Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2005-077
AITR-2005-004

November 18, 2005

# Identifying Expression Fingerprints Using Linguistic Information

Ozlem Uzuner

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

# Identifying Expression Fingerprints using Linguistic Information

by

Özlem Uzuner

Submitted to the Engineering Systems Division
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2005

© Massachusetts Institute of Technology 2005. All rights reserved.

Author .....
Engineering Systems Division
January 31, 2005

Certified by .....
Randall Davis
Professor of Computer Science
Computer Science and Artificial Intelligence Laboratory
Thesis Supervisor

Certified by .....
Boris Katz
Principal Research Scientist
Computer Science and Artificial Intelligence Laboratory
Thesis Supervisor

Certified by .....
Lee W. McKnight
Associate Professor
School of Information Studies
Syracuse University
Committee Member

Certified by .....
Frank R. Field III
Senior Research Associate
Engineering Systems Division
Committee Member

Accepted by .....
Richard de Neufville
Professor of Engineering Systems
Chair, Engineering Systems Division Education Committee

**Identifying Expression Fingerprints using Linguistic Information**

by

Özlem Uzuner

Submitted to the Engineering Systems Division
on January 31, 2005, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

## Abstract

This thesis presents a technology to complement taxation-based policy proposals aimed at addressing the digital copyright problem. The approach presented facilitates identification of intellectual property using expression fingerprints.

Copyright law protects expression of content. Recognizing literary works for copyright protection requires identification of the expression of their content. The expression fingerprints described in this thesis use a novel set of linguistic features that capture both the content presented in documents and the manner of expression used in conveying this content. These fingerprints consist of both syntactic and semantic elements of language. Examples of the syntactic elements of expression include structures of embedding and embedded verb phrases. The semantic elements of expression consist of high-level, broad semantic categories.

Syntactic and semantic elements of expression enable generation of models that correctly identify books and their paraphrases 82% of the time, providing a significant (approximately 18%) improvement over models that use tfidf-weighted keywords. The performance of models built with these features is also better than models created with standard features used in stylometry (e.g., function words), which yield an accuracy of 62%.

In the non-digital world, copyright holders collect revenues by controlling distribution of their works. Current approaches to the digital copyright problem attempt to provide copyright holders with the same kind of control over distribution by employing Digital Rights Management (DRM) systems. However, DRM systems also enable copyright holders to control and limit fair use, to inhibit others' speech, and to collect private information about individual users of digital works.

Digital tracking technologies enable alternate solutions to the digital copyright problem; some of these solutions can protect creative incentives of copyright holders in the absence of control over distribution of works. Expression fingerprints facilitate digital tracking even when literary works are DRM- and watermark-free, and even when they are paraphrased. As such, they enable metering popularity of works and make practicable solutions that encourage large-scale dissemination and unrestricted use of digital works and that protect the revenues of copyright holders, for example through taxation-based revenue collection and distribution systems, without imposing limits on distribution.

Thesis Supervisor: Randall Davis
Title: Professor of Computer Science
Computer Science and Artificial Intelligence Laboratory

Thesis Supervisor: Boris Katz
Title: Principal Research Scientist
Computer Science and Artificial Intelligence Laboratory

## Acknowledgments

I am grateful to my advisors Prof. Randall Davis, Dr. Boris Katz, Prof. Lee McKnight, and Dr. Frank Field for sharing with me their wisdom and for providing me with support. I am thankful to my friends Gregory Marton, Kate Saenko, Vineet Sinha, Jacob Eisenstein, Metin Sezgin, Mario Christoudias, Jason Rennie, Federico Mora, and Juhi Chandalia for their feedback and for always making me smile. I am thankful to Sue Felshin for all her help and for being a wonderful officemate. I am blessed with amazing parents and an amazing brother—I dedicate this thesis to them.

# Contents


**1 Introduction** **15**

1.1 Background . . . . . 15

1.2 Proposed Solutions . . . . . 20

1.3 The Promise of Expression Fingerprints . . . . . 22

**2 Digital Copyright Problem** **23**

2.1 Copyright Law and its Goals . . . . . 27

2.2 Mesh of Stakeholders and Copyright Concerns . . . . . 32

2.2.1 Copyright Holders . . . . . 32

2.2.2 Digital Copyright Problem and Reactions of Copyright Holders . . . . . 35

2.2.3 Users . . . . . 36

2.2.4 Digital Copyright Problem, DRM systems, DMCA, and Effects on Users . 38

2.2.5 Innovators and Technology Producers . . . . . 42

2.2.6 Rights Defenders . . . . . 45

2.2.7 Policy Makers: the Congress and Courts . . . . . 46

2.3 Net Effects of DRM systems and Supporting Legislation . . . . . 47

2.4 Requirements for a Promotion of Progress . . . . . 48

2.5 Conclusion . . . . . 49

2.6 Summary . . . . . 50

**3 Technology and Policy Solutions** **51**

3.1 Technologies . . . . . 53

3.1.1 Technologies for Controlling Use . . . . . 53

3.1.2 Digital Tracking Technologies . . . . . 56

3.2 Technology & Policy Solutions . . . . . 59

3.2.1 Trusted Third Party . . . 60
3.2.2 Creative Commons . . . 61
3.2.3 Taxation-Based Solutions . . . 62
3.3 Lessons Learned . . . 66
3.4 Digital Tracking Using Expression Fingerprints . . . 67
3.5 Conclusion . . . 69
3.6 Summary . . . 70

**4 Content, Expression, and Style** **71**

**5 Preliminary Experiments** **75**
5.1 Features from the Literature . . . 76
5.1.1 Surface Features . . . 76
5.1.2 Syntactic Features . . . 76
5.1.3 Semantic Features . . . 77
5.2 Methods . . . 79
5.2.1 Decision trees . . . 79
5.2.2 Boosting . . . 81
5.2.3 t-Test . . . 83
5.3 Experiments . . . 84
5.3.1 Classification Experiments . . . 84
5.3.2 Significance Testing for Feature Ranking . . . 84
5.4 Conclusion . . . 88
5.5 Summary . . . 89

**6 Verb-Based Theory of Expression** **91**
6.1 Linguistic Complexity and Syntactic Repertoire . . . 91
6.2 Tests of Independence . . . 93
6.2.1 Pearsons' Chi-Square . . . 93
6.2.2 Likelihood Ratio Chi-Square . . . 95
6.3 Expression and Meaning . . . 97
6.4 Expression as a Function of Syntax and Structure . . . 99
6.5 Expression as a Function of Sentence Structure . . . 103

6.5.1 Expressive Use of Sentence-Initial and -Final Phrase Structures . . . . . 103
6.6 Expression as a Function of Sentence Complexity . . . . . 114
6.6.1 Sentence Complexity as a Function of Clause Structure . . . . . 114
6.6.2 Clause Structure and Expression . . . . . 115
6.7 Expression as a Function of Verb Phrase Structure . . . . . 125
6.7.1 Linguistic Richness . . . . . 125
6.7.2 Semantics of Verbs . . . . . 126
6.7.3 Syntax of Verbs . . . . . 130
6.8 Summary . . . . . 142

**7 Semantic Categories and Content** **143**
7.1 Content in the Literature . . . . . 143
7.2 General Inquirer (GI) Classes . . . . . 145
7.3 Parameter Tuning . . . . . 148
7.4 Evaluation of GI Categories for Capturing Content Similarity . . . . . 151
7.5 Conclusion . . . . . 156
7.6 Summary . . . . . 157

**8 Evaluation** **159**
8.1 Data . . . . . 160
8.2 Feature Set . . . . . 163
8.3 Baseline Features . . . . . 165
8.3.1 Baseline Features . . . . . 165
8.4 Linguistic Features . . . . . 169
8.5 Classification Experiments . . . . . 170
8.5.1 Recognizing Paraphrases (Recognizing Titles) . . . . . 170
8.5.2 Recognizing Expression (Recognizing Books) . . . . . 176
8.5.3 Recognizing Authors . . . . . 179
8.6 Conclusion . . . . . 182
8.7 Summary . . . . . 183

**9 Conclusion** **185**

**10 Contributions** **189**

**11 Future Work** **191**

11.1 Implementation of Digital Tracking with Expression Fingerprints . . . . . 191

11.2 Study of Linguistic Features for Other Applications . . . . . 194

**References** . . . . . 196

**Appendix** . . . . . 209

A-1 General Inquirer Semantic Categories . . . . . 209

A-2 Tables of Novels Used in Authorship Attribution . . . . . 215

# List of Tables


5.1 Cross-validation accuracy on recognizing the translators. The corpus contains 3 distinct titles translated by a total of 7 translators. . . . . 84

5.2 Ten most useful features for distinguishing between translators who translated the same content. . . . . 86

6.1 Hypothetical example of verb classes in two books. . . . . 93

6.2 A table template for presentation of chi-square results for testing independence of use of verb classes in two books. . . . . 95

6.3 Template for presentation of chi-square and likelihood ratio test results. . . . . 95

6.4 Examples of semantically equivalent sub-excerpts from excerpts 1, 2, and 3. . . . . 98

6.5 Examples of sentence-initial and -final structures from *Robinson Crusoe* by Daniel Defoe and *Crime and Punishment* by Leo Tolstoy. . . . . 109

6.6 Sentence-initial and -final phrase structures in *20000 Leagues*. . . . . 110

6.7 Chi-square test results for sentence-final phrase structures for *20000 Leagues* and *Madame Bovary*. . . . . 111

6.8 Chi-square test results for sentence-initial and -final phrase structures for various book pairs. . . . . 112

6.9 Sample sentences broken down into their clauses and the depth of the top-level subject and predicate, measured in terms of the depth of the lowest phrase and in terms of the number of clauses each branch contains. Top-level clause is always credited to the right branch. . . . . 122

6.10 Raw counts, percentages, and chi-square test results for left- and right-heavy clauses as measured by the depth of subjects and predicates for *20000 Leagues* and *Madame Bovary*. . . . . 123

6.11 Counts, percentages, and chi-square results for left- and right-heavy clauses of *Anna Karenina*. . . . . 124
6.12 Chi-square results for left- and right-heavy clauses for pairs of various books. . . . 124
6.13 Chi-square and likelihood ratio test results for semantic classes in parallel translations. 128
6.14 Examples of paraphrases from two translations of *20000 Leagues*. . . . . 129
6.15 Chi-square and likelihood ratio test results for semantic classes in pairs of various books chapters. . . . . 130
6.16 Chi-square and likelihood ratio test results for semantic verb classes paired with their syntactic alternations in parallel translations. . . . . 132
6.17 Chi-square and likelihood ratio test results for semantic verb classes paired with their syntactic alternations in various books. . . . . 132
6.18 Syntactic formulae and examples of embedding verb classes based on Kunz and Bridgeman [1, 15]. . . . . 139
6.19 Chi-square and likelihood ratio test results for classes of embedding verbs in parallel translations. . . . . 140
6.20 Chi-square and likelihood ratio test results for classes of embedding verbs in chapters from various books. . . . . 140

7.1 Nine senses of "make" and their GI categories. See appendix for an explanation of the categories. . . . . 146
7.2 Confusion matrix for tenfold cross-validation performance on the training set of three titles using boosted decision trees. . . . . 150
7.3 Confusion matrix for test performance of boosted decision trees (30 rounds of boosting with all GI categories whose information gain value was greater than zero) using semantic category information for content modelling with accuracy of 91%. The complete corpus contains 56 chapters from each title, 32 of which were used for training and parameter tuning. Results are reported on the 22 test samples from each title. . . . . 151
7.4 Top five GI categories, their descriptions from the Inquirer dictionary. . . . . 152
7.5 Test performance of GI categories and keywords on the corpus of 45 titles. . . . . 154
7.6 Test performance of GI categories and keywords on only the paraphrased titles (parallel translations) contained in the corpus of 45 titles. . . . . 155

8.1 Linguistic elements of expression, normalized for chapter length. . . . . 169
8.2 Corpus for experiments on recognizing titles (even when they are paraphrased) and books (expression). . . . . 171
8.3 Classification results on the complete test set for recognizing titles even when some titles are paraphrased. Train on 32 chapters from each title and test on 22 chapters. 172
8.4 Some of the most useful features for recognizing titles even when some titles are paraphrased. . . . . 173
8.5 Top ten syntactic elements of expression that recognize titles—in absence of GI categories. . . . . 174
8.6 Classification results only on the paraphrased titles included in the 45-title corpus. Random chance would recognize a paraphrased title 2% of the time. . . . . 175
8.7 Classification results on the test set for expression recognition even when some books contain similar content. . . . . 177
8.8 Top ten linguistic expression features that recognize books even when some books share content. . . . . 177
8.9 Expression recognition results on the test set for paraphrased books only. . . . . 178
8.10 Results for authorship attribution. Classifier is trained on 150 chapters from each author, and tested on 40 chapters from each author. The chapters in the training and test sets come from different titles. . . . . 180

A-1 Part 1 of data used for studying the style of authors. . . . . 215
A-2 Part 2 of data used of studying the style of authors. . . . . 216

# List of Figures


6-1 Subject-predicate structure. . . . . 114

6-2 Subject and predicate of the top-level clause. . . . . 115

6-3 Subject and predicate of the embedded clause. . . . . 115

6-4 Depths of phrases. . . . . 117

6-5 Depths of phrases. . . . . 118

6-6 Ternary tree for ditransitive verb. . . . . 119

6-7 Binary branching tree for sentence in Figure 6-6. . . . . 120

6-8 Binary tree for multiple prepositional phrases with VP-shells. . . . . 121

6-9 Binary tree structure for indirect question constructs. . . . . 121

6-10 Sentence from *Robinson Crusoe*. . . . . 122

6-11 Structure for “say” with and without ‘that” complementizer. . . . . 134

6-12 Structure for “wonder” with complementizer “whether”. . . . . 135

6-13 Structure for “want” + infinitive. . . . . 136

7-1 Translations of titles plotted against the GI categories “DIST” and “PLACE”. . . . 152

7-2 Translations of titles plotted against the GI categories “PowTot” and “DIST”. . . . 153

7-3 Translations of titles plotted against the GI categories “PowTot” and “PLACE”. . . 153

# Chapter 1

# Introduction

Expressive fingerprints of documents identify copyrighted works in the absence of labels, such as watermarks, and facilitate a range of tracking-based solutions to the digital copyright problem. These fingerprints rely on linguistically grounded features in order to differentiate between literary works and effectively recognize works even when they are paraphrased. In this thesis, we describe the identification and extraction of expression fingerprints of literary works, discuss features that capture expressive elements of language, and show that we can achieve more than 80% accuracy in identifying literary works and their paraphrases from their linguistic expression.

## 1.1 Background

Clause 8 of the United States Constitution grants the Congress the power "[t]o promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries".[1] Copyright is one of the legal instruments created under this clause in order to promote progress of science [126].

Copyright law achieves its goal by providing authors with limited-time monopoly rights which enable them to collect revenues from their creative works. Through these revenues, copyright law provides authors with incentives to create and disseminate works; under this system, authors are rewarded for their works only to the extent that they are willing disseminate and share these works with the public.

> The copyright law, like the patent statutes, makes reward to the owner a secondary consideration. In Fox Film Corp. v. Doyal, 286 U.S. 123, 127, Chief Justice Hughes spoke as follows

[1] U.S. Constitution. Article. I. §8, cl. 8.

> respecting the copyright monopoly granted by Congress, "The sole interest of the United States and the primary object in conferring the monopoly lie in the general benefits derived by the public from the labors of authors." It is said that reward to the author or artist serves to induce release to the public of the products of his creative genius.
>
> United States v. Paramount Pictures, Inc. 334 U.S. 131, 158 (1948).

In addition to providing authors with incentives to create and disseminate their works, the law protects the public interest in works and enables the public to access[2], use, and build on these works. Creation and dissemination of works by authors, and use of works by the public, together promote progress of science. Copyright law achieves this goal by making each of these activities possible: it grants authors the exclusive right to reproduce, distribute, publicly perform or display, and make derivatives of their works; it allows the public to access and use these works, sometimes even when their uses interfere with copyrights of authors; thus it balances the interests of authors with the ability of the public to use works.

In this balance, control over reproduction and distribution of works plays a significant role: by having the exclusive right to distribute their works, copyright holders control the market for their works. Once copyright holders choose to distribute their works publicly, the exclusive rights of copyright holders protect their revenues without unduly restricting the public's ability to access and use these works; conversely, the privileged uses the public can make of these works are limited to those activities that do not significantly interfere with the creative incentives of copyright holders.

The non-digital world (i.e., in the absence of digital equipment, media, and networks) reinforces the control over works by limiting distribution, and supports the balance established by the privileged uses: the non-digital equipment for mass reproduction and large-scale distribution of works is expensive and not widely available; most small-scale reproduction equipment produce copies whose quality degrade with each generation, i.e., copy of a copy is lower quality than copy of the original; and in both cases, the reproduction and distribution of copies require time, effort, and resources.

The increased availability of computers, computer networks, and other digital equipment and media, and the resulting so-called "digital world", have disturbed the balance between the rights of copyright holders and the rights (and privileges) of the public by eroding the existing barriers to large-scale reproduction and distribution. In particular, availability of digital works and equipment has minimized the cost of reproduction; and the Internet has minimized the cost of distribution [86].

[2] "The public has access to versions of a work that have been published and distributed, placed in publicly accessible collections (e.g., libraries), or otherwise made available through normal channels" [86]. "Having access" means being able to obtain a copy that one can use.

These developments hold great promise for the promotion of progress: they enable copyright holders to reach wider audiences at lower cost, and they enable more people to access and use works. But these technologies also enable individual users to reproduce and distribute digital works at almost no cost—even without authorization from and compensation to copyright holders. Given the high cost of creation and the relatively low cost of reproduction and distribution, to protect creative incentives of copyright holders in the digital world, the U.S. Congress enacted the Digital Millennium Copyright Act (DMCA). The DMCA implemented the requirements of two World Intellectual Property Organization (WIPO) treaties (WIPO Copyright Treaty (WCT) [134] and WIPO Performance and Phonograms Treaty (WPPT) [135]) and it authorized use of technological protection mechanisms, called Digital Rights Management (DRM) systems, that control and limit uses of digital works. In addition, the DMCA made it illegal to circumvent DRM systems and to traffic technologies that are designed primarily for circumvention.

However, DRM systems protect copyrights at the expense of the rights and privileges of users. They can be used to limit application of the first sale doctrine in the digital world and therefore limit low-cost access to works, e.g., through secondhand markets or through libraries [86]. They enable control over access and use of works, they enable limits on fair use [19], they can limit access to uncopyrightable aspects of works, e.g., ideas and public domain aspects of works, along with copyrightable aspects, and they can be used to limit free speech (beyond the limitations imposed on free speech by copyrights) [88]. In addition, they can be used to monitor uses of works and thus enable invasion of privacy [23]. The ability of the public to use works is affected by the possible limits on:

- Low-cost access to works,
- Fair use,
- Free speech, and
- Private "intellectual consumption" [23] of works.[3]

Control of copyright holders over these privileges disrupt the balance defined by the copyright law between the rights of authors and the rights of the public.

[3]Being able to access and read works anonymously and privately is important for "intellectual consumption". When they are subjected to supervision or surveillance by outside parties, people behave differently; they may not access works they would have otherwise preferred. As a result, invasion of privacy limits "intellectual freedoms" of users [23].

For promotion of progress, we argue that we need to re-establish a balance that:

- Protects the incentives, e.g., revenues, of copyright holders to create and disseminate works, and

- Enables the public to access and use works, through alternate low-cost distribution channels as well as distribution channels that rely on copyright holders directly, for fair use and for free speech without being subjected to supervision or surveillance by copyright holders and without threats to their privacy.

Lack of a balance between the creative incentives of copyright holders and the public interest in works creates tensions between these stakeholders and results in a chain of events that can limit promotion of progress:

- Litman reports that, according to the White Paper [56], in the absence of strong protections for their works, some creators may choose not to create [76], others may choose not to disseminate their works digitally. Both underproduction of works and avoidance of digital distribution of works limit publicly distributed knowledge, and to avoid this result, we argue that we need to create incentives (and eliminate the disincentives) to create and disseminate works.

- Strong protections, e.g., DRM systems that can impose limitations on access and use in order to prevent unauthorized uses, encourage copyright holders (under current business models) to take advantage of low-cost technologies for reproduction and dissemination of works, and enable them to recoup their investments in works by maintaining control over their copyrights (especially given the relatively high cost of creation and the relative low cost of distribution of digital works);

- DRM systems used for protecting copyrights in digitally published works can conflict with the rights and privileges of users, can get in the way of secondary distribution channels of works, can limit fair use and free speech, and can limit private "intellectual consumption" [23]. When limits are imposed on the privileges of the public, the protection systems are often circumvented (with or without authorization from copyright holders) and (sometimes) unprotected copies are created;

- Unauthorized distribution of thus created unprotected copies of works threatens creative incentives of copyright holders and signals to copyright holders the need for stronger protection mechanisms;

- Overprotection of digitally published works further threatens rights and privileges of users and encourages further efforts to circumvent protection systems thus motivating more draconian protection, and so on.

This tension, and the lack of a balance between the incentives of copyright holders and the interests of the public, affect many additional stakeholder groups, including libraries and innovators. The result of this tension can limit progress by affecting both the creative incentives of copyright holders and the ability of the public to access and use works.

## 1.2 Proposed Solutions

For promotion of progress, any solution to the digital copyright problem has to provide a balance that addresses the concerns of both parties by:

- Protecting the incentives of copyright holders to create and disseminate works, and
- Enabling the public to access and use works for fair use and for free speech without being subjected to supervision or surveillance by copyright holders and without threats to their privacy.

Such a balance can be established in several ways.

- Burk and Cohen have suggested remedying the effects of overreaching DRM systems by protecting fair use through modifications to DRM systems and through trusted third parties [19];
- In separate proposals, Fisher and Netanel have argued for immunity for individuals from infringement lawsuits for distribution and use of works on digital networks in return for compensation of copyright holders through taxes imposed on digital equipment and media that facilitate such activities [39, 87];
- Lessig et al. have provided standardized "Creative Commons" [22] licenses that enable copyright holders to distribute their works to the public for use and redistribution while reserving some rights, such as proper attribution and commercial gain.

Our analysis shows that each of the proposed solutions can potentially achieve the desired balance. However, solutions such as those proposed by Fisher and Netanel are particularly promising because they render unnecessary copyright holder supervision over distribution of works but guarantee returns to copyright holders. In these proposals, revenues are allocated to copyright holders (for digitally distributed works) in proportion to the value of their works for society; popularity of works, as measured by digital tracking and copy detection mechanisms, could provide a good proxy for measurement of this value [87].

In the last decades, different kinds of watermarks [92] and labeling information have been suggested for tracking digital works. However, watermarks can be removed. Fingerprints that are extracted by analyzing the content of works can identify copies accurately without having to rely on watermarks or other labels, but to date, copies have to exhibit verbatim similarity with the original in order to be recognized.

In this thesis, we present fingerprints that capture the expression of content in literary works and that can be used to recognize even paraphrased versions of works. Mechanisms that capture expression and recognize paraphrases enable accurate metering of use and distribution of works, and they make practicable compensation systems based on metering output. With such a system in place, unrestricted circulation and use of digital works can benefit both copyright holders and society at large: copyright holders can be compensated for distribution and use of their digital works (even when their works are distributed by others), and the public can easily access and use the works. Thus, we can reach an equilibrium where copyright holders have incentives to create and distribute their works (even on digital media), and the public has access to a diverse set of works that they can benefit from not only for fair use but also for free speech, without outside supervision and surveillance, and without concerns for their privacy.

## 1.3 The Promise of Expression Fingerprints

Evaluation of text similarity in terms of expression of content is a novel approach to text similarity evaluation. Our fingerprinting method identifies the expression presented in literary works, such as novels, and uses natural language processing techniques to recognize paraphrases, as well as the expression that is unique to each of these paraphrases, making it easy for copyright holders to identify copies of their works. This method identifies syntactic elements of expression by capturing the linguistic differences exhibited in the works of different authors who write about the same content, and relies on high-level semantic categories to capture the high-level context in which these syntactic elements appear. The syntactic elements of expression and the context in which they appear define the expression in a work and can be used to automatically identify works (or components of works) that share expression.

