5-2020

On the Characterization of Natural Language Structure and Literary Stylometry - A Network Science Approach

Younis Anas Younis Al Rozz

On the Characterization of Natural Language Structure and Literary Stylometry - A Network Science Approach

by

Younis Anas Younis Al Rozz

Technical Master in
Computer Technology Engineering
Northern Technical University
2005

Bachelor of Engineering in
Computer Technology Engineering
Northern Technical University
1998

A dissertation
submitted to the College of Engineering and Science
at Florida Institute of Technology
in partial fulfillment of the requirements
for the degree of

Doctorate of Philosophy
in
Computer Engineering

Melbourne, Florida
May, 2020

We the undersigned committee hereby approve the attached dissertation

On the Characterization of Natural Language Structure and Literary Stylometry - A Network Science Approach by Younis Anas Younis Al Rozz

---

Eraldo Ribeiro, Ph.D.
Associate Professor, Computer Science
Committee Chair

Ronaldo Menezes, Ph.D.
Professor, Computer Science, University of Exeter, England
Committee Member

Susan Earles, Ph.D.
Associate Professor, Electrical Engineering
Committee Member

Veton Kepuska, Ph.D.
Associate Professor, Electrical and Computer Engineering
Committee Member

Ivica Kostanic, Ph.D.
Associate Professor, Electrical and Computer Engineering
Committee Member

William Allen, Ph.D.
Associate Professor, Computer Science
Committee Member

Philip Bernhard, Ph.D.
Associate Professor and Department Head, Computer Engineering and Sciences

ABSTRACT

Title: On the Characterization of Natural Language Structure and Literary Stylometry - A Network Science Approach

Author: Younis Anas Younis Al Rozz

Major Advisor: Eraldo Ribeiro, Ph.D.

Natural language processing (NLP) techniques have been through many advancements in recent years, linguistics and scientist utilized these techniques to solve many challenges related to written language and literary. Problems such as finding the genetic relationships among languages, attributing author of a text and categorizing text by genre have been treated throughout the years using conventional statistical methods, for instance, bag of words (BoW), N-gram, the frequency of words and the lexical distance between words. By considering written language as a complex system, network science tools and techniques can be used to address those problems. A unified methodology is proposed in this dissertation to achieve this task by (i) Propose a framework for characterizing written language as a complex system; (ii) Define three language related fields that need to be addressed by the proposed methodology; and (iii) For each field: Review related literature to get a solid background of the subject; Collect and process the data then construct the networks; Extract network measures and statistics to build the dataset; Deploy machine learning algorithms to cluster, classify the datasets; Compare and contrast results obtained with one from traditional methods.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to acknowledge the support of the Iraqi Ministry of Higher Education and Scientific research (MoHESR) for providing me this opportunity and sponsoring my Ph.D. study program in the United State. Also, the staff at the Iraqi Cultural Attache at Washington D.C. for all their help and support throughout my study.

First I would like to thank my main advisor, Dr. Ronaldo Menezes, for all the support, guidance and the knowledge that I had gained through my learning journey with him. He was not only an advisor but also a friend. Thank you for everything that you did for me.

Also, I Want to thank and express my gratitude to Dr. William Allen for advising me through the period of classwork and then became a committee member, Dr. Philip Bernhard and Dr. Eraldo Ribeiro for advising me after Dr. Menezes and provide all the help and support.

I would like also to thank my committee members, Dr. Susan Earles, Dr. Veton Kepuska, and Dr. Ivica Kostanic for their kind in providing the comments, ideas, and evaluation of my work.

I also want to thank the members of the Biocomplex Lab for all their help and support and the great times that we spent together.

I am also grateful to my father, Anas Al Rozz for his support and advise through all the stages of my life, and my mother, for her care and tenderness.

Last but not least, I would like to thank my wife, my two sons and my beloved daughter for their patient, support and understanding.

Chapter 1

Introduction

The development of society (civilization) cannot be said to be caused by the advent of writing but writing is certainly linked to modern life as it only appeared around 5,000 years ago. According to Coulmas [31], writing is the most important ”sign system” ever invented. It is quite difficult to imagine our society without writing given its ubiquity in many aspects of our lives: books, research articles, instruction manuals, lecture notes, etc. The importance of writing is even recognized by many cultures and often its invention is attributed to divine intervention such as god Ganesh in India, or the god Thoth in ancient Egypt.

Nowadays, it is quite common to have data regarding any subject of interest. In the context of text analysis, the studies range from discovering language structure [84, 53], classification of languages into families [59, 58, 11, 39, 28], word tagging problems [22], machine translation [5], summarization systems [7], to the improvement of search engines and information retrieval (IR) [82].

A deeper understanding of structural language similarities can lead to metrics to evaluate the quality of writing, translations, classification of literary styles and recognition of text genre. It is quite possible that different styles present different writing structures. An author's writing style can be considered as an example of a behavioral biometric. The words used by people and the way they structure their sentences is unique, and can frequently be used to identify the author of a certain work. The task of author attribution gained attention among researchers in the fields of statistical physics, natural language processing, and data and information science[86]. Applications of authorship attribution are not only limited to literary stylometry [9] but also expands to other fields such as social media forensic [75] and email fraud detection [24]. As researchers find complex networks a promising field in linguistic studies [3], more and more authorship attribution works based on text networks saw the light of day. Measurement from word co-occurrence network topology combined with traditional statistical methods like frequency of functional words and intermittency were used to attribute authors [6, 2]. Despite the achievements obtained from the all previous works, no agreement was accepted on how to process the text corpora, which parts of the written language and network statistics well characterize the problem under study. In this work, we attempt to address these aspects by proposing a methodology (Figure 1.1) that can be applied to different problems in the field of natural language processing and analysis. The proposed methodology does not mean to replace existing traditional methods, rather it helps to expand our understanding of natural language and written text.

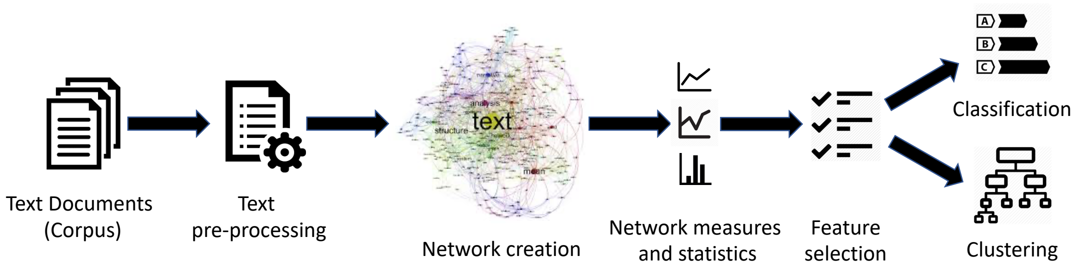


Figure 1.1: The proposed framework methodology followed throughout this dissertation.

Problem statement

Natural language processing (NLP) techniques have been through many advancements in recent years, linguistics and scientist utilized these techniques to solve many challenges related to written language and literacy. Problems such as finding the genetic relationships among languages, attributing author of a text and categorizing text by genre have been treated throughout the years using conventional statistical methods, for instance, bag of words (BoW), N-gram, the frequency of words and the lexical distance between words. By considering written language as a complex system, network science tools and techniques can be used to address those problems. A unified methodology is needed to achieve this task.

The steps taken to verify this work are:

1. Propose a framework for characterizing written language as a complex system;
2. Define three language related fields that need to be addressed by the proposed methodology;

3. For each field:

    (a) Review related literature to get a solid background of the subject;

    (b) Collect and process that data then construct the networks;

    (c) Extract network measures and statistics to build the dataset;

    (d) Deploy machine learning algorithms to cluster, classify the datasets;

    (e) Compare and contrast results obtained with one from traditional methods.

The dissertation is organized as follows:

- Chapter 2 - Introduces the concepts and the state-of-the-art research work of representing text as a complex network and its application in written language structure, classification, the attribution of author style and text categorization by genre.

- Chapter 3 - Describes the data sets used in this dissertation, preprocessing steps, and limitations, as well as how the networks created and the measures extracted. It shows that the frequency of functional words enhanced author attribution task and the use of sentence boundary in author attribution work can reflect the author's style of sentence length and reduces the density of networks.

- Chapter 4 - Characterizes language classification using features extracted from corpora of the text of 10 languages from three main families, the use of Heaps' law and structural properties of networks created from word co-occurrence of the text to find features for these languages. Next, it revealed the structure of 20 Indo-European languages belonging to three Sub-Families (Romance, Germanic, and Slavic) from a chronological perspective. It Unveils the importance of functional words in language clustering, the important features for revealing language structure, find the best language family by cluster entanglement, and observe the impact of certain language removal on the stability of the cluster.

- Chapter 5 - Describes the work on attributing authors using network motifs extracted from their text. It also describes and examines the statistical learning algorithms used to classify the motif data set. It shows the importance of sentence boundary in co-occurrence networks and the use of directed 4-node network motif in achieving a higher accuracy rate compared to a similar methodology.

- Chapter 6 - Instantiates the proposed methodology in revealing the semantic similarities among different texts to categorize them based on genre.

- Chapter 7 - Summarizes the main contributions in this work and compares the results obtained with other works. Also, point the directions for further research.

During the work on this dissertation, we published the following research papers directly related to the subject of the dissertation:

1. Characterization of Written Languages Using Structural Features from Common Corpora [3].

2. Complex Networks Reveal a Glottochronological Classification of Natural Languages [47].

3. Author Attribution Using Network Motifs [4].

In addition, we published the following work that is not closely related to the subject of this dissertation, but related to network and data science:

- On the Performance of Network Science Metrics as Long-Term Investment Strategies in Stock Markets [55].

Chapter 2

Background

This chapter provides a background for the subject of the dissertation and review for the literature related to the subject, emphasizing their contribution, gaps and how we will try to fill some of these gaps. In section 2.1 we will have a closer look at the works we found closely related to our study of natural language structure. This section consists of two parts, the first one is a review of works related to the characterization of natural language based on its network measures and the second one for works related to natural language families relation or glottochronological classification of natural languages. The literature on author attribution will be reviewed in section 2.2, and finally,in section 2.3 we will review works related to text classification and categorization based on its type or genre using textual and network features.

2.1 Characterization of Written Languages

Many researchers have investigated the possibility of using statistical and mathematical modeling to understand regularities in written languages. Chouldhury and Mukherjee [26] work discuss many ways in which networks can be created from a text but they all fall into two main categories: lexical networks and word co-occurrence networks. The first category is concerned with cognitive systems and Psycho-linguistics studies [12] and can be further classified into phonological [8], semantic [88], and orthographic networks [27]. Phonological networks can be a network of phonemes (the human speech sounds) [81] or a syllables network [83]. The second type of language network can be further categorized into co-location [62] and syntactic dependency networks [49].

The attempt to use language structure as a classification is not new. In fact, Song [85] discussed the concept of *linguistic typology* as a field which looks at the comparison of languages (search for similarities and differences) across all levels of language structure such as syntax, semantics, morphology, and phonology. Three types of linguistic typology exist [15]: qualitative, quantitative, and theoretical typology.

Liu and Xu constructed word and lemma (which is a basic form of any word or root of the word, or the headword of a dictionary entry) form dependency syntactic networks for 15 languages. They utilized and analyzed seven network parameters to classify those languages and found that word-formed networks are better than lemma networks in classifying languages [59].

Liu and Cong [58] created co-occurrence networks from a text in 14 different languages and used complex network parameters for the classification of those languages using hierarchical clustering. Ban *et al.* [11] built a co-occurrence network using text from five books for three languages and used network measures to find the similarity and differences between those three languages. Gao *et al.* [39] constructed six directed and weighted word co-occurrence networks based on 100 reports from the United Nations. Then they compared the network measures; They did not perform any clustering.

Glottochronological Classification of Natural Languages

The use of a cognate set of words to study the time of language divergence is not new. Gray *et al.* [42] studied the time separation between 87 Indo-European languages from a dataset of 2,449 cognate sets coded as discrete binary characters. They applied the likelihood models of lexical evolution to solve the problem of accuracy of tree topology and branch length estimation. Bayesian inference of phylogeny was used to enhance the estimation of tree topology and branch lengths. Also, they used rate-smoothing algorithms to reduce the rate variation across the tree. Last, they tried to examine subsets of languages using split decomposition, the result showed a strong identity for the tree when comparing a subset result with complete one. To calculate the divergence times, they sampled the trees in balance to their posterior probability, which is giving a measure in the tree topology and branch length. Then, by estimating the divergence times using the Markov Chain Monte Carlo (MCMC) sample distribution of trees, to calculate the variability in the age estimates, and hence calculate a confidence interval for the age of any node. Finally, they used penalized-likelihood rate smoothing to calculate divergence times between languages. They found the results are agreement with the Anatolian theory for the origin time of Indo-European languages.

Although several studies have been done in the history of languages and how they derived from each other, there is no unanimity on the origin of human languages because of the lack of direct evidence and empirical data [19]. Due to the difficulty to determine the specific date of language separation, scholars try to study the relationship between languages and convert the result into an estimate for when a pair diverged. However, the calculation of the distance between a pair of language is one of the most efficient methods to use it for chronological estimation. Linguistic distance—how different one language or dialect is from another [73]—can be computed by the lexical distance of the language vocabulary [43].

There are several distance measure algorithms that can be applied to a text like Hamming distance, Levenshtein distance, and Jaro-Winkler distance [90]. Levenshtein is commonly used and it is a metric for measuring the difference between two string sequences. The Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. Serva and Petroni [80] used a modification of the Levenshtein distance and normalized it by the number of longer characters of the two words, which is reasonable if two words differ by one character this is much more important for short words than it is for long words. To calculate the time distance between languages, the lexical distance between words tends to grow due to random mutations and on the contrary, can be reduced since different words may become more similar through language borrowings. To get around this, they used the following simple differential equation:

$$\bar{D} = -\alpha(1 - D) - \beta D \tag{2.1}$$

Where $\bar{D}$ is the time derivative of D, α is related to the increasing of D, β is the possibility that two words become more similar by a random mutation or by word borrowing from another language. The authors assume at the initial condition which is the time $t = 0$ two languages start to break away and the lexical distance D is zero. With this condition, equation 2.1 can be solved. The result is a relation which gives the separation time t between two languages in terms of their lexical distance D:

$$T = -\epsilon ln(1 - \gamma D) \tag{2.2}$$

Where ϵ is equal to $1/(\alpha + \beta)$ and $\gamma = (\alpha + \beta/\alpha)$ and can be solved experimentally by considering two pairs of languages whose time distance is known. Finally, using the unweighted pair group method average (UPGMA), the phylogenetic trees are constructed from the matrix. They found that the result from the method above is relatively similar to those found by glottochronologists. On the other hand, several studies have been done for the classification of languages using text characteristics without looking to the time divergence [59, 58, 3].

2.2 Author Attribution

Several studies exist that deal with the importance of network motifs in natural language networks. The first attempt to classify different networks including word co-occurrence using network motifs was made by Milo *et al.* [65]. Li *et al.* [56] extracted and studied three and four nodes directed motif structure of 72,923 two-character Chinese words network. They found that feed-forward loop (FFL) motif structure is significant in their network. Rizvić *et al.* [74] examined three nodes (triads) network motifs extracted from directed co-occurrence networks of five Croatian texts and compared their results with other languages. They realized that there is a similarity between the Croatian language networks triad significance profiles and other previously studied languages. Cabatbat *et al.* [23] compared five-nodes network motifs among other network measures of the Bible and the Universal Declaration of Human Rights (UDHR) translations in eight languages. Pearson correlation coefficient and mutual information were used to compare the metrics of real texts with random texts from other sources. Their finding is that the distribution of network motif frequency is beneficial in recognizing similar texts. Biemann *et al.* [16] realized that motif signatures serve to discriminate co-occurrence networks of natural language from artificially generated ones. To assist their finding, they present additional results on peer-to-peer streaming, co-authorship, and mailing networks. The directed motif of size 3 and undirected motif of size 4 was used in their work.

All the previous works showed the ability of various size network motifs of discriminating text from different languages and genre. They did not utilize machine learning algorithms to support their findings. On the other hand, Marinho *et al.* [63] achieved 57.5% accuracy in their best scenario of attributing eight authors of 40 novels with three nodes directed network motifs. An important aspect of author attribution task is the feature frequency [86]. To capture an author style more preciously, the feature should be more frequent. This motivates us to use the frequency of the 199 four nodes directed network motif in an attempt to attribute the authors under study.

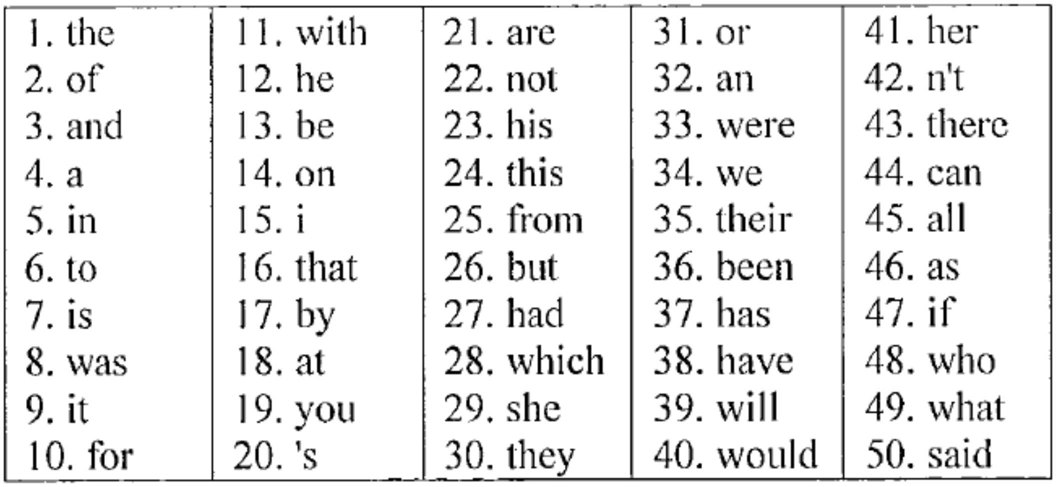
2.3 Classification of Text Genres

Several definitions for genre existed in the literature, we found that the most accurate one was that given by Finn and Kushmerick [37]. According to their definition, text genre is a type of documents that results directly from the examination of the language style and text employed in the document set. Not that they only define text genre as a reflection of a document style, but they also distinguished it from the document topic. Hence, documents from a certain genre can be related to different topics and vice versa.

