwin's works, full of theories, observations, and findings. On the other hand, the remaining books visually present more complex stories, with many intersections connecting different parts of the books.

4.2.4.3 Function word networks

As we described in Section 3.1.6, [Segarra, Eisen and Ribeiro \(2013\)](#) proposed a network representation formed only by function words. Even though their approach has proven successful to identify the authorship of texts, it is a bit demanding because many calculations have to be done, such as the entropies and the similarity measures. In order to overcome this disadvantage,

Figure 16 – Mesoscopic networks of the 20 books written by four different authors. **Charles Darwin:** (1) *Coral Reefs*, (2) *The Expression of the Emotions in Man and Animals*, (3) *Geological Observations on South America*, (4) *The Different Forms of Flowers on Plants of the Same Species*, and (5) *Volcanic Islands*. **Thomas Hardy:** (1) *A Changed Man; and Other Tales*, (2) *A Pair of Blue Eyes*, (3) *Far from the Madding Crowd*, (4) *Jude the Obscure*, and (5) *The Hand of Ethelberta*. **Edgar Allan Poe:** *The Works of Edgar Allan Poe - Volume (1) to (5)*. **Mark Twain:** (1) *Adventures of Huckleberry Finn*, (2) *The Adventures of Tom Sawyer*, (3) *The Prince and the Pauper*, (4) *A Connecticut Yankee in King Arthur's Court*, and (5) *Roughing It*. The bluish nodes represent the beginning of the book and the greenish ones represent the end of the book. The order of the nodes can be seen in the legend, where N represents the last node.

Source: [Marinho et al. \(2017\)](#).

we simplified the function word adjacency networks proposed by Segarra, Eisen and Ribeiro (2013) in order to investigate their performance for the authorship attribution task.

In our simplified version, we removed all content words from the texts and connected the remaining words (i.e. stopwords) in a directed co-occurrence fashion. As a consequence, the obtained networks have at most 127 nodes, which corresponds to the size of our stopwords list. Then, we extracted the absolute frequency of the 13 directed motifs with three nodes as explained in the Subsection 4.2.2. This decision was motivated by previous results in which stopwords played a crucial role to the success of motif-based approaches.

For this experiment, we used the texts from Dataset 3, shown in Table A.3. Because this dataset has fewer books per author – for instance, Emily Brontë and Nathaniel Hawthorne had only one book each –, the books were split into several non-overlapping partitions with L words, where L varied from 2,000 to 16,000 words. To avoid issues with imbalanced classes, we selected the same number of partitions per author. The content words were then removed from each partition and the frequencies of motifs involving 3 nodes were used as classification features. We employed four machine learning methods available in *Weka* with 10-fold cross validation. The results are presented in Table 14.

Table 14 – Percentage (%) of books correctly classified when simplified function word networks were extracted from texts of the Dataset 3.

Partition size	J48	kNN	SVM	Naïve Bayes
$L = 2,000$	27.6	29.6	32.6	32.2
$L = 3,000$	27.1	32.3	36.2	32.6
$L = 4,000$	31.1	35.5	44.8	40.0
$L = 6,000$	31.3	36.6	39.8	35.9
$L = 8,000$	33.3	34.2	41.6	33.3
$L = 12,000$	27.7	44.4	43.0	33.3
$L = 16,000$	29.6	27.7	27.7	27.7

Even though the results presented in Table 14 are not even comparable to the ones reported by Segarra, Eisen and Ribeiro (2013) and (2015), which were as high as 90%, this experiment confirms that it is still possible to achieve significant results (considerably higher than the expected chance baseline) with a simple generalization. Interestingly, these findings suggest that there is a dependency between the writing style of different authors and the way they change from one stopword to another along their texts. Moreover, given the distinct natures of the two approaches – ours uses the topology of the networks while their approach is based on Markov chains –, it is also possible to combine the two strategies.

4.3 Final Remarks

In this Chapter, we presented the methodology used and the main results achieved during this Master's research. Most of the proposed approaches are based on co-occurrence networks. In addition, we investigated alternative textual representations in authorship attribution context. The majority of the proposed techniques achieved relevant results with almost no use of NLP resources – such as taggers and parsers. As a consequence, some methods can be directly applied to many natural languages.

Taken together, the results presented in this Chapter provide enough evidence of the contribution of this Master's research in terms of methods for authorship attribution and, more generally, for the characterization of texts. Moreover, our hybrid classifier labelled motifs confirms the relevance of combining features from the topology of complex networks with traditional ones usually employed in NLP research.

In the next Chapter, we conclude this manuscript with critical reviews of our work, the main contributions and limitations of this research, as well as remarks for future work.

CONCLUSION

In this Chapter, we outline the final remarks about this Master’s research, which include a review of the results, contributions, and suggestions for future work. CN-based methods have been employed with growing success to the authorship attribution task. In recent years, several contributions have been made in this area. The state-of-the-art presented in Section 3 confirms that statistics of the measurements extracted from complex networks are able to characterize the authorship of texts and distinguish different writing styles. Despite this success, only a few works have investigated the usefulness of different representations (other than co-occurrence networks) and the appropriateness of non-traditional network measurements to characterize the networks. Furthermore, even fewer works have proposed the combination of traditional NLP techniques and networked methods. In addition, most networked techniques disregard the textual information after the networks are created.

This Master’s research extended the traditional co-occurrence model and suggested other network representations, such as mesoscopic networks, for the authorship attribution task. In addition, we unveiled the relevance of recurrent subgraphs, known as motifs, for the task. Our main contribution is the proposed hybrid classifier, called *labelled motifs*, that combines two distinct paradigms in a complementary way. In this classifier, traditional features were combined with the ones extracted from complex networks. Given the fact that most frequent words are stopwords, this hybrid classifier mainly uses the connectivity information of function words to complement the characterization achieved with the frequency of motifs. This approach was also validated in another NLP task, the translationese identification.

Different from some traditional methods that rely in many NLP resources – such as POS taggers and syntactic parsers –, the proposed techniques, in particular the hybrid classifier, do not make extensive use of such resources. As a consequence, these techniques can be easily applied to many natural languages. We are aware of the limitations of our techniques. For instance, most of our methods do not use motifs comprising more nodes and do not directly apply to texts with only a few sentences or paragraphs. In the next Section, we present our contributions,

followed by the limitations and remarks for future work. The last Section contains the list with the publications resulting from this research.

5.1 Contributions

The main contribution of this research is a hybrid classifier, called *labelled motifs*, which is the result of the combination of two distinct components. The first one, the topological component, is obtained with the frequency of motifs involving three nodes. The second component is responsible for extracting the most frequent words in the texts. Taken together, these two components extract the frequency of specific words in different subgraphs. By doing so, the disadvantages of each technique were overcome by the hybrid classification and the performance of the classifiers increased.

Most of the previous CN-based techniques for authorship attribution do not consider the textual information after the networks are created. Given that traditional techniques for the task usually achieve results as high as 90%, the textual context is a relevant information that was used to enrich the network representation in our hybrid classifier. Even though the idea of adding labels to improve the network representation has already been employed in a biological context (CHEN *et al.*, 2007), this approach was not probed in other contexts. To the best of our knowledge, this research presents the first investigation of such approach to characterize written texts.