Evaluation of our fingerprinting approach shows that models based on syntactic and semantic elements of language accurately identify chapters from individual books more than 80% of the time, even in the presence of non-verbatim copies of books. The performance of these models is significantly better than the performance of standard approaches to text similarity evaluation that use keywords, function words, sentence lengths, and word lengths.

In Chapter 2 of this thesis, we present the digital copyright problem and outline the elements of conflict between the stakeholders involved in this debate. In Chapter 3, we discuss some of the proposed technology and policy solutions to the digital copyright problem, and present the benefits of digital tracking. In Chapter 4, we conceptually define *expression* in literary works.

Our preliminary experiments (Chapter 5) with linguistic features borrowed from text classification literature showed that syntactic characteristics of sentences can be useful for capturing expression; further exploration of syntax resulted in a novel set of linguistic elements which capture the expression in a work. The language models created by combining the syntactic elements of expression discussed in Chapter 6 and the semantic elements of expression discussed in Chapter 7 are evaluated in Chapter 8 and the conclusions and contributions of this thesis are summarized in Chapters 9 and 10. The thesis concludes with a discussion of open questions and future work presented in Chapter 11.

# Chapter 2

# Digital Copyright Problem

Copyright law takes its origin from the Statute of Anne of 1710, which is "[an] act for the encouragement of learning" [115]. Enacted in 1790, U.S. copyright law aims to promote progress by providing creators with incentives—in the form of monopoly powers—to produce and disseminate works, and by enabling public access to works so that the public can use these works and new creators can build on the past [22]. Note that, in this balance, "reward to the owner[s] [is] a secondary consideration" and "the primary object in conferring the monopoly lie[s] in the general benefits derived by the public from the labors of authors."[1] As a result, copyright law gives authors the exclusive right to control dissemination and enables them to extract revenues from their works to the extent that they disseminate their works.

Copyright is a static law in a dynamic environment—it assumes a set of technologies and a set of creative works, and it balances the incentives of creators with the public interest based on these assumptions. For example, non-digital media and equipment limit the possible uses of works; lack of availability and the cost of equipment used for large-scale reproduction and distribution limit copyright infringement by individuals and reinforce copyrights. Given the possible uses of works, exceptions to copyrights protect the public interest in copyrighted works: fair use doctrine enables some otherwise infringing public uses of works and the first sale doctrine allows access to works through secondhand markets and through libraries. As a result, the public can use copyrighted works for free speech, for scholarship, etc., and "intellectually consume" [23] works privately. This balance ensures that the exclusive rights of copyright holders protect their revenues without unduly restricting public access to works and that the authorized uses the public can make of works do not

[1] United States v. Paramount Pictures, Inc. 334 U.S. 131, 158 (1948).

interfere unduly with the creative incentives of copyright holders. However, due to its static nature, with each technological change that enables new creative works and supports new modes of using these works, "Copyright laws become obsolete" [76]. As a result, application of the existing law to new technologies disturbs the existing balance between the interests of copyright holders and users, and requires legal adjustments that aim to re-establish a balance.

Widespread adoption and use of computers, computer networks, digital equipment, and information infrastructure, and the resulting so-called digital world, have facilitated low-cost reproduction and distribution of digital works, and have enabled copyright holders to reach wider audiences easily and cheaply. This decrease in cost and increase in reach promise to contribute to promotion of progress by improving public access to more and varied works.

> No single technological change in the history of the American Republic has more profoundly affected the potential for democratic speech and the spread of knowledge"[2].

However, the adoption and use of the same digital technologies as part of daily activities of users have also enabled large-scale unauthorized reproduction and distribution of works by the public, threatening the revenues of copyright holders. In order to take advantage of newly available low-cost distribution mechanisms while protecting their revenues, copyright holders (whose business models are based on revenues obtained from copies of works) have focused on protecting their copyrights in digital works. In particular, many copyright holders have used Digital Rights Management (DRM) systems that can be programmed to limit access, distribution, and use, and allow only the specific uses that copyright holders authorize. These mechanisms can limit the distribution and use of digitally-published works through technological constraints, without much regard to public interest in these works. As a result, copyright holders can extend their control over digital works to uses not protected by the copyright law, enforce their rights (and wills) through technology and can limit the ability of the public to access and use DRM protected works. For example, DRM systems can be used to limit the practice of the right of first sale, and to limit the functions of libraries which rely on this doctrine to provide low-cost public access to works [74]. In addition, DRM systems enable limits on fair use [19] and free speech [88]. Finally, they can be programmed to track use by individuals and may invade privacy of users [23, 75]. The anti-circumvention provisions of the Digital Millennium Copyright Act (DMCA) make it illegal to bypass DRM systems and thus exacerbate these problems. The conflict between the interests of the copyright holders and users, and the

[2]Kahle v. Ashcroft. Civil Complaint for Declaratory Judgement. http://cyberlaw.stanford.edu/about/ cases/Civil Complaint 203-22-04.pdf

resulting tension lie at the core of what we refer to as the digital copyright problem.

The repercussions of the resulting tension affect society at large. Without adequate protection for their revenues, copyright holders may underproduce: they may lose the incentive to publish and disseminate their works in digital format; if they fail to recoup their investments in creation of works, they may lose the incentive to create, publish, and disseminate works completely [56]. Many useful works and knowledge may not be disseminated digitally (this can also limit future creative works that would be based on the deterred digital works), other works may not be disseminated at all. As a result, we fail to make the most of the progressive potential of the digital world. But, when the creative incentives of copyright holders are protected by DRM systems, users may be limited in their ability to access and use works, and end up circumventing DRM systems (and becoming lawbreakers under the current law) in order to use these works. The result, in all cases, limits progress.

In the absence of a solution that addresses the needs and concerns of both copyright holders and the public, the ability of each of the parties, to some extent, to take matters into their own hands by utilizing technological mechanisms complicates the digital copyright problem and affects many other stakeholders that include libraries and innovators.

In order to promote progress, a solution to the digital copyright problem should strike a balance between the incentives of copyright holders to disseminate their works and the public interest in these works. We believe revenues provide many copyright holders with the incentive to create and disseminate works, and so they should be guaranteed (at least to some degree). However, in addition to this guarantee, the interests of the public should also be protected. In particular, flow of knowledge and ideas through the public should be supported; the public should have access to works through low-cost distribution mechanisms, such as secondhand markets, and they should have access to works without having to purchase copies, for example, by borrowing copies from libraries. Provided with access to works, the public should be able to use works for fair use, for free speech, and for scholarship and learning. They should be able to make all these uses in private, without having to ask for authorization from copyright holders, and without outside surveillance and supervision. Access to wide range of ideas and knowledge lies at the heart of promotion of progress. Providing copyright holders with incentives to create and distribute works, therefore, is necessary for promotion of progress. Secondhand markets and libraries widen the audience base for these works by enabling low-cost access to these works, enabling more people to benefit from them. Fair use supports some secondary markets, by allowing loans between users, and enables

distribution of ideas and speech regarding available works. Free speech, such as criticism and parody, are also supported by fair use, and some of these uses are only possible in privacy and anonymity, e.g., critiques and parody-makers may prefer to remain anonymous. Each of these mechanisms contribute to promotion of progress, by enabling dissemination of ideas to the public and by supporting the ability of the public to access and use these works.

A solution that satisfies these parameters provides a balance between the incentives of copyright holders and the rights and privileges of the public, and can promote progress by supporting and encouraging flow of knowledge and ideas to and through the public. While being the main foci of our analyses, these parameters do not comprise a comprehensive set of concerns related to the digital copyright problem worldwide. "Moral rights" of artists [99], for example, deserve attention but are beyond the scope of this thesis.

In this chapter, we present the state of affairs regarding the digital copyright problem from the perspective of major stakeholder groups that include copyright holders, users, technology producers, rights defenders, and policy makers. This presentation of stakeholders and their interests exposes the lack of balance between the incentives of copyright holders and the interests of the public due to drastic measures taken by copyright holders to protect their digitally published works and revenues. In particular, these drastic measures limit the application to digitally-published works of the first sale doctrine, invade privacy of users, and affect the public's ability to use digitally-published works for fair use and for free speech.

We follow this analysis, in Chapter 3, with a study of a set of technology and policy proposals that aim to address the digital copyright problem. We evaluate several promising proposals from the perspective of the goals we have set out to achieve: balancing copyright holder incentives with the public good in the digital world, by protecting revenues of copyright holders, by securing public access to works, by protecting fair use, by supporting freedom of speech, and by limiting invasion of privacy of users. Through our analysis, we conclude that the solution that best fits our goals is provided by digital tracking mechanisms based on expressive fingerprints of copyrighted works, and policy proposals that compensate copyright holders in proportion to popularity of their works, with revenues collected through taxes, in return for immunity to users for distribution and use of works.

## 2.1 Copyright Law and its Goals

U.S. Copyright law takes its origin from the Statute of Anne of 1710. To promote "the Progress of Science and the useful Arts"[3], the first enactment of the U.S. copyright law in 1790 reserved to authors of "maps, charts and books" the "sole right and liberty of printing, reprinting, publishing, and vending" their works for a limited time of 14 years [71]. This law protected copyright holders from competition from other publishers only if they registered their works with the Copyright Office and deposited a copy to the Library of Congress. Copyright protection thus provided did not extend to use of works by individuals, so that anyone who purchased a work was allowed to "copy, translate or make a derivative use of these maps, charts and books without the permission of the author" [71].

Since 1790, the copyright law has gone through many revisions, major ones in 1831, 1870, 1909, 1976, and 1998 [76]. To paraphrase Jessica Litman, most of these revisions were triggered by predominantly technological changes that affected fundamental assumptions about the technologies on which copyrights were based. New kinds of works, new technologies, and new media for expression necessitated amendment of the copyright law in order to provide adequate protection for the new creative works. These changes gradually changed the scope of copyrights in three main ways: they increased the kinds of works covered by copyrights, they increased the kinds of rights granted by copyrights, and they increased the term of copyrights [76]. As a result, in 2005, copyrights grant to authors of creative works the exclusive right to reproduce, distribute, publicly perform and display their works, to prepare derivatives, and to authorize others to perform these acts. These rights extend to all creative works, and persist for 70 years beyond the author's lifetime for individuals, and for 95 years for corporate copyright holders. Copyrights no longer require registration, and are granted automatically, even to unpublished works [71].

Since 1909, modifications to the copyright law have been decided through negotiations among representatives of industries with an interest in copyright, i.e., copyright industries [76]. As a result, copyright law gradually expanded the rights of copyright holders. In response to gradual and continuous expansion of the rights of copyright holders, in 1976 fair use and first sale doctrines were defined, also libraries were provided with immunity for some uses that would otherwise infringe on the copyrights of authors.[4]

[3] U.S. Constitution. Article. I §8, cl. 8.

[4] Timeline: A History of Copyright in the U.S. http://arl.cni.org/info/frn/copy/timeline.html.

In 1998, the Congress amended the copyright law in response to a new set of technological changes that collectively comprise the so-called "digital world". In the non-digital world, the nature of media and equipment limited the number of possible uses (and users) of works and reinforced copyrights. For example, users who purchased a book could read it, retype or photocopy it, and loan or sell the book to others who could perform the same activities with it. The equipment that enabled large-scale reproduction and distribution required large investments and was not widely available; most small-scale reproduction equipment produced copies that degraded with each generation; the copies were still in the form of hard copies, and their cost of reproduction was non-zero. The difficulty and cost of reproduction and distribution using these equipment limited the ability of the general public to infringe copyrights and provided barriers to large-scale infringement.

Digital equipment, media, and networks, the so-called "digital world", have enabled low-cost reproduction and distribution of copyrighted works, enabling copyright holders to reach wider audiences cheaply. This low-cost dissemination to wide audiences is promising for promotion of progress—copyright holders can disseminate their knowledge to more people cheaply. However, widespread adoption and use of variety of digital equipment, media, and networks as part of daily activities of users also eliminated most of the previous technological and financial barriers to infringement [86]. As a result, the public is capable of generating and distributing low-cost copies of digital works, without compensation to copyright holders. For works that are published digitally, perfect substitutes for, i.e., exact copies of, original works can be reproduced and distributed at zero cost; for works that are published in hard copies and later digitized, exact copies of the digital incarnation of the work can be reproduced and distributed at zero cost—beyond the one-time digitization cost, the distribution cost of this alternative is zero. In either case, unauthorized digital copies and alternatives threaten the revenues, collected from sales of individual copies, that copyright holders rely on [138]. Lack of distribution mechanisms that can disseminate digital works while protecting revenues, in ways that support current business models, put copyright holders at odds with users of their works: copyright holders desire to take advantage of low-cost distribution mechanisms to reach wider audiences but also wish to maintain control over the distribution and use of their works in order to protect their revenues, while users want to make the most of their investments in technology and works, and wish to reap the benefits of widely available, low-cost digital works even if these works come from unauthorized sources.

The response to these changes and to the disturbed balance has been internationally coordinated. The most recent series of changes to the copyright law, in response to development of digital

technologies, started with the World Intellectual Property Organization (WIPO) Copyright Treaty (WCT) and WIPO Performance and Phonograms Treaty (WPPT) of 1996 which proposed the use of Digital Rights Management (DRM) systems to prevent the deterioration of copyright holders' revenues that can result from unauthorized copying and distribution of digital works [134]. In implementing the WCT and WPPT in the United States, the Digital Millennium Copyright Act (DMCA) took drastic measures to protect the interests of copyright holders [105]. These measures included authorization of use of DRM systems for protection of digital works, prohibition of circumvention of DRM systems (with carefully defined narrow exceptions such as circumvention for achieving interoperability, and for encryption and computer security research, without much regard to most other fair uses), and prohibition of trafficking of equipment designed primarily for circumvention.

These legislative changes and the use of DRM systems to minimize copyright holders' financial losses threaten the public's ability to access and use digital works. In particular, DRM systems can be used to expand control of copyright holders over their works. They can be implemented in ways that interfere with the first sale doctrine and limit the dissemination of digital works to the public through libraries and through secondhand markets [76]. They can reduce the usability of digital works by enabling invasion of user privacy [23] and enabling limits on fair use [26]. And use of DRM systems to protect copyrights can exacerbate the effects of copyrights on free speech [88]. The result degrades the public's ability to access and use digital forms of copyrighted material, contravening the main goal of the copyright law.

The amendments proposed by the Benefit Authors without Limiting Advancement or Net Consumer Expectations (BALANCE) Act (H.R. 1066) take steps towards addressing concerns raised due to overreaching applications of DRM systems and the DMCA, and towards re-balancing the concerns (i.e., revenues) of copyright holders with socially valuable concepts, such as fair use and first sale, that enable information dissemination and use in society. In particular, the BALANCE act proposes to authorize digital first sale as long as the seller does not retain a digital copy. It also proposes to legalize circumvention of DRM systems for non-infringing uses and to make it legal to provide others with circumvention tools to enable them to make non-infringing uses, provided (in both cases) that the circumvention is performed on legally obtained copies [11].

> ‘(c) CIRCUMVENTION FOR NONINFRINGING USES-
>
> > (1) Notwithstanding any other provision in this title, a person who lawfully obtains a copy or phonorecord of a work, or who lawfully receives a transmission of a work, may circumvent a technological measure that effectively controls access to the work or protects a right of the copyright holder under this title if–

> ‘(A) such act is necessary to make a noninfringing use of the work under this title; and
>
> ‘(B) the copyright owner fails to make publicly available the necessary means to make such noninfringing use without additional cost or burden to such person.
>
> ‘(2) Notwithstanding the provisions of subsections (a)(2) and (b), any person may manufacture, import, offer to the public, provide, or otherwise make available technological means to circumvent a technological measure that effectively controls access to a work protected under this title or protects a right of a copyright holder under this title, if–
>
> ‘(A) such means are necessary to make a noninfringing use under paragraph (1)(A);
>
> ‘(B) such means are designed, produced, and marketed to make a noninfringing use under paragraph (1)(A); and
>
> ‘(C) the copyright owner fails to make available the necessary means referred to in paragraph (1)(B).’.

On the other hand, other proposed amendments to U.S. copyright law, such as the Inducing Infringement of Copyright (INDUCE) Act (S. 2560) [55], and the Limiting the Liability of Copyright Owners for Protecting their Works on Peer-to-Peer Networks Act (H.R. 5211) [73] give more power to copyright holders, and threaten to further limit the ability of the public to learn from digital works. In particular, the INDUCE Act takes a step towards minimizing infringement on peer-to-peer networks by proposing to create a new kind of secondary infringement liability called “intentional inducement of infringement” that can have adverse effects on innovation by holding technology producers responsible for infringing uses of their products. The Limiting the Liability of Copyright Owners for Protecting their Works on Peer-to-Peer Networks Act proposes to protect copyright holders from liability for some measures they may take, including removing unauthorized copies of their works from the computers of users, in order to enforce their copyrights; this act gives copyright holders control over works stored on the hard disks of users and causes privacy concerns.

Although the fate of these proposals is still unknown, there is need for balance between the rights and incentives of stakeholder groups. As stated in the findings of Congress during the introduction of the BALANCE Act in the House:

> (6) The Digital Millennium Copyright Act (‘DMCA’) was enacted as an attempt to safeguard the traditional balance in the face of these new challenges. It gave copyright holders the ability to fight digital piracy by employing technical restrictions that prevent unlawful access and copying. In practice, however, the DMCA also endangered the rights and expectations of legitimate consumers.
>
> (7) Contrary to the intent of Congress, section 1201 of title 17, United States Code, has been interpreted to prohibit all users–even lawful ones–from circumventing technical restrictions for any reason. As a result, the lawful consumer cannot legally circumvent technological restrictions, even if he or she is simply trying to exercise a fair use or to utilize the work on a different digital media device. See, e.g., Universal City Studios,

> Inc. v. Reimerdes, 111 F. Supp. 2d 294, 321-24 (S.D.N.Y. 2000) (DMCA failed to give consumers the technical means to make fair uses of encrypted copyrighted works.)
>
> (8) The authors of the DMCA never intended to create such a dramatic shift in the balance. As the report of the Committee of the Judiciary of the House of Representatives accompanying the DMCA stated: '[A]n individual [should] not be able to circumvent in order to gain unauthorized access to a work, but [should] be able to do so in order to make fair use of a work which he or she has acquired lawfully.' House Report 105-551, Part I, Section-by-Section Analysis of section 1201(a)(1).
>
> (9) It is now necessary to restore the traditional balance between copyright holders and society, as intended by the 105th Congress. Copyright laws in the digital age must prevent and punish digital pirates without treating every consumer as one.

While legal efforts to reestablish "the traditional balance between copyright holders and society" continue, the tension between the stakeholders is exacerbated by lawsuits brought against varying stakeholder groups including users of works and innovators. In the meantime, technologies that provide access to works continuously evolve to include more and varied services, and as a result raise new concerns for copyright holders, and pose new questions to copyright professionals. The digital copyright problem must be addressed under such tense and rapidly changing circumstances.

The goal of this thesis is to identify a solution that balances the rights and privileges of different stakeholder groups in a way that brings the application of copyright closer to its original goal of promotion of progress and encouragement of learning in the digital world. In order for copyrights to achieve these goals, creative artists should be provided with an incentive to produce and disseminate their works digitally. This incentive is usually provided in the form of monetary returns. On the other hand, for the encouragement of learning and promotion of progress, the public should be able to access and use digital works without having concerns for their privacy; they should be able to obtain these works through low-cost distribution channels such as libraries supported by first sale doctrine; and they should be able to make fair use and practice free speech using these works. A balance will be achieved when all of these goals are satisfied.

In the following section, we review the state of the copyright problem from the perspectives of major stakeholders (i.e., copyright holders, users, innovators, rights defenders, and decision makers) and explain the importance of these criteria for the different stakeholder groups. In Chapter 3, we review some proposed solutions to the digital copyright problem and evaluate them from the perspective of maximizing each of the parameters of this balancing act, and from the perspective of achieving the goals of progress and learning.

## 2.2 Mesh of Stakeholders and Copyright Concerns

The digital copyright problem affects a complex set of stakeholders that include users of copyrighted works, innovators and technology producers, rights defenders, copyright holders, and decision makers. Members of these groups include traditional and digital libraries; webcasters; organizations such as the Electronic Frontier Foundation that lobby for particular ideals; technology producers and service providers; authors; middlemen involved in the promotion, publication, and distribution of works; and the Congress and the courts. The individual members of these groups are intermingled; depending on the circumstance, most can belong to several different groups.

In this section, we discuss these stakeholder groups and their different concerns related to the digital copyright problem. Addressing the concerns of all parties in this debate is a very challenging task, and requires prioritization of the issues to be addressed. Our primary foci are the ability of the public to access and use works, and the protection of revenues of copyright holders as an incentive for continued generation and dissemination of works. We define access as the ability to obtain works that one can use (without necessarily having to pay the full market price, e.g., through first sale), and the ability to use works in terms of the openness of works to fair use and to practice of free speech, without invasion of privacy of users.

### 2.2.1 Copyright Holders

In the non-digital world, creation and distribution of works are both costly processes: creation requires time and resources, distribution requires sizable capital investments. Creation and distribution of books, for example, requires time and effort of the authors to create, in addition to large investments in editing, publication, promotion, and sale of books. Middlemen, e.g., publishing houses, assume the responsibilities of some of these activities with authorization from authors, enable authors to reach wider audiences by promoting works, and in return for their services, they keep part of the sales revenues. In the face of uncertainty about the level of success of a particular work, and where publishing a book requires large initial investments for a printing press, this arrangement shifts the risk associated with publication of works from authors to publishing houses [76], in return for giving some copyrights, e.g., copying and distribution, and part of the author's revenues to these middlemen. Thus, the terms "copyright holders" and "rights holders" refer to the parties who have exclusive rights to works and can mean either the authors, or the middlemen, or both.

In the non-digital world, large-scale distribution requires sizable investments in equipment and material for making copies and serves as one of the instruments of control in copyright: unable to invest in large-scale reproduction equipment, most people are limited to small-scale reproduction and uses, e.g., making copies of music for their friends or to listen in their car; moreover, in many cases, the quality of copies degrade from one generation to the next; the copies thus generated do not have a significant effect on the market for creative works and on the revenues of copyright holders.

The digital world eliminates this instrument of control by minimizing reproduction and distribution costs, and by making reproduction and distribution trivial. As a result, perfect digital copies (i.e., exact copies) of digitally-published works, whose creation still requires time and resources analogous to that required in the non-digital world, can be generated and distributed at zero marginal cost by both copyright holders and users. Moreover, the quality of copies do not degrade from one generation to the next, and anyone who obtains a digital copy can distribute that exact copy to large numbers of people.

Under their current business models which rely on sales of individual copies, unauthorized copies of works can limit the sale of originals and threaten the creative incentives of copyright holders. Therefore, given the relatively large investments associated with the creation of works, and the relatively low cost of reproduction (by anyone with access to the right technology), copyright holders need some guarantees that they will be able to recoup their investments in order to keep producing and disseminating works.

Our discussion of copyright holders has so far considered revenues to be the main source of incentives for creation and distribution of works. However, the situation is not that simple: although many believe money to be the major motivator behind creation of works [138], copyright holders vary in their interests and goals. Authors in certain communities, e.g., academia, are often more interested in gaining recognition than earning profits and are willing to give away works for free. Middlemen, on the other hand, are almost always profit-maximizing entities: they negotiate contracts with authors and compensate authors with what some consider meager royalties [87]. This heterogeneity in the interests of different copyright holders creates a tension between authors and middlemen.

Digital equipment reduces the dependency of authors on middlemen, enables authors to reach wide audiences directly, gives them control over the prices of their works, and empowers them to negotiate better contracts with middlemen in order to obtain greater royalties while keeping more of their copyrights [87]. However, despite the emergence of equipment that reduce the dependency

of authors on middlemen, so far very few authors have managed to reap revenues by reaching their audience without going through middlemen. For example, Steven King tried to electronically publish his book *The Plant* but stopped this effort when payments did not meet his pre-announced minimum [98]. Today, most authors still share both the cost and the benefit of their works with their middlemen. Therefore, in our analysis, we will represent the incentives of copyright holders with revenues.