Related works based on textual features

The work of Biber [13, 34] was the first attempt to utilize quantitative approaches in identifying the genre (types) in written/spoken text in English. A more related work by Karlgren and Cutting [52] was applied to the Brown corpus to categorize texts into genre classes using discriminant analysis. Most of the text features they extracted are structural cues that require tagging or parsing which considered as a limitation in automatic text classification. Kessler *et al.* [54] also used the Brown Corpus to classify the text into three main categories: Narrative, Genre, and Brow. The Narrative is a binary facet indicating whether a text is narrative or not, while Genre levels are: Fiction, Nonfiction, Editorial, Legal, Scitech, and Reportage. Finally, Brow classified into Middle, Upper-Middle, High, and Popular. They extracted 55 features based on Character-Level, Lexical, and Derivative generic cues which require less computation than structural cues used in [52]. Logistic Regression (LR) and Neural Networks (simple perceptron and multi-layer perceptron) were used for the classification tasks. They found that their method can detect some facet levels more accurately than the others, for instance in Genre facet, Fiction and Reportage detection accuracy is higher than that in Legal and Editorial. They justify these differences as the poor representation of those Genre levels in Brown corpus and as a result in their training set.

A more simple approach was used in [87] to classify the *Wall Street Journal* (WSJ) corpus into four categories: Reportage, Editorials, Spot news, and Letters to the editor. They utilized the frequency of the most frequent words (Figure 2.1) and punctuation marks as a feature vector with *Discriminant analysis* as the classification method. They conclude that their method was capable of handling an unrestricted text, the computational cost was minimum, and it is not bound to restricted specifications of specific domain/language.



1. the	11. with	21. are	31. or	41. her
2. of	12. he	22. not	32. an	42. n't
3. and	13. be	23. his	33. were	43. there
4. a	14. on	24. this	34. we	44. can
5. in	15. i	25. from	35. their	45. all
6. to	16. that	26. but	36. been	46. as
7. is	17. by	27. had	37. has	47. if
8. was	18. at	28. which	38. have	48. who
9. it	19. you	29. she	39. will	49. what
10. for	20. 's	30. they	40. would	50. said

Figure 2.1: Fifteen most frequent words used by Stamatatos *et al.* [87].

Related works based on network features

Most of the works on genre detection described in the previous paragraph were based on textual features that reflect the lexical and syntactic aspects of the text. Features such as the bag of words (BoW), part of speech (POS) tagging, frequency of functional words are not capable of capturing the semantic and structural aspects of the text which has been proved to enrich the classification process [50].

The graph-based approach was used in text classification and categorization problems [76, 61] and approved to be more accurate and outperformed the methods based on textual features. More related work to genre classification based on six graph features extracted from the Brown corpus [68]. A Word bigrams network was constructed for the eight genres in the corpus and then the six network measures were compared and the differences were explained, however, no classification algorithm was used.

Chapter 3

Datasets Collection and Curation

Chapter 3 is a description of the data used throughout this work and how it was curated and pre-processed. In Section 3.1 we describe the data used for revealing the structure and family relationships of natural language through network analysis. Next the data for author attribution using network-based metrics is explained in section 3.2. Finally, section 3.3 describes the steps taken in collecting and processing the data for classifying literary genre.

3.1 Data for Language Structure and Characterization

The data for characterizing written language was collected from the Leipzig corpora collection [41]. The languages chosen were English (Eng), Arabic (Ara), Russian (Rus), Italian (Ita), Spanish (Spa), French (Fra), German (Ger), Turkish (Tur), Dutch (Dut), and Danish (Dan). These languages were chosen to represent three main language families, namely Afro-Asiatic, Indo-European, and Turkic. The text corpus for each language was constructed from Wikipedia and news pages to ensure some vocabulary diversity and a good representation for each specified language. The size of the corpus for each language is consistently made of one million sentences. The entire text was converted to lower case, then punctuation and special characters were removed. This work looks at language structure in *meaningful* words and sequences of such words which means that stop words (e.g. prepositions, articles, etc.) were removed from the text. These so-called functional words can skew the statistical representation of the words in particular in the context of network science (described later).

The languages chosen for revealing the family relationships among natural languages were Romanian (Ron), French (Fra), Catalan (Cat), Italian (Ita), Spanish (Spa), Portuguese (Por), German (Ger), Dutch (Dut), Danish (Dan), Norwegian (Nor), Swedish (Swe), English (Eng), Slovenian (Slv), Bulgarian (Bul), Polish (Pol), Russian (Rus), Ukrainian (Ukr), Croatian (Cro), Czech (Ces), and Slovak (Slk). These languages give good representations of three large sub-families of the Indo-European family, which are Italic, Germanic, and Slavic. The text corpus for each language was also constructed from Wikipedia and news pages. After the entire text was converted to lower case, and the punctuation and special characters were removed, we used 100,000 words from each corpus for this work.

The second type of data we used relates to the languages tree topology, branches length, and divergence period between languages (year the languages separate), which we reconstructed from several works [43, 42, 80] in linguistics. This hierarchy was done for the 20 languages we deal with in this work and is used as the ground truth (see Figure 3.1).

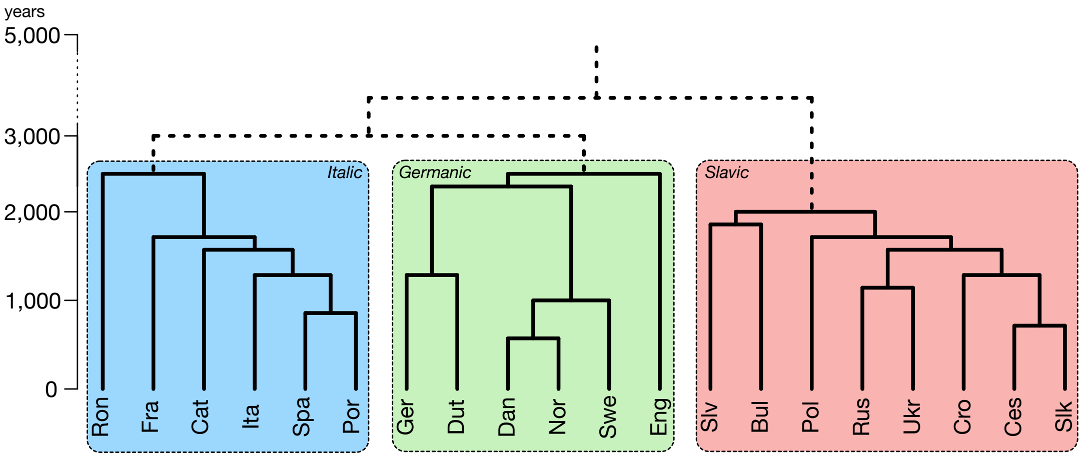


Figure 3.1: A dated phylogenetic tree of 20 Indo-European languages with three sub-families Italic, Germanic, and Slavic. The dates on the y-axis are approximations for when these languages split from a common language.

3.2 Data for Author Attribution

The dataset used in this part of the work comprised of 100 literature books authored by 10 different authors; 10 books for each individual author. The books are listed in Table 3.1, and were collected from the Project Gutenberg website[1]. Each book was limited to 20 thousand words which is the length on the shortest book in the set. Text pre-processing steps were applied to remove punctuation, numbers and non-Latin alphabets, and all letters were converted to lowercase. We preserved functional words (stop words) in the text as their frequency has been proven to reflect stylistic aspects of the text and improve authorship attribution task [67, 14, 79]. A sample text from Charles Dickens's "A Christmas Carol" novel and the resulted pre-processed text are shown in (Figure 3.2(a) and (b) respec-

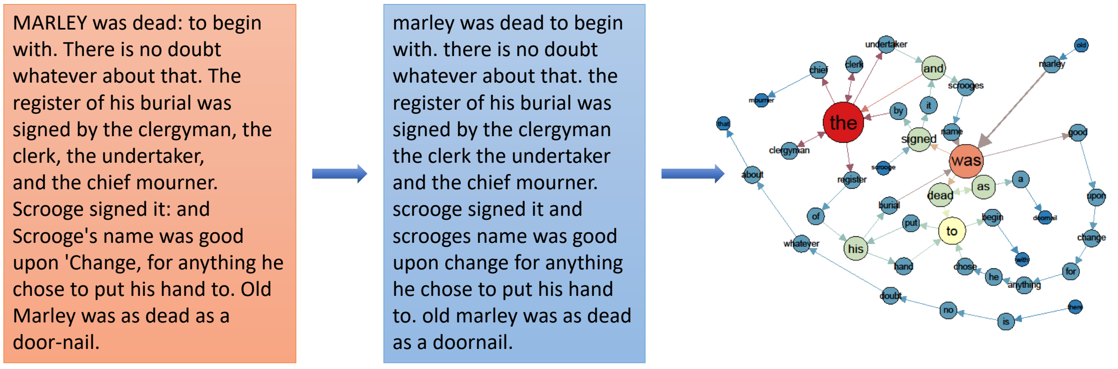
[1] http://www.gutenberg.org

tively) to illustrate this process.

Table 3.1: Authors used in our experiments and their book titles.

Authors	Book Titles
Bernard Shaw 1856-1950	Man and Superman, Candida, Arms and the Man,The Philanderer, Caesar and Cleopatra, Pygmalion, Major Barbara, Heartbreak House, The Devil's Disciple, Cashel Byron's Profession.
Charles Dickens 1812-1870	A Christmas Carol, A Tale of Two Cities, The Pickwick Papers, Oliver Twist, Great Expectations, David Copperfield, Little Dorrit, Our Mutual Friend, The Life and Adventures of Nicholas Nickleby, Dombey And Son.
George Eliot 1819-1880	The Essays of George Eliot, Impressions of Theophrastus Such, Silas Marner, Scenes of Clerical Life, The Mill on the Floss, Adam Bede, Romola, Daniel Deronda, Felix Holt The Radical, Middlemarch.
Herbert George Wells 1866-1946	Tales of Space and Time, The Food of the Gods and How It Came to Earth, The Country of the Blind, And Other Stories, The Invisible Man, The First Men in The Moon, The Island of Doctor Moreau, The War of the Worlds, The Time Machine, In the Days of the Comet, Ann Veronica.
Jack London 1876-1916	The Call of the Wild, White Fang, The Iron Heel, Before Adam, Martin Eden, The People of the Abyss, The Night-Born, The Sea Wolf, South Sea Tales, The Valley of the Moon.
Mark Twain 1835-1910	The Adventures of Tom Sawyer, Adventures of Huckleberry Finn, Life on The Mississippi, The Mysterious Stranger and Other Stories, A Tramp Abroad, Following the Equator, The Innocents Abroad, Roughing It, The Prince and The Pauper, A Connecticut Yankee in King Arthur's Court.
Oscar Wilde 1854-1900	A House of Pomegranates, The Duchess of Padua, Vera, Lady Windermere's Fan, A Woman of No Importance, Intentions, An Ideal Husband, Lord Arthur Savile's Crime and Other Stories, The Importance of Being Earnest, The Picture of Dorian Gray.
Sir Arthur Conan Doyle 1859-1930	Rodney Stone, The Adventures of Sherlock Holmes, A Duet, The Tragedy of The Korosko, The Refugees, Uncle Bernac, The Valley of Fear, The Hound of the Baskervilles, Sir Nigel, The Lost World.
William Henry Giles Kingston 1814-1880	Hendricks the Hunter, The Three Lieutenants, The Three Midshipmen, The Three Commanders, Peter the Whaler, Ben Burton, The Three Admirals, Adventures in Africa, In the Wilds of Florida, Peter Trawl.
William Shakespeare 1564-1616	Hamlet, Prince of Denmark, The Life of Henry the Fifth, The Merchant of Venice, The Tragedy of Antony And Cleopatra, The Tragedy of Coriolanus,The Tragedy of Julius Caesar, The Tragedy of King Lear, The Tragedy of Othello, Moor of Venice, The Tragedy of Romeo And Juliet, The Winter's Tale.

Next, we created the directed co-occurrence networks from the result of the pre-processed text of the 100 books. Co-occurrence networks can be constructed based on the sentence, paragraph, or the whole text boundary. We chose the sentence boundary as it produces less dense network hence, reduces the amount of time required to extract network motifs. Sentence boundary is defined by period, exclamation point, and question mark [71]. The network constructed from the pre-processed text is depicted in (Figure 3.2(c)).



MARLEY was dead: to begin with. There is no doubt whatever about that. The register of his burial was signed by the clergyman, the clerk, the undertaker, and the chief mourner. Scrooge signed it: and Scrooge's name was good upon 'Change, for anything he chose to put his hand to. Old Marley was as dead as a door-nail.

marley was dead to begin with. there is no doubt whatever about that. the register of his burial was signed by the clergyman the clerk the undertaker and the chief mourner. scrooge signed it and scrooges name was good upon change for anything he chose to put his hand to. old marley was as dead as a doornail.

(a) Original Text

(b) Pre-processed Text

(c) Resulted Network

Figure 3.2: Sample text from Charles Dickens's "A Christmas Carol" novel showing the stages of text pre-processing and the co-occurrence network created from the text. (a) Original text. (b) Resulted text from the pre-processing stage which removed punctuation, numbers and non-Latin alphabets, converting letters to lowercase. Functional words (stop words) and end-of-sentence punctuation were preserved as explained in the text. The word 'Change is a shortened form of Exchange (the stock exchange).(c) Directed co-occurrence networks for the sample text, where each word in the text represented as a node in the network and the direction of the edges reflects the co-occurrence of words.

3.3 Data for Text Genre Classification

As we saw in section 2.3 of chapter two, most of the works on genre classification were done using well-known corpora, such as the Brown corpus and corpus from newspaper articles that do not reflect the true meaning of genre. For this work, by genre, we refer to the literary genre or the fiction genre where the events and characters are invented by the author of the work and describes the style of this artistic work and the particularized target of readers. On the other hand, the non-fiction genre describes factual events and have sub-genres such as Textbook, Biography, Speech, Journalism to name a few [36].

Based on that, we decided to build our dataset from the Project Gutenberg online library as it contains most of the well-known fiction sub-genres. The sub-genres we chose are Adventure, Detective, Fantasy, Gothic, Historical, Horror, Humor, Mystery, Science Fiction, and Western. Ten books are assigned for each sub-genre with a total of 100 books for the complete dataset that is listed in table 3.2 which shows the associated book titles and their authors for each sub-genre. The book's titles were carefully chosen according to the following criteria:

1. Book titles should closely related to the designated sub-genre as much as possible and if the book is related to more than one sub-genre, the designated sub-genre should be first mentioned in the book metadata.

2. To eliminate the bias of author style on classification accuracy, authors should not have more than one book title in the same sub-genre.

3. Comprehensive reviews were conducted on the author's biographies and bibliographies to make sure the author and the chosen book title are well-known for the designated sub-genre.

The same steps in section 3.2 were taken to curate the text data and create co-occurrence networks. For the text classification part, we test the algorithms with two scenarios, first, we kept functional words and didn't stem the words and then we repeated the experiments by removing the functional words and stemming the text. The classification accuracy was better in the second scenario, so we decided to go in that route for curating our text classification data.

Table 3.2: Book titles and author names for the genre used in the work.

Genre	Book Titles	Author
Adventure	Around the World in 80 Days	Jules Verne
	Bones	Edgar Wallace
	Kim	Rudyard Kipling
	King Solomon's Mines	H. Rider Haggard
	Moby Dick	Herman Melville
	Tarzan of the Apes	Edgar Rice Burroughs
	The Black Arrow	Robert Louis Stevenson
	The Call Of The Wild	Jack London
	The Count of Monte Cristo	Alexandre Dumas
	The Prisoner Of Zenda	Anthony Hope
Detective	Cleek/ the Man of the Forty Faces	Thomas W. Hanshew
	Dead Men's Money	J. S. Fletcher
	The Adventures of Sherlock Holmes	Arthur Conan Doyle
	The Clue of the Twisted Candle	Edgar Wallace
	The Gloved Hand	Burton E. Stevenson
	The Moon Rock	Arthur J. Rees
	The Mysterious Affair at Styles	Agatha Christie
	The Mystery of the Hasty Arrow	Anna Katharine Green
	The Red House Mystery	A. A. Milne
	The Works of Edgar Allan Poe Volume 1	Edgar Allan Poe
Fantasy	Don Rodriguez: Chronicles of Shadow Valley	Edward Plunkett
	Gulliver of Mars	Edwin L. Arnold
	Irish Fairy Tales	James Stephens
	Jurgen: A Comedy of Justice	James Branch Cabell
	The House on the Borderland	William Hope Hodgson
	The Legends Of King Arthur And His Knights	James Knowles
	The Merry Adventures of Robin Hood	Howard Pyle
	The Wallet of Kai Lung	Ernest Bramah
	The Well at the World's End	William Morris
	The Wonderful Wizard of Oz	L. Frank Baum
Gothic	Caleb Williams: Things As They Are	William Godwin
	Carmilla	J. Sheridan LeFanu
	Northanger Abbey	Jane Austen
	The Castle of Otranto	Horace Walpole
	The History of the Caliph Vathek	William Beckford
	The Master of Ballantrae	Robert Louis Stevenson
	The Monk	M. G. Lewis
	The Talisman	Sir Walter Scott
	Wieland	Charles Brockden Brown
	Wuthering Heights	Emily Bronte
Historical	Ben-Hur: A Tale of the Christ	Lew Wallace
	Rob Roy	Sir Walter Scott
	The Black Arrow	Robert Louis Stevenson
	The Last Days of Pompeii	Edward George Bulwer-Lytton
	The Prince and The Pauper	Mark Twain
	The Refugees	Arthur Conan Doyle
	The Three Musketeers	Alexandre Dumas
	Vittoria	George Meredith
	War and Peace	Leo Tolstoy
	Wulf the Saxon	G. A. Henty

Table 3.2: Continued.