Furthermore, we proposed some methods to extend the co-occurrence networks and also investigated the relevance of other network representations for the task, such as mesoscopic networks and named entity networks. The findings related to those alternative representations suggest that not only the four categories of stylometric features (i.e. lexical, character, syntactic, and semantic) are relevant for the authorship attribution task. For instance, the way the named entities are connected along the text captures information about its authorship. Furthermore, there is also a dependency between authors' writing styles and the story flow of their texts, which can be associated with the discourse level of the text.

We also investigated the relevance of recurrent interconnection patterns, known as *motifs*, as classification features for the task. The accuracy rates obtained with those features were relatively high considering the simplistic approach, it uses only 13 classification features. To our knowledge, we presented one of the first experiments in which network motifs were used as stylometric features to identify the authorship in texts.

Due to the smaller dependency in NLP resources, most of the techniques proposed in this research could be easily applied to other natural languages, as we already showed in the experiments for the translationese identification. Our methods may lead to new discussions and contributions for future research on authorship attribution using complex networks. Moreover, we believe that the contributions of this Master's research go beyond the authorship attribution

task because most of our methods could be applied in other related tasks, such as the analysis of text complexity, and the identification of plagiarism and stylistic inconsistencies.

5.2 Limitations and Future Work

We outline the following limitations of our research and briefly discuss some approaches that can be employed to fill these gaps:

- The optimal value of J in the *further neighborhood extension*:** As suggested by the results presented in Section 4.2.1, the best value of the parameter J may vary for different datasets. Therefore, the choice of a single and optimal parameter J is impractical. A possible solution would be to combine the features obtained with different values of J , as it was done for the average degree in Section 4.2.4.2. Another option would be to use different edge weights for distinct values of J , in which the immediate neighbors would have the strongest connections.
- Application to short texts:** Texts with only a few sentences or paragraphs represent a challenge to our techniques because the statistics extracted from those texts might not be significant. In addition, some words from the set of frequent words might be absent in those texts, which would insert a lot of zeros in the classification features for the hybrid classifier. In this scenario, the obtained co-occurrence networks are likely to present a chain-like topology, with many words occurring only once. Other approaches could be used to minimize some of these problems. One solution, as suggested by Santos *et al.* (2017), is to enrich the network with connections representing the semantic relationships among words. One technique usually employed to obtain such relationships is called *word embeddings* (MIKOLOV *et al.*, 2013a; MIKOLOV *et al.*, 2013b). For instance, the similarity between two words can be obtained as the inverse of the distance of their word embedding vectors.
- Characterization with motifs comprising more than three nodes:** Most experiments conducted during this Master's research included motifs with only 3 nodes because the computational cost to extract motifs with more nodes is high. In fact, we extracted the frequency of the 199 possible directed motifs with 4 nodes from the texts of Dataset 4. However, it would be unfeasible to extract the *labelled motifs* in those subgraphs. For instance, if we wanted to extract the *labelled motifs* for the 20 most frequent words, this would lead to almost 4,000 features. A simple solution would be to select a subset of motifs with 4 nodes, and then extract the labelled motifs in that subset.

Other remarks for future work would be the extension of the hybrid classifier to include other labels, such as the POS tags of the words. By doing so, the network would capture how the words are connected according to their categories. Assuming that there are some common

patterns, such as a determiner (e.g. an article) is likely to be followed by a noun, different connection patterns found in the texts might be related to the writing style.

5.3 Publications

The main contributions of this Master's research are reported in the following research papers:

- Marinho, V. Q., Hirst, G., Amancio, D. R. (2016). **Authorship attribution via network motifs identification**. In *Proceedings of the 5th Brazilian Conference on Intelligent Systems (BRACIS)*. Recife, Brazil.
- Marinho, V. Q., de Arruda, H. F., Lima, T. S., Costa, L. da F., Amancio, D. R. (2017). **On the "Calligraphy" of Books**. In *Proceedings of the 2017 Workshop on Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics, Vancouver, Canada.**
- Marinho, V. Q., Hirst, G., Amancio, D. R. (in press). **Labelled network subgraphs reveal stylistic subtleties in written texts**. *Journal of Complex Networks*.

In addition, the following publications were developed in collaboration with other researchers:

- Corrêa Jr, E. A., Marinho, V. Q., Santos, L. B. (2017). **NILC-USP at SemEval-2017 Task 4: A Multi-view Ensemble for Twitter Sentiment Analysis**. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval'17)*. Association for Computational Linguistics, Vancouver, Canada.
- Corrêa Jr, E. A., Marinho, V. Q., Santos, L. B., Bertaglia, T. F. C., Treviso, M. V., Brum, H. B. (2017). **PELESent: Cross-domain polarity classification using distant supervision**. In *Proceedings of the 6th Brazilian Conference on Intelligent Systems (BRACIS)*. Uberlândia, Brazil.
- de Arruda, H. F., Silva, F. N., Marinho, V. Q., Amancio, D. R., Costa, L. da F. (2017). **Representation of texts as complex networks: a mesoscopic approach**. *Journal of Complex Networks*.

** This paper was listed as one of the most provoking ones from the Physics section of arXiv released during the week ending in June 10th. See the link: <https://www.technologyreview.com/s/608057/the-best-of-the-physics-arxiv-week-ending-june-10-2017/>.

BIBLIOGRAPHY

ABBASI, A.; CHEN, H. Applying authorship analysis to extremist group web forum messages. **IEEE Intelligent Systems**, IEEE Educational Activities Department, v. 20, n. 5, 2005. Citations on pages [24](#), [25](#) e [38](#).

AHA, D. W.; KIBLER, D.; ALBERT, K., M. Instance-based learning algorithms. **Mach. Learn.**, Kluwer Academic Publishers, v. 6, n. 1, Jan. 1991. Citation on page [60](#).

AKIMUSHKIN, C.; AMANCIO, D. R.; OLIVEIRA JR., O. N. Text authorship identified using the dynamics of word co-occurrence networks. **PLOS ONE**, Public Library of Science, v. 12, n. 1, p. 1–15, 01 2017. Citations on pages [20](#), [51](#) e [53](#).

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. **Rev. Mod. Phys**, 2002. Citations on pages [23](#), [27](#) e [28](#).

ALVAREZ-LACALLE, E.; DOROW, B.; ECKMANN, J. P.; MOSES, E. Hierarchical structures induce long-range dynamical correlations in written texts. **PNAS**, v. 103, n. 21, p. 7956–7961, May 2006. Citation on page [57](#).

AMANCIO, D. R. Authorship recognition via fluctuation analysis of network topology and word intermittency. **Journal of Statistical Mechanics: Theory and Experiment**, 2015. Citations on pages [20](#), [48](#) e [53](#).

_____. A complex network approach to stylometry. **PloS One**, 2015. Citations on pages [20](#), [43](#), [49](#) e [53](#).

_____. Probing the topological properties of complex networks modeling short written texts. **PLoS ONE**, 2015. Citations on pages [20](#), [50](#) e [53](#).

AMANCIO, D. R.; ALTMANN, E. G.; OLIVEIRA JR, O. N.; COSTA, L. F. Comparing intermittency and network measurements of words and their dependence on authorship. **New Journal of Physics**, v. 13, n. 12, p. 123024, 2011. Citations on pages [20](#), [23](#), [24](#), [25](#), [34](#), [36](#), [45](#), [53](#), [57](#) e [58](#).