**Authors/Creators**

Authors are writers, painters, song writers, composers, and other creators and artists who invest their creative energy and time to produce original works. Technological advancements have provided authors with tools for creation of new kinds of and varied works, making it easier to create works based on existing material, e.g., sampling and remixes.[5] and in some cases making it easier and cheaper to produce works.

Technological advancements have also progressively reduced the reproduction and distribution costs associated with creative activities; digital media reduced the distribution costs to a minimum and thus enabled authors to reach the public directly, without going through middlemen.[6] If they can take advantage of these technological advancements, authors can potentially create more and diverse set of works, obtain higher net revenues for their works, and/or reduce the sale price of their works in order to enable the dissemination of their works to wider audiences.[7]

Despite the promise of digital technologies, lack of mechanisms that support conventional methods consistent with existing business models, for collection of compensation in return for digital distribution of works, makes it difficult for most authors to break free from the traditional publishing methods of middlemen. In addition, many creators still need the endorsement of middlemen

[5]The new trends of sampling and remixing create tensions between copyright holders as remixing or sampling other's work may infringe on their copyrights. Not surprizingly, on this point also, copyright holders have different preferences and attitudes. While some release their works with licenses, e.g., through Creative Commons (see http://www.wired.com/wired/archive/12.11/sample.html), that allow sampling and remixes, others limit and deter it (see, http://www.newmediamusings.com/blog/2004/02/copyright_vs_re.html).

[6]Some artists, e.g., Courtney Love, have explicitly complained about being cheated out of their copyrights and revenues by middlemen [77].

[7]Forty machine learning professors resigned from the editorial board of the Machine Learning Journal, and started a free electronic publication. The letter of resignation and the reasons can be found at http://mail.cs.uiuc.edu/pipermail/colt/2001October/000553.html. However, academics and journal publications are a special case because journal publications bring authors recognition but not money (although middlemen obtain financial return on the journals). Besides, the peer-review process involved with most academic publications endorses the quality of the publications and gives academic authors the freedom to publish the same high-quality works without having to go through middlemen and without losing their audience. By expanding their audience base, these creators incur no financial losses but gain more recognition.

in order to market their works; middlemen endorse creative works simply by investing in these works—which indicates that the quality of the work promises them enough returns to recoup their investments. Although they can directly publish, promote, and distribute their works, this endorsement helps authors distinguish themselves among the large volume of works that become available everyday. As a result, those authors without the means to commission middlemen and those who have already established their own brand name, e.g., Steven King[8], are more likely to take the initiative to work independently of middlemen.

#### Middlemen: Publishing Houses, Recording Studios, and their Equivalents

Middlemen such as publishing houses and recording studios review, edit, publish, advertise, promote, and distribute works. These services require large investments, which middlemen, as copyright holders, recoup during the copyright period by taking advantage of their exclusive rights to reproduce and distribute works.

Technological advancements have reduced the dependency of authors on middlemen; the digital world has put middlemen into a difficult situation by enabling authors, in some industries more than others, to potentially perform some of the functions of middlemen at no cost. However, authors still rely on middlemen to reach wide audiences, for endorsement of the works, and for review, editing, advertisement of their works, for discovering new talent, for packaging and marketing works, and for financing production and promotion [87].[9]

### 2.2.2 Digital Copyright Problem and Reactions of Copyright Holders

Reactions of copyright holders to threats to their businesses due to development of digital world have followed two general patterns: they have tried to control unauthorized distribution of their works, and they have tried to take advantage of low-cost reproduction and distribution mechanisms in order to provide services that meet the demand for digital works.

To limit and control unauthorized reproduction and distribution, copyright holders have employed a variety of strategies, including participating in educational campaigns about infringement, using DRM systems, and taking legal action against infringers. For example, the Recording Industry Association of America (RIAA), representing the middlemen for the music industry, provides edu-

[8] And even those with established brand names have not yet been successful.

[9] For example, some Internet publishing companies, such as "iUniverse" (http://iuniverse.com), offer editing, distribution, advertising, and marketing aid to authors, and endorse the submissions that pass the editorial review.

cational information about what constitutes infringement and how to recognize counterfeit products, and offers "up to $10,000 in reward" for people who report "information regarding CD manufacturers illegally producing RIAA member company sound recordings" [95]. Similarly, the Association of American Publishers (AAP), which is "the principal national trade organization of the U.S. book publishing industry and represents more than 300 corporate members including most of the major commercial book publishers in the United States as well as smaller and medium-sized houses, non-profit publishers, university presses, and scholarly societies," [2] has an anti-piracy program aimed at eliminating infringement and leads efforts to develop DRM systems for e-books [113].

To strengthen the message conveyed by DRM systems, copyright holders have also brought lawsuits against parties involved in unauthorized uses of digital works [68]. The music industry, in particular, has made efforts to deter infringement through lawsuits [39, 68, 87, 95]: they sued the providers of technologies, such as Napster and Grokster, with the goal of eliminating innovations and innovators that enable infringement by individuals, and also brought lawsuits against individual file-swappers.

In addition to limiting unauthorized distribution of their works, in order to take advantage of low-cost distribution mechanisms, to meet the demand for digital works, and to provide users with low-cost alternatives to unauthorized copies of their works, copyright holders have found ways to reach their audiences in the digital world. For example, iTunes sells DRM-protected music for $0.99 a song without requiring subscription, these songs are available "in digital quality and can be burned onto CDs for personal use, played on up to three computers, and listened to in an unlimited number of portable players such as iPods". Similar service by Napster 2.0 offers "unlimited listening rights to its entire library for a $10.00 monthly fee; users of the subscription service then pay a discounted rate of as little as $0.80 for the rights to burn, download, and use on portable devices" [42].

### 2.2.3 Users

Content users consist mainly of the public and organizations that provide the public with low-cost access to works, e.g., libraries, as well as creators that rely on existing works to produce new ones. In the non-digital world, the public had very limited capability to engage in activities that could hurt the revenues of copyright holders. As a result, potential infringements by individuals at this time were too insignificant to worry about, were too costly to pursue [68], usually required purchase of at least one legitimate copy, and in most cases were covered by fair use [76]. As a result, the public enjoyed many small-scale personal uses, even if these uses technically infringed copyrights.

In the digital world, use of technologies to control access, use, and distribution of works enable enforcement of copyrights more strictly, and limit many small-scale personal uses, including those covered by fair use.

The digital world expanded the set of possible uses the users can make of works. Given the greater range of uses they can make of works, users are interested in making the most of their investments in equipment and works. However, strict enforcement of copyrights through use of DRM systems can limit these capabilities and raise concerns regarding fair use, freedom of speech, privacy, and first sale. These systems can also affect the functions of libraries and limit activities of creators who wish to build new works based on existing ones.

### The Public

Despite being the largest stakeholder (in terms of population) in the digital copyright dilemma, the public has so far had very little say in the evolution of the copyright law [76].

As digital equipment enters homes, the members of the public are no longer limited to small-scale uses of works; they can also act as large-scale distributors.

Polls indicate that a large section of the public does not see the immorality of downloading works, e.g., music, off the Internet [39, 74], and disagrees with the stringent regulations imposed on them by the technological protection mechanisms. In general, "[u]sers are frustrated by the restrictive technical mechanisms that are now applied because their intended (legitimate) uses, while predictable and acceptable for the most part, are inadequately supported technologically" [113]. As a result, many of those with the means circumvent the protection technologies and disobey technical protection mechanisms that limit accessibility and usability of works.

### Libraries

"Libraries are among the largest single concentration of customers and users of" digital works [113]. They consider themselves the "voice for the public good" and their participation is "often sought in 'friend of the court' briefs in important intellectual property cases" [3]. Libraries provide the public with access to works that may otherwise be unaffordable. They acquire copies of books, journals, articles, music, and other works, and loan them out to the public at low membership costs.

First sale and fair use doctrines play important roles for the proper function of libraries. First sale doctrine makes it legal to loan works to others, while fair use allows libraries to make copies of works for backup or for preservation. DRM systems can limit the implementation and applications

of these doctrines in the digital world and interfere with the functions of libraries. In particular, possible limitations in first sale doctrine raise concerns regarding "interlibrary loans, off-site accessibility, archiving/preservation, availability of works, and use of donated copies" [113].

Statements by the American Library Association indicate that libraries find the current DRM systems inadequate and would like DRM systems to evolve to support fair use and other library and education exceptions without applying predecided and over-broad restrictions that disable use. They want DRM systems to support many different sets of activities taken on by higher education institutions and libraries, including preservation and archiving [4].

### 2.2.4 Digital Copyright Problem, DRM systems, DMCA, and Effects on Users

Technological and legal developments aimed at addressing the digital copyright problem have enabled copyright holders to control and limit some fair uses, to monitor the way individuals use digitally-published copyrighted works, to limit the ability to exercise privileges given by the first sale doctrine, and even to inhibit freedom of speech (to a greater extent than already limited by copyrights).

#### Copyright and Fair Use

Fair use allows works to be used in ways that violate copyrights but encourage learning and promote the progress of useful arts and sciences. Fair use is codified in section 107 of the Title 17 of the United States code as follows:

> §107. Limitations on exclusive rights: Fair use
>
> Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include –
>
> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
>
> (2) the nature of the copyrighted work;
>
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
>
> (4) the effect of the use upon the potential market for or value of the copyrighted work.
>
> The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors.

Thus defined, fair use covers many mainstream small-scale uses of works by individuals. Most of these uses can now be controlled by DRM systems. The anti-circumvention provisions of the DMCA exacerbate the limitations on fair use: "a comprehensive ban on circumvention negates the fair use doctrine" [26]. In this environment, in order for fair use to remain effective "law should afford users of digital works rights of electronic self-help" [24].

### Copyright and Consumer Privacy

In many activities in the non-digital world, individuals assume a certain level of anonymity and privacy. Especially in their homes, individuals expect to have maximal privacy.

Regardless of their physical location, individuals cannot assume the same level of privacy while using digital works. In particular, DRM systems control many aspects of works and can monitor the activities of users [23, 24, 25, 27, 28]. This kind of close monitoring enables copyright holders to profile their users, to identify the products that they can successfully market to targeted individuals, and to measure the demand for their products in general [23].

A "fundamental assumption underlying our discourse about the activities of reading, thinking, and speech is that individuals in our society are guaranteed the freedom to form their thoughts and opinions in privacy, free from intrusive oversight by governmental or private entities" [23]. Close surveillance and profiling through DRM systems threaten this "breathing space" by imposing limits on and keeping track of user activities, eliminating privacy in "intellectual consumption" and limiting some fair uses that are most effectively practiced in privacy and anonymity, e.g., criticism and parody.

Privacy is necessary for people to define their personal identity without having to consider how they may be perceived by others. "Persistent, fine-grained observation subtly shapes behavior, expression and ultimately identity" [28].

Due to heightened awareness of invasions of privacy in the digital world, the last few decades have seen development of technologies that enable private communications and anonymity over the Internet. For example, GNUnet is a peer-to-peer network that consists of nodes that use keys to identify, authenticate, and encrypt communications [66]. These nodes hide the communication between nodes by routing communications through intermediaries and use cover traffic to make it difficult to link individuals with network activities. Many other anonymizing and privacy-enhancing technologies exist, e.g., [46, 54, 96, 41] and many rely on encryption and rerouting techniques. Analysis and study of these technologies is beyond the scope of this thesis.

### Copyright and First Sale Doctrine

First sale doctrine limits the control of copyright holders over copies of works that have already been sold.

> §109. Limitations on exclusive rights: Effect of transfer of particular copy or phonorecord
>
> (a) Notwithstanding the provisions of section 106(3), the owner of a particular copy or phonorecord lawfully made under this title, or any person authorized by such owner, is entitled, without the authority of the copyright owner, to sell or otherwise dispose of the possession of that copy or phonorecord...
>
> (b)
>
> (1)
>
> (A) Notwithstanding the provisions of subsection (a), unless authorized by the owners of copyright in the sound recording or the owner of copyright in a computer program (including any tape, disk, or other medium embodying such program), and in the case of a sound recording in the musical works embodied therein, neither the owner of a particular phonorecord nor any person in possession of a particular copy of a computer program (including any tape, disk, or other medium embodying such program), may, for the purposes of direct or indirect commercial advantage, dispose of, or authorize the disposal of, the possession of that phonorecord or computer program (including any tape, disk, or other medium embodying such program) by rental, lease, or lending, or by any other act or practice in the nature of rental, lease, or lending. Nothing in the preceding sentence shall apply to the rental, lease, or lending of a phonorecord for nonprofit purposes by a nonprofit library or nonprofit educational institution. The transfer of possession of a lawfully made copy of a computer program by a nonprofit educational institution to another nonprofit educational institution or to faculty, staff, and students does not constitute rental, lease, or lending for direct or indirect commercial purposes under this subsection.

Secondhand markets allow consumers to resell works they have legally purchased and no longer want. Sale of works in secondhand markets allows consumers with lower budgets to have access to works by purchasing secondhand instead of a brand-new versions. First sale doctrine enables transfer of works in secondhand markets and has, as such, played an important role in aiding continuous flow of works through society in the non-digital world. Libraries, which loan copies of works to the public at low costs, also rely on the first sale doctrine for their functions.

In the non-digital world, the loan or exchange of works deprives the lending party of that object, i.e., between the lender and the borrower there is a total of one original copy. The digital world changed this constraint, because making a loan or a sale of a digital product does not automatically terminate the original, raising questions about applicability of the first sale doctrine in the digital world.

Limitations on application of the first sale doctrine may not be detrimental to accessibility of digital works at lower cost; after all DRM systems allow availability of different versions of the

same products at different prices, enable price discrimination, and reduce the need for secondhand markets. However, uncertainties about applicability of the first sale doctrine to the digital world adversely affect the functions of libraries, which improve public access to works.

### Copyright and First Amendment

Since the evolution of the DMCA and DRM systems, there have been concerns related to use of DRM systems and the DMCA to inhibit free speech. In particular, protection of expression by copyrights leaves ideas used in works available for free speech. In addition, public domain components included in copyrighted works are not protected by copyrights and can be used for free speech. Finally, fair use enables some free speech by protecting against infringement claims criticism and parody.

However, DRM systems can limit fair use and access to the public domain components of works as well as access to original expression. The anticircumvention provisions of the DMCA do not allow circumvention of DRM systems even for free speech. Netanel [88] writes that:

> The First Amendment provides that 'Congress shall make no law ... abridging the freedom of speech.'... Courts have almost never imposed First Amendment limitations on copyright, and most have summarily rejected First Amendment defenses to copyright infringement claims. Courts have recognized that copyright can abridge speech and thus that it raises First Amendment concerns. But in almost every instance, courts have assumed that First Amendment values are fully and adequately protected by limitations on rights within copyright doctrine itself. Some have even posited that 'copyrights are categorically immune from challenges under the First Amendment'.

In the case of Eldred v. Ashcroft,[10] the Supreme Court held that "copyrights are categorically immune" from first amendment challenges, stating that the copyright law has built-in exceptions (namely fair use and the idea–expression dichotomy,[11] and time limits on copyrights) that protect first amendment rights while enforcing copyrights. The Supreme Court also ruled that, as long as the "contours" of the copyright law do not change, first amendment scrutiny is not necessary for copyright infringement claims.

However, to paraphrase Netanel, the DMCA has arguably changed the "contours" of the copyright law. More specifically, in the last decade, limited-time monopolies, the idea–expression dichotomy, and fair use, which collectively kept copyrights and freedom of speech in balance, have

[10] Eldred v. Ashcroft 537 US 186 (2003)

[11] The idea–expression dichotomy limits the scope of copyrights to expression, allowing ideas to be used without control by copyright holders.

been significantly changed to the detriment of free speech. The Sonny Bono Copyright Term Extension Act (CTEA) extended the duration of limited-time monopolies by 20 years; the DMCA, by authorizing use of DRM systems and providing anti-circumvention provisions, expanded the control of copyright holders to fair uses and to uncopyrightable elements of works. As a result, "today's expanded copyright regularly chills creative and pointed criticism, commentary, artistic insight, and self-expression" and enables copyright holders to control others' speech (for longer periods) [88].

#### Digital Copyright and Moral Intuitions

The limitations imposed on commonplace uses of works through DRM systems and anti-circumvention provisions create a discrepancy between applications of the copyright law to non-digital works and their digital counterparts. These discrepancies lead to conflicts between some applications of the copyright law and conventional wisdom about copyrights, leading to the conclusion that the law does not make sense [76].

One result of this conclusion is civil disobedience which stems largely from the belief that "law by and large tracks [our] sense of justice" [39]. If a law does not make sense, many believe that it cannot be true; what is more, according to Litman, disobeying such a law expresses public displeasure with the law and can force it to be changed [76].

As a result, infringing activities such as file-swapping over the Internet and over peer-to-peer networks has become quite popular. Many users believe that personal uses of works are covered by fair use [76], even if these works are obtained on peer-to-peer networks. Fisher reviews the results of polls taken in the year 2000 that show that 40-55% of respondents, most of whom were American Internet users, did not consider it immoral to download music from the Internet [39].

### 2.2.5 Innovators and Technology Producers

In order to prevent reproduction and distribution of works among users, copyright holders brought lawsuits against providers of technologies that facilitate infringing activities of individuals—instead of suing millions of users one at a time, they tried to shut out all direct infringers at once by removing the services they make use of in order to infringe. Despite the Supreme Court ruling in Sony v. Universal[12] that protected innovations with substantial non-infringing uses even if they can also

[12]Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417 (1984)

be used to infringe, many innovators have been sued for secondary infringement,[13] regardless of non-infringing uses of their technologies.

For example, lawsuits have been brought against innovators like Napster whose works facilitate infringement[14], software producers like Grokster whose technologies facilitate file sharing[15], search engines like Arriba Soft that help people locate infringing material[16], and Internet connectivity providers like universities [68]. Similar lawsuits have been brought against companies such as Streambox[17] and 321 Studios[18] whose products could be used for circumvention.

Some of these cases have been resolved and the outcomes have been mixed—Napster, Streambox, and 321 have been resolved in favor of copyright holders while Grokster and Arriba Soft have so far been resolved in favor of defendants. Regardless of their outcome, these kinds of lawsuits deter many technology producers from developing potentially controversial products and have a chilling effect on innovation.

DRM systems have also been problematic for technology producers because of lack of widely adopted standards and interoperability. The diversity of the DRM systems currently available and lack of standards for DRM systems requires electronics producers to design equipment that interoperates with all existing DRM systems in order to be able to market their equipment to larger audiences. Lack of interoperability between DRM systems affects marketability of electronics: consumers do not want to invest in equipment which will fail to play some of the DRM-protected works they legally purchase. Standardization of DRM systems, and standards that all stakeholder representatives agree on, may ameliorate this problem.

Last but not least, DRM systems can be used to extend the reach of copyright holders to secondary markets, raising concerns about anti-competitive uses of copyrights. For example, "Apple's use of a proprietary DRM standard allows it to control secondary markets. Currently, only iTunes and Quicktime Software can play FairPlay files, and the iPod is the only compatible portable player" [42]. Because FairPlay is a proprietary DRM system, any competitor to iPod would have to obtain a license from Apple to implement the FairPlay DRM system.

[13] Secondary infringement refers to contributory infringement and vicarious infringement. Contributory infringement takes place when the secondary infringer knows specific cases of and knowingly contributes to directly infringing activity, e.g., by providing tools that facilitate infringement. Vicarious infringement takes place when the secondary infringer has the right and the ability to control or supervise direct infringing activity, and obtains financial gain from letting it happen. Peer-to-peer network providers have been targets of both kinds of secondary infringement claims.

[14] A&M Records Inc., v. Napster, Inc., 239 F. 3d 1004 (9th Cir. 2001)

[15] MGM Studios, Inc., v. Grokster Ltd., 259 F. Supp. 2d 1029 (C.D. Cal 2003)

[16] Kelly v. Arriba Soft Corp., 226 F. 3d 811 (9th Cir. 2003)

[17] Real Networks v. Streambox, 2000 WL 127311 (W.D. Wash. Jan. 18 2000)

[18] 321 Studios v. MGM Studios, Inc., et. al., 307 F. Supp. 2d 1085 (N.D. Cal. 2004)

### Copyrights and Innovation

The increase in the number of people with Internet access increased the demand for technologies that connect people and enable quick data exchange. Peer-to-peer networks addressed this demand, reduced the dependency of users on central data sources and distribution nodes, and offered a different method of collaboration between users. While facilitating easy communication and data exchange, and enabling transfer of educational materials between groups, these technologies also enabled unauthorized distribution and exchange of digital copyrighted material.

To deter copyright infringement on such networks, copyright holders have taken infringers to court. They have also brought secondary infringement lawsuits against innovators of technologies that facilitate or enable infringement. Lemley argues that producers of technologies sued for secondary infringement, such as Napster and Grokster, cover a continuum that represents different levels of infringing and non-infringing activities. The technologies whose uses are limited to infringing activities form clear cases of liability. However, many technologies also have non-infringing uses, some of which have significant value for society. Convicting innovators for infringement when their technologies have non-infringing uses discourages many valuable technologies and services, and stifles innovation [68].

### DRM systems and Competition

DRM systems, anti-circumvention provisions, and license agreements enforced by DRM systems have all been used to deter competition. For example, in the case of Blizzard v. BnetD[19], Blizzard sued BnetD for violation of the DMCA as well as violation of the end user license agreement (EULA). On the other hand, DRM systems can be used to limit the kind of software that can execute on particular operating systems; hardware DRM systems are particularly useful for such anti-competitive strategies, because they enable technology producers and operating system producers to dictate the kind of software that can execute on a device [5]. Further, proprietary DRM systems limit use of works to only some media, e.g., iTunes files protected by FairPlay can only play on iPod [42].

[19] Davidson & Associates, Inc., et al., v. Internet Gateway, et al., 334 F. Supp. 2d 1164 (E.D. Mo. 2004)

#### DRM systems, Standards, and Electronics

Since the enactment of the DMCA, in the absence of standards that satisfy the concerns of all copyright holders, the market has provided several alternative DRM systems.

Lack of interoperability between available DRM systems limits the ability of users of copyrighted works to benefit from DRM-protected works and adversely affects electronics producers. In the absence of standardized DRM systems, electronics producers are limited in the range of products and markets they can reach; their products will likely be compatible with only a subset of the available DRM systems. As a result, consumers may not be able to use the works they have legitimately purchased on some digital equipment.

So far, many different DRM systems have been developed and adopted by different industry sectors. For example, ContentGuard's Extensible Rights Markup Language (XrML) has been adopted for the MPEG Rights Expression Language (REL) and Microsoft has been using XrML for its own DRM technology.[20] The Open Digital Rights Language (ODRL) has been adopted by the Open Mobile Alliance (OMA) for OMA DRM. Other popular DRM technologies include Electronic Book Exchange (EBX) Working Group[21], Open eBook Forum (OeBF)[22] and Secure Digital Music Initiative (SDMI)[23]. Until standards are chosen and enforced, many more DRM systems will appear and compete for market share.

### 2.2.6 Rights Defenders

The actions taken to resolve the digital copyright problem, the lawsuits against both users and producers of technologies, and the tendency of current laws to favor copyright holders, have motivated some groups to take an active role in educating the public about the state of affairs in the digital copyright problem. One such vocal group is the Electronic Frontier Foundation which consists of lawyers, technologists, and other volunteers, and aims to defend fair use, privacy, free speech, and "the vast wealth of digital information, innovation, and technology that resides online" [38]. According to the EFF, use of DRM systems reduce "freedom of expression", reduce innovation, erode privacy, "freeze fair use", undermine libraries and archives, and reduce competition [38].

[20] http://www.drmwatch.com/standards/article.php/3295291
[21] http://www.ebxwg.org/
[22] http://www.openebook.org/
[23] http://www.sdmi.org/

Other organizations with similar interests include the Personal Technology Freedom Coalition which includes major library associations, high tech firms, universities, and public interest groups; the Center for Democracy and Technology (CDT); Computer Professionals for Social Responsibility (CPSR); and the Electronic Privacy Information Center (EPIC).