Genre	Book Titles	Author
Horror	Dr. Jekyll And Mr. Hyde	Robert Louis Stevenson
	Dracula	Bram Stoker
	Frankenstein	Mary Wollstonecraft Shelley
	The Damned	Algernon Blackwood
	The Great God Pan	Arthur Machen
	The House of the Vampire	George Sylvester Viereck
	The King in Yellow	Robert W. Chambers
	The Sorcery Club	Elliott O'Donnell
	Varney the Vampire	Thomas Preskett Prest
	Widdershins	Oliver Onions
Humor	Adventures of Bindle	Herbert George Jenkins
	Baboo Jabberjee, B.A.	F. Anstey
	Further Foolishness	Stephen Leacock
	Mr. Dooley in Peace and in War	Finley Peter Dunne
	My Man Jeeves	P. G. Wodehouse
	Once a Week	Alan Alexander Milne
	Queen Lucia	E. F. Benson
	The Diary of a Nobody	George Grossmith
	The Idiot	John Kendrick Bangs
	Three Men in a Boat	Jerome K. Jerome
Mystery	After Dark	Wilkie Collins
	Alice	Edward Bulwer-Lytton
	K	Mary Roberts Rinehart
	The Abandoned Room	Wadsworth Camp
	The Mysteries of London	George W. M. Reynolds
	The Mysterious Key And What It Opened	Louisa May Alcott
	The Mystery of Cloomber	Arthur Conan Doyle
	The Mystery of the Sea	Bram Stoker
	The Return of Dr. Fu-Manchu	Sax Rohmer
	The Wrong Box	Robert Louis Stevenson
Science Fiction	In Search of the Unknown	Robert W. Chambers
	Islands of Space	John W Campbell
	Masters of Space	Edward Elmer Smith
	People Minus X	Raymond Zinke Gallun
	Stand by for Mars!	Carey Rockwell
	The Blind Spot	Austin Hall
	The Time Machine	H. G. Wells
	Twenty Thousand Leagues under the Sea	Jules Verne
	Two Thousand Miles Below	Charles Willard Diffin
	Warlord of Mars	Edgar Rice Burroughs
Western	A Deal in Wheat	Frank Norris
	Bucky O'Connor	William MacLeod Raine
	Bull Hunter	Max Brand
	Heart of the Sunset	Rex Beach
	Ride Proud, Rebel!	Andre Alice Norton
	The Happy Family	Bertha Muzzy Bower
	The Hidden Children	Robert W. Chambers
	The Mysterious Rider	Zane Grey
	The Outlet	Andy Adams
	The Virginian	Owen Wister

Chapter 4

On the Structure of Languages

The first section of this chapter focus on regularities of 10 languages from Afro-Asiatic, Indo-European, and Turkic families. In order to find features for these languages we use *(1)* Heaps' law, which models the number of distinct words in a corpora as a function of the total number of words in the same corpora, and *(2)* structural properties of networks created from word co-occurrence in large corpora of 10 written languages. Using clustering approaches we show that despite differences from years of being used separately, the cluster of languages still seem to respect the organization based on historical families.

In the next section we constraint on revealing the structure of 20 Indo-European languages belonging to three Sub-Families (Romance, Germanic, and Slavic) from a chronological perspective. We show that even without lexical distance analysis or word-pair relations, and focusing merely on the structure built from syntax, we can detect useful structure of language families. More specifically, we used statistical measures of a word co-occurrence networks as well as regularities extracted from Heaps law parameters to classify and detect the temporal separation of 20 world languages. The classification process was performed using hierarchical clustering while the comparison of clusters done using entanglement [40, 45].

4.1 The Effect of Vocabulary Growth on the Characterization of Written Languages

One of the best-known characteristics of vocabulary is the Heaps' law (also known as Herdan's law) introduced in the 1960s [48] which describe the vocabulary growth in texts [60, 38]. The law is defined as:

$$V_R(n) = k n^{\beta}, \tag{4.1}$$

where V_R is the number of vocabulary words in the text of size n, k and β are parameters determined experimentally.

Because Heaps' law represents the vocabulary richness of a certain language, a large text corpus of ten million words was used for the fitting of the two Heaps' law parameters as depicted in Figure 4.1. These parameters are used as a part of the features vector that later is used to characterize the ten languages we have selected for this study.

Table 4.1 shows the values of k and β for the languages in Figure 4.1. For English, the values of k are expected to be between 10 and 100 and the values β between 0.4 and 0.6. Our results agree with this expectation but the values of k for Arabic and Russian are greater than 100, which could be resulted from the morphological nature of these two languages.

After the fitting of Heaps' law to the corpora, we set to create co-occurrence

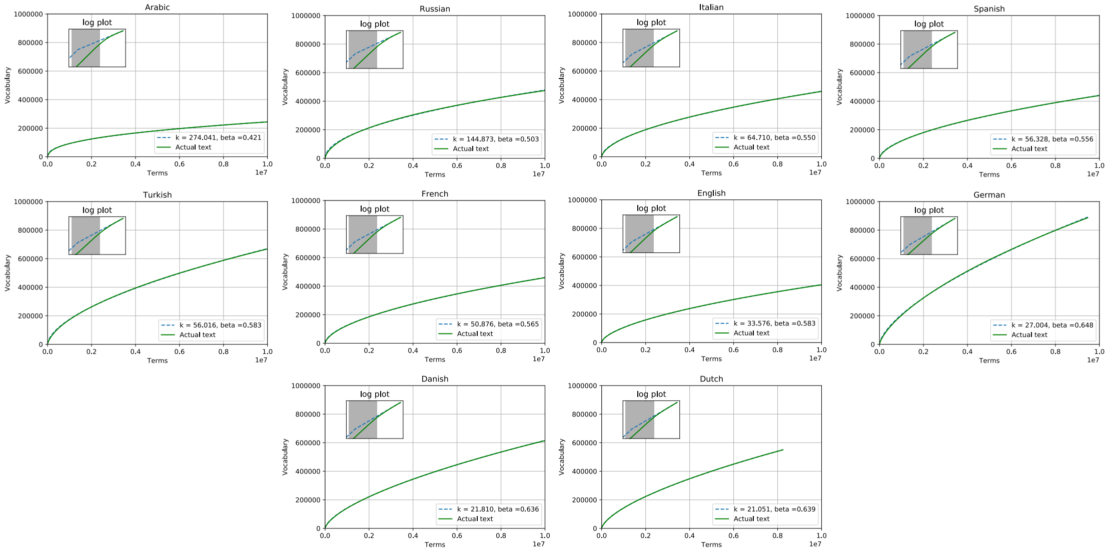


Figure 4.1: Heaps' law for the 10 languages used in this study (and the value of k and β respectively).

Table 4.1: The languages in Figure 4.1. The values of k and β from Equation 4.1 is shown, sorted by the value of k.

	Arabic	Russian	Italian	Spanish	Turkish	French	English	German	Danish	Dutch
k	274.04	144.87	64.71	56.32	56.01	50.87	33.57	27.00	21.81	21.05
β	0.42	0.50	0.55	0.55	0.58	0.56	0.58	0.64	0.63	0.63

word networks. Our networks are simple and link words in each corpus if they are adjacent to each other in text. Hence, nodes represent unique words and edges represent the connection between each two consecutive words. The edges' weights represent the frequency in which the two words appear next to each other. Table 4.2 shows the size of each network in terms of number of nodes n and number of edges m.

Table 4.2: Size of the word co-occurrence networks for all 10 languages.

	English	Arabic	Russian	Italian	Dutch	French	German	Turkish	Danish	Spanish
n	18,986	29,995	37,341	31,361	30,475	30,248	39,098	34,945	30,329	29,999
m	77,989	81,046	93,587	94,494	94,427	94,611	95,774	89,385	88,985	94,919

The generation of the networks gives us the structure and the values for n and m. Note however from Table 4.2 that for all languages the values of n and m are very similar which indicates they are not good features to let us characterize the languages. However, there are many structural characteristics that can be computed from the networks. We focus on the metrics described below.

The average degree $\langle k \rangle$ is generally provided as an information item given that word co-occurrence networks are not well represented by averages. These networks tend to display a power-law degree distribution and the average degree does not represent the distribution. The highest average degree was 8.21 for English, whereas the lowest was 4.89 for German. The reason for this is because the German language's vocabulary is much bigger than that of English, hence the fewer word connections [17].

The clustering coefficient of a network (C) is given by the average of the clustering coefficients of each node (C_i). In turn, C_i captures a ratio of the number of triangles that exist in the graph over how many triangles could possible exist in the graph given that a triple is already present. More formally:

$$C_i = \frac{2E_i}{k_i(k_i - 1)}, \tag{4.2}$$

where, E_i is the number of links that exist between the neighbors of node i, and the denominator number of possible links that could exist between nodes i.

The network clustering coefficient is the average of all C_i. The Russian and Arabic languages have the lowest clustering coefficient, 0.012 and 0.019 respectively, while English and Danish score the highest with 0.047 and 0.041 respectively. This is due to the fact that Russian and Arabic are morphological languages, which means that they have more word forms than analytic languages such as English and Danish [1].

Another important characteristic for networks analysis is the average path length. We know that social networks have high C and low average path length (ℓ) computed as:

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}, \tag{4.3}$$

where d_{ij} is the distance between nodes i and j. Russian has the longest value for ℓ of 4.91 steps, while the shortest one was 3.82 for English. Again, this is because morphological languages like the Russian and Arabic tend to have a longer path than analytic languages like English and Dutch [1].

Networks can be divided into consistent groups of nodes called communities [32] whose density of edges within the community is higher than outside it. There are many algorithms in the literature proposed to find these communities but one of the classical ways is to calculate the modularity of the network (Q). We computed the value of Q for all 10 networks using the approach proposed by Newman [70]. Based on this metrics, Russian has the largest modularity value of 0.481, while the lowest value was 0.379 scored by English.

The last two parameters, α_d and α_s were obtained by fitting weighted degree distribution of the network and size distribution of communities of words. As shown later in Table 4.3 the values of α_d are quite close to what is expected for real-world networks ($2 \leq \alpha \leq 3$). We believe the reason for the lower exponent values was the removal of the functional words. Figure 4.2 shows that a power law distribution (i.e. $P(k) \sim k^{\alpha}$, where k represents the node degrees) has the best fit when compared to other common distributions of real-world networks.

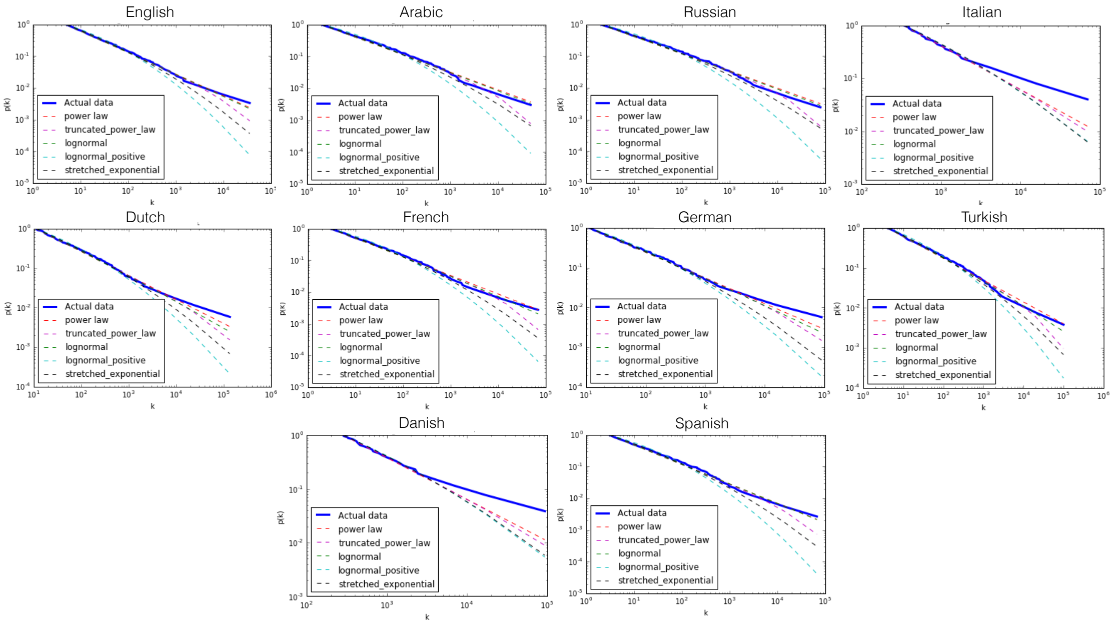


Figure 4.2: Fitting of the degree distribution in the power law package.

Similarly, the α_s value for the distribution of community size shows a good fit with a power law distribution, which is expected also in real-world networks with community structure; according to Arenas *et al.* [10] the distribution of community sizes in real network appear to have a power law form $P(s) \sim s^{\alpha}$. Both values have been used as part of the feature vector representing the languages. Figure 4.3 shows the fitting for the community size for all 10 languages and Table 4.3 shows the values for α_s.

For each of the networks we built, we generated random networks with the same size and edge creation probability using Erdős-Rényi model. The purpose was to analyze the clustering of our word networks. The average clustering coefficient values for the random networks were much smaller than those in the word networks. For example, in Italian, the average clustering coefficient for our network is 0.022 while in the random network was 0.00019. Also, the average path length (ℓ) for

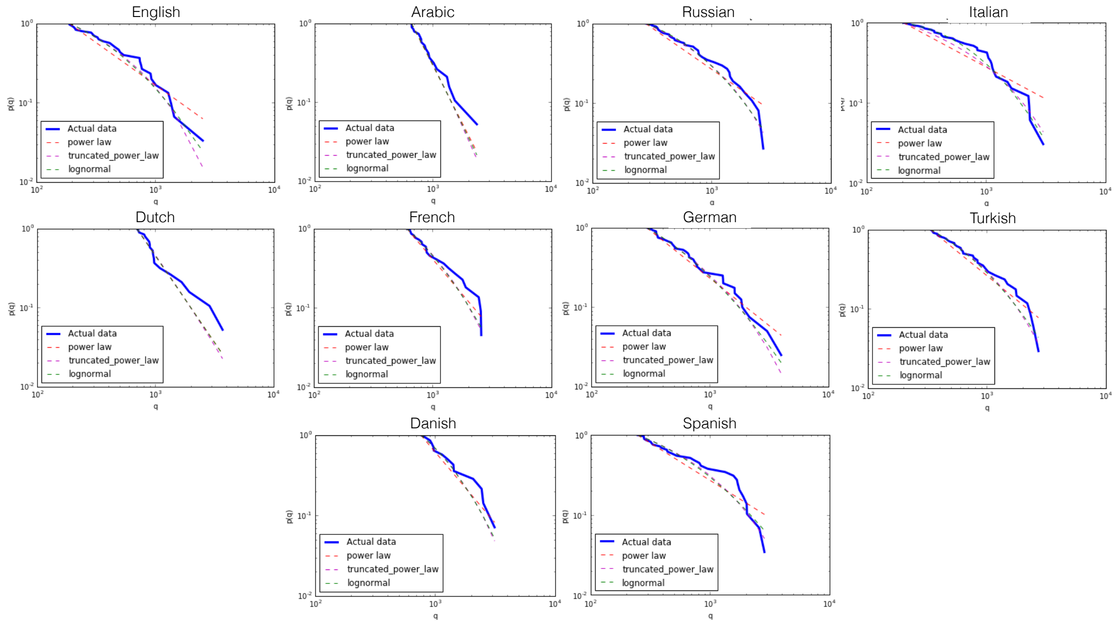


Figure 4.3: Fitting of the size distribution in the power law package.

the ten languages was between 3.8 and 4.9 which means our networks appear to be small-world [91].

4.2 Clustering Results

After all the analysis we had an 8-dimension feature vector for each language as depicted in Table 4.3. Next, we will use these vectors to do a clustering of the languages and hence characterize them based on their structural similarities.

We have performed clustering using two known algorithms: K-means and hierarchical clustering with the aim of finding similarities between languages using only the structural characteristics of the languages. K-means is a fast and widely used clustering algorithm that works by minimizing the sum-of-squares distance of the data points within the cluster. The number of clusters must be specified in advance, so two methods were used to find the optimal number of clusters. The first

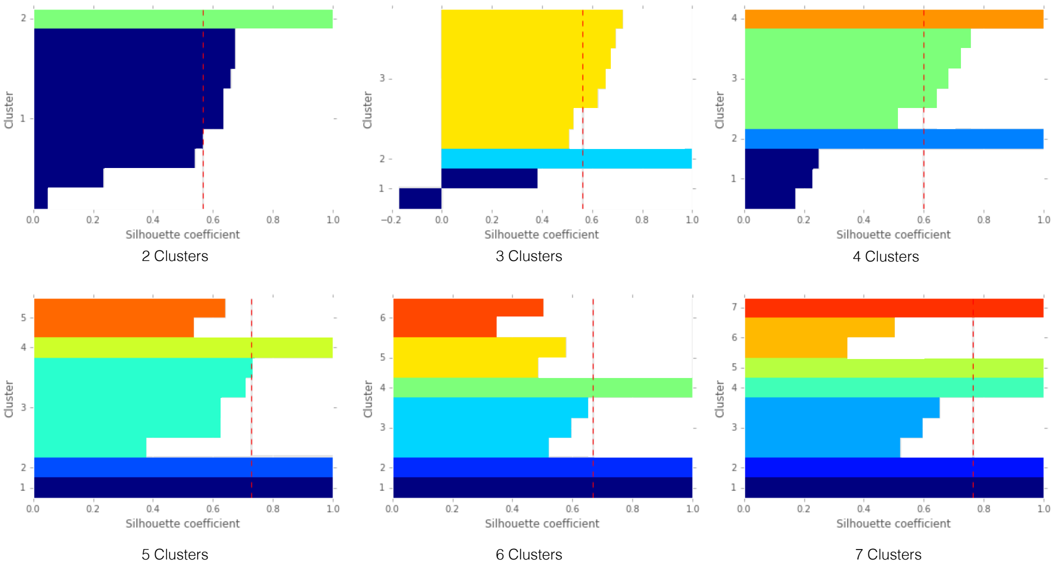
Table 4.3: Each line in this table represent 8-dimension feature vector for the language shown in the first column.