AMANCIO, D. R.; ALTMANN, E. G.; RYBSKI, D.; OLIVEIRA JR, O. N.; COSTA, L. F. Probing the statistical properties of unknown texts: Application to the voynich manuscript. **PLoS ONE**, Public Library of Science, v. 8, p. e67310, 07 2013. Citations on pages [25](#), [34](#), [35](#) e [43](#).

AMANCIO, D. R.; ALUISIO, S. M.; OLIVEIRA JR, O. N.; DA, L. Complex networks analysis of language complexity. **EPL (Europhysics Letters)**, v. 100, n. 5, p. 58002+, Feb. 2013. Citation on page [43](#).

AMANCIO, D. R.; NUNES, M. G. V.; OLIVEIRA JR, O. N.; COSTA, L. F. Extractive summarization using complex networks and syntactic dependency. **Physica A: Statistical Mechanics and its Applications**, v. 391, n. 4, p. 1855 – 1864, 2012. ISSN 0378-4371. Citations on pages [23](#), [43](#) e [58](#).

AMANCIO, D. R.; OLIVEIRA JR, O. N.; COSTA, L. F. Identification of literary movements using complex networks to represent texts. **New Journal of Physics**, v. 14, n. 4, p. 043029, 2012. Citations on pages 23, 36, 43 e 44.

_____. Structure-semantics interplay in complex networks and its effects on the predictability of similarity in texts. **Physica A: Statistical Mechanics and its Applications**, 2012. Citations on pages 20, 46 e 53.

AMANCIO, D. R.; SILVA, F. N.; COSTA, L. da F. Concentric network symmetry grasps authors' styles in word adjacency networks. **EPL (Europhysics Letters)**, v. 110, n. 6, 2015. Citations on pages 20, 50, 51 e 53.

ANTIQUERA, L.; PARDO, T. A. S.; NUNES, M. G. V.; OLIVEIRA JR, O. N.; COSTA, L. F. Some issues on complex networks for author characterization. In: **Fourth Workshop in Information and Human Language Technology (TIL'06) in the Proceedings of International Joint Conference IBERAMIA-SBIA-SBRN**. Ribeirão Preto, Brazil: ICMC-USP, 2006. Citations on pages 20, 23, 24, 44 e 53.

ARGAMON, S.; WHITELAW, C.; CHASE, P.; HOTA, S. R.; GARG, N.; LEVITAN, S. Stylistic text classification using functional lexical features. **Journal of the American Society for Information Science and Technology**, v. 58, n. 6, p. 802–822, 2007. Citation on page 40.

ARRUDA, H. F. de; COSTA, L. d. F.; AMANCIO, D. R. Topic segmentation via community detection in complex networks. **Chaos: An Interdisciplinary Journal of Nonlinear Science.**, v. 26, n. 6, 2016. Citations on pages 58 e 61.

ARRUDA, H. F. de; COSTA, L. F.; AMANCIO, D. R. Classifying informative and imaginative prose using complex networks. 2015. Citations on pages 23 e 58.

ARRUDA, H. F. de; SILVA, F. N.; MARINHO, V. Q.; AMANCIO, D. R.; COSTA, L. F. Representation of texts as complex networks: a mesoscopic approach. **Journal of Complex Networks**, p. cnx023, 2017. Available: <<http://doi.org/10.1093/comnet/cnx023>>. Citations on pages 13, 58, 59 e 75.

AVNER, E. A.; ORDAN, N.; WINTNER, S. Identifying translationese at the word and sub-word level. **Digital Scholarship in the Humanities**, v. 31, n. 1, p. 30–54, 2016. Citation on page 70.

BAAYEN, H.; HALTEREN, H. van; TWEEDIE, F. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. **Literary and Linguistic Computing**, v. 11, n. 3, p. 121–132, Sep. 1996. Citation on page 40.

BARABASI, A.-L. **Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life**. [S.l.]: Plume Books, 2003. Paperback. Citation on page 27.

BARABÁSI, A.-L. **Network Science**. [s.n.], 2014. Available: <<http://barabasi.com/networksciencebook/>>. Citation on page 27.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science**, v. 286, n. 5439, p. 509–512, 1999. Citations on pages 29 e 30.

BARONI, M.; BERNARDINI, S. A new approach to the study of translationese: Machine-learning the difference between original and translated text. **Literary and Linguistic Computing**, v. 21, n. 3, p. 259–274, 2006. Citation on page 70.

BECKNER, C.; BLYTHE, R.; BYBEE, J.; CHRISTIANSEN, M. H.; CROFT, W.; ELLIS, N. C.; HOLLAND, J.; KE, J.; LARSEN-FREEMAN, D.; SCHOENEMANN, T. Language is a complex adaptive system: Position paper. **Language Learning**, v. 59, 2009. Citation on page 30.

BIEMANN, C. **Structure Discovery in Natural Language**. Heidelberg: Springer, 2012. (Theory and Applications of Natural Language Processing). ISSN 2192-032X. ISBN 978-3-642-25922-7. Citation on page 31.

BIEMANN, C.; ROOS, S.; WEIHE, K. Quantifying semantics using complex network analysis. In: **International Conference on Computational Linguistics (COLING)**. [S.l.: s.n.], 2012. Citations on pages 34 e 35.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python**. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2009. ISBN 0596516495, 9780596516499. Citation on page 56.

BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D.-U. Complex networks : Structure and dynamics. **Phys. Rep.**, v. 424, n. 4-5, p. 175–308, Fervier 2006. Citations on pages 27 e 34.

BREIMAN, L. Random forests. **Mach. Learn.**, Kluwer Academic Publishers, Hingham, MA, USA, v. 45, n. 1, p. 5–32, Oct. 2001. ISSN 0885-6125. Citation on page 60.

BRENNAN, M. R.; GREENSTADT, R. Practical attacks against authorship recognition techniques. In: HAIGH, K. Z.; RYCHTYCKYJ, N. (Ed.). **IAAI**. [S.l.]: AAAI, 2009. Citation on page 39.

BURROWS, J. ‘delta’: a measure of stylistic difference and a guide to likely authorship. **Literary and Linguistic Computing**, v. 17, n. 3, p. 267–287, 2002. Citation on page 38.

BURROWS, J. F. Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style. **Literary and Linguistic Computing**, v. 2, p. 61–70, 1987. Citations on pages 24, 38 e 39.

CABATBAT, J. J. T.; MONSANTO, J. P.; TAPANG, G. A. Preserved network metrics across translated texts. **International Journal of Modern Physics C**, v. 25, n. 02, p. 1350092, 2014. Citations on pages 34 e 35.

CANCHO, R. Ferrer i; SOLÉ, R. V. Least Effort and the Origins of Scaling in the Human Language. **Proceedings of the National Academy of Science (USA)**, v. 100, 2003. Citation on page 44.

CANCHO, R. Ferrer i; SOLÉ, R. V.; KÖHLER, R. Patterns in syntactic dependency networks. **Phys. Rev. E**, American Physical Society, v. 69, p. 051915, May 2004. Citations on pages 23, 30, 43, 44 e 57.