### 2.2.7 Policy Makers: the Congress and Courts

The solution to the digital copyright problem is not straightforward and policy makers have a difficult task at hand. In order to address this problem, the policies to be developed should protect the incentives of copyright holders and satisfy the concerns of the disparate groups of users and technology producers. In particular, any solution should encourage learning and promote progress, while protecting economic incentives. As court cases reveal unintended consequences of the DMCA and overreaching uses of DRM systems, the Congress has been working to amend the copyright law, to eliminate the overreaching interpretations and over-broad implementations of the law and technologies and to protect the ability of the public to access, use, and learn from creative works. It has also focused on protecting the incentives of copyright holders to produce and disseminate works.

## 2.3 Net Effects of DRM systems and Supporting Legislation

Use of DRM systems by copyright holders is becoming common practice. Many copyright holders, such as the American Association of Publishers, claim that concerns regarding DRM systems are mostly misconceptions, that DRM systems "do not remove original works from the public domain because users are free to copy the work with whatever means" they have available, such as "by hand, re-keying, copying and pasting," etc.; that DRM systems do not limit first sale doctrine because this doctrine still applies to printed works, as well as CDs, floppies, etc.; that DRM systems "are not implemented for the purpose of preventing reverse engineering"; that the time limits of DRM systems are determined based on market effects; that DRM systems no longer suffer from "lack of standards for software and hardware"; and that DRM systems are beyond the stage where many vendors tried to develop proprietary systems to gain market dominance [113].

For copyright owners, DRM systems provide security from infringement, encourage dissemination of works to the public, and also "change consumer expectations about what they are entitled to do with digital [works]" [106]. These mechanisms also enable business opportunities based on price-discrimination schemes that make use of accurate information about user preferences regarding differing entertainment products, creating opportunities for new business models [23, 28, 39] that reach wider audiences and serve their needs more accurately. They also enable copyright holders to extend their reach to secondary markets [42], such as electronics markets.

On the other hand, many scholars have observed that the effects of the copyright law's application to the digital world and the provisions of the DMCA are far-reaching [23, 39, 74, 104]. For example, although price-discrimination through DRM systems can be used to provide consumers with legal access to more works, more easily, and cheaply, their use for copyright protection also imposes significant costs on society by deterring spontaneous and unpredictable creative uses. These systems can limit accessibility of digital works by blocking distribution channels justified by first sale, and can be used to limit free speech and invade consumer privacy. In addition, lack of standards among DRM systems affect the market for many electronics and can affect innovation [42]. Other undesirable aspects of DRM systems include: "increased concentration in the industries and associated price increases; impediments to the ability of consumers to reshape the entertainment products they receive; ... and, perhaps most seriously, a brake on the pace and range of innovation in computer equipment and software" [39]. The result limits the ability of the public to access and use digital works, and inhibits progress.

## 2.4 Requirements for a Promotion of Progress

Availability of digital equipment and the low-cost reproduction and distribution of works on digital networks promises to promote progress by enabling wider dissemination of works to the public at lower cost (compared to non-digital world). However, in order for this potential to become reality, the solution to the digital copyright problem should encourage continuous creation and dissemination of digital works and also enable the public to access, use, and learn from these works. Parameters that contribute to such a solution include:

- Protection of the revenues of copyright holders in order to encourage copyright holders to produce and disseminate works;
- Protection of flow of works through society through channels analogous to those supported by the fist sale doctrine in the non-digital world;
- Protection of the privacy of consumers while they use digital works;
- Protection of fair use privileges; and
- Protection of First Amendment rights.

The digital copyright problem is a complex one, and many different solutions can be arrived at depending on the general goal. For example, solutions that focus on the protection of moral rights of authors would differ from those focused on elimination of anti-competitive uses of DRM systems. These goals, while important, are beyond the scope of this thesis.

For our analysis of the digital copyright problem, we focus on a subset of parameters relevant to this problem that are important for the progress of society, and that improve accessibility and usability of works. A solution that focuses on these parameters enables users to make use of digital creative works without sacrificing from their rights and privileges, and creates incentives for copyright holders to keep producing and disseminating digital works. In Chapter 3, we review a set of proposals that take different approaches to achieving these goals.

## 2.5 Conclusion

This chapter's discussion of stakeholders and their interests revealed the problems associated with use of digital protection systems in order to address the digital copyright problem. DRM systems and their supporting legislation prohibit unauthorized uses and, at least to some extent, protect the revenues of copyright holders. However, these systems have side-effects that range from limitations on free speech to invasion of privacy, and raise concerns for many stakeholders including users and technology innovators.

Focusing on the goal of promotion of progress, we argue that a solution to the digital copyright problem should ensure dissemination of digital works to the public, accessibility of these works by the public, and usability of these works for fair use, for free speech, without raising concerns for privacy, and without threatening the creative incentives of copyright holders. In Chapter 3, we will evaluate several proposals from this perspective.

## 2.6 Summary

The digital copyright problem involves a complex set of stakeholders that include copyright holders, users, rights defenders, innovators, and decision makers. The concerns and interactions of these stakeholders are complicated: many of them share members, others include members with conflicting interests.

The rights and privileges of these stakeholder groups have been identified and balanced in the copyright law. Development of new technologies frequently disturbed the balance between the rights and privileges of stakeholders, and the law was amended to address the concerns of affected stakeholders. Recently, the digital world required some changes. While promising to promote progress by enabling copyright holders to each wider audiences at low reproduction and distribution costs, and enabling the public to access a wider range of works to learn from, the digital world also enabled unauthorized reproduction and distribution of works and threatened the royalty-based businesses of copyright holders. The amendments aimed at addressing the concerns related to unauthorized reproduction and distribution of works failed to re-establish the balance between the rights and incentives of different stakeholder groups. One of the most drastically affected groups is the public whose ability to access and use digital works can be controlled and limited by DRM systems. For example, the DRM systems used by copyright holders can limit dissemination of works even when this dissemination is justified by first sale. These systems can also eliminate some uses that are traditionally covered by fair use, and that are considered free speech. Moreover, the DRM systems control these uses by monitoring the activities of users and invading their privacy, while the anti-circumvention provisions exacerbate these problems.

Many academics have argued that in order for the copyright law to promote progress of science, a balance between the rights of copyright holders and the rights of users, i.e., the public, has to be established [76, 104]. This balance can be established by enabling unrestricted distribution and use of digital works for fair use and free speech, without raising privacy concerns; and protecting the revenues of copyright holders in order to encourage them to create and disseminate their works.

# Chapter 3

# Technology and Policy Solutions

The digital copyright problem can be addressed through technology and/or policy solutions. Given our goal of promoting progress and encouraging learning, different technology and/or policy solutions can be implemented that encourage copyright holders to create and disseminate their works, and that enable the public to access and use these works, without necessarily having to purchase a new copy, for fair use, and for free speech, without threats to their privacy.

Technology provides many alternative ways of protecting intellectual property, each with different limitations on use and access. For example, DRM systems can be programmed to encode and enforce authorized uses of works, to limit access to authorized parties, e.g., proprietary data, or to limit use, e.g., play but not copy. Digital tracking can be implemented, for example, to track individual copies, to pinpoint sources of infringement, or to meter use of works.

In the absence of policies that control the reach of technological protection systems, technology has been used to protect the revenues of copyright holders without much regard to the ability of the public to access and use works. In the case of intellectual property, too much protection can be as bad as not enough protection [69]. Therefore, use of technological mechanisms for intellectual property protection requires policies that balance the incentives, rights, and privileges of different stakeholders, in order to provide a solution that promotes progress in the digital world.

In Chapter 2, we argued that for the promotion of progress in the digital world, copyright holders need to have incentives to produce digital works, and that these incentives can be provided in the form of revenues. At the same time, the public needs to have easy access to digital works (without necessarily having to pay the full price for the product, e.g., through secondhand markets and libraries) and be able to use these works for fair use and for free speech, without any threats to

their privacy. In the first part of this chapter, we review technologies currently used for protecting copyrights in digital works, focusing on two kinds of popular approaches: technologies based on controlling use of works, e.g., hardware and software DRM systems, and technologies that support digital tracking, e.g., watermarks. In the second part of this chapter, we discuss several policy proposals that support the use of these technologies and balance their reach from the perspective of revenues, first sale, fair use, free speech, and privacy.

Our analysis of technology and policy proposals suggests that solutions based on digital tracking mechanisms, combined with taxation-based compensation systems, enable continuous flow of digital creative works through society while also protecting the revenues of copyright holders. Digital tracking can be implemented to meter popularity of works and to minimize fair use concerns, free speech and digital first sale limitations, and privacy implications, balancing the incentives and interests of copyright holders and users. To facilitate implementation of such solutions and to improve accuracy of digital tracking technologies, we propose tracking works using *expressive fingerprints*.

The term "fingerprints" has been used in the literature in two ways: it has been used to refer to digital labels that encode the identity of the lawful users of works [14, 92]; and it has been used to refer to information extracted from the content of works [110, 111, 121, 122]. The fingerprints described in this thesis are of the second kind: they capture characteristics of content of literary works and are extracted directly from the works. The characteristics we focus on capture expression and help meter use and distribution even when works are not properly labeled (as would be the case after an attack that rendered any labels, such as watermarks, ineffective), and even when works are paraphrased [121, 122]. In order for a fingerprint of a copy of a work to not match the fingerprint of the original, the contents and the expression of the copy itself, as opposed to its label, would have to be modified significantly. The details of this fingerprinting mechanism are discussed in the remaining chapters of this thesis.

## 3.1 Technologies

### 3.1.1 Technologies for Controlling Use

DRM systems exist in many different forms and take different approaches to securing works. "The first is 'containment' of the wrapper, an approach where the [work] is encrypted in a shell so that it can only be accessed by authorized users. The second is 'marking' or using an encrypted header, such as the practice of placing a ... flag, XML or XrML tag on [works] as a signal to a device that the media is copy protected. The third is the secure container, such as a dedicated reading device" [113].

In general, DRM systems are "trusted" [120] systems that can be programmed to reliably limit access to works and control their use. In different contexts, DRM systems have been described as "business processes that for legal and commercial purposes track rights, copyright holders, licenses, sales, agents, royalties, and associated terms and conditions, using digital technology" [100] and that apply to digital works; as "a technology that enables the secure distribution, promotion, and sale of digital media [works] on the Internet"[1]; and as "code as code" [70].

DRM systems have changed the way copyright holders do business and have opened doors to new business opportunities based on price-discrimination schemes. As a result, the market for DRM systems and DRM-related products developed very quickly; this market was valued at $96 million in 2000, and is expected to reach $3.5 billion in 2005 [100].

Most DRM systems are software-based; however, "Trusted Computing" and similar initiatives combine software DRM systems with hardware to provide "trusted platforms" on which DRM-protected works can be used [120]. Below, we discuss software DRM systems and trusted computing platforms.

#### DRM systems

DRM systems encode and enforce authorized uses of works. Given a set of instructions about authorized uses of works, DRM systems execute these instructions, and grant users access to works in return for payments or information about the way they use works [44]. In a typical DRM system, "DRM network server software wraps the digital [work]. DRM client software unwraps it or otherwise makes it accessible in accordance with its rights" [89]. Some DRM systems can be programmed to [23]:

[1] http://www.microsoft.com/windows/windowsmedia/drm.asp

- Recognize what the user is doing with a work, i.e, read, copy, paste, etc.;
- Allow or disallow any particular use by checking it against the license stored in the DRM system; and
- Remove the work from the user's machine if the license is violated.

Copyright holders can use DRM technologies to maintain control over their works and to price-discriminate by making different versions of their works available to different users, potentially extending their audience base even to those users who would not purchase their works at full market price. However, DRM systems enforce the use limitations that they impose on works very strictly—if the author does not authorize copying of his work, users cannot make any copies, even if their act would be justified under fair use. As a result, DRM systems have been widely criticized for being over-broad [104, 105]: they can be used to prevent non-infringing uses of works along with infringing ones.

To enable non-infringing, fair uses of works, and in order to minimize the conflicts between such uses and the incentives of copyright holders, some have proposed mandating relaxed implementations of DRM systems. For example, DRM systems should allow access to uncopyrightable aspects of works, should enable fair uses, and should not gather information about consumer behavior. However, design of a general purpose DRM that satisfies the concerns of all stakeholders, e.g., that can satisfy all possible fair uses of a work and sufficiently prevent infringement[2], is generally recognized as infeasible [19].

Even in the presence of DRM systems that allow a broad set of non-infringing, socially valuable uses, lack of interoperability between DRM systems provides a barrier to use of many works and also limits marketability of equipment (and works) that fail to interoperate. In order to maximize the market reach of their works, equipment producers need to design their equipment to include mechanisms that interoperate with all DRM systems; similarly, copyright holders need to protect their works with DRM systems that do not interfere with use of works on existing media and that can be used on a wide range of equipment. In the absence of standards, many DRM systems are not yet interoperable with each other; proprietary DRM systems are used to extend the reach of copyright holders to secondary markets, e.g., iTunes uses a proprietary DRM which is only compatible with iPod. Over the years, several DRM technologies and potential standards have appeared that include

[2]Fair use is not limited to a clear set of uses; it is complex and is usually determined on a case-by-case basis.

the Digital Object Identifier (DOI) and the eXtensible Rights Markup Language (XrML). However, these technologies are still far from becoming broad standards.

### Trusted Computing

DRM systems impose certain limitations on the use of works. Despite their overreaching implementations and the resulting side-effects on some socially valuable uses, these DRM systems can only be effective if devices in fact execute the license stored in DRM systems. This observation has led to "trusted computing initiatives" [120] which embed DRM systems in the platform infrastructures [106].

The trusted, DRM-based computing platforms have been the product of the Trusted Computing Group [120] which includes Microsoft, Intel, IBM, HP, and AMD. Trusted computing "provides a platform on which [users] can't tamper with the application software, and where these applications can communicate securely with their authors and with each other" [5].

The most prominent example of a trusted computing platform is Microsoft's Next Generation Secure Computing Base (NGSCB). This platform uses a smartcard chip along with secure memory, operating system, and security kernels in applications, and "a back-end infrastructure of online security servers maintained by hardware and software vendors" [5].

The trusted computing platform performs authenticity checks on devices, monitors the state of the device at each stage of application execution, and compares the observed state with the expected state to confirm that the platform and the applications have not been tampered with. If the expected and the observed states do not match, the platform can prevent applications from running [120]. In protecting the integrity of the system and applications, the computing platform respects the wishes of the software and hardware vendors, as well as copyright holders. Authenticity of works can be checked with the vendor automatically, and works that do not pass the authenticity checks can be prevented from running on the device.

Trusted computing promises to transfer the control of personal computers from the owners of the machines to copyright holders. As such, the development of this platform has been controversial. Given that the owners of computers can be opposed to limitations imposed on their devices by outside parties, the level of success of these systems to protect copyrights is unknown. Moreover, these devices can be used to limit the software that can run on certain devices and have implications for innovation as well as competition [5].

#### Costs and Benefits of DRM-based Technologies

Trusted computing initiatives and DRM systems give copyright holders control over their works and revenues. However, these systems also threaten to give the control of personal computers and legitimately purchased works to copyright holders instead of the owners of equipment and legitimate users of works.

In general, DRM systems have been widely criticized for enabling limits on distribution channels based on first sale, for enabling limits on uses based on fair use and freedom of speech, and for enabling invasion of user privacy. So far, the public response to overreaching use of DRM systems has been civil disobedience: those with the means bypass these protection mechanisms to use and manipulate works. As a result, through this approach, copyright holders fail to protect their revenues adequately, and also limit the public interest in their works.

### 3.1.2 Digital Tracking Technologies

Another technology frequently used for intellectual property protection is digital tracking based on copy recognition. The most prevalent forms of copy recognition rely on digital labels, such as watermarks, that facilitate identification of works.

#### Digital Tracking Using Watermarks

Watermarks, i.e., imperceptible identifying information embedded in works [136], label works and enable tracking them.[3] Watermarks can be used to label each work with a unique identifier so that each digital copy can be recognized. The unique identifiers can encode any information, including information about the identity of the person who purchased the particular copy, who the copyright holder is, the terms of use, and any other information the copyright holders desire.[4]

In general, in order to protect intellectual property, digital watermarks need to be [92]:

1. Reliable, so that both false negative and bit error rates are low,

2. Robust when the works are used normally,

3. Tamper resistant, and

[3] One open-source labeling system for digital music files is provided by MusicBrainz. http://www.musicbrainz.org.
[4] E.g., Giovanni Digital Watermarking, http://www.bluespike.com/giovanni/gdigmark.html.

4. Based on public algorithms—their strength should come from the nature of the algorithm and not from keeping it secret.

Watermarks do not interfere with usability of works; this technology does not limit distribution or use of works, allows fair use and free speech without any limitations. If used to link users with works, watermarks can enable invasion of privacy of users. However, this particular implementation is not necessary for watermarks to be useful; these technologies can be used to implement different proposals that can address the digital copyright problem. One proposal that protects the revenues of copyright holders without limiting the public's ability to access and use works focuses on tracking works without tracking users, so that information about the popularity of works can be used for proper compensation of copyright holders. Where copyright holders are compensated proportionally to popularity of their works, widespread distribution and use of works benefits both copyright holders and users; copyright holders gain revenues based on distribution and use of their works, and users have easy access to works and can use works without limitations of fair use, free speech, or concerns for their privacy.

Following in this vein, Gerovac and Solomon [43] suggested protecting "revenues not bits" by tracking works in order to compensate copyright holders without interfering with the rights and privileges of the public. However, a solution that relies on identification of works in order to compensate copyright holders has to provide some guarantee of authenticity of the identifying information used for tracking works, i.e., that the identifying information has not been tampered with, so that works can be identified accurately.

The last decades have seen an increase in the number and quality of available watermarking schemes [109]. However, mechanisms for removing watermarks and for disabling watermark readers have also become widely available [93].

### Digital Tracking Using Content Fingerprints

An alternative to digital tracking using labels, such as watermarks, relies on recognizing works from their content—an approach that has become popular recently, e.g., Audible Magic[5], Relatable[6], and SCAM [110, 111]. These technologies focus on different kinds of content that varies from music to text. SCAM is a text-based approach that identifies copies that exhibit verbatim similarity to their

[5]http://www.audiblemagic.com
[6]http://www.relatable.com

original (or to each other) and has been shown to successfully identify copies that have overlaps with each other [110, 111].

### Costs and Benefits of Digital Tracking

Digital tracking mechanisms do not interfere with socially valuable uses of works, they do not limit distribution of works, they do not interfere with fair use or free speech. However, some implementations of digital tracking can raise privacy concerns for users. If addressing the digital copyright problem indeed required giving copyright holders control over all copies of their works, digital labels and tracking mechanisms could achieve this goal by providing a unique label for each individual copy of a work, at the cost of potential privacy concerns, e.g., if the unique label can be linked to an individual. Alternatively, digital labels and tracking can be designed and used to measure popularity of works, without tracking individual copies; content-based tracking mechanisms can also be used for this purpose. In the absence of appropriate compensation mechanisms however, digital tracking by itself does not guarantee any revenues to copyright holders.

By combining digital tracking with payment systems based on metering popularity of works, we can protect the revenues of copyright holders without deterring or inhibiting privileged and legal uses and without raising privacy concerns. Indeed, as long as works can be tracked, revenues can be protected even if copies are not [43, 121, 122].

Solutions based on digital tracking can be most effective if tracking is accurate and reliable. Unfortunately, when tracking systems rely on digital labels such as watermarks, unlabeled copies of works, e.g., copies that have never been labeled and copies whose labels have been tampered with, cannot be tracked.

Content-based tracking systems can recognize copies of works that have verbatim overlap with the original even in the absence of digital labels. In order to further improve the accuracy of these systems, we propose tracking works using fingerprints that are based on the expression of content in works and that recognize paraphrased as well as verbatim copies [121, 122]. We discuss the expression fingerprints in Section 3.4, and provide the details of this fingerprinting system in the rest of this thesis.

## 3.2 Technology & Policy Solutions

Both DRM and digital tracking mechanisms require supporting policies in order to address the concerns of stakeholders adequately. The anti-circumvention provisions of the DMCA, for example, aim to strengthen the DRM systems by making it illegal to circumvent these technologies. In the absence of this policy enforcement, many users would break the DRM systems and use of DRM systems to protect revenues of copyright holders would be meaningless. Given DRM systems and anti-circumvention provisions, a different set of policies are necessary to protect the ability of the public to access and use works. Similarly, digital tracking mechanisms, by themselves, do not protect the revenues of copyright holders, although their use does not interfere with access to and use of works.

In this section, we discuss four policy proposals that support DRM systems and digital tracking mechanisms in addressing the digital copyright problem:

- Burk and Cohen try to address the overreaching fair use and privacy concerns of DRM systems through obligatory relaxed DRM systems that enable a default set of fair uses and through a trusted intermediary that can protect the privacy of the users and distribute keys to those who need access to works in order to make fair use [19].

- Fisher proposes compensating copyright holders through taxes imposed on equipment and media used for copying, using, and distributing works. He suggests a system where copyright holders register their works with the Copyright Office in order to qualify for compensation for use of their works. The Copyright office assigns unique identifiers to works, and uses this information to meter popularity of registered works [39].

- Netanel focuses on the effect of peer-to-peer networks on the revenues of copyright holders and argues for immunity of users from infringement claims for noncommercial uses of works on peer-to-peer networks in return for noncommercial use levies (NUL) imposed on the services and equipment whose values are enhanced by these large-scale noncommercial uses [87]. Similar to Fisher's proposed compensation mechanism, Netanel also proposes to compensate copyright holders proportional to the value of their works for the society. He claims that popularity of works as measured by digital tracking systems that can keep track of different uses, not just downloads, provides a good proxy for this value.

- Lessig et al. try to promote and foster the development of the public domain through the Creative Commons initiative that to some extent liberates the less revenue-oriented authors from profit-maximizing middlemen. Under the Creative Commons project, authors can retain their copyrights while allowing certain uses of their works; the initiative includes standardized licences and tools to support this [22].

All of these proposals try to balance the incentives of copyright holders to create with the ability of the public to access and use digital works. In this chapter we discuss these proposals from the perspective of enhancing (or at least not inhibiting) fair use and free speech, protecting privacy, enabling digital first sale (or an analogous mechanism that provides the public with low-cost access to works), and protecting revenues of copyright holders.

### 3.2.1 Trusted Third Party

Cohen and Burk focus on the effect of DRM systems on fair use and suggest mitigating this problem by first requiring all DRM systems to enable commonplace fair use privileges by default. Once DRM systems are programmed to enable the default set of fair uses, to further reduce the effects of DRM systems on unforeseen fair uses or those fair uses that are not covered by the default set, they propose to supplement the slightly relaxed DRM systems with a trusted third party that can use a key-escrow system to issue keys that can unlock, i.e., decrypt, the DRM systems in order to support fair uses that are out of the ordinary [19]. According to this proposal, the keys that allow access to DRM-protected works would be deposited to the trusted intermediary in advance. The copyright holders would be granted copyright protection for their works in return for depositing their keys. Only those works for which keys have been deposited to the intermediary would be eligible for copyright protection and copyright holders who fail to deposit their keys would not be able to bring lawsuits for circumvention of technological protection mechanisms used in their works.

Such a system reduces the problem associated with the inability of DRM systems to encode all possible fair uses of a work; fair use is situation-specific and is determined on a case-by-case basis. However, a system that determines whether a use is fair before authorizing the use is problematic both because it deters spontaneous fair uses and because it puts a burden on the intermediary, namely that the intermediary would have to justify its decisions to the courts and could be sued by copyright holders in case of misjudgements.