Languages	β	$\textit{\k{k}}$	$\langle k \rangle$	C	ℓ	Q	α_d	α_s
English	0.583	33.576	8.215	0.047	3.824	0.379	1.827	2.070
Arabic	0.421	274.041	5.404	0.019	4.454	0.466	1.508	3.937
Russian	0.503	144.873	5.012	0.012	4.910	0.481	1.660	2.037
Italian	0.550	64.710	6.026	0.022	4.280	0.405	1.751	1.800
Dutch	0.639	21.051	6.197	0.026	4.194	0.388	1.725	3.186
French	0.565	50.876	6.255	0.023	4.213	0.385	1.745	2.774
German	0.648	27.004	4.899	0.023	4.471	0.464	1.689	2.194
Turkish	0.583	56.016	5.115	0.023	4.430	0.471	1.716	2.223
Danish	0.636	21.810	5.868	0.041	4.200	0.438	1.740	2.761
Spanish	0.556	56.328	6.328	0.023	4.239	0.389	1.730	1.934

one is the silhouette method; it provides a visual aid in determining the number of clusters. The silhouette coefficient which ranged between -1 and 1 indicates the closeness of each data point in a cluster to other points in the neighboring clusters. After that, we used the elbow method to validate the number of clusters found in the silhouette method.

Due to the high dimensionality of the feature vectors, we run Principle Component Analysis (PCA) to reduce the dimensionality of the features vector into two dimensions so that the resulting K-mean clusters can be visualized. We also wanted to independently check whether the parameters extracted from the Heaps' law were providing extra information to the clustering of the feature vectors. The silhouette method was applied with and without the two Heaps' law parameters ($\textit{\k{k}}$ and β). In the first case, the optimal number of clusters was three. When the Heaps' parameters were added, the silhouette plot suggests the number of clusters between four and five as a good choice (Figure 4.4). These results indicate the importance of the Heaps' parameters to the process of the language classification.

The elbow method was used to validate the optimal number of clusters found



2 Clusters

3 Clusters

4 Clusters

5 Clusters

6 Clusters

7 Clusters

Figure 4.4: Silhouette analysis on K-Means clustering where the value of the Heaps' law parameters were included and after PCA.

by the silhouette method. The elbow plot suggests an optimal number of 3 clusters when removing the two Heaps' parameters, which agreed with the results of the silhouette method. The result of the K-means clustering for this case was that Italian, Spanish, German, Russian, and Turkish clustered together. The second cluster contains French, Danish, Dutch, and English, while Arabic appeared in its own cluster. When adding the parameters of Heaps' law, the elbow of the curve indicates an optimal number of 4 clusters (Figure 4.5(right)). In this case, Italian, Spanish, French, Danish, and Dutch were clustered together. The second cluster contains Russian, German, and Turkish, while English and Arabic separated into their own clusters (Figure 4.5(left)), which also supports the results of the silhouette method indicating the importance of Heaps' parameters to the classification process and the fact that the complete set of parameters offers a higher granularity for the clustering. These results match, to a certain degree, the linguistic typology classification of languages into genetic families as the Arabic language belongs to the Afro-Asiatic family, while the rest of the languages belong to the Indo-European Family.

An interesting finding from the clustering process is Turkish, which belongs to the Turkic family, was clustered with the Indo-European Family. As the aim of this work is to classify languages based on lexical rather than syntactical perspective, the removal of the functional words (stop words) has affected the structure of the languages networks [25]. This in turn has reduced the syntactic barriers between languages belonging to different families. The addition of the Heaps' law parameters enforced the separation of the languages based on their vocabulary richness and lexical structure represented by the network statistics.

In light of the previous assumption, the development of languages seen in the modern age, caused by the effects of technology, globalization, and migration among other factors, has had on effect on languages classification. For the case of the Turkish language, as of the year 2011, three million Turkish people were living in Germany, representing 3.6% of the German population [33].

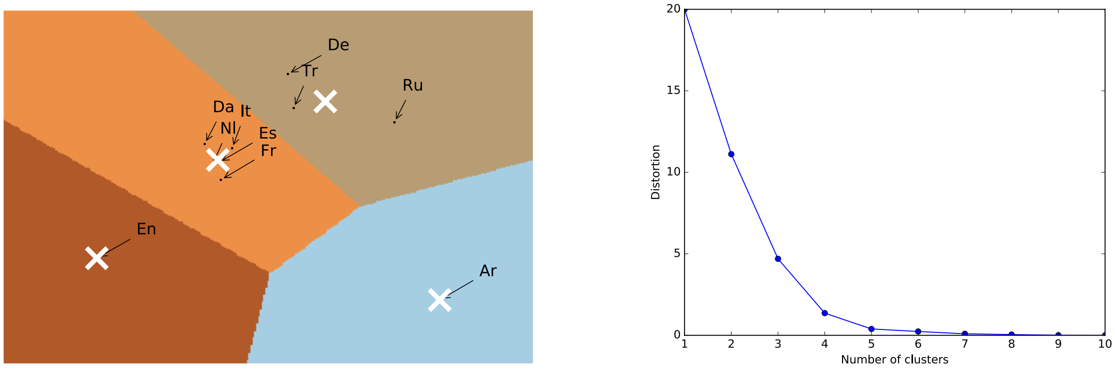


Figure 4.5: K-means clustering after PCA and using Heaps' law parameters and network parameters. The elbow rule (on the right) shows that 4 clusters appear to be the best choice for the K-means.

The results of K-means clustering can only classify languages from the top level of the family tree. To find the relationships between languages in a more structured way we applied a hierarchical clustering to the feature vectors we have for each language. In this case, we decided to also test whether the heaps' law features alone would provide a similar classification to the classification based on network features alone. Figure 4.6(b) show the classification using only the Heaps' parameters while Figure 4.6(c) shows the same results using only network parameters. Although both classifications have interesting characteristics that resemble traditional language classifications, the combination of both features (Heaps' and network parameters) yields a classification that appears to be enhanced. For instance, the distance between the Turkish and German languages was increased.

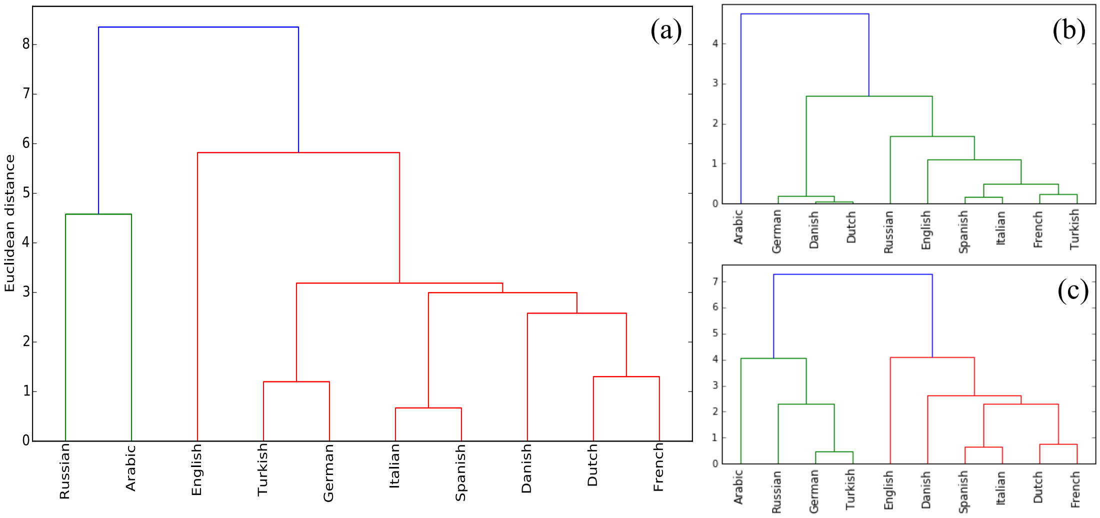


Figure 4.6: Hierarchical clustering of the 10 languages used in our study. (a) Shows the classification using the network parameters as well as the Heaps' law parameters while (b) shows the classification using Heaps' law parameters and (c) network parameters separately.

4.3 Glottochronological Classification of Natural Languages

We extracted a set of 19 features for each language; we want to demonstrate that one could use these features (or some of them) to unveil a structure similar to the ground truth. The first two features represent the vocabulary richness of the language as expressed by Heaps' Law [48]. The parameters $\Bbbk$ and β describes the vocabulary growth (distinct words) in texts as a function of the total number of words seen [60, 3]. More formally, $V_R(n) = \Bbbk n^{\beta}$. where V_R is the number of vocabulary words in the text of size n, $\Bbbk$ and β are parameters determined experimentally from the fitting of Heaps' Law.

Table 4.4: Each line in this table represent 19-dimension feature vector for the language shown in the first column.

Languages	$\Bbbk$	β	α_d	α_s	n	m	$\langle k \rangle$	C_4	C	$\langle C_d \rangle$	$\langle C_b \rangle$	$\langle C_c \rangle$	D	$trans$	$\eta_{\triangledown}$	ℓ	r	Q	com
Portuguese	6.40	0.702	2.302	1.343	20,641	70,816	6.86	0.044	0.186	0.00033	0.00010	0.305	11	0.0103	11.729	3.331	-0.135	0.392	47
Spanish	7.63	0.694	2.323	1.462	22,258	73,026	6.56	0.059	0.241	0.00030	0.00010	0.315	14	0.0088	12.972	3.217	-0.227	0.351	111
Italian	8.28	0.689	2.291	1.399	22,885	77,693	6.79	0.035	0.170	0.00030	0.00010	0.302	13	0.0113	11.721	3.357	-0.223	0.363	55
Catalan	7.69	0.686	2.324	1.335	20,856	68,005	6.52	0.073	0.277	0.00030	0.00010	0.322	10	0.0084	13.551	3.151	-0.210	0.364	44
French	7.41	0.690	2.289	1.324	20,700	73,241	7.08	0.051	0.257	0.00030	0.00010	0.322	09	0.0109	16.628	3.146	-0.245	0.336	58
Romanian	8.91	0.683	2.307	1.252	22,821	75,361	6.60	0.043	0.175	0.00028	0.00010	0.305	10	0.0106	11.306	3.325	-0.185	0.371	33
Dutch	6.54	0.700	2.175	3.529	20,485	72,745	7.10	0.081	0.320	0.00030	0.00010	0.326	11	0.0157	26.030	3.102	-0.219	0.304	31
German	0.23	1.008	2.214	1.427	24,296	73,841	6.08	0.088	0.260	0.00020	0.00009	0.317	10	0.0120	16.121	3.200	-0.195	0.352	112
Danish	5.70	0.720	2.217	4.804	22,234	71,612	6.44	0.066	0.246	0.00020	0.00010	0.311	10	0.0130	16.535	3.259	-0.183	0.358	34
Norwegian	6.13	0.706	2.231	4.456	20,571	63,997	6.22	0.090	0.298	0.00030	0.00010	0.322	10	0.0108	15.349	3.143	-0.210	0.364	30
Swedish	4.65	0.743	2.186	1.330	24,071	70,887	5.89	0.081	0.278	0.00020	0.00010	0.316	11	0.0086	11.808	3.209	-0.199	0.386	44
English	9.88	0.650	2.368	1.404	17,448	68,762	7.88	0.074	0.318	0.00040	0.00010	0.339	09	0.0107	22.913	2.994	-0.193	0.291	47
Bulgarian	5.41	0.728	2.449	1.854	23,655	58,746	4.97	0.061	0.185	0.00020	0.00009	0.306	17	0.0034	5.091	3.323	-0.189	0.503	496
Slovenian	7.58	0.716	2.343	1.791	28,669	83,470	5.82	0.037	0.122	0.00020	0.00008	0.286	11	0.0105	8.593	3.558	-0.117	0.396	62
Russian	7.51	0.719	2.334	4.502	29,333	81,405	5.55	0.045	0.123	0.00010	0.00008	0.285	10	0.0057	5.415	3.576	-0.112	0.428	57
Ukrainian	4.41	0.765	2.345	2.629	29,363	78,155	5.32	0.054	0.147	0.00018	0.00008	0.289	15	0.0066	5.654	3.543	-0.159	0.438	36
Czech	4.71	0.765	2.387	1.878	31,486	83,320	5.29	0.041	0.101	0.00016	0.00008	0.274	12	0.0057	4.298	3.726	-0.086	0.438	64
Slovak	7.07	0.733	2.288	2.305	32,542	87,625	5.39	0.029	0.086	0.00016	0.00008	0.270	13	0.0075	4.896	3.775	-0.081	0.431	65
Croatian	7.31	0.716	2.317	2.003	27,369	63,826	4.66	0.039	0.144	0.00017	0.00010	0.267	14	0.0040	2.693	3.819	-0.134	0.550	132
Polish	5.92	0.734	2.390	3.155	27,390	72,721	5.31	0.048	0.122	0.00019	0.00009	0.277	16	0.0082	5.123	3.678	-0.130	0.470	70

The other 17 features were obtained from the word co-occurrence network for each language. The network is simple and built having words as nodes and linking words if they appear in the corpus consecutively. The edges' weights represent the frequency in which the two words appear next to each other. The networks follow a power-law distribution and have community structures (we used Louvain

Table 4.5: Best 10 Entanglement with its combinations.

Entanglement	k	β	α_d	α_s	n	m	$\langle k \rangle$	C_4	C	$\langle C_d \rangle$	$\langle C_b \rangle$	$\langle C_c \rangle$	D	$trans$	$\eta_\triangledown$	ℓ	r	Q	com
0.0602616					✓		✓	✓				✓		✓		✓	✓		
0.0602616					✓		✓	✓				✓		✓		✓	✓		✓
0.0604673			✓			✓	✓							✓	✓	✓	✓	✓	✓
0.0604673			✓		✓	✓	✓					✓		✓	✓		✓	✓	✓
0.0604673			✓		✓	✓	✓							✓	✓	✓	✓	✓	✓
0.0604673			✓			✓	✓							✓		✓	✓	✓	✓
0.0604673			✓			✓								✓	✓	✓	✓	✓	✓
0.0653795			✓		✓				✓	✓		✓		✓		✓	✓		✓
0.0663400			✓				✓		✓		✓	✓	✓		✓		✓		
0.0687276					✓		✓	✓						✓		✓	✓		✓

modularity [18]); the number of communities is an important feature (com). The features α_d and α_s represent respectively the scaling of the degree distribution and the distribution of community sizes. The size of the network is given by the number of nodes n and number of edges m.

There are many other structural characteristics that can be computed from the networks. For this work we exhaustively added many features without too much concern for an exact number. The purpose is to make sure we are capturing as many uncorrelated metrics as possible. Later we worked on reducing the dimensions and identifying the most significant parameters. The degree k of a node is the number of edges connected to it. The higher average degree $\langle k \rangle$ the network has, the more density it is [21]. From Table 4.4, we can clearly see that the Slavic languages have a lower $\langle k \rangle$ compared to all other languages in the dataset, while the English language has the higher one.

In addition to the network clustering coefficient (C), a measure of the degree to which nodes in a graph tend to cluster together, we calculate the square clustering (C_4) which is the quotient between the number of squares and the total number of possible squares [57].

Similar to the concept of clustering C is the concept of transitivity ($trans$) [78] of the network. Moreover, both C and $trans$ depend on the number of triangles (cliques of 3 nodes) in the network, so we have also included these features ($trans$ and $\eta_\bigtriangledown$ respectively) as part of our set of metrics. Another important feature of networks is the average path length (ℓ) between two nodes which also included in our list. Croatian has the longest value for $\ell = 3.81$ steps, while the shortest one was English with $\ell = 2.99$. This is likely because morphological languages like most of Slavic languages tend to have longe sentences than analytic languages like English and Dutch [1]. The diameter of the network D is the largest shortest path and another important feature we included here. Note that at this point the idea is to have an exhaustive list of features that could represent a language.

Related to community detection algorithm is the modularity of the network given by Q which is designed to measure the strength of a division of a network into groups; a measure of the community structure[32]. The value of Q for all 20 networks was calculated using the approach proposed by Newman [70]. Based on this metric, Croatian has the largest modularity value of 0.550, while the lowest value was 0.291 scored by English.

Centrality measures are used to identify the important nodes within a network, here we used degree-centrality (C_d) which is highly correlated to $\langle k \rangle$, betweenness (C_b), and closeness (C_c) as defined by Borgatti [20]. However since we want a network feature, we represent the average of all these values given by $\langle C_d \rangle$, $\langle C_b \rangle$, and $\langle C_c \rangle$. Last, we compute the degree assortativity of the network which is given by r [69].

For each of the networks we built, we generated random networks with the same size and edge creation probability using Erdős-Rényi model. The purpose was to analyze the clustering of our networks. The average clustering coefficient values for the random networks were much smaller than for our networks. For example, in Polish, the average clustering coefficient for our network is 0.122 while in the random network was 0.00037. Given that ℓ is also small, we argue that the networks have small-world characteristics [91].