CANCHO, R. Ferrer i; SOLÉ, R. V. The small world of human language. **Proceedings of The Royal Society of London. Series B, Biological Sciences**, v. 268, p. 2261–2266, 2001. Citations on pages 23, 30, 31, 33, 44 e 57.

CHASKI, C. E. Who’s At The Keyboard? Authorship Attribution in Digital Evidence Investigations. **International Journal of Digital Evidence**, v. 4, 2005. Citation on page 38.

CHEN, J.; HSU, W.; LEE, M. L.; NG, S.-K. Labeling network motifs in protein interactomes for protein function prediction. In: IEEE. **23rd International Conference on Data Engineering**. [S.l.], 2007. p. 546–555. Citation on page 82.

CLAUSET, A.; SHALIZI, C. R.; NEWMAN, M. E. J. Power-law distributions in empirical data. **SIAM Rev.**, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, v. 51, n. 4, p. 661–703, Nov. 2009. ISSN 0036-1445. Citations on pages 28 e 29.

COLIZZA, V.; PASTOR-SATORRAS, R.; VESPIGNANI, A. Reaction–diffusion processes and metapopulation models in heterogeneous networks. **Nature Physics**, v. 3, p. 276–282, Jan. 2007. Citation on page 28.

CONG, J.; LIU, H. Approaching human language with complex networks. **Physics of life reviews**, Elsevier, v. 11, n. 4, p. 598–618, 2014. Citations on pages 30, 43 e 58.

CONSTANTOUDIS, V.; KALIMERI, M.; DIAKONOS, F.; KARAMANOS, K.; PAPADIMITRIOU, C.; CHATZIGEORGIOU, M.; PAPAGEORGIOU, H. Long-range correlations and burstiness in written texts: universal and language-specific aspects. **International Journal of Modern Physics B**, p. 1541005, 2015. Citation on page 61.

COSTA, L. F.; OLIVEIRA JR, O. N.; TRAVIESO, G.; RODRIGUES, F. A.; Villas Boas, P. R.; ANTIGUEIRA, L.; VIANA, M. P.; ROCHA, L. E. C. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. **Advances in Physics**, v. 60, n. 3, p. 329–412, 2011. Citation on page 27.

COSTA, L. F.; RODRIGUES, F. A.; TRAVIESO, G.; BOAS, P. R. V. Characterization of complex networks: A survey of measurements. **Advances in Physics**, v. 56, n. 1, p. 167–242, January 2007. Citations on pages 28, 29, 31, 32, 33 e 41.

CSARDI, G.; NEPUSZ, T. The igraph software package for complex network research. **Inter-Journal**, Complex Systems, p. 1695, 2006. Citation on page 41.

DOROGOVTSSEV, S. N.; MENDES, J. F. F. Language as an evolving word web. **Proceedings of the Royal Society of London. Series B: Biological Sciences**, v. 268, n. 1485, p. 2603–2606, 2001. Citations on pages 30 e 31.

DUCH, J.; ARENAS, A. Community detection in complex networks using extremal optimization. **Physical Review E**, v. 72, p. 027104, 2005. Citation on page 28.

EL-FIQI, H.; PETRAKI, E.; ABBASS, H. A. A computational linguistic approach for the identification of translator stylometry using Arabic-English text. In: **IEEE International Conference on Fuzzy Systems**. [S.l.]: IEEE, 2011. p. 2039–2045. Citations on pages 34 e 35.

ELSON, D. K.; DAMES, N.; MCKEOWN, K. R. Extracting social networks from literary fiction. In: **Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics**. Stroudsburg, PA, USA: [s.n.], 2010. (ACL '10), p. 138–147. Available: <<http://dl.acm.org/citation.cfm?id=1858681.1858696>>. Citation on page 73.

ERDŐS, P.; RÉNYI, A. On random graphs i. **Publicationes Mathematicae Debrecen**, v. 6, p. 290, 1959. Citation on page 28.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. [S.l.]: LTC, 2011. Citation on page 60.

FINKEL, J. R.; MANNING, C. D. Nested named entity recognition. In: **Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1**. [S.l.: s.n.], 2009. p. 141–150. ISBN 978-1-932432-59-6. Citation on page 73.

FRANTZESKOU, G.; STAMATATOS, E.; GRITZALIS, S.; KATSIKAS, S. Effective identification of source code authors using byte-level information. In: **Proceedings of the 28th International Conference on Software Engineering**. New York, NY, USA: ACM, 2006. (ICSE '06), p. 893–896. Citations on pages 25 e 38.

GAMON, M. **Linguistic correlates of style: authorship classification with deep linguistic analysis features**. 2004. Citation on page 40.

GARCÍA, A. M.; MARTÍN, J. C. Function words in authorship attribution studies. **Literary and Linguistic Computing**, v. 22, n. 1, p. 49, 2007. Citations on pages 38, 47 e 48.

GELLERSTAM, M. Translationese in Swedish novels translated from English. In: **Translation Studies in Scandinavia**. [S.l.: s.n.], 1986. p. 88–95. Citation on page 70.

GIRVAN, M.; NEWMAN, M. E. J. Community structure in social and biological networks. **Proceedings of the National Academy of Sciences**, v. 99, n. 12, p. 7821–7826, 2002. Citation on page 28.

GRABSKA-GRADZINSKA A. KULIG, J. K. I.; DROZDZ, S. Complex network analysis of literary and scientific texts. **International Journal of Modern Physics C**, v. 23, 2012. Citation on page 43.

GRANT, T. D. Quantifying evidence for forensic authorship analysis. **International journal of speech, language and the law**, 2007. First publication by 'International Journal of Speech, Language and the Law' and Equinox. Citations on pages 24, 25 e 38.

GRIEVE, J. Quantitative authorship attribution: An evaluation of techniques. **Literary and Linguistic Computing**, v. 22, n. 3, p. 251, 2007. Citations on pages 24, 38, 39, 47 e 48.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The weka data mining software: An update. **SIGKDD Explorations Newsletter**, ACM, New York, NY, USA, v. 11, n. 1, p. 10–18, Nov. 2009. ISSN 1931-0145. Citation on page 60.

HALTEREN, H. van. Source language markers in europarl translations. In: **Proceedings of the 22nd International Conference on Computational Linguistics**. [s.n.], 2008. (COLING '08), p. 937–944. ISBN 978-1-905593-44-6. Available: <<http://dl.acm.org/citation.cfm?id=1599081.1599199>>. Citation on page 70.

HAVLIN, S. The distance between zipf plots. **Physica A: Statistical Mechanics and its Applications**, v. 216, n. 1, p. 148–150, 1995. Citation on page 38.

HIRST, G.; FEIGUINA, O. Bigrams of syntactic labels for authorship discrimination of short texts. **Literary and Linguistic Computing**, v. 22, n. 4, p. 405–417, 2007. Citation on page 40.

HOLMES, D. I. Authorship attribution. **Computers and the Humanities**, v. 28, n. 2, p. 87–106, 1994. Citation on page 37.

ILISEI, I.; INKPEN, D.; PASTOR, G. C.; MITKOV, R. Identification of translationese: A machine learning approach. In: **11th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)**. [S.l.]: Springer, 2010. v. 6008, p. 503–511. Citation on page 70.