To minimize the cost of this process, both in terms of time and the risk of litigation against the intermediary, Burk and Cohen go on to suggest limiting the responsibility of the intermediary to

distribution of keys so that the public can make fair use of works anonymously and without revealing themselves to copyright holders. The issuance of keys by the trusted third party would not put any responsibility on this agent to enforce copyrights or to make a judgement about the nature of the use, i.e., whether it is in fact fair use. The decision (and therefore the risk) regarding whether the use is fair or not would still remain the responsibility of the consumer, as is the case in the non-digital world. In the case of issuance of keys for infringing uses, the dispute would take place between the user and the copyright holder. After infringement is proven, the trusted party might have to reveal (after a subpoena or court order from the copyright holder) the identity of the person responsible for the use or the key request [19]. This approach would protect the privacy of fair users and limit the ability of copyright holders to control fair uses, especially those such as parody and criticism that are of value to society but are not always in the best interest of copyright holders.

Addressing fair use and privacy concerns of the public through an intermediary, in fact, also alleviates concerns regarding freedom of speech. This mechanism can also address some digital first sale concerns, if at least one time resale of works is allowed by default by the DRM systems. Further, DRM systems used to protect copies distributed by libraries can be adjusted to implement digital first sale, although details of this implementation need to be worked out. Through the use of DRM systems, this proposal would also protect the revenues of copyright holders. As a result, this proposal would encourage copyright holders to create and disseminate their works; it would also enable the public to access works through digital first sale, and use works for free speech, for fair use, and in private.

However, the difficulty of identifying a set of fair uses that can be built into DRM systems complicate the implementation of this solution. Considering the already existing interoperability concerns with DRM systems, the difference in the level of protection each copyright holder desires for their works, and difference in the nature of fair uses which depend on the work (along with other concerns), mandating a standard level of fair use privileges on all creative works is not straightforward. As a result, a range of different use privileges may be allowed on different sets of works, and the diversity of the available creative works may make it unclear, both for copyright holders and consumers, which fair use privileges by default should be allowed for which works.

### 3.2.2 Creative Commons

Groups that oppose the stringent enforcement of copyrights have proposed more relaxed rights control systems. Creative Commons [22] is one such initiative whose goal is to promote and protect

the public domain. To achieve this goal, Creative Commons provides licences that enable authors and copyright holders to retain their copyrights and reserve some rights while allowing large-scale distribution and use of their works.

Creative Commons offers copyright holders four basic rights: the right to ask for attribution, the right to control commercial use, the right to prevent modifications, and the right to demand that people share under similar conditions any derivatives they create. For each of their works, copyright holders are free to reserve any subset of these four basic rights. Creative Commons offers licences in machine-readable form, so that machinery can be developed that can automatically identify the license conditions under which works are released [22].

Creative Commons achieves many of the goals necessary for the public to access and use works. It enables large-scale dissemination of works, it raises no concerns for fair use and free speech, and it does not affect privacy of users. Because dissemination of works is free, there is no concept of secondhand markets and restrictions on their functions. However, the ability of this solution to protect the creative incentives of copyright holders is not clear. While some creators embrace the ability to reach larger audiences through Creative Commons, for attribution even if not for compensation, many profit-driven creators may not find the terms of Creative Commons licences adequate—they may not be willing to relinquish profits they could collect from large-scale distribution and use of their works if these uses are protected by DRM systems or digital tracking mechanisms.

### 3.2.3 Taxation-Based Solutions

Netanel [87] and Fisher [39] propose protection of revenues of copyright holders through taxes imposed on media, products, and services used for sharing, copying, distributing, and even modifying works in return for immunity for users and innovators from infringement claims for such uses. The goal is to use these taxes to offset the displacement of the revenues of copyright holders due to use of digital media without restricting use and distribution of works by users, encouraging dissemination of works by both users and copyright holders, and reducing conflicts between stakeholders regarding free speech, fair use, and privacy.

In order to provide copyright holders with adequate compensation without limiting access and use capabilities of the public, Fisher proposes a copyright registry system and taxation-based compensation of copyright holders. According to this proposal, a copyright holder who wishes to protect the copyrights in a work would register it with a registry office, such as the Copyright Office, and clearly state the copyrights reserved in the work as well as the components of the work to which

these rights applied. In the case of works that combine expression of different creators, the copyright holder would attribute certain percentages of the work to relevant parties. This registration process would be considered a prerequisite for copyright holders to be eligible to receive any compensation, and the registry and the supporting government offices would be responsible for raising enough funds to compensate the registered copyright holders.

Fisher's preferred source of funds is taxes imposed on digital equipment and media that enable large-scale distribution and use of works; the level of these taxes would be calculated so that the collected revenues would offset the reduction in revenues of copyright holders due to digital, currently unauthorized, distribution of their works. Metering use of works, so that the revenues would be distributed justly among relevant parties, would be achieved through a system similar to that used for television rating services. In other words, popularity of different works among a sample of households would be used to estimate popularity of works. If the sample of households were carefully selected, this method would provide a way of evaluating popularity of works by closely observing how much these households use, not just download, works.

However, no sampling mechanism is perfect and the samples selected for metering use of works can be biased. Therefore, the sample selection should be done very carefully so that the sample selected is to be representative of the population. In addition, taxes collected from the sales of equipment distribute the cost among the users unjustly—low-volume users and high-volume users share the bill equally. Netanel's taxation-based solution to the digital copyright problem suffers from the same drawbacks.

Netanel presents a similar taxation-based proposal that gathers revenues from "Noncommercial Use Levy" (NUL), i.e., levies imposed on peer-to-peer network services and products whose value is significantly increased by widespread use and dissemination of works on these networks. Netanel points to existence of taxation schemes aimed at compensating copyright holders: before the advent of peer-to-peer systems that allowed such large-scale sharing and distribution of works, equipment that enabled personal copying was taxed for copyright purposes. Similarly, compulsory licensing regimes imposed on entities that broadcast and distribute works, e.g., TV operators and webcasters, protected the revenues of copyright holders. A NUL imposed on service providers, such as ISPs, in return for unlimited noncommercial distribution of works among their subscribers, would provide an analogous compensation system for the use of works on peer-to-peer networks. Similarly, providers of other equipment and media whose value is enhanced by unlimited use, reproduction, and distribution of works would be subjected to the levy.

Netanel extends noncommercial distribution privileges to modified material, effectively arguing for immunity for network users against unauthorized copying, distribution, and creation of derivative works, in return for revenues raised through levies. Netanel estimates that a levy equivalent to 4% of the retail value of peer-to-peer products and services would raise ample funds to offset the displacement of the revenues of copyright holders due to use of these products and services. These funds would be distributed among copyright holders proportionally to the popularity of their works as determined by digital tracking mechanisms employed for metering distribution and use of works. The tracking mechanisms aimed at this purpose, such as those provided by Relatable[7], Audible[8], or Entriq[9], would possibly reside either on the ISP gateways or on user devices [87].

Perfect metering of distribution of works is possible only if every copy of every work is accessible and identifiable by a tracking system. However, at the current state of technology, digital tracking mechanisms do not have access to many media that are used for experiencing works, e.g., mp3 players. But, the distribution of works via the Internet, and their level of use on individual machines, could be tracked, and could provide a large enough sample so that leaks due to untrackable uses do not appreciably change the measurement. However, this kind of tracking assumes absence of encryption technologies and privacy enhancing technologies that can encrypt the works and reroute them through random intermediaries and that also flood the systems with cover traffic in order to make it very difficult to recognize works and meter their downloading and use.

Both of the above tax-based solutions suffer from privacy concerns raised due to close metering of use of works. If metering is performed on a sample of households and if the sample is chosen from volunteers, the sampled population is likely to be relatively insensitive to privacy issues and this may introduce a bias into the sample. This bias can be minimized by metering use of works among the general population (so as to get as large a sample size as possible) and policies that allow the gathered data to be used only in aggregate can alleviate the privacy concerns associated with metering.

When revenues are distributed proportionally to the relative popularity of works, artists and copyright holders have incentives to game the system by artificially increasing the popularity of their works, for example by repeatedly playing or downloading their own songs. Based on studies on metering use of works when people choose from a large number of options, i.e., in this case digitally

[7] http://www.relatable.com
[8] http://www.audiblemagic.com
[9] http://www.entriq.com

available works, Netanel states that popularity of works over the Internet is likely to follow a "power law distribution" where most people use a small number of especially popular works, meaning that artists that try to game the system by inflating the use of their works artificially would only be able to displace mostly those works that are of "middle to marginal popularity", and receive only a small fraction of the revenues. Such cheating would not cripple the taxation-based compensation system. To further minimize the effect of such cheating, the tracking mechanisms could track a "constantly changing, random sample of uses", minimizing the effects of potential abuse of the system [87].

Common criticisms to tax- and license-based systems revolve around the administrative cost of maintaining these systems. Indeed, there are significant administrative costs in the imposition of extra taxes, the collection of these taxes, the collection of necessary information that can be used for proper allocation of revenues among copyright holders, and the distribution of these revenues. In addition, the public pays by sharing the taxes imposed on entities that provide the products and services that enable distribution and use of works; inevitably, low-volume users of such products and services unjustly subsidize the activities of high-volume users. Netanel suggests that most low-volume users will not mind paying "a surcharge for the possibility of unlimited file sharing even if they don't actually engage in much file sharing"; however, Netanel argues that, the taxation schemes have to be designed so as to avoid the non-user subsidy of low-volume users [87]. Odlyzko's research [90] suggests that consumers of information products prefer flat rates over differential unit pricing, even if that requires some users to pay somewhat more overall.

Overall, both taxation-based proposals encourage copyright holders to create and disseminate their works to the public by providing them with revenues that reflect the value of their works for society. These solutions do not interfere with fair use, free speech, or even distribution of copyrighted works between users. Copyright holders benefit from the distribution and use of their works, even when distribution is initiated by individual users. However, depending on the implementation, the digital tracking mechanisms used for metering distribution and use can raise privacy concerns. In order to eliminate these concerns, the collected information has to be used only in aggregate, and this policy should be strictly enforced.

## 3.3 Lessons Learned

The policy proposals reviewed in this section address the same general problem in different ways but share one common goal: to encourage the copyright holders to create and disseminate works, and to enable the public to access and use works, without necessarily having to obtain a copy at full price, to make fair use, to practice free speech, and to perform all these activities without any concerns for their privacy.

Solutions based on digital tracking and tax-based revenue collection are promising because they do not rely on DRM systems that can limit spontaneous fair uses [19]; they also guarantee revenues to copyright holders in proportion to popularity of their works in order to encourage them to create and disseminate more works.

In these mechanisms, popularity of works is determined through technological tools. Fisher suggests a simple sampling-based system where a representative sample of the population is watched closely, through technological means, in order to determine popularity of different works. Netanel suggests digital tracking technologies for metering distribution and use of works even when they are modified. As we mentioned in Section 3.1.2, digital labels used for quick identification of works can be absent: in some cases the works may have been created without the watermarks and in some cases the watermarks may be tampered with. Inadequacy and inaccuracy of metering mechanisms undermines the solutions that rely on digital tracking technologies.

The main goal of this thesis is to provide a method for fingerprinting novels so that literary works can be tracked and their use can be metered accurately even when they are paraphrased, in the absence of any other labelling information such as watermarks. Similar content-based fingerprinting attempts are under development mostly by commercial parties for different digital works; for example, Relatable claims to recognize music files from the acoustic properties of the music itself by using pattern analysis[10] and Audible Magic fingerprints audio files based on perceptible properties of their content[11]. Our approach to fingerprinting text files not only recognizes written works from patterns that distinguish them from others, but actually learns elements of expression that remain unchanged even when works are paraphrased. Our fingerprinting method also has applications in plagiarism detection, which is outside the scope of this thesis.

[10] http://www.relatable.com
[11] http://www.audiblemagic.com

## 3.4 Digital Tracking Using Expression Fingerprints

Many technology and policy solutions to the digital copyright problem rely on identifying and labelling intellectual property. Watermarks have so far been the dominant method for labelling data: they label works with information about the works' content, the identity of the copyright holders, authorized uses, and other similar information, in machine-readable form. An alternative to such labels is recognition of works from their content—an approach that has become popular recently, e.g., Audible Magic, Relatable, Scam [110, 111], and mostly focuses on identification of copies that exhibit some level of verbatim similarity.

Our approach to fingerprinting also uses the actual content of works, but focuses on the expression of content by identifying and studying the expressive elements of works instead of studying verbatim components. In the case of text-based works, expression is related to language and linguistic elements of works. Therefore, our fingerprints focus on these elements of expression; we use natural language processing techniques such as low-cost syntactic analysis and study the elements of the language used in a work that differentiate the work from others. This mechanism relies on a repository of works and builds models representing the expression of the documents in this repository; it uses these models to recognize the identity of works that are presented to the system thereafter. Details of the elements of expression studied in this thesis will be presented in the remaining chapters.

When fingerprints are extracted directly from the works, making the fingerprint of a work unrecognizable requires substantial changes to the work itself. This allows identification of works that are derived from each other or from a common source, sometimes without proper attribution, as well as identification of works that are superficially modified, without any real change to the expression of the original content. Even when works are circulated without identifying digital labels, and even when such labels are removed, the expression of the content in works remains intact, and the works can be recognized by studying the expression they contain.

Digital tracking using expression fingerprints, whether or not they are verbatim, is naturally more computationally expensive than tracking with labels such as watermarks. However, development of networks of devices, such as computational grids and wireless computational grids [79], and the ability of networked devices to share computational and storage resources, enable cost-effective and efficient fingerprinting on these networks. These developments indicate that the relative computational expense of the content-based fingerprinting mechanisms will soon be insignificant for

large-scale deployment.

Availability of a fingerprinting mechanism that can recognize works in the absence of identifying labels and even when works are modified improves the accuracy of digital tracking mechanisms, and allows more works (even works that are not labeled) to be tracked. Increased accuracy of digital tracking due to availability of such a system strengthens policy proposals that rely on determination of popularity of works. We predict that the fingerprinting mechanism we propose will be an integral part of policy solutions such as those proposed by Netanel and Fisher, and will be used for metering use of works.

Independent of such policy proposals, availability of mechanisms that recognize works, their copies, and derivatives gives copyright holders the ability to understand the demand for and use of their works, and to prepare future works that address that demand.

## 3.5 Conclusion

We believe that in order to encourage learning and to promote progress in the digital world, any solution to the digital copyright problem should provide copyright holders with incentives to create and disseminate their works, and it should enable the public to access and use these works, sometimes through alternative distribution channels enabled by first sale, for fair use, to practice free speech, and without imposing any limitations on their privacy.

Digital tracking, when implemented to track popularity of works rather than copies of works or individuals, can achieve this balance and can also protect the revenues of copyright holders.

Most work related to digital tracking has used digital labels, such as watermarks, that mark individual copies of works and make tracking easy. However, these labels can be removed or altered, preventing identification of works. As an alternative to watermarks, we propose using fingerprints that are directly generated from the expression of the content in works and that can identify even paraphrased copies. Our proof-of-concept mechanism for identifying novels uses fingerprints that are based on analysis of linguistic elements of works and that represent the expressive elements of works in order to recognize both verbatim and paraphrased copies. These fingerprints cannot be removed or altered because they are generated from the works themselves on demand. Thus, they make it possible to implement proposals that rely on accurate metering of use of works (even when works are paraphrased and even when watermarks and other identifying information are removed). When copyright holders are compensated proportionally to popularity of their works, large-scale dissemination and use of works benefits both copyright holders and users, encourages copyright holders to disseminate more of their works to the public, enables users to use works for fair use and free speech without any threats to their privacy, and brings us closer to an equilibrium that satisfies all stakeholders.

## 3.6 Summary

Digital copyright problem can be addressed through technology and/or policy. DRM systems, and digital tracking are the two most popular forms of digital rights enforcement currently in use. Both of these technologies have their benefits and costs:

- DRM systems prevent at least the general public from infringing copyrights and therefore provide some protection for the revenues of copyright holders. However, their overreaching use aggravates the public, limits dissemination as well as fair use and free speech and can be invasive of privacy. This results in circumvention of DRM systems.

- Digital tracking mechanisms track works without interfering with their distribution and use; however, they do not protect the revenues of copyright holders—one can only use tracking to recognize infringement after it has taken place.

Solutions that take advantage of the strengths of these technologies and propose policies that can minimize their shortcomings, to both copyright holders and society, can lead to a successful resolution of the issues regarding digital copyrights. A trusted third party that reduces DRM systems' effects on privacy and fair use provides one example [19]. Policy proposals that benefit from digital tracking and that reduce the effects of its overreaching aspects, e.g., privacy concerns, provide another example [39, 87]; they can balance the copyrights with social interests and provide a solution that promotes progress.

Ideally, all copyright holders should be fairly compensated for use and distribution of their digital works. The value of their works to the society should be used as a guide for this purpose. Solutions based on digital tracking are most suitable for such a solution. Netanel and Fisher both present policy solutions that take advantage of digital tracking to meter use and determine popularity of works in order to provide such fair compensation. Netanel's proposal is dependent on the creation of digital tracking mechanisms that recognize not only literal but modified copies of works. In this thesis, we present a system for fingerprinting novels that achieves this goal by capturing the expression of content in a work.

# Chapter 4

# Content, Expression, and Style

> §102. Subject matter of copyright: In general
>
> (a) Copyright protection subsists, in accordance with this title, in original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device. Works of authorship include the following categories:
>
> (1) literary works;
> (2) musical works, including any accompanying words;
> (3) dramatic works, including any accompanying music;
> (4) pantomimes and choreographic works;
> (5) pictorial, graphic, and sculptural works;
> (6) motion pictures and other audiovisual works;
> (7) sound recordings; and
> (8) architectural works.
>
> (b) In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work.
>
> United States Code, Title 17, Chapter 1, §102.

Thus defined, the subject matter of copyright includes an author's expression of content, but not ideas or facts.[1] Evaluating the similarity of documents for copyright purposes requires identifying expression of particular content.

For literary works, we distinguish between idea (i.e., content) and expression as follows: By content, we mean the story or the information contained in documents. By expression, we mean the linguistic choices made by authors in presenting content, such as authors' choices of particular vocabulary items from a set of synonyms (e.g., words "clever" vs. "smart" in sentences shown

[1] United States Code, Title 17, Chapter 1, §102.

in (1) below), whether they use direct or indirect speech (e.g., sentences in (2)), passive or active voice (e.g., sentences in (3)), direct statements or rhetorical questions, or whether they prefer complex sentences with embedded relative clauses to simple sentences with independent clauses (e.g., sentences in (4)), as well as combinations of such choices.

1. (a) Jill is very clever.

   (b) Jill is very smart.

2. (a) Jill said: "I have to go!"

   (b) Jill said that she had to go.

3. (a) The pirates sank the boat.

   (b) The boat was sunk by the pirates.

4. (a) The woman carrying the red umbrella walked in the rain without getting wet.

   (b) The woman walked in the rain without getting wet. She was carrying a red umbrella.

In this thesis, we rely predominantly on experiments with a corpus of parallel translations. Parallel translations of literary works are interesting because they differ in expression but convey similar content. Consider the following short excerpts taken from three different translations of *Madame Bovary*, written by Gustave Flaubert.

> Excerpt 1: "Behind the door hung a coat with a short cape, a bridle, and a black leather cap. And on the floor in a corner lay a pair of gaiters still caked with mud." (Written by Flaubert, translated by Unknown1.)
>
> Excerpt 2: "Behind the door hung a cloak with a small collar, a bridle, a black leather cap, and on the floor, in a corner, were a pair of leggings, still covered with dry mud." (Written by Flaubert, translated by Aveling.)
>
> Excerpt 3: "Hanging up behind the front door were a cloak with a narrow collar, a bridle, a black leather cap, and, in the corner, on the floor, a pair of leggings covered with stale mud." (Written by Flaubert, translated by Unknown2.)

Despite being semantically similar and based on the same book, the three translations differ in expression. For example, instead of the word `covered` used in excerpt 2, the translator of excerpt 1

uses `caked`. Similarly, the adjective `dry` in excerpt 2 is replaced with `stale` in excerpt 3. The translator of excerpt 1 also chose to break the content into two simpler sentences, unlike the long sentence of short phrases connected with punctuation and conjunctions in excerpts 2 and 3. Finally, excerpt 3 starts with a gerundive verb while the other excerpts start with the prepositional phrase "behind the door", resulting in a different flavor of literary expression.

Expression, as defined in this thesis, focuses on the linguistic choices of the authors, such as those described above, and does not include layout or generic genre characteristics of documents because neither layout characteristics (such as use of titles, tables, and figures) nor genre characteristics (e.g., all poems consist of stanzas) represent linguistic choices of the writers.

Expression and style both refer to linguistic elements of an author's writings, but differ in their goal. Style refers to linguistic elements that persist over the works of an author; style has been widely studied in the literature for authorship recognition [12, 53, 65, 85, 91, 128]. Expression involves linguistic elements of a single work and can be used in identifying potential cases of copyright infringement. Similarities in the expression of the same content in two different works signal potential copying and require further scrutiny under copyright.

In text classification literature, the study of style focused on elements of language that are independent of content, that stay consistent in the works of authors, and that differentiate the language of an author from the language of other authors. For example, Mosteller and Wallace have used function words to differentiate between the styles of Madison and Hamilton in the Federalist Papers [85]. Peng used the same function words to differentiate between Austen, Dickens, and other authors [91]. Koppel used bigrams of part-of-speech tags in addition to function words to predict the genders of authors [65]. Many other researchers used measures of vocabulary richness and sentence length to recognize authors [53].

Expression refers to how an author phrases particular content. This means that a unique fingerprint that can identify a work for copyright purposes has to capture the expression of content that is unique to that work, and that differentiates it from the expression of other content by the same author, as well as the expressions of other authors when they write about similar content.

Expression and style are both based on linguistic elements of authors' writings. However, different linguistic elements are useful for capturing each. For example, if an author always uses long sentences, his style can partly be described in terms of the length of his sentences; however, his expression in a particular work would not be captured by this information, because sentence length—being part of his style and being consistent in his works—would not be useful for dis-

tinguishing among his works, and would not be useful for creating the unique fingerprints that would help identify potential copies of each of these works. On the other hand, the author may use different forms of embedded sentences in different works; one work might have predominantly left-embedded sentences while a second one might have predominantly right-embedded sentences. This information can be used to capture the difference between the expressions of his works, but would not help define his style.

Which linguistic features are more useful for identifying expression and which are more useful for identifying style depends on the group of authors that are studied. If the goal were to distinguish the style of Shakespeare from that of Austen, looking at their word choices would reveal much about their style, simply because the two authors lived in different time periods and used different vocabularies. If the goal were to differentiate between Twain and London, vocabulary features would not be as useful and we would need to identify other linguistic elements. When we are comparing works by one person and want to capture his expression in a particular work, we need to include even more linguistic information, because most of the high-level linguistic elements will not vary between his works. For example, function words, used frequently in studying style, cannot be expected to differentiate between the works of the same author as successfully as they differentiate between authors. The reason they are useful for recognizing authors is because authors use them consistently even when they write about different content. Deeper linguistic analysis might reveal differences in the structure of phrases, e.g., noun phrases, that contain these function words in different works of a single author and might reveal his expression in each work.

Deeper linguistic elements, such as sentence and verb phrase structure, provide a general set of features that can differentiate between the works of authors of the same time period and who write in the same genre (see Chapter 8). Simpler features, such as distribution of word lengths [80], differentiate between genres, allowing us to divide the expression fingerprinting task into subtasks, i.e., the task of capturing expression in a genre.

In this thesis, we set out to identify syntactic elements of expression. We demonstrate this capability on novels, i.e., narrative fiction, using a methodology that can be applied to other genres. In Chapter 5 we experiment with some of the existing language-oriented and content-independent features from literature to identify features that are useful for differentiating between translators of similar content. Inspired by the findings of these experiments, in Chapter 6 we study deeper linguistic elements and identify a novel set of syntactic features that reveal expression in the genre of narrative fiction.

# Chapter 5

# Preliminary Experiments

The study of expression requires identification of linguistic elements that capture the differences in the way people convey content. The goal of this thesis is to devise a novel set of features that serve this purpose (see Chapter 6). By analyzing features obtained from text classification literature, we can evaluate the promise of each feature for capturing expression, and identify features that need to be studied in more detail to define a novel set of elements that capture expression.