4.4 Effect of Feature selection and Language Removal on Clustering Entanglement

After all the analysis we had a 19-dimension feature vector for each language as depicted in Table 4.4. This vector is used in clustering the networks but we will also try to identify the most significant features and reduce the dimension.

We have performed a clustering using a hierarchical clustering algorithm, which provides the relationships between languages in a more structured way as compared to the original hierarchy. The result is presented as a tree diagram called a dendrogram. Hierarchical clustering algorithm begins with each object in a separate cluster, at each step, the two clusters that are most similar are joined into a single new cluster [51]. The aim is to find similarities and the time separation topology between languages using only the structural characteristics of the languages.

We applied a clustering to the complete feature vectors in Table 4.4 (19 parameters for each language). In order to compare the tree resulted from the hierarchical clustering with the ground truth tree (Figure 3.1), we measured the quality of the alignment of the two trees by calculating the entanglement function. Entanglement is a measure between 1 (full entanglement) and 0 (no entanglement). A lower entanglement coefficient corresponds to a good alignment [45, 40]. The entanglement result from the two trees is 0.27 (Figure 4.7) which represents about 73% of agreement between two trees. In general, this number is good, but when looking at the relation of languages in Figure 4.7, one can easily determine the distortion in branches. Moreover, the result from the complete matrix does not catch even the three sub-families structure, specifically, the intersection between Germanic and Romance languages make it difficult to separate them.

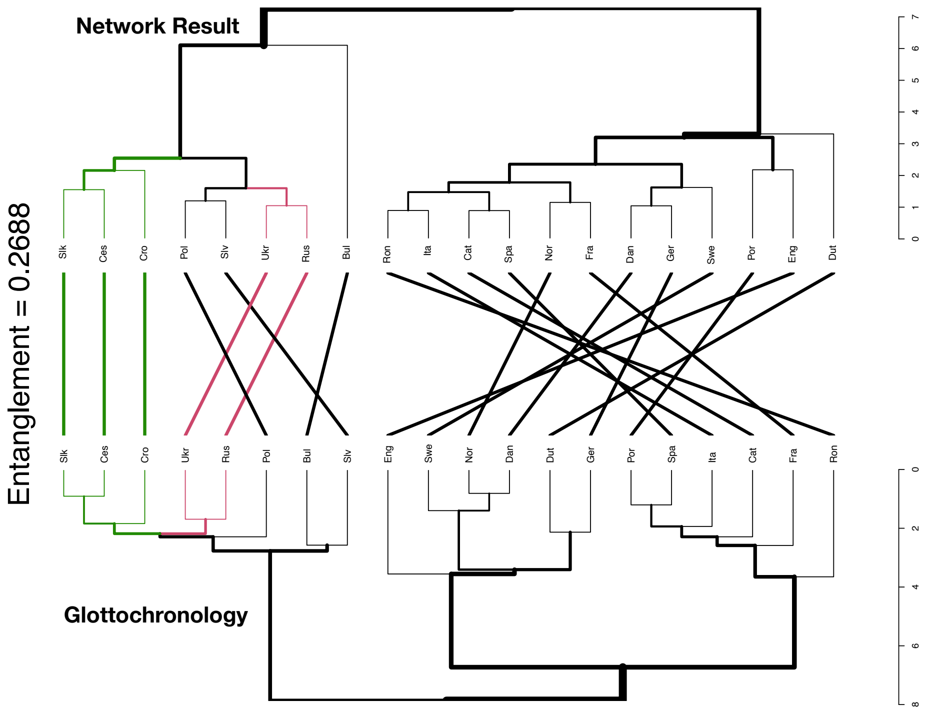


Figure 4.7: Entanglement between two trees using complete feature vectors (19 parameter).

For the sake of enhancing the entanglement results, we took all the possible combination of the 19 parameters in the matrix (from 1 to 19 parameters). For each combination, we constructed a tree and compared it with the ground truth in order to find the entanglement. As a result, we found many combinations that give better results compared to the complete feature vectors results. Table 4.5 contains the best 10 entanglement from all combinations. Furthermore, we can evoke which features that have high influential on the results like transitivity, degree assortativity coefficient, average shortest path length, and average degree which they appeared in the most cases. In contrast, there are some parameters useless for this kind of work like Heaps' law parameters and betweenness centrality (Table 4.5).

The best combination between all cases has the entanglement value of 0.06 (first case in Table 4.5), this case has only 7 parameters, which is the smallest combination parameters that gives better values. The parameters are number of nodes, average degree, square clustering, closeness centrality, transitivity, average shortest path length, and degree assortativity coefficient (Figure 4.8 shows the dendrogram result of the 7 parameters combination). The hierarchical clustering was able not only to distinguish the Slavic languages from the non-Slavic language but also to capture the branches relation and distances for this sub-family with one exception which is the Bulgarian language (discussed later). Moreover, it was ambidextrous to recognize the Germanic from Romance languages with some differences in the branches relation like Germany with Norwegian instead of the Dutch language.

In order to check the consistency of result, we tested the sensitivity of removing languages. First, we remove one language each time and calculate the average entanglement for all cases. Secondly, we remove two languages and calculate the average entanglement, and so on (Figure 4.9(b)). The average entanglement increased until the 6th language removed and then starts to be constant at a high level, which means that the topology of the tree is completely destroyed and the removal of more languages does not affect the result.

To test for certain language impact on the average entanglement and tree topology, we removed one language each time and recalculated the average entangle-

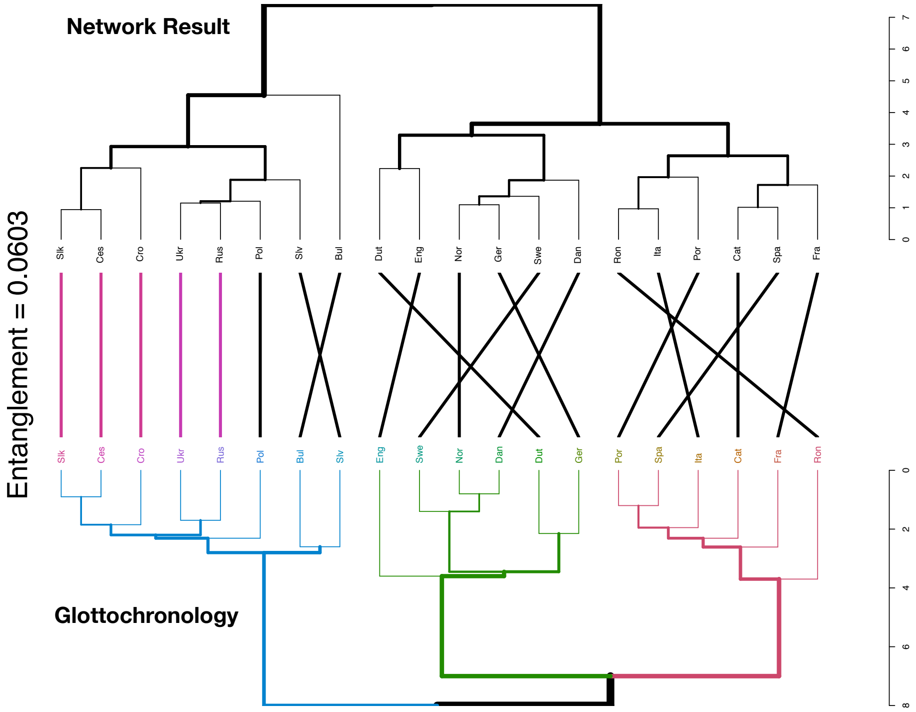
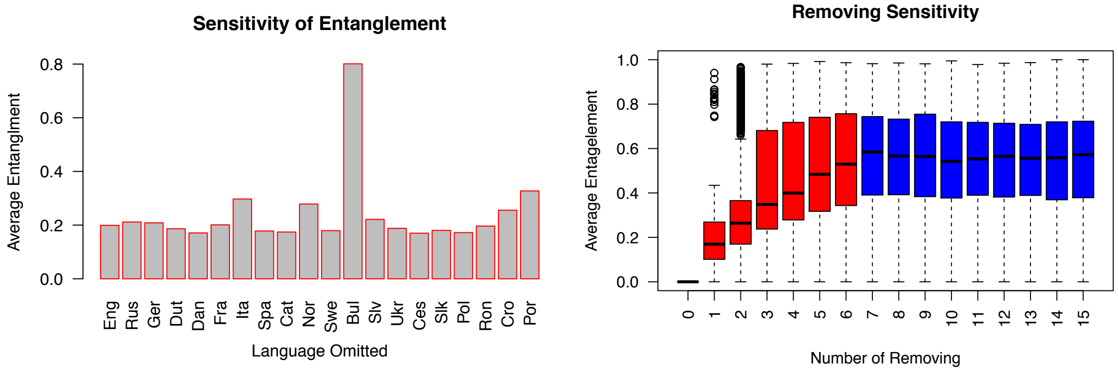


Figure 4.8: Entanglement between two trees using the best entanglement case (7 parameters).

ment. The language with high average entanglement in Figure 4.9(a) means the most effective language on the tree topology. In our languages set, when we removed the Bulgarian language, the average entanglement became very high (0.79) which means the branches relation is very tangled. The unpredictable behavior of the Bulgarian language may be due to several reasons, first, the number of unique words (nodes) is less than others Slavic languages. Also, words in the Bulgarian language are most likely to connect with another word several times which describes the reason why the language has a number of connections less than all other language networks in the dataset. On the other hand, several important dissimilarities exist between the Bulgarian language and other Slavic languages. For instance, Bulgarian is an analytic language and its unique morphological fea-



(a) The effect of language removal to the average entanglement.

(b) Average entanglement sensitivity as a function of languages removal.

Figure 4.9: Validating data and entanglement sensitivity as a function of removing languages.

tures tend toward the Balkan family of languages. The Bulgarian language roots back to the Proto-slavic branch of the Indo-European language family which have common features with the Indo-Iranian languages, more specifically, the Germanic family but it was much similar to the Baltic family of languages. Finally, a lot of the words in the Bulgarian language were borrowed from the Turkish and Greek languages [72].

Chapter 5

Author Attribution

Network motif defined by Milo *et al.* [66] as a statically significant subgraphs pattern occurred in real-world networks compared to random ones, has gained a lot of attention because of its ability in discriminating networks from different discipline [89]. In this chapter, we utilized network motifs as a fingerprint to attribute authors by their writing style. More precisely, we extract network motifs from directed co-occurrence networks of 100 books by 10 well-known authors and then we use 5 machine learning algorithms to classify the authors by their network motif signature. We show that 4-nodes directed network motifs alone can be utilized to attribute authors of different books.

The chapter is organized as follows. Section 5.1 is an overview of the steps taken place to extract the network motif from the text networks. In Section 5.2 we describe the standardization and feature selection methods utilized to obtain our feature vector, finally discussion of the results obtained is in section 5.3 with a road-map for future work.

5.1 Extraction of Network Motif

A plethora of network motif extraction tools exist, each one has its pros and cons related to the number of motif's nodes count and the algorithm speed. We chose the iGraph [1] implementation for its flexibility and fast execution time. The iGraph's motifs function returns a vector of the number of occurrence of each connected motif in the graph ordered by their isomorphism classes. For a directed network, as in our case, the VF2 algorithm by Cordella *et al.* [30, 29] will be used to find the motifs. The VF2 algorithm works as follows:

Let $G1 = (V1, E1)$ and $G2 = (V2, E2)$ two graphs, where:

$V1, V2$ are the set of vertices in $G1$ and $G2$ consecutively and $E1, E2$ denotes the set of edges for the two graphs.

A matching process between the two graphs comprises of finding a mapping M that associates vertices of $G1$ with vertices of $G2$ and vice versa, based on some predetermined constraints. In general, the mapping is expressed as the set of pairs (u, v) (with $u \in G1$ and $v \in G2$) each referring to the mapping of a vertex u of $G1$ with a vertex v of $G2$. A mapping $M \subset V1 \times V2$ is said to be an isomorphism iff M is a bijective function that maintains the edge structure of the two graphs. A mapping $M \subset V1 \times V2$ is said to be a graph-subgraph isomorphism iff M is an isomorphism between $G2$ and a subgraph of $G1$.

Tran *et al.* [89] suggested that small undirected network motifs cannot reveal differences among networks from different disciplines, while large ones do. Based on this argument and the importance of feature frequency explained earlier, we chose

[1] http://igraph.org

the directed 4-node network motifs shown in Figure 5.1. For each book in the dataset, we extracted the 199 motifs from their directed network and then a data frame contains the motifs frequencies was created. Figure 5.2 illustrate a sample 4-node directed motif extracted from the example network of (Figure 3.2(c)). The frequency distribution of the extracted 4-node motifs from the books of Bernard Shaw, H. G. Wells, Jack London and William Shakespeare shown in Figure 5.3.

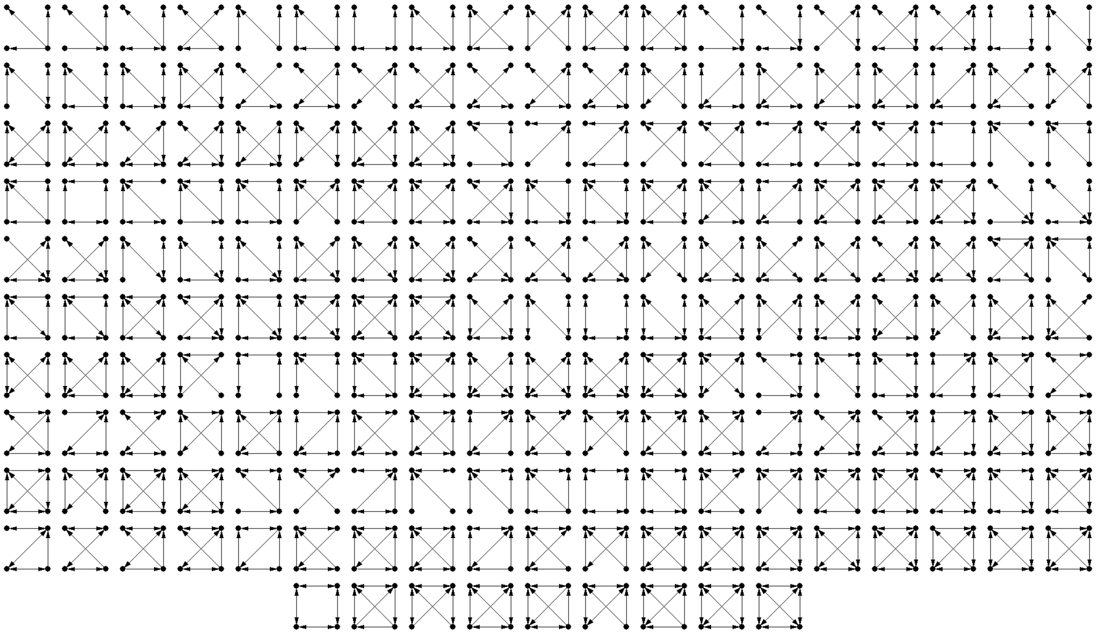


Figure 5.1: 199 different orientation of the directed 4-node network motif.

5.2 Feature Selection and Classification

We utilized five supervised machine learning classification algorithms namely K nearest neighbors (KNN), decision trees, random forests, support vector machines

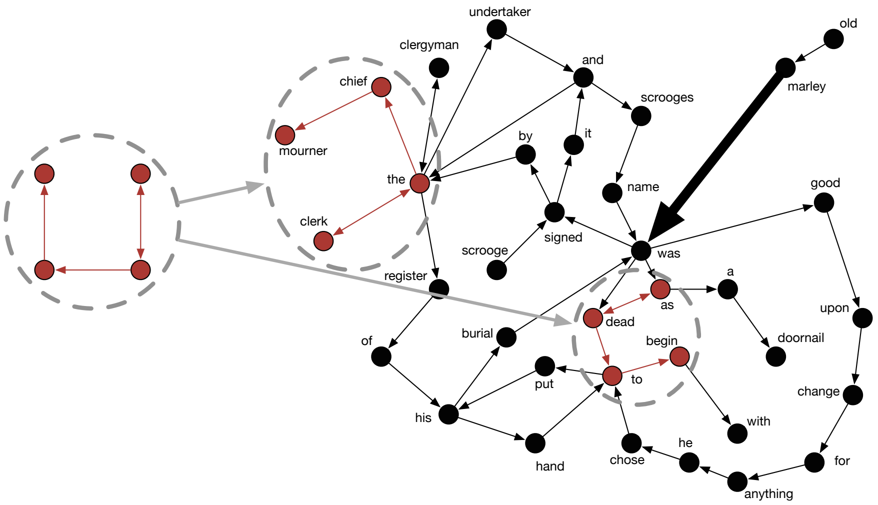


Figure 5.2: 4-node directed network motif sample from the network of Figure 3.2.