JANKOWSKA, M.; MILIOS, E.; KESELJ, V. Author verification using common n-gram profiles of text documents. In: **Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers**. [S.l.]: Dublin City University and Association for Computational Linguistics, 2014. Citation on page [39](#).

JOLLIFFE, I. **Principal component analysis**. New York: Springer Verlag, 2002. Citation on page [64](#).

JUOLA, P. Authorship attribution. **Found. Trends Inf. Retr.**, Now Publishers Inc., Hanover, MA, USA, v. 1, n. 3, p. 233–334, Dec. 2006. ISSN 1554-0669. Citations on pages [24](#) e [38](#).

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. 1st. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2000. ISBN 0130950696. Citation on page [40](#).

KAISER, M.; HILGETAG, C. C. Edge vulnerability in neural and metabolic networks. **Biol. Cybern.**, Springer-Verlag New York, Inc., Secaucus, NJ, USA, v. 90, n. 5, p. 311–317, May 2004. ISSN 0340-1200. Citation on page [33](#).

KASHTAN, N.; ITZKOVITZ, S.; MILO, R.; ALON, U. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. **Bioinformatics**, Oxford University Press, Oxford, UK, v. 20, n. 11, p. 1746–1758, Jul. 2004. ISSN 1367-4803. Citation on page [63](#).

_____. Topological generalizations of network motifs. **Phys. Rev. E**, American Physical Society, v. 70, p. 031909, Sep 2004. Citations on pages [28](#) e [34](#).

KOEHN, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In: **Conference Proceedings: the Tenth Machine Translation Summit**. Phuket, Thailand: [s.n.], 2005. p. 79–86. Citations on pages [70](#) e [100](#).

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. **Artificial Intelligence**, v. 97, n. 1-2, p. 273–324, 1997. ISSN 0004-3702. Special issue on relevance. Citation on page [60](#).

KONG, J. S.; REZAEI, B. A.; SARSHAR, N.; ROYCHOWDHURY, V. P.; BOYKIN, P. O. Collaborative spam filtering using e-mail networks. **Computer**, IEEE Computer Society, Los Alamitos, CA, USA, v. 39, n. 8, p. 67–73, 2006. ISSN 0018-9162. Citation on page [43](#).

KOPPEL, M.; ORDAN, N. Translationese and its dialects. In: **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1**. [S.l.: s.n.], 2011. (HLT '11), p. 1318–1326. ISBN 978-1-932432-87-9. Citations on pages [25](#), [71](#) e [73](#).

KOPPEL, M.; SCHLER, J.; ARGAMON, S. Computational methods in authorship attribution. **Journal of the American Society for Information Science and Technology**, John Wiley & Sons, Inc., New York, NY, USA, v. 60, n. 1, p. 9–26, Jan. 2009. ISSN 1532-2882. Citations on pages [24](#), [38](#), [39](#), [47](#) e [48](#).

KOPPEL, M.; SCHLER, J.; MUGHAZ, D. Text categorization for authorship verification. **Eighth International Symposium on Artificial Intelligence and Mathematics**, Fort Lauderdale, Florida, 2004. Citation on page [66](#).

KRUMOV, L.; FRETTER, C.; MÜLLER-HANNEMANN, M.; WEIHE, K.; HÜTT, M. Motifs in co-authorship networks and their relation to the impact of scientific publications. **The European Physical Journal B: Condensed Matter and Complex Systems**, v. 84, n. 4, p. 535–540, 2011. Citations on pages [34](#) e [35](#).

KUMAR, E. **Natural language processing**. New Delhi: I K International, 2012. Citation on page [40](#).

KUROKAWA, D.; GOUTTE, C.; ISABELLE, P. Automatic detection of translated text and its impact on machine translation. In: **Proceedings of MT Summit XII**. [S.l.: s.n.], 2009. p. 81–88. Citation on page [73](#).

LAHIRI, S.; MIHALCEA, R. Authorship attribution using word network features. **arXiv preprint arXiv:1311.2978**, 2013. Citations on pages [20](#), [24](#), [25](#), [47](#), [53](#) e [58](#).

LARSEN-FREEMAN, D.; LYNNE, C. Complex systems and applied linguistics. v. 92, n. 4, 2008. Citation on page [30](#).

LIU, H. Statistical properties of chinese semantic networks. **Chinese Science Bulletin**, SP Science in China Press, v. 54, n. 16, p. 2781–2785, 2009. ISSN 1001-6538. Citation on page [23](#).

LUDUEÑA, G. A.; BEHZAD, M. D.; GROS, C. Exploration in free word association networks: models and experiment. **Cognitive Processing**, Springer Berlin Heidelberg, v. 15, n. 2, p. 195–200, 2014. ISSN 1612-4782. Citation on page [23](#).

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge, MA, USA: MIT Press, 1999. ISBN 0-262-13360-1. Citation on page [58](#).

MARINHO, V. Q.; ARRUDA, H. F. de; LIMA, T. S.; COSTA, L. da F.; AMANCIO, D. R. On the “calligraphy” of books. **Proceedings of the 2017 Workshop on Graph-based Methods for Natural Language Processing**, Association for Computational Linguistics, 2017. Citations on pages [58](#), [59](#), [76](#), [77](#) e [78](#).

MARINHO, V. Q.; HIRST, G.; AMANCIO, D. R. Authorship attribution via network motifs identification. In: **Proceedings of the 5th Brazilian Conference on Intelligent Systems (BRACIS)**. Recife, Brazil: [s.n.], 2016. Citations on pages [34](#) e [35](#).

_____. Labelled network subgraphs reveal stylistic subtleties in written texts. **Journal of Complex Networks**, 2017. Citations on pages [35](#) e [66](#).

MASUCCI, A. P.; KALAMPOKIS, A.; EGUÍLUZ, V. M.; HERNÁNDEZ-GARCÍA, E. Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. **PLoS ONE**, Public Library of Science, v. 6, 2011. Citation on page [43](#).

MATHIESEN, J.; YDE, P.; JENSEN, M. H. Modular networks of word correlations on twitter. **Scientific Reports**, Macmillan Publishers Limited. All rights reserved, v. 2, Nov. 2012. Citation on page [74](#).

MATTHEWS, R. A. J.; MERRIAM, T. V. N. Neural computation in stylometry i: An application to the works of shakespeare and fletcher. **Literary and Linguistic Computing**, v. 8, n. 4, p. 203–209, 1993. Citations on pages [24](#), [25](#) e [38](#).

MCCARTHY, P. M.; LEWIS, G. A.; DUFTY, D. F.; MCNAMARA, D. S. Analyzing writing styles with coh-metrix. In: SUTCLIFFE, G.; GOEBEL, R. (Ed.). **FLAIRS Conference**. [S.l.]: AAAI Press, 2006. p. 764–769. Citation on page 40.

MEHRI, A.; DAROONEH, A. H.; SHARIATI, A. The complex networks approach for authorship attribution of books. **Physica A: Statistical Mechanics and its Applications**, v. 391, n. 7, p. 2429 – 2437, 2012. Citations on pages 20, 23, 24, 25, 45, 46 e 53.