In this chapter, we describe a set of features from the literature that are more focused on language than content [53] and we evaluate the relative strengths of these features in recognizing the differences in the linguistic choices of translators, even when they translate the same original. Our results show that:

- The features selected from the literature can be used to recognize translators accurately 75.6% of the time, and

- The standard deviation of sentence length and syntactic features of sentences such as the frequency of use of "get-passives" and "be-passives" are among the top five most useful features for differentiating between the translators of the same content.

The relatively high informativeness of features related to sentence structure, e.g., frequency of alternative passive constructs, warrants deeper analysis of sentence structure for the study of expression. The relatively high importance of the standard deviation of the sentence lengths in different books indicates that information about the length of sentences within a work is also useful for the study of expression; the difference in the lengths of sentences may provide information about the levels of linguistic complexity of sentences in different books or throughout a book. Based on

these results, in Chapter 6 we study sentence structure and length in more detail and identify a novel set of syntactic elements that capture expression.

## 5.1 Features from the Literature

Different sets of features are appropriate for different text classification tasks. Some text classification tasks require features that are more focused on content; others require features that are more dependent on linguistic choices. To gain insight into what differentiates the language of two translators of the same content, we considered a set of features that focus more on language, and less on content. These features are grouped into surface, syntactic, and semantic features.

### 5.1.1 Surface Features

Surface features are those that can be extracted from text without any tagging (i.e., assigning part of speech tags to works) or parsing (i.e., determining the sentence structure). Because they do not have information from a tagger or a parser, these features are usually limited to information that can be captured by the distributions of keywords and string lengths. In this chapter, we focus on five surface features:

- Number of words in the document;
- Type–token ratio, i.e., the ratio of the total number of unique words in the document to the number of words in the document;
- Average and standard deviation of the lengths of words (in characters) in the document;
- Average and standard deviation of the lengths of sentences (in words) in the document; and
- Number of sentences in the document.

### 5.1.2 Syntactic Features

Syntactic features are extracted from text after it is parsed, and after the syntax tree that best matches each sentence is identified. The syntax tree of each sentence contains information about the structure of the sentence that, for example, shows whether the sentence is interrogative, imperative, or declarative, whether the sentence is in active or passive voice, or what type of genitive constructs

it contains. The information obtained from the syntax tree can be combined with lexical information to further differentiate between the underlying structures. For example, the use of passive constructs can be investigated in conjunction with lexical information to identify whether a passive voice sentence contains "be-passives" or "get-passives".

We parsed our corpus with the Collins parser [29] and collected information about the following syntactic features:

- Sentence type:
  - Frequency of declarative sentences, i.e., constructs that follow the subject–verb–object pattern;
  - Frequency of interrogatives, i.e., constructs that exhibit subject–auxiliary inversion, sometimes accompanied by wh-phrases, e.g., what, which, who, why, etc., and appropriate punctuation, as well as wh-questions that do not exhibit subject–auxiliary inversion;
  - Frequency of imperatives, i.e., constructs that start with an imperative verb and do not have an explicit subject;
  - Frequency of fragmental sentences;
- Voice:
  - Frequency of active voice;
  - Frequency of be-passives, i.e., passive constructs that use "be", e.g., I was robbed;
  - Frequency of get-passives, i.e., passive constructs that use "get", e.g., I got robbed;
- Genitive use:
  - Frequency of 's-genitives, i.e., possessive " 's" observed in the "*noun*'s *noun*" construct;
  - Frequency of of-genitive, i.e., possessive "of" observed in the "*noun* of *noun*" construct; and
  - Frequency of phrases that lack of genitives, i.e., noun phrases that do not include genitives.

### 5.1.3 Semantic Features

Thorough semantic analysis of documents requires creating a representation of meaning by analyzing the interactions and relations of the concepts present in a document. Short of tools that can

generate such a representation, researchers work with keywords and assign them to semantic categories to capture as much of the meaning in a document as possible. In the stylometry literature, in order to capture semantic features that are unique to the use of language rather than the content of the document, researchers study semantic classes of communicative words. These categories of words are very different from the semantic categories that capture the content of documents. For example, some communicative semantic categories can help identify whether the document has a negative or positive tone; others indicate the level of confidence reflected in the writings of the author. In particular, in our studies we focused on two semantic classes of words:

- Frequency of overt negations, i.e., explicit negations such as “not”, “no”, “nowhere”, “no one”, “none”, and several others [52], and
- Frequency of uncertainty markers, i.e., words like “can”, “could”, “maybe”, “may”, “might”, “should”, “kinda”, “sorta”, “probably”, “possibly”, etc.

We used the features discussed in this section to build models for recognizing translators in a corpus that contains one book from each translator.

## 5.2 Methods

Throughout this thesis, we used boosted decision trees. In this section, we introduce boosting and decision trees. In this chapter, we used boosted decision trees in combination with statistical methods, in particular t-tests, that reveal whether a particular feature contributes significantly to the classification performance.

### 5.2.1 Decision trees

Decision trees use a coordinate space to represent documents. In this coordinate space, each feature of the document is a dimension, and each document is a single data point. Given a set of documents that belong to different classes, decision trees divide up the coordinate space into regions that contain homogeneous groups of data points, i.e., data points that belong to the same class. In order to achieve this separation, the decision tree algorithm chooses both the coordinate axes and the values along these axes that are most useful for separating the data points with respect to their class. For example, the hypothetical test $x > 5$, if chosen by the algorithm, indicates that splitting the coordinate space where $x = 5$ will best separate the data points with respect to their classes. Each chosen test generates a branch in the decision tree. Sample points are split between these two branches depending on whether they satisfy the condition or not. The goal of this separation is to collect members of different classes under different branches of the tree. If this separation is successful, then given any sample point from the same distribution, we can follow the branches of the tree, evaluate the tests along a branch, and determine where in the tree, and therefore to which class, the sample belongs.

However, no single division guarantees to separate the data points perfectly, i.e., so that the data points that belong to each class will be separated from other classes. However, by iteratively breaking the coordinate space into progressively smaller regions, the decision trees improve the homogeneity of the data points in each region of the coordinate space. At each iteration, we use the feature (i.e., coordinate axis) and the test that best separate the classes of data points. This process continues until either all the regions of the coordinate space contain only homogeneous data points, or until no tests are left that can improve the separation of the data points. In both of these terminal cases, each of the final regions is assigned the label of the majority class represented in that region. Each of the final regions is defined by the cascaded tests along a branch of the tree, and the final node along a branch is referred to as the leaf.

The decision tree algorithm uses a measure of average disorder to evaluate each feature and each test at every iteration. This measure of average disorder is given by the formula:

Average disorder = $\sum_b \left(\frac{n_b}{n_t}\right) \times \sum_c -\frac{n_{bc}}{n_b} log_2 \frac{n_{bc}}{n_b}$

where

$n_b$ is the number of samples in branch $b$,

$n_t$ is the total number of samples in all branches, and

$n_{bc}$ is the number of samples in branch $b$ that belong to class $c$ [131].

Some implementations of the decision tree algorithm, including the one we use in this thesis, maximize the information gain instead of minimizing average disorder. Information gain is given by:

$$G(X) = I(n_{bc}, n_b) - EI(X)$$

where

EI(X) is Average Disorder as calculated above, and

$$I(n_{bc}, n_b) = \sum_c -\frac{n_{bc}}{n_b} log_2 \frac{n_{bc}}{n_b}$$

when

$n_{bc}$ and $n_b$ are not zero [137].

Because they choose the feature and the test that best separate the data in a particular region of the coordinate space at that point in time, the decision trees make only local optimizations. These optimizations continue until either the overall average disorder reaches a minimum or no features can be found that can further reduce the disorder. Given the requirement to find a model that brings the average disorder to zero, decision trees are susceptible to overfitting. In other words, the algorithm can generate a tree that accommodates only one sample point in each leaf (so that the average disorder of the tree is zero) and would give 100% accuracy in classifying the training points. However, the resulting decision tree classifier would not generalize to data points outside of the training set. One can minimize this risk of overfitting by stopping the division of the coordinate

space when:

1. Cross-validation, i.e., multiple rounds of training and testing on different subsets of the data set, shows that additional splits do not reduce the overall classification error, i.e., the percentage of data points that are misclassified;

2. The best candidate split reduces the average disorder less than a pre-set threshold amount;

3. The average disorder of the overall tree reaches a minimum that is not zero;

4. There are at least a threshold number of samples in each leaf of the tree; or

5. The information gain due to a split is not significantly different from information gain due to a random split.

Controlling the complexity of the tree by enforcing one of these constraints helps reduce the risk of overfitting [36].

### 5.2.2 Boosting

Boosting is an algorithm for improving the accuracy of any given classifier [108, 107]. In general, boosting works by calling a classifier such as a decision tree multiple times. At each round of boosting, a weak learner, i.e., a classifier that performs slightly better than chance, is trained on a different subset of the sample data. These learners are combined to form the final classifier, which performs better than its components.

Boosting achieves its performance by maintaining a distribution of weights over the samples in the training set. These weights are a key part of boosting: they are used for selecting the next subset of the samples that will be used for training the next weak learner. Initially, all weights are set to the same value; however, these weights are adjusted at each round of boosting, i.e., each time a weak learner is trained. Based on the performance of the most recently trained weak learner, the weights of the correctly classified samples are decreased and the weights of the misclassified samples are increased. Thus adjusted, the weight of a sample indicates the difficulty of correctly classifying that sample with the weak learners already trained. Using the weights of samples, boosting biases the subsample selection towards samples with higher weights, and thus forces future weak learners to focus on the "harder" samples [107].

Given a subset of the sample data points, each weak learner finds a "weak hypothesis" [107], i.e., a model that performs slightly better than chance on this set. The goodness of each generated weak hypothesis is measured by its error rate on the samples it is trained on, and is denoted by $\epsilon_t$.

Each generated weak hypothesis contributes to the final hypothesis, i.e., the output of the final classifier. The final hypothesis is created by combining the weak hypotheses linearly, and using the measure of goodness of each weak hypothesis to weigh its contribution. In general, the weight of each weak hypothesis is calculated as:

$$\alpha_t = \frac{1}{2} ln \frac{(1-\epsilon_t)}{\epsilon_t}$$

The final hypothesis is denoted as:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

where

$T$ is the number of rounds of boosting,

$t$ is the iteration index indicating the current round of boosting,

$H(x)$ is the final hypothesis,

$h_t(x)$ is the weak hypothesis generated at round $t$,

$\epsilon_t$ is the error rate of the weak hypothesis,

and

$\alpha_t$ is the weight of the weak hypothesis.

At each round of boosting, the weights of the samples is adjusted based on the formula:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times e^a$$

where

$a = -\alpha$ if $h_t(x_i) = y_i$,

$a = \alpha$ if $h_t(x_i) \neq y_i$,

$Z_t$ is a normalization factor chosen so that $D_{t+1}$ will be a distribution,

and

$(x_i, y_i)$ are pairs of inputs and their labels.

Boosting usually continues either for a user-determined number of rounds, or until the training error of the final classifier reaches a desired low [36].

### 5.2.3 t-Test

In this chapter, we use boosted decision trees on pairs of translations to obtain the distribution of average cross-validation accuracies on the classification of translators of the same title. To determine whether a particular feature has a significant effect on the distribution of average cross-validation accuracies, we use the t-test.

The t-test is a method for testing the significance between the differences of the means of two paired sets, $X_i$ and $Y_i$, of $n$ measured values, when the two sets are independently and identically normally distributed. For data that satisfies the independence and the normality assumptions, the t-statistic is given by:

$$\text{t} = (\bar{X} - \bar{Y}) \sqrt{\frac{n(n-1)}{\sum_{i=1}^{n} (\hat{X}_i - \hat{Y}_i)^2}}$$

where

$$\hat{X}_i = X_i - \bar{X}$$
$$\hat{Y}_i = Y_i - \bar{Y}$$

and $\bar{X}$ and $\bar{Y}$ are the means of the distributions.

This statistic has $(n - 1)$ degrees of freedom. The number of degrees of freedom is used in conjunction with the t-statistic and confidence intervals to determine the significance of the differences between the two distributions [127].

The null hypothesis for this test is that the two distributions are the same and that the differences can be explained by chance. The null hypothesis is rejected when the p-value associated with the t-statistic is less than the confidence interval, $\alpha$. For our purposes, $\alpha$ is set to $0.05$.

## 5.3 Experiments

To test the ability of the surface, syntactic, and semantic features to recognize translators, and to identify the features that are most useful for differentiating between translators of the same content, we ran two experiments on a corpus of parallel translations. This corpus contained chapters from each of two translations of *20000 Leagues Under the Sea* (Verne), three translations of *Madame Bovary* (Flaubert), and two translations of *The Kreutzer Sonata* (Tolstoy). In the first experiment, we classified seven translations of these three titles with respect to their translators. In the second experiment, we ran significance tests, i.e., two-tailed t-tests, to rank the features based on their ability to differentiate between translators of the same original.

### 5.3.1 Classification Experiments

To capture the differences in the way different translators convey content, we used boosted decision trees and tenfold cross-validation. These models successfully recognized the translator of any given chapter 75.6% of the time. In comparison, randomly selecting a translator label would have given 14% accuracy (see Table 5.1).

| | Features from the Literature | Random Baseline | No. of samples |
|---|---|---|---|
| Identifying Translators | 75.6 % | 14% | 7 |

Table 5.1: Cross-validation accuracy on recognizing the translators. The corpus contains 3 distinct titles translated by a total of 7 translators.

These results indicate that the features we have selected indeed contain information that can be used for recognizing the linguistic choices of different translators. These features differentiate between translators of the same original due to the differences in the linguistic choices of the translators for that particular content. However, the distinction between the translators of different titles, i.e., translators of *Madame Bovary* and *20000 Leagues*, could be due to the difference in the content of the titles as well as the differences in linguistic choices of the translators.

### 5.3.2 Significance Testing for Feature Ranking

In order to identify the features that capture the differences that are solely due to translators, we studied these features on corpora of pairs of parallel translations of the same title. We ran t-tests on these corpora, and ranked the features with respect to their ability to recognize translators.

The t-tests in this section were run on the distribution of average cross-validation accuracies of pairwise classification experiments which included one experiment for modelling the two versions of *The Kreutzer Sonata*, one for modelling the two versions of the *20000 Leagues under the Sea*, and three for modelling the three versions of *Madame Bovary*, i.e., compare version 1 against version 2, version 1 against version 3, version 2 against version 3. For each of these binary classification tasks, we obtained the distribution of average cross-validation errors in the presence of all features. Next, we reran the same experiments with subsets of the feature set. Each of these subsets contained all but one of the features in the set. The significance of the differences of the resulting distributions of average cross-validation accuracies from the original distribution were tested using t-test.

The null hypothesis was that the distributions of the average cross-validation performances would stay the same when the same binary classification experiments were run with and without a particular feature. If the null hypothesis could be rejected, then the feature contributed significantly to the classification accuracy. Otherwise, it did not contribute significantly to classification and could be dropped from the feature set.

To obtain a ranking of the features, we started with the complete set of $n$ features, compared $n$ distributions with the original distribution, and obtained significance values. Using these significance values, we permanently eliminated the feature whose absence least affected the distribution of average cross-validation accuracies. We repeated this process until only one feature was left, effectively ranking the features based on how much they contributed to recognizing translators of the same content.

This method for evaluating the contribution of features first eliminated, one at a time, the features for which we could not reject the null hypothesis, i.e., the features that did not contribute to the cross-validation performance. Once we ran out of such features, we iteratively eliminated the features with the least significant contribution to classification, i.e., the features with the highest p-value.

When run on the distributions of cross-validation accuracies of pairs of parallel translations, the significance tests identified standard deviation of sentence lengths as the most useful feature for recognizing the differences between pairs of translators who translated the same original title. The second most useful feature was the frequency of use of "get-passives", followed by the frequency of use of " 's-genitives". These and the remaining top ten most useful features thus identified are shown in Table 5.2.

| Rank | Feature |
|---|---|
| 1 | Standard deviation of sentence lengths |
| 2 | Frequency of use of “get-passives” |
| 3 | Frequency of use of “ ’s-genitives” |
| 4 | Standard deviation of word lengths |
| 5 | Frequency of use of “be-passives” |
| 6 | Frequency of active voice sentences |
| 7 | Frequency of use of declaratives |
| 8 | Frequency of overt negations |
| 9 | Type–token ratio |
| 10 | Number of sentences in the document |

Table 5.2: Ten most useful features for distinguishing between translators who translated the same content.

In general, the high rank of the more syntactic features, such as passives and genitives, when differentiating between the translators is expected: given the content, the contribution of the translators to the work comes from the way they write, and we can learn the way different people write by analyzing syntax.

## 5.4 Conclusion

The classification experiments show that the features presented in this chapter differentiate between translators; on a 7-way classification experiment, they correctly recognize the translator of a chapter from a given title 75.6% of the time.

The significance tests used for ranking the features for their usefulness in differentiating between translators of the same content highlight the importance of syntax for this task; three of the top five most useful features are related to sentence structure and syntax.

These results indicate that further analysis of syntactic elements of language and deeper analysis of sentence structure can help us capture linguistic features of expression.

## 5.5 Summary

In this section, we described surface, syntactic, and semantic features that focus on the language of novels and we evaluated these features for their ability to recognize translators using boosted decision trees and significance tests. Our results indicate that the features related to structure of sentences, such as use of passive structures, are particularly useful for differentiating between translators of the same original.

Inspired by these results, in the next chapter, we describe a syntactic theory of expression and identify a novel set of features that capture linguistic elements of expression.

# Chapter 6

# Verb-Based Theory of Expression

## 6.1 Linguistic Complexity and Syntactic Repertoire

The relatively high ranking of syntactic features of sentences in the preliminary experiments presented in Chapter 5 indicates the presence of structural information that can be extracted from sentences and exploited for modelling expression. We, therefore, use information about structure of sentences to identify a novel set of features that represent different aspects of the syntactic repertoire of authors and the linguistic complexity of their works. By *syntactic repertoire*, we refer to the variety of syntactic choices made by authors when conveying a particular content. By *linguistic complexity*, we refer to a measurement of the average complexity of the syntax tree of individual sentences in a document.

In this chapter, we introduce and describe a novel set of features that capture syntactic repertoire and linguistic complexity. The features we present have been selected as potential elements of expression based on our observations related to the differences in their uses by different authors. For each of these linguistic features, after discussing the differences in their use in different contexts and by different authors, we present a validation of their ability to discriminate between authors and books through significance tests. Our significance tests follow the same template for all features. More precisely, for each feature:

- We hypothesize that the observed differences in its use comprise differences in expression, i.e., authors use the feature in question differently in different contexts.

- Setting this as our alternate hypothesis, and using chi-square and/or likelihood ratio tests, we test the null hypothesis that these features are used similarly by all authors and that the

observed differences are due to chance. We test this hypothesis in three different settings: on parallel translations of the same novels (i.e., titles) which represent similar content but different expression, on different books by different authors which represent different content and different expression, and finally on disjoint sets of chapters from the same book which represent similar content and same expression.

We reject the null hypothesis for almost all of the features when comparing literary works that contain different expression, indicating that regardless of content, these features can capture expression. For all of the features, we are unable to reject the null hypothesis when we compare chapters from the same book, indicating a certain consistency in the distributions of these features throughout a work.

In the next section, we first briefly describe the chi-square and likelihood ratio tests. Next, we define linguistic features and, following the template described above, test each one for independence from authors and content.

## 6.2 Tests of Independence

Chi-square and likelihood ratio tests are widely used to test the independence of categorical variables. We use these methods to test the independence of use of particular features from either authors or content.

### 6.2.1 Pearsons' Chi-Square

Generally speaking, Pearson's chi-square test identifies the effects of two categorical variables on each other. Categorical variables used for this test are unordered discrete classes, such as genders of a group of people, the college a student goes to, classes of verbs, etc. Chi-square tests whether one categorical variable is independent of the other by first calculating the individual expected values of all cells in a contingency table and then comparing the expected values with the observed values to determine whether their differences can be explained by chance. For example, consider the following hypothetical example.

For two sets of fiction and history books, we are interested in finding whether the distributions of the kinds of verbs are different, i.e., whether the genre of the book has an effect on the kinds of verbs that are used. Let's assume that a verb can belong to only one of two mutually exclusive classes, i.e., classes 1 and 2. We start by classifying every verb in the corpus and keeping track of the frequencies of the kinds of verbs we see in each of the genres, as shown in Table 6.1. Each cell in this table represents a joint event, e.g., there are 30 verbs that are in a fiction book and that belong to class 1.

| **Book** | **Verb Class 1** | **Verb Class 2** | **Row sum** |
|---|---|---|---|
| Fiction | 30 | 10 | 40 |
| History | 15 | 25 | 40 |

Table 6.1: Hypothetical example of verb classes in two books.

In this population, the probability that a random verb will belong to row *i* is given by $p_i$, and the probability that a random verb will belong to column *j* is given by $p_j$.

If the kind of verbs used in a book are independent of the book then the probability that a random verb belongs to a particular verb class and a particular genre should be given by the product of the probabilities of each of the events, as shown below:

$$p_{ij} = p_i \times p_j$$

The hypothesis that the verb classes and the books are independent from each other is the null hypothesis, denoted by $H_0$. The alternate hypothesis, $H_1$, simply says that $H_0$ is not correct. If the verb classes are distributed differently over the books, we can reject the null hypothesis in favor of the alternate hypothesis [47].

Under the null hypothesis, the expected value of each cell is given by $p_{ij}$. The difference, over all cells, of the observed values from the expected values is captured by the $\chi^2$ (chi-square) statistic as follows:

$\chi^2 = \sum_{cells} \frac{(O-E)^2}{E}$

where

$O$ is the observed value of a cell,

$E$ is the expected value of a cell.

For the degrees of freedom of a given contingency table, we can reject the null hypothesis if the calculated chi-square value is greater than the chi-square value at the desired confidence interval, e.g., $\alpha = 0.05$. The degrees of freedom can be calculated as:

$df = (R-1) \times (C-1)$ where R is the number of rows and C is the number of columns.

Chi-square test has three fundamental assumptions:

1. That the observations are probabilistically independent of each other, i.e., each categorical entry in the contingency table is independent of the other entries. In our case, each verb can be either from a fiction or a history book and it can belong to either class one or two.

2. That each observation represents one and only one joint event, i.e., each observation falls into exactly one of the cells in the contingency table.

3. That the number of observations is large. This condition ensures good approximations of $p$ values. Where one or more cells have small expected values, the chi-square statistic is not reliable. In these cases, the likelihood ratio chi-square test should be used.

For our experiments, we will present the results of chi-square tests in tables such as Table 6.2.

Chi-square tests are widely used in corpus-based studies of linguistic phenomena. For example, Roeck et al. use this test to show that the distribution of function words are significantly different

| Book | Degrees of Freedom | Chi-Square | p-value |
|---|---|---|---|
| Books in population | $df$ | $chi$ | significance |

Table 6.2: A table template for presentation of chi-square results for testing independence of use of verb classes in two books.

in different partitions of the same data set, defeating a common assumption that these features are uniformly distributed throughout corpora and contain no information [97].

### 6.2.2 Likelihood Ratio Chi-Square

Likelihood ratio chi-square, also referred to as likelihood ratio test and g-test, is an alternative way of testing the independence of two variables, again represented as the rows and columns of a contingency table. As with Pearson's chi-square, the likelihood ratio test also assumes probabilistic independence of observations and that each observation corresponds to only one joint event. Under these assumptions, the likelihood ratio is calculated as:

$$G = 2 \sum O \times \ln(O/E)$$

where

$O$ is the observed value of a cell,

$E$ is the expected value of a cell.

When the row and column variables are independent, $G$ has an asymptotic chi-square distribution whose degrees of freedom is given by $(R-1) \times (C-1)$. For adequate sample sizes, chi-square and likelihood ratio test lead to the same conclusions. When $|O - E| > E$, likelihood ratio test gives a better approximation of the theoretical chi-square distribution than Pearson's chi-square, and is therefore preferred.