(SVM), and multi-layer perceptrons (MLP). They are all part of the scikit-learn[2] machine learning package for Python. As we try to attribute 10 authors, we have a multi-class classification problem with the number of samples ($N = 100$) which represents the number of books and the dimension of the feature set ($D = 199$) was relatively high. We used two cross-validation methods, the first one is to split our dataset into 75% training set and 25% testing set and then shuffle the dataset and repeat the operation for 100 times. The second method was leave-one-out, where the dataset is split into 99 sample for training and one sample for testing then iterate through the remaining samples. The average classification accuracy was calculated with both methods for all the algorithms used in the work as follows:

$$Accuracy = \frac{1}{n}\sum_{i=0}^{n-1} 1(\hat{y}_i = y_i) \tag{5.1}$$

[2] http://scikit-learn.org

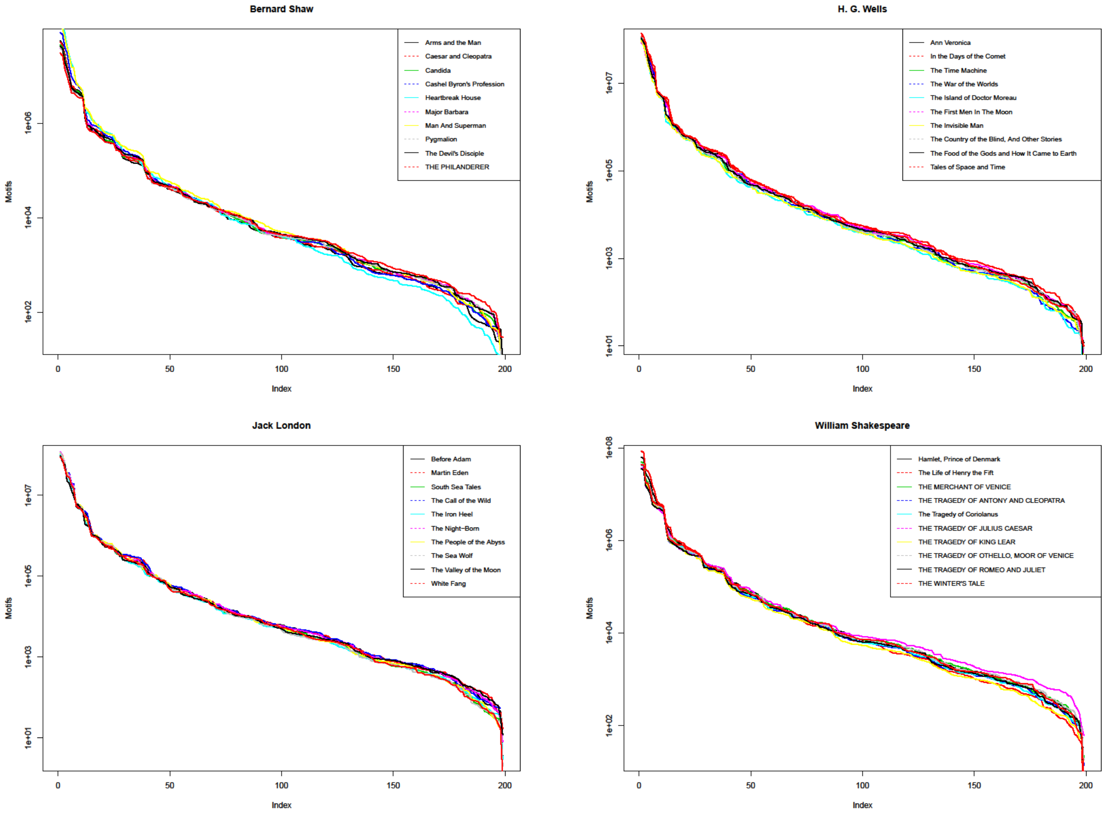


Figure 5.3: 4-node network motif sorted frequency for the networks created from the books by Bernard Shaw (upper left), H. G. Wells (upper right), Jack London (bottom left) and William Shakespeare (bottom right).

where n is the total number of samples, $\hat{y}_i$ is the predicted label of the i-th sample, and y_i is its corresponding actual label.

The dataset was standardized by scaling to unit variance and removing the mean. The classification was performed on all the feature sets, that is the whole 199 4-nodes directed motifs and then recursive feature elimination (RFE) [46] feature selection method used to find the best 75%, 50%, 25%, and 10% features respectively. An alternative method mostly used in the literature is to choose significant motifs based on the highest Z-scores, but we preferred to collect the whole set of motifs and then use feature selection methods to choose the best set. This magnificently reduced the time required to generate multiple random networks and the time of searching for the statistically significant motifs.

5.3 Discussion

The results of classification using the first cross-validation method of shuffling and splitting the dataset are listed in Table 5.1, while Table 5.2 lists the results of the leave-one-out cross-validation method. As can be seen from both tables, the two basic classification algorithms KNN and the decision trees did not perform well compared to more sophisticated algorithms. Although KNN gives us an average accuracy of 60% when using 25% of the dataset and the leave-one-out cross-validation method, it is still lower than the accuracy obtained from the other classification methods. The best classification accuracy of 77% was obtained when the MLP classifier used with leave-one-out validation method.

Table 5.1: Average classification accuracy results for the four nodes directed motifs when splitting the dataset into 75% for training set and 25% for testing set with 100 times random shuffling. First column represents complete dataset, while consecutive columns presents the classification results when 75%, 50%, 25%, 10% of the features were selected. Best accuracy obtained was 70% (shown in boldface) in the case were MLP classifier was used with 25% and 50% of the best features.

Method	**Features selected**				
	Complete set	75%	50%	25%	10%
KNN	0.42	0.42	0.48	0.53	0.52
Decision Tree	0.41	0.45	0.46	0.45	0.50
Random Forest	0.56	0.58	0.58	0.59	0.64
SVM	0.53	0.56	0.62	0.63	0.67
MLP	0.66	0.68	**0.70**	**0.70**	0.68

Table 5.2: Average classification accuracy results for the four nodes directed motifs with leave-one-out cross-validation method. Again, the first column represents complete dataset, while consecutive columns presents the classification results when 75%, 50%, 25%, 10% of the features were selected. Best accuracy obtained was 77% (shown in boldface) in the case were MLP classifier was used with 25% of the best features.

Method	**Features selected**				
	Complete set	75%	50%	25%	10%
KNN	0.41	0.42	0.51	0.60	0.54
Decision Tree	0.44	0.55	0.47	0.55	0.57
Random Forest	0.61	0.56	0.62	0.60	0.65
SVM	0.61	0.66	0.70	0.66	0.71
MLP	0.72	0.73	0.75	**0.77**	0.72

Chapter 6

Classification of Text into Literary Genre

In chapter 5 of this dissertation we show that the frequency of network motif was capable of attributing different authors of literary books. As we explained, network motif reflects the author's writing style as it represents a signature of repetitive sentence structure or part of speech combinations used frequently by an author. We want to take this assumption further and test it on a more challenging problem of classifying literary genre and see if it will compete against well known and widely used traditional text and topic classification methods.

We will start this chapter by describing our method of obtaining the features that will be used in the classification process (section 6.1) and the classification algorithms utilized to obtain the results (section 6.2). Finally, the results will be explained and discussed (section 6.3)

6.1 Feature Extraction

To justify our approach of classifying literary genre using network motif, we decided to compare the obtained accuracy with the ones from traditional text classification methods. We used numerous well known and state-of-the-art text feature extraction methods so we can choose the one which yields the best results. For motif feature, we used the same method described in section 5.2 of chapter 5. Thus, we created 100 co-occurrence networks from the text corpus obtained from the Project Gutenberg online library and all 199 directed 4-node network motifs were extracted using iGraph motif extraction function. The motif frequency distribution for adventure, fantasy, mystery, and science-fiction genre is shown in figure 6.1, we can easily notice the difficulty of classifying these genres due to the variation of the motif frequency for individual books in a certain genre.

Four text features extraction methods were used for the traditional methods of text classification part of the work including two state-of-the-art word vector based and two classical methods; bag-of-words (BoW) and term frequency-inverse document frequency (TF-IDF).

The well-known Vector Space Model (VSM) will be used here to formalize the text classification task. Let $D = \{d_1, d_2, ..., d_n\}$ represents the set of n text documents in our data set and $T = \{t_1, t_2, ..., t_m\}$ denotes the set of m terms or words in D. Now each text document $d_i \in D$ can be represented in the vector space as $d_i = \{w_{i,1}, w_{i,2}, ..., w_{i,m}\}$, where $w_{i,m}$ refers to the weight of term m in text document d_i.

For the bag-of-words feature extraction model, which considered the simplest model, the term frequency $tf(t, d)$ within each document in the data set will represents the weight of the term or word in that particular document and can be

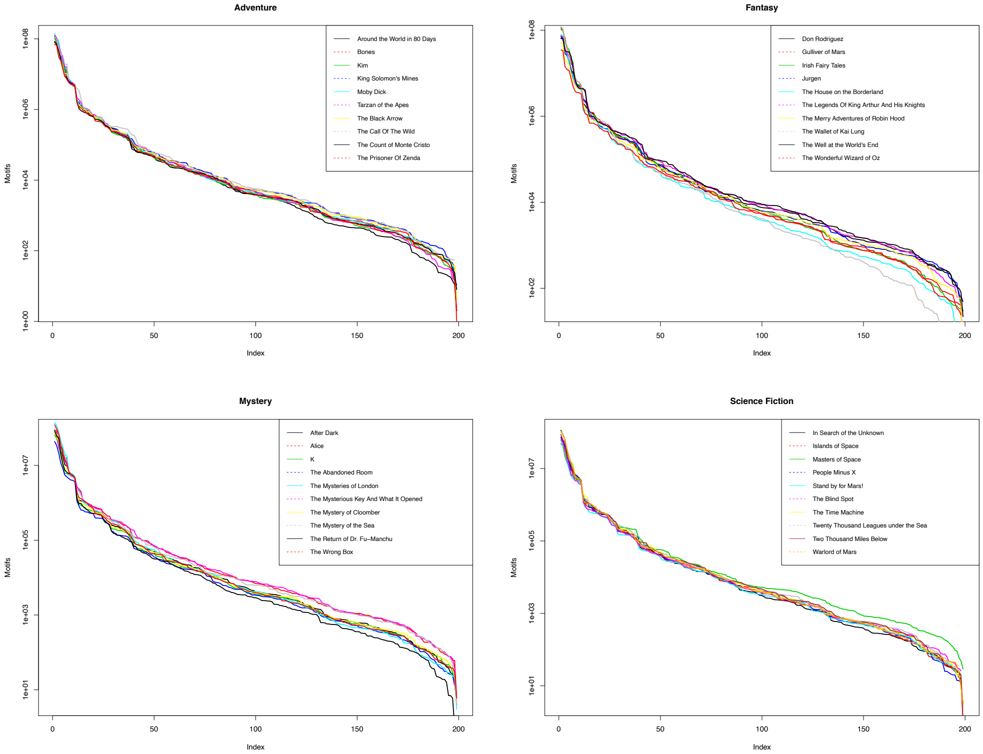


Figure 6.1: 4-node directed motif frequency distribution of adventure, fantasy, mystery, and science-fiction genre. Each graph represents the 199-motif frequency obtained from the co-occurrence network of 10 books for each genre. The motif frequency for the Masters of Space (green) book in science-fiction genre (lower-right) has noticeable variation from the other book's motif frequency in the same genre, which will have a negative effect on the classification accuracy.

mathematically represented as:

$$tf(t,d) = f_{t,d} \tag{6.1}$$

where $f_{t,d}$ is the the frequency of term t in a document d.

The inverse document frequency is a measure of how much a term is important in a document compared to its weight in the rest of the documents in the corpus. this can be calculated as:

$$idf(t, D) = \log \frac{n}{1 + |\{d \in D : t \in d\}|} \tag{6.2}$$

where n denotes to the total number of documents in the corpus D and the denominator represents the number of documents that contain the term t plus 1 to avoid division by zero if the term does not show in any document inside the corpus. Then the term frequency-inverse document frequency (TF-IDF) can be computed by taking the product of both measures as follows:

$$tfidf(t, d, D) = tf(t, d).idf(t, D) \tag{6.3}$$

Word2vec is a neural network-based word distributed representation introduced by Google in 2013 to learn features from huge text data-set consists of billions of words [64]. The averaged word vector ($AvgWV$) used so that the dimensions of the resulting feature vector are equal and can be represented as:

$$AvgWV(d_i) = \frac{\sum_{i=1}^{n} wv(t)}{n} \tag{6.4}$$

where $wv(t)$ represents the word vector for term t and n is the total number of terms in a document d. TF-IDF weighted averaged word vector ($AvgWV$) is an advanced word vector representation method where for each term in the document, the word vector is matched by that term TF-IDF score as follows:

$$TwWV(d_i) = \frac{\sum_{i=1}^{n} wv(t) \times tfidf(t)}{n} \tag{6.5}$$

here n denotes the sum of all TF-IDF scores of the matched term t in document d [77].

The Gensim[1] library implementation of word2vec was used in this work for extracting averaged word vector ($AvgWV$) and TF-IDF weighted averaged word vector ($AvgWV$) text features.

6.2 Classification Methods Selection

To classify literary genre based on network motifs, multi-layer perceptrons (MLP) algorithm was chosen based on the results obtained in chapter (5) where it scores the best classification accuracy of 77% for the author attribution task. But we only got an accuracy of 30% in literary genres classification work. To justify our results, we have to compare our methodology with traditional text classification methods and answer the question of whether network motifs are incapable of classifying literary genres or is the fine granularity of the genre categories we chose is what causes the low classification accuracy results. To do that, Multinomial Naïve Bayes (MNB) and Support vector machines (SVM) algorithms were used in addition to MLP classifier. A combination of the feature extraction methods described in section 6.1 was used with each classification algorithm to find the best match which yields the highest accuracy possible. The two word-vector methods produce negative and positive feature values which cannot be used with MNB algorithms because it expects word frequencies as an input which does not have negative values. Also, the MLP algorithm does not perform well when BoW, TF-IDF, and TwWV features were used, hence they were discarded from the results. Table 6.1 summarize the results of the classification performed using combinations of

[1] https://radimrehurek.com/gensim/models/word2vec.html

different algorithms and features with text data. Based on these results, we choose MNB with BoW features and MLP with AvgWV features which they have the highest accuracy as our baseline classifiers.

We conducted two experiments regarding text processing to extract the features, in the first one we kept functional words and did not lemmatized the text as we did in author attribution work. And in the second experiment we removed the functional words and performed text lemmatization which yields better classification accuracy and used to obtain our results shown next.

Table 6.1: Classification accuracy of different algorithms and text features extracted from the text corpus. MNB+BoW (first value) and MLP+AvgWV (last value) scored the highest accuracy and hence these two algorithms/features combination considered our baseline for the next results.

Method	Accuracy
MNB + BoW	0.48
MNB + TF-IDF	0.17
SVM + BoW	0.42
SVM + TF-IDF	0.44
SVM + AvgWV	0.41
SVM + TwWV	0.33
MLP + AvgWV	0.51

6.3 Results and Discussion

The results presented in table 6.1, show that MNB classifier with BoW features and MLP classifier with AvgWV score the best classification accuracy so they are considered our baseline classification methods for the rest of the work. We will only show the results of MNB classifier with BoW features because it maintain higher accuracy than MLP with AvgWV features as we will further investigate next. It can be easily noticed from the results that text classification methods have slightly better accuracy than the network motifs method but they are not statistically significant. Based on this, we concluded that the fine granularity of the chosen genre categories is the cause of the low accuracy obtained in both methods. We also need more metrics besides classification accuracy to compare and contrast the performance of the classification methods. These metrics are:

- Precision: defined as the number of correct predictions produced by the classifier for the positive class out of all predictions or the ability of the classifier not to label a negative class as a positive one.

$$Precision = \frac{TP}{TP + FP} \tag{6.6}$$

- Recall: also known as (hit rate) or (sensitivity) and defined as the number of correct predictions of the positive class.

$$Recall = \frac{TP}{TP + FN} \tag{6.7}$$

- F1-score: defined as the harmonic mean of precision and recall.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6.8}$$

where TP (True Positive) the number of correct predictions for the positive class, FP (False Positive) the number of incorrect predictions of negative class as a positive one, and FN (False Negative) the number of incorrect predictions of positive class as a negative one.

To overcome the low accuracy due to the fine granularity of the genre classes, we decided to eliminate genres with the lowest recall score one at a time and repeat the classification process for both network motifs and text features methods. The process of genre elimination will aid in revealing the similarities between genres which prevent achieving a high classification accuracy rate. Starting with 10 genres shown in the confusion matrix (figure 6.2) and the corresponding classification report (table 6.2), we notice that the lowest recall was for mystery genre when network motifs features were used and humor genre recall in the case of text features. By further investigation, we examined the book: The abandoned room which belongs to the mystery genre and misclassified as a detective one, we found that the word "detective" occurred 141 times in the book, word "murder" 115 times, and word "crime" 26 times. This should give us an indication of how hard is to classify such closely related genres.

By removing these two genres from the set and repeat the classification process, we obtained a slightly better classification performance as indicated by the confusion matrix in figure 6.3 and its related classification report (table 6.3). Here we have adventure genre for network motif features and mystery for text features have the lowest recall so they were removed (figure 6.4 and table 6.4).

Next, we continue our elimination of lowest recall genes process to get 7 genres (figure 6.5 and table 6.5), 6 genres (figure 6.6 and table 6.6), and finally, the 5 genres (figure 6.7 and table 6.7) where we stopped the elimination process by obtaining a relatively high classification accuracy of 62% for network motif-based classification, and 84% for text-based one.

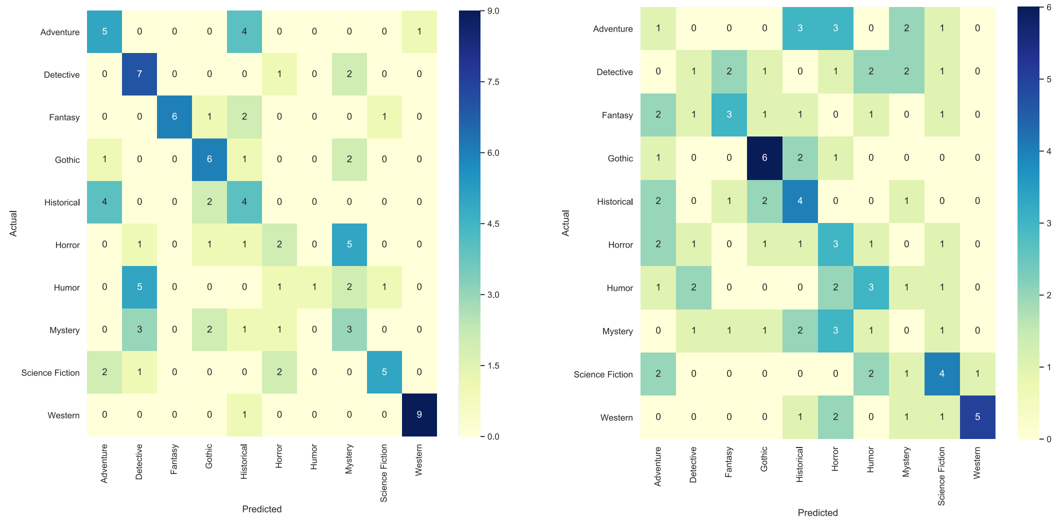


Figure 6.2: Confusion matrix for 10 genre, MLP classifier with network motif features (right) and MNB classifier with BoW features(left).