MENDENHALL, T. C. The characteristic curves of composition. **Science**, ns-9, n. 214S, p. 237–246, 1887. Citation on page 38.

MESGAR, M.; STRUBE, M. Graph-based coherence modeling for assessing readability. In: **Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, *SEM 2015, June 4-5, 2015, Denver, Colorado, USA**. [s.n.], 2015. p. 309–318. Available: <<http://aclweb.org/anthology/S/S15/S15-1036.pdf>>. Citations on pages 34 e 35.

MIHALCEA, R.; RADEV, D. **Graph-based natural language processing and information retrieval**. Cambridge; New York: Cambridge University Press, 2011. ISBN 9780521896139 0521896134. Citations on pages 23, 24 e 43.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013. Citation on page 83.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2013. p. 3111–3119. Citation on page 83.

MILLER, G. A. Wordnet: A lexical database for english. **Commun. ACM**, ACM, New York, NY, USA, v. 38, n. 11, p. 39–41, Nov. 1995. ISSN 0001-0782. Citation on page 40.

MILO, R.; ITZKOVITZ, S.; KASHTAN, N.; LEVITT, R.; SHEN-ORR, S.; AYZENSHTAT, I.; SHEFFER, M.; ALON, U. Superfamilies of evolved and designed networks. **Science**, v. 303, n. 5663, p. 1538–1542, March 2004. Citations on pages 34 e 35.

MILO, R.; SHEN-ORR, S.; ITZKOVITZ, S.; KASHTAN, N.; CHKLOVSKII, D.; ALON, U. Network motifs: simple building blocks of complex networks. **Science**, v. 298, n. 5594, p. 824–827, October 2002. Citations on pages 28, 34 e 63.

MITCHELL, T. **Machine Learning**. McGraw-Hill, 1997. (McGraw-Hill International Editions). ISBN 9780071154673. Available: <<https://books.google.com.br/books?id=EoYBngEACAAJ>>. Citation on page 60.

MOSTELLER, F.; WALLACE, D. L. **Inference and Disputed Authorship: The Federalist Papers**. Reading, Mass.: Addison-Wesley, 1964. Citations on pages 24 e 37.

MOURA, A. P. de; LAI, Y.-C.; MOTTER, A. E. Signatures of small-world and scale-free properties in large computer programs. **Physical Review E**, v. 68, 2003. Citation on page 43.

NEWMAN, M. **Networks: An Introduction**. New York, NY, USA: Oxford University Press, Inc., 2010. Citations on pages 23, 27, 28 e 34.

NEWMAN, M. E. Assortative mixing in networks. **Phys. Rev. Lett.**, v. 89, n. 20, p. 208701, 2002. Citations on pages 31 e 32.

NEWMAN, M. E. J. The structure and function of complex networks. **SIAM REVIEW**, v. 45, p. 167–256, 2003. Citations on pages 28 e 29.

PASTOR-SATORRAS, R.; VÁZQUEZ, A.; VESPIGNANI, A. Dynamical and correlation properties of the Internet. **Physical Review Letters**, APS, v. 87, n. 25, p. 258701, 2001. Citation on page 32.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. *et al.* Scikit-learn: Machine learning in python. **Journal of Machine Learning Research**, v. 12, n. Oct, p. 2825–2830, 2011. Citation on page 60.

PENG, F.; SCHUURMANS, D.; WANG, S. Augmenting naive bayes classifiers with statistical language models. **Information Retrieval**, Kluwer Academic Publishers, Hingham, MA, USA, v. 7, n. 3-4, p. 317–345, Sep. 2004. ISSN 1386-4564. Citation on page 39.

POPESCU, M. Studying translationese at the character level. In: **Recent Advances in Natural Language Processing**. [S.l.: s.n.], 2011. p. 634–639. Citation on page 70.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0. Citation on page 60.

RABINOVICH, E.; WINTNER, S. Unsupervised identification of translationese. **Transactions of the Association for Computational Linguistics**, v. 3, p. 419–432, 2015. Citation on page 70.

ROXAS, R. M.; TAPANG, G. Prose and Poetry Classification and Boundary Detection Using Word Adjacency Network Analysis. **International Journal of Modern Physics C**, v. 21, p. 503–512, 2010. Citations on pages 23 e 57.

SANDERSON, C.; GUENTER, S. Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In: **Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. (EMNLP '06), p. 482–491. Citation on page 41.

SANTOS, L. B. d.; JR, E. A. C.; JR, O. N. O.; AMANCIO, D. R.; MANSUR, L. L.; ALUÍSIO, S. M. Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics**, 2017. Citations on pages 58, 61 e 83.

SAPKOTA, U.; BETHARD, S.; MONTES-Y-GÓMEZ, M.; SOLORIO, T. Not all character n-grams are created equal: A study in authorship attribution. In: **The 2015 Conference of the North American Chapter of the Association for Computational Linguistics**. [S.l.: s.n.], 2015. p. 93–102. Citation on page 39.

SEGARRA, S.; EISEN, M.; RIBEIRO, A. Authorship attribution using function words adjacency networks. In: **2013 IEEE International Conference on Acoustics, Speech and Signal Processing**. [S.l.: s.n.], 2013. p. 5563–5567. ISSN 1520-6149. Citations on pages 20, 47, 48, 53, 73, 77 e 79.

_____. Authorship attribution through function word adjacency networks. **IEEE Transactions on Signal Processing**, IEEE, v. 63, n. 20, p. 5464–5478, 2015. Citations on pages 20, 47 e 48.

SILVA, E.; STUMPF, M. P. H. Complex networks and simple models in biology. **Journal of the Royal Society Interface**, 2005. Citation on page 35.

SILVA, F. N.; COMIN, C. H.; PERON, T. K.; RODRIGUES, F. A.; YE, C.; WILSON, R. C.; HANCOCK, E. R.; COSTA, L. da F. Concentric network symmetry. **Information Science**, Elsevier Science Inc., New York, NY, USA, v. 333, p. 61–80, 2016. ISSN 0020-0255. Citation on page 36.

SILVA, T. C.; AMANCIO, D. R. Discriminating word senses with tourist walks in complex networks. **The European Physical Journal**, 2013. Citation on page 43.

STAMATATOS, E. A survey of modern authorship attribution methods. **Journal of the American Society for Information Science and Technology**, John Wiley & Sons, Inc., New York, NY, USA, v. 60, n. 3, p. 538–556, Mar. 2009. ISSN 1532-2882. Citations on pages 24, 37, 38, 39, 40, 41, 47 e 48.

_____. On the robustness of authorship attribution based on character n-gram features. **Jornal of Law and Policy**, v. 21, p. 421, 2012. Citation on page 39.

STEIN, B.; LIPKA, N.; PRETTENHOFER, P. Intrinsic plagiarism analysis. **Lang. Resour. Eval.**, Springer-Verlag New York, Inc., Secaucus, NJ, USA, v. 45, n. 1, p. 63–82, Mar. 2011. ISSN 1574-020X. Citations on pages 25 e 38.

TRAVENÇOLO, B. A. N.; VIANA, M. P.; COSTA, L. F. Border detection in complex networks. **New Journal of Physics**, v. 11, n. 6, p. 063019, 2009. Citation on page 35.