For some of our experiments, we will present the results of likelihood ratio test along with Pearson's chi-square results as shown in the template in Table 6.3.

| | | Chi-Square | | Likelihood Ratio | |
|---|---|---|---|---|---|
| **Book** | **#df** | $\chi^2$ | p-value | LR | p-value |
| Books in population | $df$ | $chi$ | $p$ | $LR$ | $p$ |

Table 6.3: Template for presentation of chi-square and likelihood ratio test results.

We interpret the test results for chi-square and for likelihood ratio with their respective p-values. If the p-values are less than $\alpha = 0.05$, we reject the null hypothesis, concluding that the categorical variables are not independent. When the test results disagree, and $|O - E| > E$, our conclusions favor the likelihood ratio test.

We use Pearson's chi-square and likelihood ratio tests to validate the ability of proposed linguistic features to differentiate between books and authors. For the features that are ideal for modelling expression, we would like to reject the null hypothesis with $\alpha = 0.05$ when we compare different books or different authors; the same features should not be able to differentiate between the groups of chapters taken from the same work of the same author, i.e., we should not be able to reject the null hypothesis.

In the rest of this chapter, we will show that most of our features satisfy these criteria. These features are linguistic features that are inspired by our observations on different syntactic presentations of semantically similar content.

## 6.3 Expression and Meaning

Evaluating works for expressive similarities and capturing expression requires looking at paraphrases of the same content and analyzing the ways in which people differ in the expression of this particular content. The main goal of this thesis is identifying the expressive fingerprint that is unique to a work and that can help us identify works, even when they are paraphrased by substitution of keywords and superficial rewrites.

Authors of creative works rely on elements of language to create a particular expression. Translated literary works provide examples for a wealth of diverse linguistic choices that differ in expression but manage to convey similar content. As an example, consider the following semantically equivalent excerpts taken from three different translations of *Madame Bovary* by Gustave Flaubert.

> Excerpt 1: "Where should he practice? At Tostes. In that town there was only one elderly doctor, whose death Madame Bovary had long been waiting for, and the old man had not yet breathed his last when Charles moved in across the road as his successor." (Written by Flaubert, translated by Unknown1.)
>
> Excerpt 2: "Where should he go to practice? To Tostes, where there was only one old doctor. For a long time Madame Bovary had been on the look-out for this death, and the old fellow had barely been packed off when Charles was installed, opposite his place, as his successor." (Written by Flaubert, translated by Aveling.)
>
> Excerpt 3: "And now where was he to practice? At Tostes, because at Tostes there was only one doctor, and he a very old man. For a long time past Madame Bovary had been waiting for him to die, and now, before the old fellow had packed up his traps for the next world, Charles came and set up opposite, as his accredited successor." (Written by Flaubert, translated by Unknown2.)

The translators of these excerpts differ in their use of vocabulary and syntax in several ways. For instance, the three translators choose to explain the presence of the old doctor by using the expressions: "there was only one elderly doctor", "there was only one old doctor" and "there was only one doctor, and he a very old man". Other examples of semantically equivalent but syntactically different text units from excerpts 1, 2, and 3 are shown in Table 6.4.

Observations such as those presented in Table 6.4 inspire us to analyze syntax to capture the expressive differences between works. In particular, we tie the observed differences of expression into

| Excerpt 1 | Excerpt 2 | Excerpt 3 |
|---|---|---|
| Where should he practice? | Where should he go to practice? | Where was he to practice? |
| In that town there was only one elderly doctor | Tostes, where there was only one old doctor | at Tostes there was only one doctor, and he a very old man. |
| doctor, whose death Madame Bovary had long been waiting for | For a long time Madame Bovary had been on the look-out for his death | For a long time past, Madame Bovary had been waiting for him to die |

Table 6.4: Examples of semantically equivalent sub-excerpts from excerpts 1, 2, and 3.

elements of syntax and semantics that capture the patterns in phrase structure, linguistic complexity, and overall sentence structure.

In the following sections, we present these syntactic and semantic elements of expression in detail and validate our intuition about the expressive nature of each one through significance tests on a small corpus.

## 6.4 Expression as a Function of Syntax and Structure

Language allows authors some flexibility in the vocabulary and syntax they use to convey content. Given their particular expressive goals and style, different authors make different lexical and syntactic choices even when trying to express the same content. Their syntactic choices usually result in semantically equivalent paraphrases and their particular choices with respect to these paraphrases constitute their "syntactic repertoire". Syntactic repertoire, coupled with assessments of the complexity of clause structures (i.e., linguistic complexity), can help us identify elements of language that constitute expression.

We determine the linguistic complexity of literary works by analyzing sentence structure along several dimensions, including the clause structure of the sentences. For the syntactic repertoire, we rely on the syntactic structure of verb phrases and the tendencies of authors to use different classes of verbs, including both embedding verbs, i.e., verbs that take sentences and clauses as complements [1], and nonembedding verbs [72].

More precisely, our syntax-driven definition of expression considers the following features, which capture the sentence structure in increasingly more depth, starting from the top-level ordering of the syntactic phrases in a sentence and concluding with an analysis of their internal structure:

1. Choices of authors with respect to positions of different clauses and phrases within a sentence:

    (a) Sentence-initial phrase structure: The use of particular syntactic phrases in a sentence-initial position.

    (b) Sentence-final phrase structure: The use of particular syntactic phrases as well as stranded prepositions, modals, and auxiliary verbs in sentence-final positions.

2. The level of complexity of sentences and their top-level constituents in the works of an author in terms of:

   (a) The average and the standard deviation of the depths of the top-level left and right branches in sentences in terms of phrase depth.

   (b) The average and the standard deviation of the number of prepositional phrases in sentences, as well as the average and the standard deviation of the depths of the deepest prepositional phrases in sentences.

   (c) The percentage of left-heavy, right-heavy, and equal-weight sentences, e.g., sentences where the top-level right branch of the syntax tree is deeper than the top-level left branch are considered right-heavy.

   (d) The average and the standard deviation of the number of embedded clauses in the top-level left and right branches in sentences.

   (e) The percentage of left-embedded, right-embedded, and equally-embedded sentences, e.g., sentences where the top-level right branch of the syntax tree embeds more clauses than the top-level left branch are considered right-embedded.

3. Syntactic choices of authors as determined by:

   (a) Classes of verbs the authors use, classified with respect to their ability to take particular embeddings [1] as their arguments and the distributions of these embedding classes in documents.

   (b) The distribution of semantic classes of verbs that do not take sentential or clause complements and the distribution of their particular alternations [72], i.e., alternate syntactic configurations that verbs can appear in, over the documents as well as the tendencies of authors to use them with particular groups of embedding constructs.

Given these parameters, studying the semantically similar excerpts by different translators presented in Section 6.3 focuses our attention on the fact that:

- All of excerpts 1, 2, and 3 contain question constructs.

- Only excerpt 3 starts with the conjunction *and*.

- The majority of sentences in all of these excerpts start with prepositional phrases, e.g., *"To Tostes", "At Tostes"* and *"To Tostes"* as well as *"In that town", "For a long time"* and *"For a long time past"*.

- All three excerpts contain one sentence (an interrogative sentence) that ends with a verb phrase, i.e., *"practice"*; however, the majority of the sentences in all of the excerpts end with noun phrases, i.e., *"Tostes", "his successor", "only one old doctor", "very old man", "his accredited successor"*.

- Excerpt 1 includes three sentences. One of these sentences is a fragment and does not include any subject-predicate pairs, i.e., *"To Tostes."*. Of the other two sentences, one contains only one pair, i.e., *he–practice*, and the other contains four pairs, i.e., *there–was...* where the subject is the existential "there" and the predicate is the verb phrase starting with "was", *Madame Bovary–had...been...*, *the old man–had...breathed...*, and *Charles–moved...*. All of these clauses have deeper right branches than left branches. Similar analysis for the other two excerpts reveal the absence of a fragmental sentence; however, the rest of the structures look similar to that of excerpt 1.

- Excerpts 1 and 2 have relative clauses marked with wh-words, e.g., *"...whose death Madame Bovary had been waiting for..."* and *"...where there was only one doctor."*

- Excerpt 2 uses the passive voice, e.g., *"...been packed off..."*, *"...Charles was installed..."*

- Most of the verbs used in these excerpts are nonembedding verbs, meaning that they do not take clausal complements. These verbs are *"practice", "be", " breathe", "move", "pack off", "install", "die", "pack up", "come"*, and *"set up"*.

- All excerpts use copular *be*; excerpts 1 and 2 use it only once while excerpt 3 uses it twice.

- Some of the nonembedding verbs in excerpt 1 are intransitive, i.e., verb phrase structure is denoted by "V", e.g., *"practice"*. Others are observed in structures denoted by "V+NP" which indicates a verb followed by a direct object noun phrase, e.g., *was only one elderly doctor"*; "V+prep" which indicates a verb followed by a preposition and no noun phrase, and typically occurs at the end of a sentence, or is separated from the following clause by punctuation—this structure is observed usually with relative clauses and topicalization, and is a result of the movement of the noun phrase subsumed by the prepositional phrase, e.g., *"waiting for"*; and

"V+PP" which indicates a verb followed by a prepositional phrase, e.g., *"moved in across the road..."*.

- Excerpt 2 contains intransitive verbs, e.g., *"practice"*, as well as instances of each of the structures "V+NP" and "V+PP", which correspond to *"was only one old doctor."* and *"been on the lookout for..."* respectively.

- Excerpt 3 contains verb phrase structures that are similar to the other two excerpts but with different frequencies.

- The differences in sentence structures of excerpts 1 and 3 result in different observations associated with the use of the verb *"wait"* in these two excerpts. In excerpt 1, presence of the relative clause and movement of the object noun phrase of the verb results in the structure "V+prep"; whereas in excerpt 3, there is no relative clause or movement, and therefore the verb appears with a prepositional phrase in the configuration "V+PP".

In the following sections, we discuss the elements of expression exhibited by these examples in more detail and describe them in terms of phrase-internal structure.

## 6.5 Expression as a Function of Sentence Structure

The structure of a sentence plays a role in the overall message conveyed by the sentence. Authors may choose to order the facts in a sentence in a particular way to highlight or to downplay them. They may arrange these elements for ease of comprehension or for artistic effect. For example, consider the following sentences, which convey the same basic facts but with different flavors of expression:

- At five o'clock, we will meet at the train station.
- We will meet at the train station at five o'clock.
- We will meet at five o'clock at the train station.

We can analyze the syntax of these sentences in terms of their constituents. In this chapter, adopting terminology and examples from Baker [8], we discuss phrase-internal syntax in terms of *heads*, *complements*, and *modifiers*, and we show that we can capture a significant part of expression by analyzing sentence-initial and sentence-final uses of complements and modifiers.

### 6.5.1 Expressive Use of Sentence-Initial and -Final Phrase Structures

#### Features

Phrases consist of *heads*, *complements*, and *modifiers*. The head of a phrase is the word that determines the syntactic characteristics of the phrase. For example, the head of a verb phrase is a verb, and it determines the structure of the phrase by requiring and allowing only certain kinds of *complements*.

Baker refers to a head and its complements as a *minimal phrase*. For example, the verb *put* requires two objects as its complements; the absence of these complements from the verb phrase results in loss of grammaticality. If elimination of a phrase from the sentence does not result in loss of grammaticality, we can generally (though not always) conclude that the phrase is a modifier, not a complement. In the examples taken from Baker and shown below, the prepositional phrase "on Friday" is a modifier, whereas the noun phrase "some money" and the prepositional phrase "in the bank" are complements. Sentences marked with a (*) are considered ungrammatical.

1. Martha put some money in the bank on Friday.

2. * Martha put some money — on Friday.

3. * Martha put — in the bank on Friday.

4. Martha put some money in the bank —. She did this on Friday.

Different modifiers can be used in various sentence-initial, -medial or -final positions. For example, adverb phrase modifiers can be in sentence-initial, -medial or -final positions, although when used in the sentence-medial positions they have to be placed between the verb and its auxiliary (if the auxiliary is present):

1. Martha can put some money in the bank, finally.

2. Finally, Martha can put some money in the bank.

3. Martha can finally put some money in the bank.

Prepositional phrase modifiers, on the other hand, can be in sentence-initial and -final positions:

1. Martha put some money in the bank on Friday.

2. On Friday, Martha put some money in the bank.

3. * Some money, Martha put in the bank on Friday.[1]

4. Some money *is what* Martha put in the bank on Friday.

5. * In the bank, Martha put some money on Friday.[2]

6. In the bank *is where* Martha put some money on Friday.

While the movement of prepositional phrase modifiers does not require any other changes to the sentence structure, movement of prepositional phrase complements requires clefting, i.e., insertion of "be" as shown above, in order to be syntactically acceptable. These constructs are used in cases of heavy topicalization, where the speaker really wants to emphasize the complement that is moved to the sentence-initial position.

[1] Grammatical in some dialects.
[2] Grammatical in some dialects.

These examples show that authors have some expressive freedom in their use and placement of complements and modifiers in a sentence. The resulting semantic change due to such movements is usually minimal. In the case of omission of modifiers, the semantic change is due solely to the left-out information. In the case of movement of complements, the semantic change involves a shift in emphasis; however, the facts remain the same.

The expressive freedom in the positioning of complements and modifiers can result in different overall sentence structure. In the simplest terms, as a result of repositioning of modifiers and complements, the sentence-initial and -final phrase structures change.

Our goal is to find a set of easily extractable features that can help us capture the expressive differences due to repositioning of complements and modifiers in sentences. For this, we use a part-of-speech tagged text, and for each sentence in this text we identify its first and last constituents and their syntactic categories.

The syntactic phrase types of initial and final constituents tell us about the high-level expressive differences between sentences without getting into the details of whether these constituents are heads, complements, or modifiers. This level of analysis considers two constructs to be of the same type if they have similar phrase structures in matching positions. For example, this analysis does not detect any differences between two sentences when one of them has a prepositional phrase complement while the other a prepositional phrase modifier, as long as the same phrase types appear in matching positions. Thus, the sentences "Jane put the book on the table" and "Jane left the office in a rush" are equivalent from the perspective of phrase structure analysis.

There are no surface features that can tell us definitively whether the syntactically-analyzed constituents are modifiers or complements. Although this information would enrich our study of expression, extraction of this information is very costly, and even with full syntactic parsing, not always accurate. Therefore, we only use sentence-initial and -final structures to capture variances in the structure of semantically equivalent content, independent of the role these phrases play in the argument structure of the sentence.

Despite its inability to detect the structural changes that do not affect the sentence-initial and -final phrase types, this approach is a quick and efficient way of capturing most phrase-level expressive differences between semantically equivalent content.

In addition to capturing expressive differences due to movement of modifiers and complements, this level of analysis helps us capture different sentential structures, including question constructs, imperatives, and coordinating and subordinating conjuncts. Other structural differences can be cap-

tured by analyzing the uses of stranded prepositions, sentence-final adjectival phrases, intransitive verb constructs, and adverb phrases. These phrase types can be used in different positions in sentences, resulting in different sentence-initial and -final constructs and expressions:

1. Imperatives

   (a) “Come here!” she said.

   (b) She said to come here.

2. Verbal Ellipsis

   (a) Jane did not run the marathon but Jackie did.

   (b) Jane did not run the marathon but Jackie ran the marathon.

3. Stranded Prepositions

   (a) I do not know which bank she deposited her money in.

   (b) I do not know in which bank she deposited her money.

4. Conjunctions

   (a) She went to the bank. And, she deposited some money.

   (b) She went to the bank and she deposited some money.

To capture this level of difference in expression, we consider the following phrase structures:

- Sentence-initial and -final noun phrases. Sentence-initial noun phrases include simple noun phrases as well as gerundives (e.g., “Going to the school was his favorite pastime”), the noun phrases of the wh-words *what, who, which*, and *whose* when they are not used in question constructs (e.g., “What he bought, we did not see.”), and existentials (e.g., “There is a flower in the vase”). Sentence-final noun phrases include existentials and wh-words *what, who, which*, and *whose*.

- Sentence-initial and -final prepositional phrases. We theorize that the use of this syntactic structure will differ between authors, even if we study it without differentiating the modifiers from complements. Sentence-initial and -final wh-phrases using *where*, when not used in question constructs, usually correspond to prepositional phrases, except when they stand for

*here* and *there*, which themselves stand for prepositional phrases. As it is impossible to tell the exact semantics behind the "where" constructs, we will consider them to be prepositional phrases.

- Sentence-initial and -final verb phrases. Excluding the gerundive verbs that appear in sentence-initial positions, most of the sentence-initial uses of verbs are due to imperative constructs and can be recognized by their unique part-of-speech tag sequences. Inevitably, constructs such as "Had it not been for John, she would have..." become confused with the imperative construct. Sentence-final verb phrases indicate use of intransitive verbs, as well as verbs whose complements have been moved from the end of the sentence.

- Sentence-initial and -final adverb phrases: As mentioned earlier, adverb phrases can be placed in different positions for slightly different expressions.

- Sentence-initial conjuncts: We consider coordinating and subordinating conjuncts, as they can be used in sentence-medial or -initial positions, and result in a change in expression.

- Question Constructs: These include constructs with subject–auxiliary inversion, some of which also use wh-phrases.

- Sentence-final adjective phrases: These phrases usually indicate predicative use of adjective phrases and small clauses.

- Verbal ellipsis: Whether authors choose to repeat previously stated information can depend on several factors, including their choice of audience and their expressive preference.

In this analysis, it is hard to place sentence-initial and -final wh-phrases formed with "why", "when", and "how" in any of the phrase categories; their phrase category is more ambiguous than that of sentence-initial and -final "where". In cases where the wh-phrases with "why", "when", and "how" are used in question constructs, we only note the existence of a question. However, when these phrases do not appear in question constructs, each of these wh-word phrases in sentence-initial and -final positions stand for any of two or more of adverb phrases, prepositional phrases, noun phrases, infinitive clauses, or subordinate clauses. We show this with examples of question forms of these wh-phrases, which demonstrate the structure of the phrases they replace:

1. Place holder for prepositional phrase

   (a) They told us that Jane quit her job due to health problems on Friday.

   (b) When did Jane quit her job? On Friday.

   (c) Why did Jane quit her job? Due to health problems.

2. Place holder for adverb phrase

   (a) Jane quickly wrote a resignation letter.

   (b) How did Jane write the letter? Quickly.

3. Place holder for noun phrase

   (a) Jane quit her job this Friday.

   (b) When did Jane quit her job? This Friday.

4. Place holder for subordinate clause

   (a) After she talked to her superiors, Jane quit her job for a better position by submitting a resignation letter.

   (b) When did Jane quit her job? After she talked to her superiors.

   (c) How did Jane quit her job? By submitting a resignation letter.

5. Place holder for infinitive clauses

   (a) Jane quit her job to take a better position.

   (b) Why did Jane quit her job? To take a better position.

From their syntactic constructs, we are unable to tell the difference between the wh-phrases that fall into different syntactic classes; therefore we keep track of these wh-phrases in a class of their own. Admittedly, this approach provides less information than full parsing; however, the information we gather is sufficient for capturing expression.

**Examples and Preliminary Evaluation of Sentence-Initial and -Final Phrase Structures**

Examples of some sentence-initial and -final phrase structures are shown, with examples, in Table 6.5. We obtain the frequency distributions of these phrase structures over all chapters in a book using a shallow parser that takes advantage of part-of-speech tags.

| Sentence-Initial and -Final Phrases | Examples |
|---|---|
| Sentence-initial conjunctions | And though I did not finish them very handsomely yet I made them sufficiently serviceable for my purpose. |
| Sentence-initial existential phrases | There was a sound of a heavy thud. |
| Question constructs | Can God spread a table in the wilderness... |
| Sentence-initial noun phrases | The rainy season sometimes held longer or shorter as the winds happened to blow but this was the general observation I made. |
| Sentence-initial prepositional phrases | In the midst of the greatest composures of my mind this would break out upon me like a storm and make me wring my hands and weep like a child. |
| Sentence-initial subordinate clauses | After I had found by experience the ill consequence of being abroad in the rain I took care to furnish myself with provisions beforehand that I might not be obliged to go out... |
| Imperative constructs | Stop stop! |
| Sentence-initial wh-phrases | When I had passed the vale where my bower stood as above I came within view of the sea to the west... |
| Sentence-final adjective phrases | ...it was exceedingly heavy. |
| Sentence-final prepositional phrases | But Raskolnikov was already stepping into the street. |
| Sentence-final adverb phrases | “I was sitting up” he said still more timidly. |
| Sentence-final verb phrases | ...because you are kinder than any one clever I mean and can judge. |
| Sentence-final stranded prepositions | What were they beating the landlady for? |

Table 6.5: Examples of sentence-initial and -final structures from *Robinson Crusoe* by Daniel Defoe and *Crime and Punishment* by Leo Tolstoy.

We hypothesize that these features can capture the expression of an author in a work; that the use of these features by each author follows a distributional pattern that is consistent, at least within one work; and that different authors, having different expressive goals and styles, will differ in their use of these phrase types. As a result, the distributions of phrase types in sentence-initial and -final positions should be different between the translations of a title and between different books.

Looking at several translations of Jules Verne's *20000 Leagues under the Sea* and Gustave Flaubert's *Madame Bovary*, we find that, despite having translated the same content, the translators have used different sentence-initial and -final phrase structures at different rates. The frequency distributions of each of these phrase types in translations of *20000 Leagues* are shown in Table 6.6(a) and Table 6.6(b).

| | **20000 Leagues v.1** | | **20000 Leagues v.2** | |
|---|---|---|---|---|
| **Feature** | frequency | % | frequency | % |
| Adverb Phrase | 444 | 6.3% | 719 | 7.8% |
| Coordinating Conjunction | 924 | 13.1% | 1123 | 12.2% |
| Imperatives | 326 | 4.6% | 451 | 4.9% |
| Noun Phrases | 4060 | 57.9% | 5142 | 56.7% |
| Prepositional Phrase | 624 | 8.9% | 817 | 8.8% |
| Questions | 155 | 2.2% | 141 | 1.5% |
| Speech Fragment | 125 | 1.8% | 171 | 1.9% |
| Subordinating Conjunction | 294 | 4.2% | 579 | 6.3% |
| Wh-phrases | 60 | 0.9% | 98 | 1.1% |
| Total | 7012 | 100% | 10241 | 100% |

(a) Raw counts and percentages of sentence-initial phrase structures in translations of *20000 Leagues under the Sea*.

| | **20000 Leagues v.1** | | **20000 Leagues v.2** | |
|---|---|---|---|---|
| **Feature** | frequency | % | frequency | % |
| Adjective Phrase | 540 | 7.7% | 636 | 6.9% |
| Adverb Phrase | 503 | 7.2% | 698 | 7.6% |
| Ellipsis | 6 | 0.0008% | 8 | 0.0008% |
| Noun Phrase | 1651 | 23.6% | 1936 | 21% |
| Prepositional Phrase | 3308 | 47.3% | 4535 | 49.1% |
| Speech Fragment | 19 | 0.3% | 19 | 0.2% |
| Stranded Preposition | 31 | 0.4% | 66 | 0.7% |
| Subordinating Conjunction | 1 | 0.0001% | 3 | 0.0003% |
| Verb Phrase | 939 | 13.4% | 1340 | 14.5% |
| Total | 6998 | 100% | 9241 | 100% |

(b) Raw counts and percentages of sentence-final phrase structures in two translations of *20000 Leagues under the Sea*.

Table 6.6: Sentence-initial and -final phrase structures in *20000 Leagues*.

In order to validate the intuition that sentence-initial and -final phrase structures depend on the expression of the author/translator, we test (using chi-square) whether the differences in 6.6 can be explained by chance alone.

Our null hypothesis is that the use of sentence-initial and -final structures are independent of the author/translator. As the translations of a title are similar in content, the difference in the distribution

of phrase structures among the translations of each title should be due to the translator. In order to accept this alternate hypothesis, we would like to reject the null hypothesis at a confidence level of $\alpha = 0.05$.