The five genres that remained after eliminating low recall ones are Fantasy, Gothic, Historical, Science Fiction, and Western when using network motifs as features and Detective, Gothic, Historical, Science Fiction, and Western when using text features. We conclude from these results that network motifs is not as good in classifying genres or text type with fine granularity as it did in attributing authors. that assists the assumption that network motifs reflect the writing style which is more related and applicable to author attribution than to literary genres. Finally, figure 6.8 summarize the results of genres elimination method for network motif features and text feature for both Multinomial Naïve Bayes (MNB) and multi-layer perceptrons (MLP) classifiers which we didn't show its results in the text as it performed below the MNB classifier.

Table 6.2: Classification report for 10 genre with both MNB classifier with BoW and MLP classifier with network motif features.

Genre	**Precision**		**Recall**		**F1-score**	
	Text	Motif	Text	Motif	Text	Motif
Adventure	0.42	0.09	0.50	0.10	0.45	0.10
Detective	0.41	0.17	0.70	0.10	0.52	0.12
Fantasy	1.00	0.43	0.60	0.30	0.75	0.35
Gothic	0.50	0.50	0.60	0.60	0.55	0.55
Historical	0.29	0.29	0.40	0.40	0.33	0.33
Horror	0.29	0.20	0.20	0.30	0.24	0.24
Humor	1.00	0.30	**0.10**	0.30	0.18	0.30
Mystery	0.21	0.00	0.30	**0.00**	0.25	0.00
Science Fiction	0.71	0.36	0.50	0.40	0.59	0.38
Western	0.90	0.83	0.90	0.50	0.90	0.62
Average	0.57	0.32	0.48	0.30	0.48	0.30

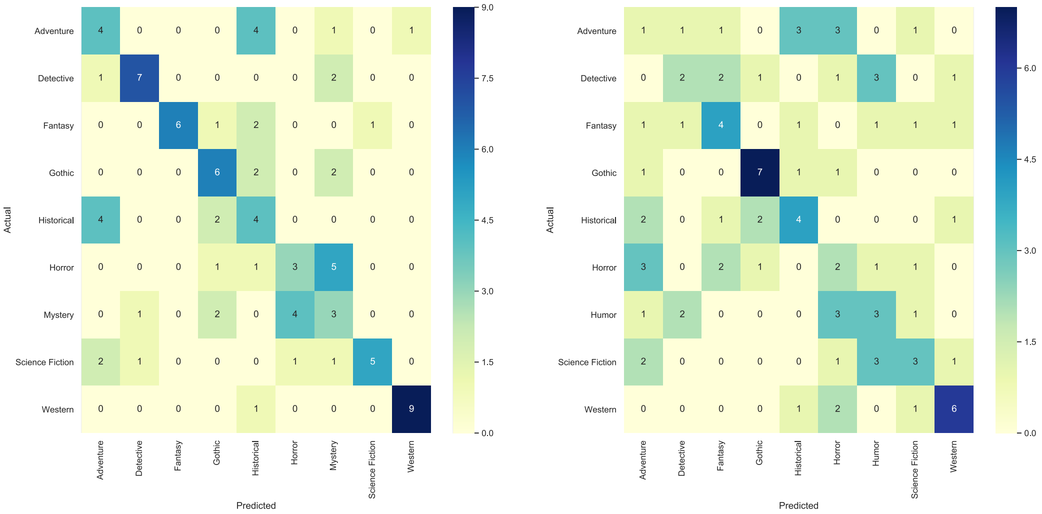


Figure 6.3: Confusion matrix for 9 genre, MLP classifier with network motif features (right) and MNB classifier with BoW features(left).

Table 6.3: Classification report for 9 genre with both MNB classifier with BoW and MLP classifier with network motif features.

Genre	**Precision**		**Recall**		**F1-score**	
	Text	Motif	Text	Motif	Text	Motif
Adventure	0.36	0.09	0.40	**0.10**	0.38	0.10
Detective	0.78	0.33	0.70	0.20	0.74	0.25
Fantasy	1.00	0.40	0.60	0.40	0.75	0.40
Gothic	0.50	0.64	0.60	0.70	0.55	0.67
Historical	0.29	0.40	0.40	0.40	0.33	0.40
Horror	0.38	0.15	0.30	0.20	0.33	0.17
Humor	-	0.27	-	0.30	-	0.29
Mystery	0.21	-	**0.30**	-	0.25	-
Science Fiction	0.83	0.38	0.50	0.30	0.62	0.33
Western	0.90	0.60	0.90	0.60	0.90	0.60
Average	0.58	0.36	0.52	0.36	0.54	0.36

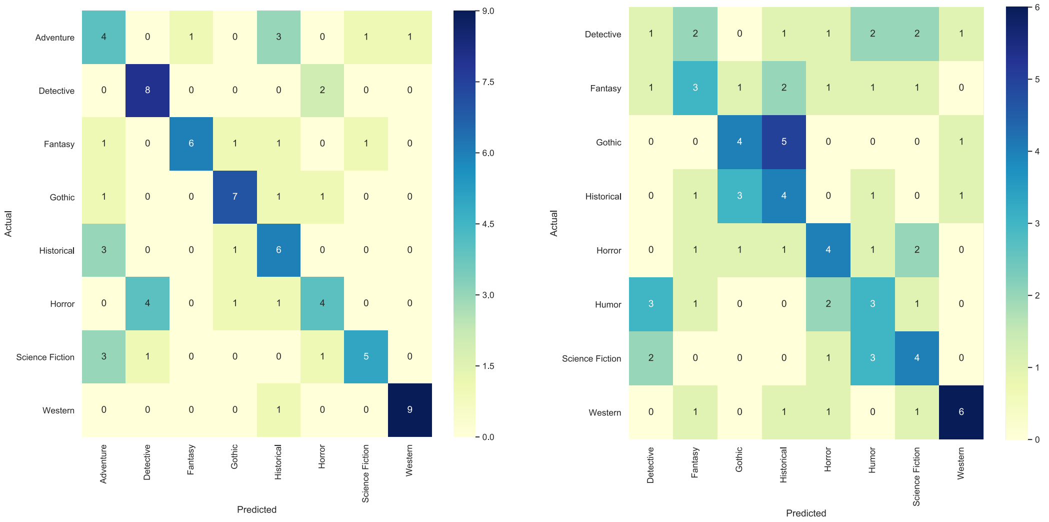


Figure 6.4: Confusion matrix for 8 genre, MLP classifier with network motif features (right) and MNB classifier with BoW features(left).

Table 6.4: Classification report for 8 genre with both MNB classifier with BoW and MLP classifier with network motif features.

Genre	**Precision**		**Recall**		**F1-score**	
	Text	Motif	Text	Motif	Text	Motif
Adventure	0.33	-	**0.40**	-	0.36	-
Detective	0.62	0.14	0.80	**0.10**	0.70	0.12
Fantasy	0.86	0.33	0.60	0.30	0.71	0.32
Gothic	0.70	0.44	0.70	0.40	0.70	0.42
Historical	0.46	0.29	0.60	0.40	0.52	0.33
Horror	0.50	0.40	0.40	0.40	0.44	0.40
Humor	-	0.27	-	0.30	-	0.29
Mystery	-	-	-	-	-	-
Science Fiction	0.71	0.36	0.50	0.40	0.59	0.38
Western	0.90	0.67	0.90	0.60	0.90	0.63
Average	0.64	0.36	0.61	0.36	0.61	0.36

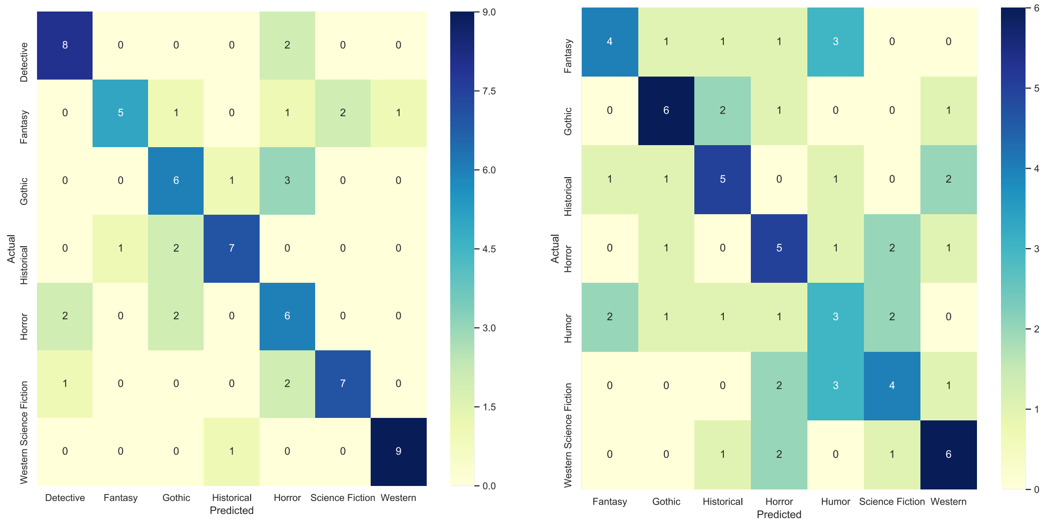


Figure 6.5: Confusion matrix for 7 genre, MLP classifier with network motif features (right) and MNB classifier with BoW features(left).

Table 6.5: Classification report for 7 genre with both MNB classifier with BoW and MLP classifier with network motif features.

Genre	**Precision**		**Recall**		**F1-score**	
	Text	Motif	Text	Motif	Text	Motif
Adventure	-	-	-	-	-	-
Detective	0.73	-	0.80	-	0.76	-
Fantasy	0.83	0.57	**0.50**	0.40	0.62	0.47
Gothic	0.55	0.60	0.60	0.60	0.57	0.60
Historical	0.78	0.50	0.70	0.50	0.74	0.50
Horror	0.43	0.42	0.60	0.50	0.50	0.45
Humor	-	0.27	-	**0.30**	-	0.29
Mystery	-	-	-	-	-	-
Science Fiction	0.78	0.44	0.70	0.40	0.74	0.42
Western	0.90	0.55	0.90	0.60	0.90	0.57
Average	0.71	0.48	0.69	0.47	0.69	0.47

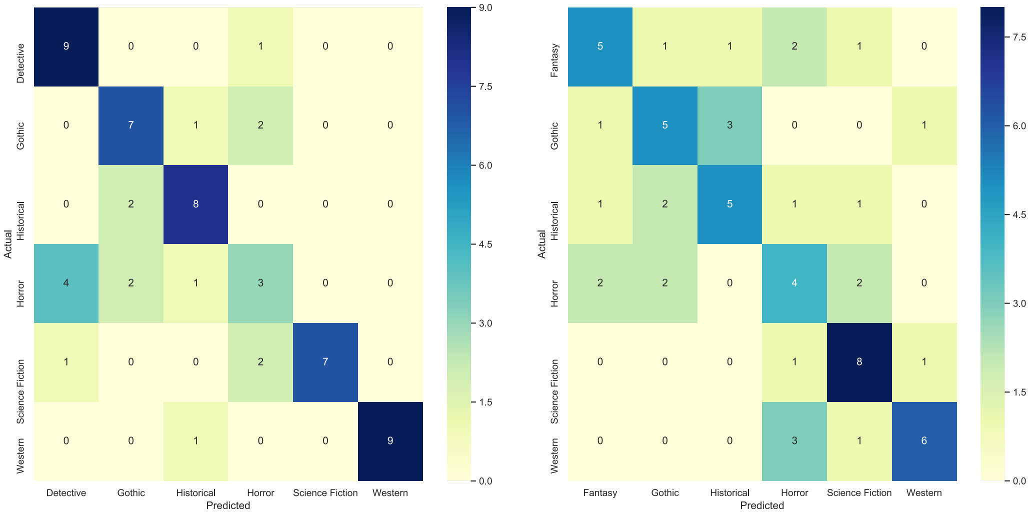


Figure 6.6: Confusion matrix for 6 genre, MLP classifier with network motif features (right) and MNB classifier with BoW features(left).

Table 6.6: Classification report for 6 genre with both MNB classifier with BoW and MLP classifier with network motif features.

Genre	**Precision**		**Recall**		**F1-score**	
	Text	Motif	Text	Motif	Text	Motif
Adventure	-	-	-	-	-	-
Detective	0.64	-	0.90	-	0.75	-
Fantasy	-	0.56	-	0.50	-	0.53
Gothic	0.64	0.50	0.70	0.50	0.67	0.50
Historical	0.73	0.56	0.80	0.50	0.76	0.53
Horror	0.38	0.36	**0.30**	**0.40**	0.33	0.38
Humor	-	-	-	-	-	-
Mystery	-	-	-	-	-	-
Science Fiction	1.00	0.62	0.70	0.80	0.82	0.70
Western	1.00	0.75	0.90	0.60	0.95	0.67
Average	0.73	0.56	0.72	0.55	0.71	0.55

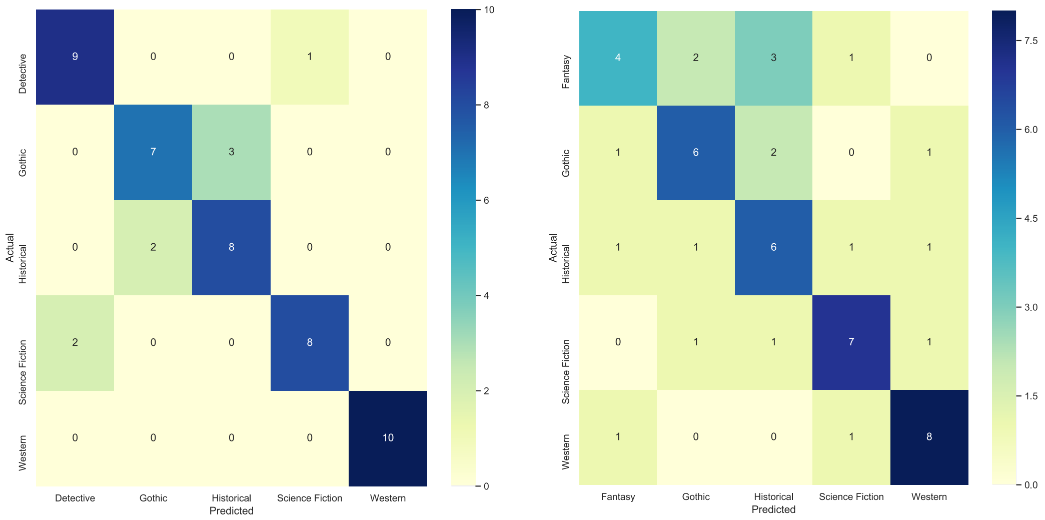


Figure 6.7: Confusion matrix for 5 genre, MLP classifier with network motif features (right) and MNB classifier with BoW features(left).

Table 6.7: Classification report for 5 genre with both MNB classifier with BoW and MLP classifier with network motif features.

Genre	**Precision**		**Recall**		**F1-score**	
	Text	Motif	Text	Motif	Text	Motif
Adventure	-	-	-	-	-	-
Detective	0.82	-	0.90	-	0.86	-
Fantasy	-	0.57	-	**0.40**	-	0.47
Gothic	0.78	0.60	**0.70**	0.60	0.74	0.60
Historical	0.73	0.50	0.80	0.60	0.76	0.55
Horror	-	-	-	-	-	-
Humor	-	-	-	-	-	-
Mystery	-	-	-	-	-	-
Science Fiction	0.89	0.70	0.80	0.70	0.84	0.70
Western	1.00	0.73	1.00	0.80	1.00	0.76
Average	0.84	0.62	0.84	0.62	0.84	0.62

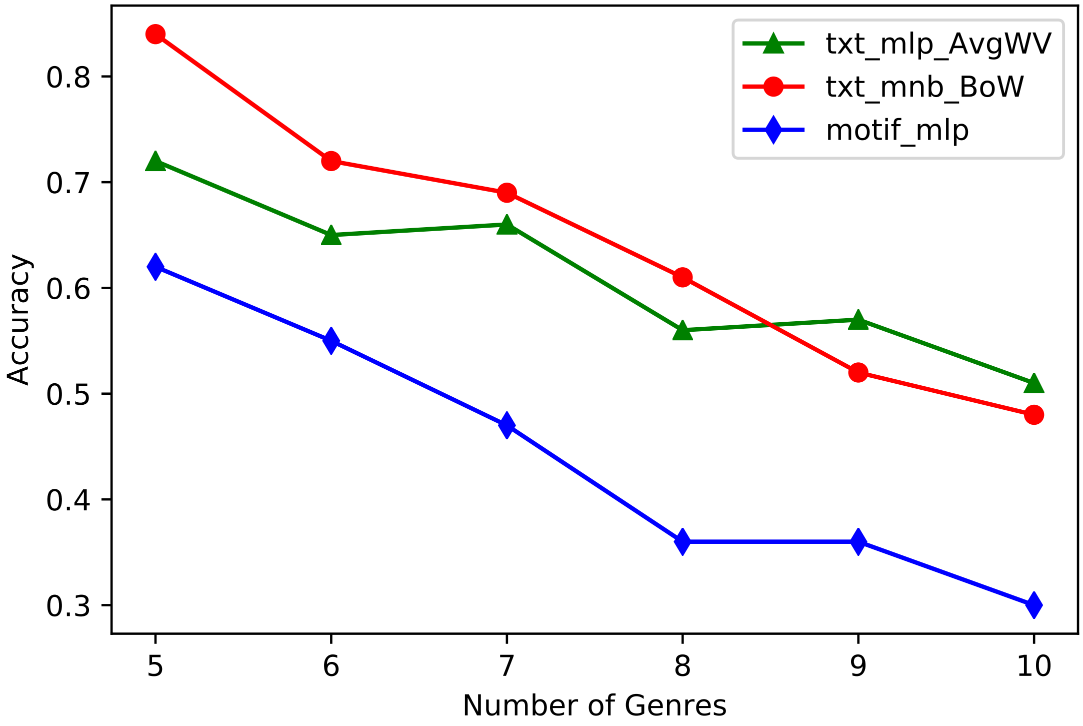


Figure 6.8: Accuracy vs Genres, multi-layer perceptrons (MLP) with network motif features (blue) and text feature for both Multinomial Naïve Bayes (MNB) with BoW features (red) and MLP with AvgWV features (green) classifiers.