TWEEDIE, F. J.; SINGH, S.; HOLMES, D. I. Neural network applications in stylometry: The federalist papers. **Computers and the Humanities**, v. 30, n. 1, p. 1–10, 1996. Citations on pages 25 e 38.

UZUNER, O.; KATZ, B. A comparative study of language models for book and author recognition. In: **Proceedings of the Second International Joint Conference on Natural Language Processing**. Berlin, Heidelberg: Springer-Verlag, 2005. p. 969–980. Citation on page 38.

VIANA, M. P.; BATISTA, J. L. B.; COSTA, L. F. Effective number of accessed nodes in complex networks. **Phys. Rev. E**, American Physical Society, v. 85, p. 036105, Mar 2012. Citations on pages 35 e 36.

WATTS, D.; STROGATZ, S. Collective dynamics of 'small-world' networks. **Nature**, n. 393, p. 440–442, 1998. Citations on pages 28 e 29.

WHITE, D. R.; JOY, M. S. Sentence-based natural language plagiarism detection. **J. Educ. Resour. Comput.**, ACM, New York, NY, USA, v. 4, n. 4, Dec. 2004. ISSN 1531-4278. Available: <<http://doi.acm.org/10.1145/1086339.1086341>>. Citation on page 38.

ZIPF, G. Human behaviour and the principle of least-effort. In: . Cambridge, MA: Addison-Wesley, 1949. Citation on page 44.

DATASETS USED FOR AUTHORSHIP ATTRIBUTION

The datasets used in the authorship attribution studies are listed below.

Table A.1 – Dataset 1 - List of 40 books written by 8 different authors.

Author	Books
Arthur Conan Doyle	<i>The Adventures of Sherlock Holmes</i> (1892), <i>The Tragedy of the Korosko</i> (1897), <i>The Valley of Fear</i> (1914), <i>Through the Magic Door</i> (1907), <i>Uncle Bernac - A Memory of the Empire</i> (1896).
Bram Stoker	<i>Dracula's Guest</i> (1914), <i>Lair of the White Worm</i> (1911), <i>The Jewel Of Seven Stars</i> (1903), <i>The Man</i> (1905), <i>The Mystery of the sea</i> (1902).
Charles Dickens	<i>A Tale of Two Cities</i> (1859), <i>American Notes</i> (1842), <i>Barnaby Rudge: A Tale of the Riots of Eighty</i> (1841), <i>Great Expectations</i> (1861), <i>Hard Times</i> (1854).
Edgar Allan Poe	<i>The Works of E. A. P - Volume 1-5</i> (1835).
Hector Hugh Munro (Saki)	<i>Beasts and Super Beasts</i> (1914), <i>The Chronicles of Clovis</i> (1912), <i>The Toys of Peace</i> (1919), <i>When William Came</i> (1913), <i>The Unbearable Bassington</i> (1912).
P. G. Wodehouse	<i>Girl on the Boat</i> (1920), <i>My Man Jeeves</i> (1919), <i>Something New</i> (1915), <i>The Adventures of Sally</i> (1922), <i>The Clicking of Cuthbert</i> (1922).
Thomas Hardy	<i>A Pair of Blue Eyes</i> (1873), <i>Far from the Madding Crowd</i> (1874), <i>Jude the Obscure</i> (1895), <i>Mayor Casterbridge</i> (1886), <i>The Hand of Ethelberta</i> (1875)
William Makepeace Thackeray	<i>Barry Lyndon</i> (1844), <i>The Book of Snobs</i> (1848), <i>The History of Penden-nis</i> (1848), <i>The Virginians</i> (1859), <i>Vanity Fair</i> (1848).

Table A.2 – Dataset 2 - List of 100 books written by 20 different authors.

Author: Books	Author: Books
Andrew Lang: <i>The Arabian Nights Entertainments; The Blue Fairy Book; The Pink Fairy Book; The Violet Fairy Book; The Yellow Fairy Book</i>	Herman Melville: <i>Moby Dick, Or, The Whale; The Confidence-Man: His Masquerade; The Piazza Tales; Typee: A Romance of the South Seas; White Jacket, Or, The World on a Man-of-War</i>
Arthur Conan Doyle: <i>The Tragedy of the Korosko; The Valley of Fear; The War in South Africa; Through the Magic Door; Uncle Bernac - A Memory of the Empire</i>	Horatio Alger: <i>Adrift in New York: Tom and Florence Braving the World; Brave and Bold, Or, The Fortunes of Robert Rushton; Fame and Fortune or, The Progress of Richard Hunter; Ragged Dick, Or, Street Life in New York with the Boot-Blacks; The Errand Boy, Or, How Phil Brent Won Success</i>
B. M. Bower: <i>Cabin Fever; Lonesome Land; The Long Shadow; The Lookout Man; The Trail of the White Mule</i>	Jane Austen: <i>Emma; Mansfield Park; Persuasion; Pride and Prejudice; Sense and Sensibility</i>
Bram Stoker: <i>Dracula's Guest; Lair of the White Worm; The Jewel Of Seven Stars; The Lady of the Shroud; The Man</i>	Mark Twain: <i>A Connecticut Yankee in King Arthur's Court; Adventures of Huckleberry Finn; The Adventures of Tom Sawyer; The Prince and the Pauper; Roughing It</i>
Charles Darwin: <i>Coral Reefs; Geological Observations on South America; The Different Forms of Flowers on Plants of the Same Species; The Expression of the Emotions in Man and Animals; Volcanic Islands</i>	Nathaniel Hawthorne: <i>Mosses from an Old Manse, and Other Stories; The Blithedale Romance; The House of the Seven Gables; The Scarlet Letter; Twice Told Tales</i>
Charles Dickens: <i>American Notes; A Tale of Two Cities; Barnaby Rudge: A Tale of the Riots of Eighty; Great Expectations; Hard Times</i>	P. G. Wodehouse: <i>My Man Jeeves; Tales of St. Austin's; The Adventures of Sally; The Clicking of Cuthbert; The Man with Two Left Feet</i>
Edgar Allan Poe: <i>The Works of Edgar Allan Poe (Volume 1 - 5)</i>	Richard Harding Davis: <i>Cinderella, and Other Stories; Notes of a War Correspondent; Real Soldiers of Fortune; Soldiers of Fortune; The Congo and Coasts of Africa</i>
Hector H. Munro (Saki): <i>Beasts and Super Beasts; The Chronicles of Clovis; The Toys of Peace; The Unbearable Bassington; When William Came</i>	Thomas Hardy: <i>A Changed Man and Other Tales; A Pair of Blue Eyes; Far from the Madding Crowd; Jude the Obscure; The Hand of Ethelberta</i>
Henry James: <i>The Ambassadors; The American; The Portrait of a Lady - Volume 1; The Real Thing and Other Tales; The Turn of the Screw</i>	Washington Irving: <i>Chronicle of the Conquest of Granada, from the mss. of Fray Antonio Agapida; Knickerbocker's History of New York; Tales of a Traveller; The Alhambra; The Sketch-Book of Geoffrey Crayon</i>
H. G. Wells: <i>A Short History of the World; Tales of Space and Time; The First Men in the Moon; The War of the Worlds; The World Set Free</i>	Zane Grey: <i>Riders of the Purple Sage; The Call of the Canyon; The Lone Star Ranger: A Romance of the Border; The Mysterious Rider; To the Last Man</i>

Table A.3 – Dataset 3 - List of 19 books written by 9 different authors.