Chapter 7

Conclusions, Limitations, and Future Works

The understanding of languages and their characterization has again become a topic of interest for the scientific community. Studies using a large amount of data may be able to provide a different view of how languages related to one another as well as see possible trends of influences of one over the other.

In this dissertation first we look at the possibility of characterizing written language solely from the point of view of structural features. We concentrated on two class of features: Heaps' law, that looks at richness of vocabulary in a language, and Network Science features extracted from the construction of word co-occurrence networks. In the process of extracting network features we also demonstrated that these networks exhibit both scale-free and small-world properties.

We then used K-means and hierarchical clustering together with the silhouette and elbow methods to identify the optimal number of language clusters to the dataset we have. We showed that the hierarchical clustering distinguish relationships between languages sub-families, while K-Means clusters languages based on their main genetic families (Proto-Families). It has been shown that the Heaps' law parameters enhanced the classification process by distinguishing languages based on their vocabulary richness.

Then, we went deeper in the characterization of of languages by augmenting the number of languages we use from 10 to around 20 languages. The difficulty is to find good corpora that includes this number of languages. We used the topological measurements extracted from word co-occurrence networks of 20 Indo-European languages along Heaps' law parameters to construct the hierarchical cluster that represent the chronological distance between those languages. The comparison that we made of our results with the glottochronological classification based on the lexical distance between word fluctuation among different languages, shows a strong agreement between the two methods. In order to support this finding, we test the tolerance of the cluster against languages variation. We did this by removing one language at a time and calculate the entanglement. Also, we extracted the best features that give the lowest entanglement, these features we believe they best describe the chronological difference between languages.

The results we get from this work open the door for many future works, for instance, we could expand our study to include languages from different main families. Also, it is possible to apply our method to find the closest translation of document to the original text in order to assets the quality of translation.

We believe structural analysis of written language could be used in identification of literary styles or even author analysis. We performed a similar analysis for several authors and tried to understand if authors have a structural fingerprint in their writing style that can be identified. We attempted to attribute 10 authors of

100 books using 4-nodes directed network motifs. Functional words (stop words) were kept during text pre-processing as they proven by many previous works to reflect author style and increase the accuracy of attributing authors. The results we obtained herein outperformed other works when network motifs were the only feature used in attributing authors. Also, the number of 100 books used in this work are much higher than other works, which statistically means if we used the same smaller dataset, we will get better classification accuracy. This proves the importance of network motifs in recognizing the variety of writing styles among different authors.

This opens the door for future work to generalize this method in translation assessment and whether authors fingerprint resist the translations of their texts. Other possibilities are to study the effect of extracting higher motif order on the accuracy of classification.

Finally, we generalize this method in attributing text from 10 different literature genres. The 100 books from Gutenberg online library project were converted into co-occurrence text network and then 4-nodes directed network motifs were extracted and used as feature vectors to perform classification. The classification accuracy was lower than the one obtained from author attribution due to the lower granularity of the genres and the different writing style of the book's authors chosen for our dataset. To further justify our results, we used traditional text classification methods with different feature extraction methods to compare and contrast with results obtained from the network motif method. We observed slight enhancements in classification accuracy, so our next step was to remove genres with low classification recall and perform the classification process until we get a statistically significant accuracy to assist our hypothesis of the degradation in classification accuracy due to the lower granularity of the genres. The results assist our assumption that network motifs reflect the author and their text style. In the future, we can further investigate the effect of genres granularity on the classification accuracy by building the dataset from higher granular genres.

Bibliography

[1] Olga Abramov and Alexander Mehler. Automatic language classification by means of syntactic dependency networks. *Journal of Quantitative Linguistics*, 18(4):291–336, 2011.

[2] Camilo Akimushkin, Diego Raphael Amancio, and Osvaldo Novais Oliveira Jr. Text authorship identified using the dynamics of word co-occurrence networks. *PloS one*, 12(1):e0170527, 2017.

[3] Younis Al Rozz, Harith Hamoodat, and Ronaldo Menezes. Characterization of written languages using structural features from common corpora. In *Workshop on Complex Networks CompleNet*, pages 161–173. Springer, 2017.

[4] Younis Al Rozz and Ronaldo Menezes. Author attribution using network motifs. In *International Workshop on Complex Networks*, pages 199–207. Springer, 2018.

[5] Diego R Amancio, Lucas Antiqueira, Thiago AS Pardo, Luciano da F. Costa, Osvaldo N Oliveira Jr, and Maria GV Nunes. Complex networks analysis of manual and machine translations. *International Journal of Modern Physics C*, 19(04):583–598, 2008.

[6] Diego Raphael Amancio. A complex network approach to stylometry. *PloS one*, 10(8):e0136076, 2015.

[7] Lucas Antiqueira, Osvaldo N Oliveira, Luciano da Fontoura Costa, and Maria das Graças Volpe Nunes. A complex network approach to text summarization. *Information Sciences*, 179(5):584–599, 2009.

[8] Samuel Arbesman, Steven H Strogatz, and Michael S Vitevitch. The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos*, 20(03):679–685, 2010.

[9] Ahmed Shamsul Arefin, Renato Vimieiro, Carlos Riveros, Hugh Craig, and Pablo Moscato. An information theoretic clustering approach for unveiling authorship affinities in shakespearean era plays and poems. *PloS one*, 9(10):e111445, 2014.

[10] Alex Arenas, Leon Danon, Albert Diaz-Guilera, Pablo M Gleiser, and Roger Guimera. Community analysis in social networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):373–380, 2004.

[11] Kristina Ban, Ana Meštrović, and A Martinčić-ipšić. Initial comparison of linguistic networks measures for parallel texts. In *5th International Conference on Information Technologies and Information Society (ITIS), 97104*. Citeseer, 2013.

[12] Nicole M Beckage and Eliana Colunga. Language networks as models of cognition: Understanding cognition through language. In *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, pages 3–28. Springer, 2016.

[13] Douglas Biber. Spoken and written textual dimensions in english: Resolving the contradictory findings. *Language*, pages 384–414, 1986.

[14] Douglas Biber. *Variation across speech and writing*. Cambridge University Press, 1991.

[15] Balthasar Bickel. Typology in the 21st century: major current developments. *Linguistic Typology*, 11(1):239–251, 2007.

[16] Chris Biemann, Lachezar Krumov, Stefanie Roos, and Karsten Weihe. Network motifs are a powerful tool for semantic distinction. In *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, pages 83–105. Springer, 2016.

[17] Christian Biemann, Stefan Bordag, Gerhard Heyer, Uwe Quasthoff, and Christian Wolff. Language-independent methods for compiling monolingual lexical data. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 217–228. Springer, 2004.

[18] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[19] Johan J Bolhuis, Ian Tattersall, Noam Chomsky, and Robert C Berwick. How could language have evolved? *PLoS biology*, 12(8):e1001934, 2014.

[20] Stephen P Borgatti. Centrality and network flow. *Social networks*, 27(1):55–71, 2005.

[21] Amiangshu Bosu and Jeffrey C Carver. How do social interaction networks influence peer impressions formation? a case study. In *IFIP International Conference on Open Source Systems*, pages 31–40. Springer, 2014.

[22] Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4):543–565, 1995.

[23] Josephine Jill T Cabatbat, Jica P Monsanto, and Giovanni A Tapang. Preserved network metrics across translated texts. *International Journal of Modern Physics C*, 25(02):1350092, 2014.

[24] Xiaoling Chen, Peng Hao, Rajarathnam Chandramouli, and KP Subbalakshmi. Authorship similarity detection from email messages. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 375–386. Springer, 2011.

[25] Xinying Chen and Haitao Liu. Function nodes in chinese syntactic networks. In *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, pages 187–201. Springer, 2016.

[26] Monojit Choudhury and Animesh Mukherjee. The structure and dynamics of linguistic networks. In *Dynamics on and of Complex Networks*, pages 145–166. Springer, 2009.

[27] Monojit Choudhury, Markose Thomas, Animesh Mukherjee, Anupam Basu, and Niloy Ganguly. How difficult is it to develop a perfect spell-checker? a cross-linguistic analysis through complex network approach. *TextGraphs-2: Graph-Based Algorithms for Natural Language Processing*, page 81, 2007.

[28] Jin Cong and Haitao Liu. Approaching human language with complex networks. *Physics of life reviews*, 11(4):598–618, 2014.

[29] Luigi P Cordella, Pasquale Foggia, Carlo Sansone, and Mario Vento. A (sub) graph isomorphism algorithm for matching large graphs. *IEEE transactions on pattern analysis and machine intelligence*, 26(10):1367–1372, 2004.

[30] Luigi Pietro Cordella, Pasquale Foggia, Carlo Sansone, and Mario Vento. An improved algorithm for matching large graphs. In *3rd IAPR-TC15 workshop on graph-based representations in pattern recognition*, pages 149–159, 2001.

[31] Florian Coulmas. *The writing systems of the world.* B. Blackwell, 1989.

[32] Henrique F. de Arruda, Luciano da F. Costa, and Diego R. Amancio. Topic segmentation via community detection in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(6):063120, 2016.

[33] Deutschland and Statistisches Bundesamt Deutschland. *Statistisches Jahrbuch Deutschland und Internationales.* Statistisches Bundesamt, 2012.

[34] Douglas Douglas. The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, 26(5-6):331–345, 1992.

[35] Isidore Dyen, Joseph B Kruskal, and Paul Black. File ie-data1, 1997.

[36] Geir Farner. *Literary fiction: The ways we read narrative literature.* Bloomsbury Publishing USA, 2014.

[37] Aidan Finn and Nicholas Kushmerick. Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57(11):1506–1518, 2006.

[38] Francesc Font-Clos, Gemma Boleda, and Álvaro Corral. A scaling law beyond zipf's law and its relation to heaps' law. *New Journal of Physics*, 15(9):093033, 2013.

[39] Yuyang Gao, Wei Liang, Yuming Shi, and Qiuling Huang. Comparison of directed and weighted co-occurrence networks of six languages. *Physica A: Statistical Mechanics and its Applications*, 393:579–589, 2014.

[40] Ali Hashemi Gheinani, Bernhard Kiss, Felix Moltzahn, Irene Keller, Rémy Bruggmann, Hubert Rehrauer, Catharine Aquino Fournier, Fiona C Burkhard, and Katia Monastyrskaya. Characterization of mirna-regulated networks, hubs of signaling, and biomarkers in obstruction-induced bladder dysfunction. *JCI insight*, 2(2), 2017.

[41] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, pages 759–765, 2012.

[42] Russell D Gray and Quentin D Atkinson. *Language-tree divergence times support the Anatolian theory of Indo-European origin*, volume 426. Nature Publishing Group, 2003.

[43] Russell D Gray, Quentin D Atkinson, and Simon J Greenhill. Language evolution and human history: what a difference a date makes. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 366(1567):1090–1100, 2011.

[44] Simon J Greenhill, Robert Blust, and Russell D Gray. The austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary bioinformatics online*, 4:271, 2008.

[45] D Gunlycke, VM Kendon, V Vedral, and S Bose. Thermal concurrence mixing in a one-dimensional ising model. *Physical Review A*, 64(4):042302, 2001.

[46] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[47] Harith Hamoodat, Younis Al Rozz, and Ronaldo Menezes. Complex networks reveal a glottochronological classification of natural languages. In *International Workshop on Complex Networks*, pages 209–219. Springer, 2018.

[48] Gustav Herdan. *Type-token mathematics*, volume 4. Mouton, 1960.

[49] Ramon Ferrer i Cancho. The structure of syntactic dependency networks: insights from recent advances in network theory. *Problems of quantitative linguistics*, pages 60–75, 2005.

[50] Chuntao Jiang, Frans Coenen, Robert Sanderson, and Michele Zito. Text classification using graph mining-based feature extraction. In *Research and Development in Intelligent Systems XXVI*, pages 21–34. Springer, 2010.

[51] Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[52] Jussi Karlgren and Douglass Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th conference on Computational linguistics-Volume 2*, pages 1071–1075. Association for Computational Linguistics, 1994.

[53] Jinyun Ke. Complex networks and human language. *arXiv preprint cs/0701135*, 2007.

[54] Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. Automatic detection of text genre. *arXiv preprint cmp-lg/9707002*, 1997.

[55] Andre Leone, Marcello Tomasini, Younis Al Rozz, and Ronaldo Menezes. On the performance of network science metrics as long-term investment strategies in stock markets. In *International Conference on Complex Networks and their Applications*, pages 1053–1064. Springer, 2017.

[56] Jianyu Li, Feng Xiao, Jie Zhou, and Zhanxin Yang. Motifs and motif generalization in chinese word networks. *Procedia Computer Science*, 9:550–556, 2012.

[57] Pedro G Lind, Marta C Gonzalez, and Hans J Herrmann. Cycles and clustering in bipartite networks. *Physical review E*, 72(5):056127, 2005.

[58] Haitao Liu and Jin Cong. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58(10):1139–1144, 2013.

[59] Haitao Liu and Chunshan Xu. Can syntactic networks indicate morphological complexity of a language? *EPL (Europhysics Letters)*, 93(2):28005, 2011.

[60] Linyuan Lü, Zi-Ke Zhang, and Tao Zhou. Zipf's law leads to heaps' law: Analyzing their relation in finite-size systems. *PloS one*, 5(12):e14139, 2010.

[61] Fragkiskos D Malliaros and Konstantinos Skianis. Graph-based term weighting for text categorization. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1473–1479. ACM, 2015.

[62] Nuno Mamede, José Correia, and Jorge Baptista. Syntax deep explorer. In *Computational Processing of the Portuguese Language: 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016, Proceedings*, volume 9727, page 189. Springer, 2016.

[63] Vanessa Queiroz Marinho, Graeme Hirst, and Diego Raphael Amancio. Authorship attribution via network motifs identification. In *Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on*, pages 355–360. IEEE, 2016.

[64] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[65] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

[66] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[67] Frederick Mosteller and David L Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

[68] Ahmed Ragab Nabhan and Khaled Shaalan. A graph-based approach to text genre analysis. *Computación y Sistemas*, 20(3):527–539, 2016.

[69] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[70] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[71] G. Nunberg. *The Linguistics of Punctuation*. CSLI lecture notes. Cambridge University Press, 1990.

[72] Petya Osenova. Bulgarian. *Revue belge de philologie et d'histoire*, 88(3):643–668, 2010.

[73] Colin Renfrew, April McMahon, and Robert Lawrence Trask. *Time depth in historical linguistics*. The Macdonald Institute for Archaelogical Research, 2000.

[74] Hana Rizvić, Sanda Martinčić-Ipšić, and Ana Meštrović. Network motifs analysis of croatian literature. *arXiv preprint arXiv:1411.4960*, 2014.

[75] Anderson Rocha, Walter J Scheirer, Christopher W Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne RB Carvalho, and Efstathios Stamatatos. Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, 12(1):5–33, 2017.

[76] François Rousseau, Emmanouil Kiagias, and Michalis Vazirgiannis. Text categorization as a graph classification problem. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1702–1712, 2015.

[77] Dipanjan Sarkar. *Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from your Data.* Apress, 2016.

[78] Thomas Schank and Dorothea Wagner. *Approximating clustering-coefficient and transitivity.* Universität Karlsruhe, Fakultät für Informatik, 2004.

[79] Santiago Segarra, Mark Eisen, and Alejandro Ribeiro. Authorship attribution through function word adjacency networks. *IEEE transactions on signal processing*, 63(20):5464–5478, 2015.

[80] Maurizio Serva and Filippo Petroni. Indo-european languages tree by levenshtein distance. *EPL (Europhysics Letters)*, 81(6):68005, 2008.

[81] Cynthia SQ Siew. Community structure in the phonological network. *Frontiers in psychology*, 4:553, 2013.

[82] Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[83] M Medeiros Soares, G Corso, and LS Lucena. The network of syllables in portuguese. *Physica A: Statistical Mechanics and its Applications*, 355(2):678–684, 2005.

[84] Ricard V Solé, Bernat Corominas-Murtra, Sergi Valverde, and Luc Steels. Language networks: Their structure, function, and evolution. *Complexity*, 15(6):20–26, 2010.

[85] Jae Jung Song. *The Oxford Handbook of Linguistic Typology*. Oxford University Press, 2010.

[86] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, 60(3):538–556, 2009.

[87] Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 808–814. Association for Computational Linguistics, 2000.

[88] Mark Steyvers and Joshua B Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78, 2005.

[89] Ngoc Tam L Tran, Luke DeLuccia, Aidan F McDonald, and Chun-Hsi Huang. Cross-disciplinary detection and analysis of network motifs. *Bioinformatics and Biology insights*, 9:49, 2015.

[90] Mark PJ Van der Loo. The stringdist package for approximate string matching. *The R*, page 2, 2014.

[91] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.