Author	Books
Anne Brontë	<i>Agnes Grey</i> (1847), <i>The Tenant of Wildfell Hall</i> (1848)
Jane Austen	<i>Emma</i> (1815), <i>Mansfield Park</i> (1814), <i>Sense and Sensibility</i> (1811)
Charlotte Brontë	<i>Jane Eyre</i> (1847), <i>The Professor</i> (1857)
James Fenimore Cooper	<i>The Last of the Mohicans</i> (1826), <i>The Spy</i> (1821), <i>The Water Witch</i> (1831)
Charles Dickens	<i>Bleak House</i> (1853), <i>Dombey and Son</i> (1848), <i>Great Expectations</i> (1861)
Ralph Waldo Emerson	<i>The Conduct of Life</i> (1860), <i>English Traits</i> (1853)
Emily Brontë	<i>Wuthering Heights</i> (1847)
Nathaniel Hawthorne	<i>The House of the Seven Gables</i> (1851)
Herman Melville	<i>Moby Dick</i> (1851), <i>Redburn</i> (1849)

Table A.4 – Dataset 4 - List of 66 books written by 4 different authors split into time periods.

Author	Books
Agatha Christie	Early-period: <i>The Mysterious Affair at Styles, The Secret Adversary, The Murder of Roger Ackroyd, Murder on the Orient Express, Appointment with Death, Curtain, Towards Zero, A Murder is Announced, Destination Unknown, Ordeal by Innocence</i> Transition-period: <i>The Clocks, Endless Night</i> Late-period: <i>Nemesis, Elephants Can Remember, Postern of Fate</i>
Iris Murdoch	Early-period: <i>Under the Net, The Flight from the Enchanter, The Bell, A Severed Head, An Unofficial Rose, The Unicorn, The Italian Girl, The Time of the Angels, The Nice and the Good</i> Transition-period: <i>Bruno's Dream, A Fairly Honorable Defeat, The Black Prince, The Sacred and Profane Love Machine, Henry and Cato, The Sea, the Sea, The Philosopher's Pupil, The Good Apprentice, The Book and the Brotherhood</i> Late-period: <i>The Green Knight, Jackson's Dilemma</i>
P. D. James	Early-period: <i>Cover Her Face, A Mind to Murder, Unnatural Causes, Shroud for a Nightingale, An Unsuitable Job for a Woman, The Black Tower, Death of an Expert Witness</i> Transition-period: <i>Innocent Blood, Taste for Death, The Children of Men, A Certain Justice</i> Late-period: <i>Death in Holy Orders, The Murder Room, The Lighthouse, The Private Patient</i>
Ross Macdonald	Early-period: <i>The Moving Target, The Drowning Pool, The Way Some People Die, The Ivory Grin, Meet Me at the Morgue, Find a Victim, The Barbarous Coast, The Doomsters, The Galton Case</i> Transition-period: <i>The Wycherly Woman, The Zebra-Striped Hearse, Black Money, The Instant Enemy, The Goodbye Look</i> Late-period: <i>Sleeping Beauty, The Blue Hammer</i>

CANADIAN HANSARD AND EUROPARL

The Canadian Hansard and Europarl are the two parallel corpora used during this Master's research for the identification of translationese. They are briefly described below.

B.1 Canadian Hansard

The Canadian Hansard comprises transcripts of the debates occurring at the House of Commons of the Parliament of Canada. The debates are available online¹ in an Extensible Markup Language (XML) format. During the debates, the members are allowed to speak in the two official languages of the country, English and French. We collected 463 sessions from the 39th to 41st Parliaments, spanning the years 2006-2013. In these debates, the tag **<FloorLanguage>** has an attribute **language** that can assume the values "FR" or "EN" in order to indicate the original language of the subsequent blocks of sentences (each block is indicated with a **<ParaText>** tag). An extract of a session is presented below.

```

1  <FloorLanguage language="FR">
2    [
3      <I>Translation</I>
4    ]
5  </FloorLanguage>
6  <ParaText id="4923738">
7    Mr. Speaker, pursuant to Standing Orders 104 and 114, I
    have the honour to present, in both official languages,
    the 32nd report of the Standing Committee on Procedure and
    House Affairs regarding the membership of committees of
    the House.
```

¹ <http://www.ourcommons.ca/en/open-data#ChamberDebates>

```

8  </ParaText>
9  <ParaText id="4923740">
10   Does the hon. member have the unanimous consent of the
    House to move this motion?
11 </ParaText>
12 ...
13 <FloorLanguage language="EN">
14   [
15   <I>English</I>
16   ]
17 </FloorLanguage>
18 <ParaText id="4923744">
19   Mr. Speaker, if you seek it, I believe you will find
    consent for the following motion: That at the conclusion
    of today's debate on the opposition motion in the name of
    the member for Chilliwack-Hope, all questions necessary to
    dispose of the motion be deemed put and a recorded
    division requested and deferred until Tuesday, June 6, at
    the expiry of the time provided for oral questions.
20 </ParaText>
21 <ParaText id="4923745">
22   Does the hon. member have the unanimous consent of the
    House to propose the motion?
23 </ParaText>

```

B.2 Europarl

The Europarl ([KOEHN, 2005](#)) is a parallel corpus extracted from the Proceedings of the European Parliament. It provides versions of the debates in more than 20 European languages. We used the 5th version of the corpus. As in the Canadian Hansard, the blocks of texts are annotated with their source language. In these debates, the tag **<SPEAKER>** has an attribute **LANGUAGE** in order to indicate the original language of the subsequent blocks of sentences. In our experiments, a few sentences were discarded because they had inconsistent source language tags. An extract of the corpus is presented below.

```

1 <SPEAKER ID=14 LANGUAGE="ES" NAME="Berenguer Fuster">
2 Madam President, Mrs Díez González and I had tabled questions
   on certain opinions of the Vice-President, Mrs de Palacio

```

, which appeared in a Spanish newspaper. The competent services have not included them in the agenda on the grounds that they had been answered in a previous part-session. I would ask that they reconsider, since this is not the case. The questions answered previously referred to Mrs de Palacio's intervention, on another occasion, and not to these comments which appeared in the ABC newspaper on 18 November.

3 ...

4 <SPEAKER ID=23 LANGUAGE="DE" NAME="Poettering">

5 Madam President, I can hear a ripple of laughter from the Socialists. I was told that large sections of the Socialist Group were also keen to have this item taken off the agenda, because at the vote in the Conference of Presidents no vote was received from the working group of Members of the Socialist Group responsible for this matter. I do not know whether this information is correct, but the PPE-DE Group would, in any case, be grateful if this item were removed because Parliament has addressed this issue several times already. Decisions have also been adopted against a tax of this kind. That is why my Group moves that this item be taken off the agenda.

LIST OF STOPWORDS

The list below contains the 127 stopwords in English that can be removed during pre-processing steps.

i, me, my, myself, we, our, ours, ourselves, you, your, yours, yourself, yourselves, he, him, his, himself, she, her, hers, herself, it, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very, s, t, can, will, just, don, should, now