As can be seen in Table 6.7, for both sets of translations, we reject the null hypothesis with at least 99% confidence and conclude that the sentence-initial and -final phrase structures capture expression differences of translators when they translate the same content.

| Book | Degrees of Freedom | Chi-Square | p-value |
|---|---|---|---|
| Two translations of *20000 Leagues* | 8 | 63 | $\leq 0.001$ |
| Three translations of *M. Bovary* | 16 | 199.3 | $\leq 0.001$ |

(a) Chi-square results for sentence-initial phrase structures for the translations of each of the two titles

| Book | Degrees of Freedom | Chi-Square | p-value |
|---|---|---|---|
| Two translations of *20000 Leagues* | 8 | 29.3 | $\leq 0.01$ |
| Three translations of *M. Bovary* | 16 | 149 | $\leq 0.001$ |

(b) Results of Chi-Square for sentence-final phrase structures for the translations of each of the two titles

Table 6.7: Chi-square test results for sentence-final phrase structures for *20000 Leagues* and *Madame Bovary*.

In addition to being dependent on the translator, the expression within a book should be similar throughout, if the expression of the translator indeed stays consistent within a book. To test this hypothesis, we separated odd-numbered chapters of *Anna Karenina* by Leo Tolstoy from the even-numbered chapters and tested the similarity of the distributions of sentence-initial and -final phrase structures in the two sets. If the expression of the translator indeed stays constant throughout this book, and if we have indeed captured expression, then the distributions of the different phrase structures in the two groups should be independent of the group.

As can be seen in Table 6.8, the distribution of sentence-initial features are similar across odd- and even-numbered chapters with 99% confidence. For sentence-final structures, also, we cannot reject the null hypothesis that their distributions in the two groups are similar. However, these features are less indicative of expression when compared with sentence-initial structures, as shown by the lower p-values in Table 6.8.

To test the effectiveness of sentence-initial and -final phrase structures for capturing expression, two tests remain. Can these features differentiate between two works of the same author and can they differentiate between two unrelated works of two different authors? We hypothesize that they can differentiate both between the works of independent authors and different works of the same

author, i.e., their distributions in the different works are significantly different (at $\alpha = 0.05$) even if the works are written by the same author.

Chi-square test results in *Emma* and *Sense and Sensibility*, both from Jane Austen, show that we can indeed reject the null hypothesis with $p < 0.001$, for these two books (see Table 6.8).

Similarly, results in Table 6.8 show that we can indeed reject, with very high confidence, the null hypothesis of independence when we compare different books by different authors.

| Book | Degrees of Freedom | Chi-Square | p-value |
|---|---|---|---|
| Odd and even chapters from *Anna Karenina* | 8 | 1.62 | $p = 0.99$ |
| *Emma* v. *Robinson Crusoe* | 8 | 2028.86 | $p \leq 0.001$ |
| *Emma* v. *Sense and Sensibility* | 8 | 177.0 | $p \leq 0.001$ |

(a) Chi-square test results for sentence-initial phrase structures for various books.

| Book | Degrees of Freedom | Chi-Square | p-value |
|---|---|---|---|
| Odd and even chapters from *Anna Karenina* | 8 | 9.8 | $p = 0.31$ |
| *Emma* v. *Robinson Crusoe* | 8 | 384.43 | $p \leq 0.001$ |
| *Emma* v. *Sense and Sensibility* | 8 | 83.39 | $p \leq 0.001$ |

(b) Chi-square test results for sentence-final phrase structures for various books.

Table 6.8: Chi-square test results for sentence-initial and -final phrase structures for various book pairs.

The results of significance tests imply the following:

- Rejecting the null hypothesis for *Emma* and *Robinson Crusoe* indicates that the sentence-initial and -final features can capture the differences between two independent works by two independent authors. The captured differences here can either be due to the authors or the content.

- Rejecting the null hypothesis for *Emma* and *Sense and Sensibility* indicates that the sentence-initial and -final features can capture the differences between two independent works by the same author. The captured differences can either be due to the change in the expression of the author from one work to the other, or due to content change.

- Rejecting the null hypothesis for translations of *20000 Leagues* and *Madame Bovary* indicates that the sentence-initial and -final features can capture the differences between two translators when they write about the same content. In this case, as content is the same, the differences in the use of sentence-initial and -final features are due to expressive differences of the translators.

- Not being able to reject the null hypothesis for two subsets of chapters from *Anna Karenina* indicates that the authors are in fact mostly consistent in their use of sentence-initial and sentence-final features throughout a book and these features can be used to recognize the expression in a book.

In addition to adequately capturing expression in a work and distinguishing an author's works from each other, these features also seem to capture elements of style of the authors. This fact is shown by the difference in the relative significance of the differences in the distributions when we test two books by the same author and two books by different authors. The chi-square values for books by different authors are much higher than the chi-square values for books by the same author. This implies that certain elements of the use of sentence-initial and -final phrase structures vary more markedly between authors than within the works of one author.

## 6.6 Expression as a Function of Sentence Complexity

In addition to high-level phrase ordering and sentence structure, we consider the syntactic structure of sentences in greater depth for capturing more of an author's expression. To capture the sentence-level linguistic complexity of the expression of an author, we analyze the number of clauses per sentence, and the depth of the syntax tree representing the subject and the predicate of the clauses within the sentence.

Ejerhed [37] claims that clauses are "building blocks in the construction of a richer linguistic representation that encompasses syntax as well as semantics and discourse structure." For our task, we use Quirk's definition of a clause as a text unit that contains a subject and a predicate [94]. In some cases the subject can be covert.

In this section, we first describe the basics of our analysis of clause structure and sentence complexity. Assuming a binary branching syntax tree structure, we break sentences into subject–predicate pairs and determine the depth of each subject and predicate, as well as keeping track of the number of subject–predicate pairs encountered in each of the sentences. To evaluate the potential contribution of each of these features to the unique fingerprint of literary works, we perform significance tests that show that features related to sentence complexity and clause structure can be useful for capturing the unique expression present in each work.

### 6.6.1 Sentence Complexity as a Function of Clause Structure

Sentences differ in their linguistic complexity: Simple sentences contain a single independent clause, whereas complex sentences contain two or more clauses [94].

For example, as shown in Figure 6-1, the simple sentence "I will feed my cat" contains a single independent clause whose subject is *"I"* and whose predicate is *"will feed my cat"*.

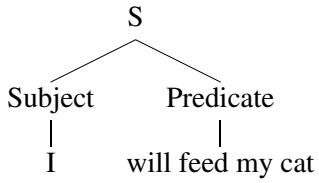


Figure 6-1: Subject-predicate structure.

On the other hand, the complex sentence "I will feed my cat before I leave home" contains two clauses, namely *"before I leave home"* which is embedded in the clause *"I will feed my cat before I leave home"*. The subject of the top-level clause is *"I"* and the predicate is *"will feed my cat before*

*I leave home"* (see Figure 6-2). Note that the clause *"before I leave home"* is a modifier for the verb *give* and it has a subject *"I"* and predicate *"leave home"*, as shown in Figure 6-3.

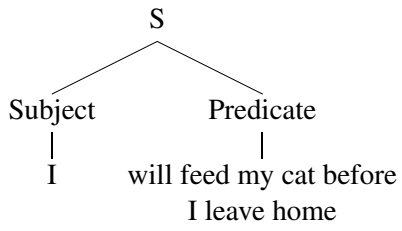


Figure 6-2: Subject and predicate of the top-level clause.

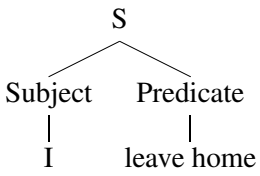


Figure 6-3: Subject and predicate of the embedded clause.

Similarly, the complex sentence *"The woman whom I met on the subway was my new roommate"* contains two clauses. The top-level clause has the subject *"The woman whom I met on the subway"* and the predicate *"was my new roommate"*. The subject of the top-level clause, in turn, contains the clause *"I met — on the subway"* which has its own subject and predicate.

### 6.6.2 Clause Structure and Expression

We theorize that the level of linguistic complexity in a work is partly due to a conscious effort on the part of the author and can depend of factors that range from the choice of the audience for the work to the style of the author. We define complexity of a sentence in terms of the number of clauses contained in the sentence and their depth. The number of clauses in a sentence (and in each branch) gives an indication of the level of embedding in the sentence and branch while the depth along a branch of the syntax tree reveals whether the structure of the syntax tree is right- or left-heavy. We observe that in most cases, the level of linguistic complexity of sentences within a work follows similar distributional patterns throughout the work.

For example, the following sentences from *Robinson Crusoe*, taken at random from different parts of the novel, contain five to twelve clauses each:

1. It was a great comfort to me afterwards that I did so, for not one grain of that I sowed this time came to anything, for the dry months following, the earth having

had no rain after the seed was sown, it had no moisture to assist its growth, and never came up at all till the wet season had come again, and then it grew as if it had been but newly sown.

2. During this time, I made my round in the woods for game every day, when the rain admitted me, and made frequent discoveries in these walks of something or other to my advantage; particularly I found a kind of wild pigeons, who built, not as wood pigeons in a tree, but rather as house pigeons, in the holes of the rocks.
3. The excessive hardness of the wood, and having no other way, made me a long while upon this machine, for I worked it effectually, by little and little, into the form of a shovel or spade, the handle exactly shaped like ours in England, only that the broad part having no iron shod upon it at bottom, it would not last me so long.

Sentences below taken at random from *Anna Karenina* contain three to four clauses each and demonstrate a different distributional pattern.

1. She had called him "Stiva," and he glanced at her with gratitude, and moved to take her hand, but she drew back from him with aversion.
2. But I have nothing very particular, only a few words to say, and a question I want to ask you, and we can have a talk afterwards.
3. Levin obediently helped himself to sauce, but would not let Stepan Arkadyevitch go on with his dinner.

Analysis of these small sample of sentences shows that the sentences taken from *Robinson Crusoe* are more complex, contain more embedded clauses, and show larger variance in the number of clauses they contain, while those from *Anna Karenina* have mostly coordinated simple clauses, fewer embedded clauses, and smaller variance. We expect that parallel translations of these titles will also exhibit expressive differences, indicating that given the same content, the mean and variance of the level of linguistic complexity of expression vary with author/translator.[3]

[3]Contrary to our initial intuitions, many translators do not follow the author's expression closely, but instead improvise and add their own artistic expression to works.

Evaluating linguistic complexity of sentences requires recursive analysis of sentences and their clauses. For example, Figure 6-4 shows the analysis of the sentence "I would not think that this was possible", in terms of its subject, i.e., "I", and predicate, "would not think that this was possible". The top-level subject (i.e., the left branch) and predicate (i.e., the right branch) have depths one and seven respectively, which is given by the distance from their lowest governing phrase node to the top-level S-node. Each of the left and right branches can also be analyzed in terms of the clauses they contain. In this case, the left branch, i.e., the subject, contains no further clauses, while the right branch which corresponds to the predicate, contains the clause "that this was possible".

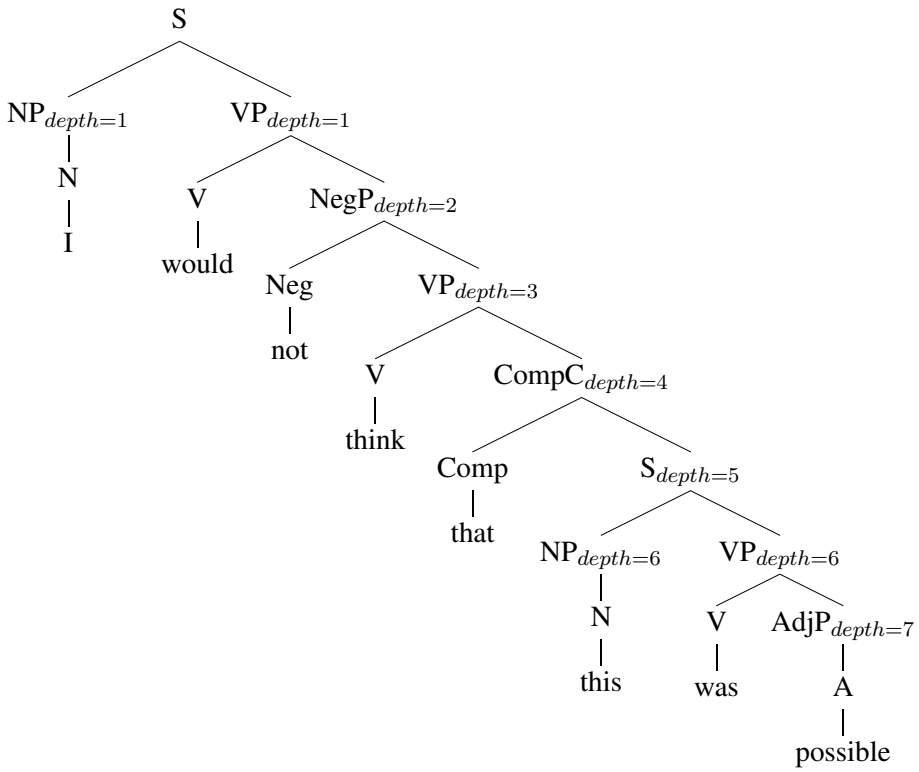


Figure 6-4: Depths of phrases.

Similarly, Figure 6-5 shows that the left branch of the sentence "That she would give such a violent reaction was unexpected" is the complementizer clause "That she would give such a violent reaction" of depth 8 and the right branch corresponds to the simple verb phrase "was unexpected" of depth 2. The complementizer clause, in turn, contains the clause "she would give such a violent reaction" with the subject of "she" and predicate of "would give such a violent reaction".

Note that the sentence "I would not think that this was possible" (in Figure 6-4) has a top-level right branch that is deeper than the top-level left branch, while the reverse is true for the sentence "That she would give such a violent reaction was unexpected" (in Figure 6-5). Therefore, although

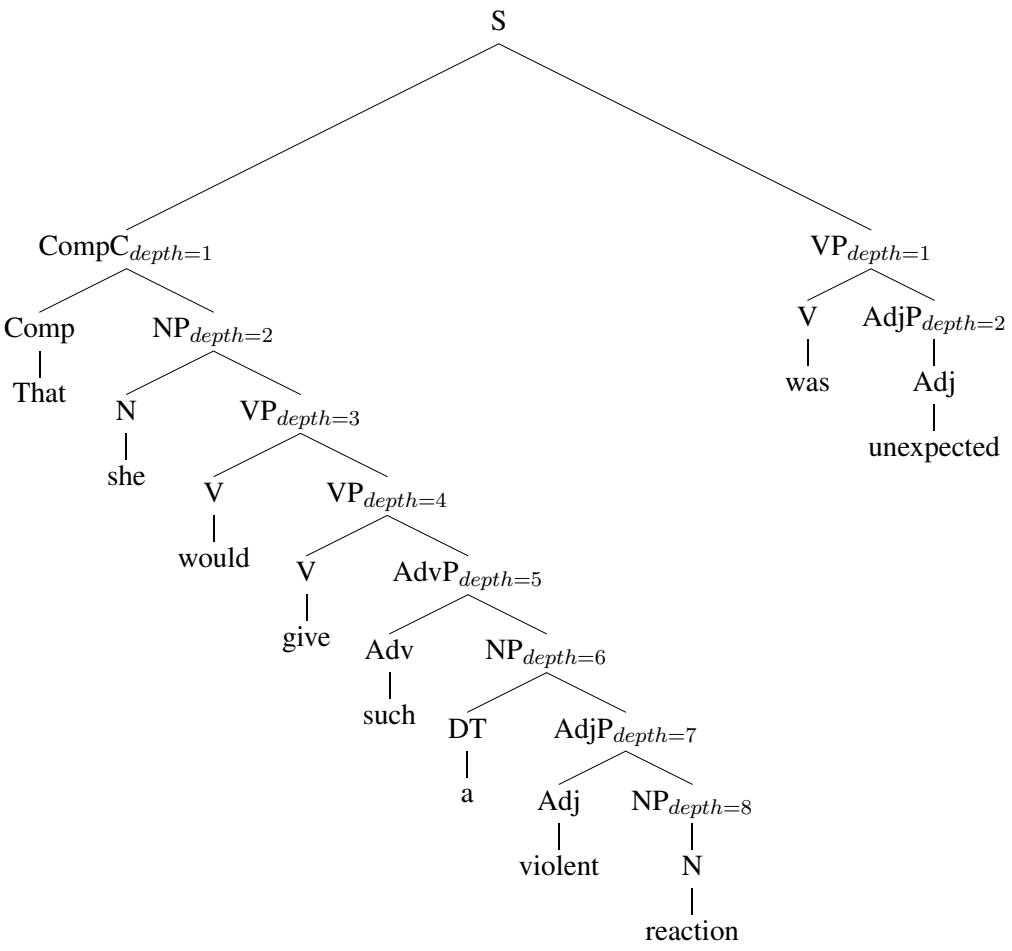


Figure 6-5: Depths of phrases.

they contain almost the same number of words, the structures of the two sentences are different.

To calculate the complexity of the sentence in terms of the number and depth of clauses, we used a corpus which is part-of-speech tagged using the Brill tagger [16]. Based on the part-of-speech tags, we wrote a rule-based shallow parser for breaking sentences into the correct left and right branches (which do not always correspond to subject–predicate pairs, e.g., see Figure 6-9) and into subject–predicate pairs. We assumed a binary branching syntax tree and represented the depth of a branch by the depth of the lowest phrase node along that branch. We also calculated the level of embedding in the left and right branches of the tree by counting the number of clauses included in each. Putting these statistics together, we represented the structure of a tree in terms of the depth of the top-level left branch, the depth of the top-level right branch, and the depth of the longest prepositional phrase, all measured by the distance of the deepest phrase node to the closest governing sentential node in that sub-tree; and the level of embedding of the top-level left and right

branches measured by the number of the clauses they contain. Additionally, we kept track of the number of clauses in which the left and right branches were equally heavy; where they were equally embedded; where the left branch was deeper than the right branch in terms of phrase depth (i.e., left-heavy), and vice versa (i.e., right-heavy); and where the left branch included more embeddings than right branch (i.e., left-embedded), and vice versa (i.e., right embedded).

Representation of binary syntax trees is not straightforward because many verbs take multiple complements and can produce trees with higher branching factors (Figure 6-6).

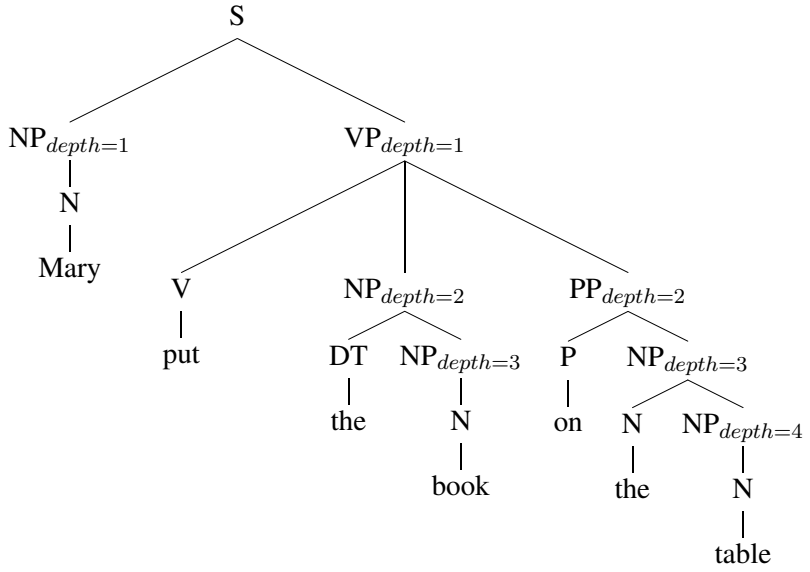


Figure 6-6: Ternary tree for ditransitive verb.

Therefore, commitment to binary branching requires special handling of multiple complements and modifiers. For ditransitive verbs this is especially important. We address this problem by using VP-shells, which wrap around lower level VP structures and allow attachments to the encapsulating shell instead of the encapsulated VP. In addition, we assume that the prepositional phrases are attached to the closest governing noun phrase or the verb phrase. For example, the sentence "Mary put the book on the table" is represented by the tree shown in Figure 6-7. The higher level VP at depth 1 in this tree acts as a shell for the VP at depth 2 and allows complements or modifiers to be attached to the VP without violating the binary branching requirement.

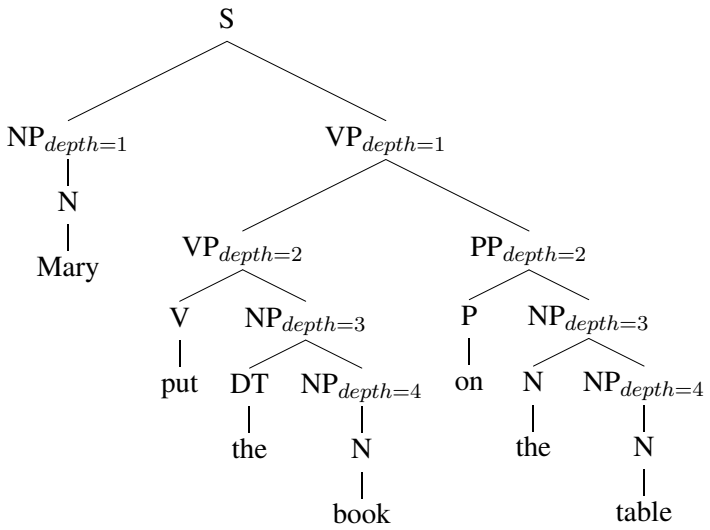


Figure 6-7: Binary branching tree for sentence in Figure 6-6.

Similarly, the sentence “Mary put the book on the table on Friday” adds one more VP-shell to the structure resulting in the tree in Figure 6-8.

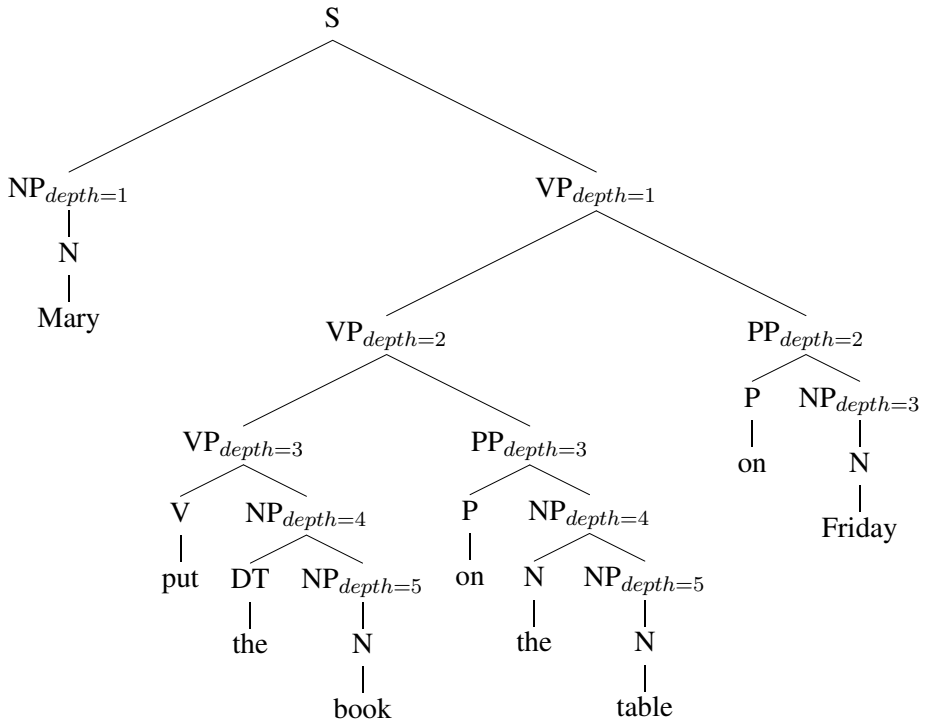


Figure 6-8: Binary tree for multiple prepositional phrases with VP-shells.

We also handle indirect questions and calculate the depth of the branches in the sentence “which book John bought I do not know” as shown in Figure 6-9. More examples of identified subject–predicate pairs in a random selection of sentences are shown in Table 6.9.

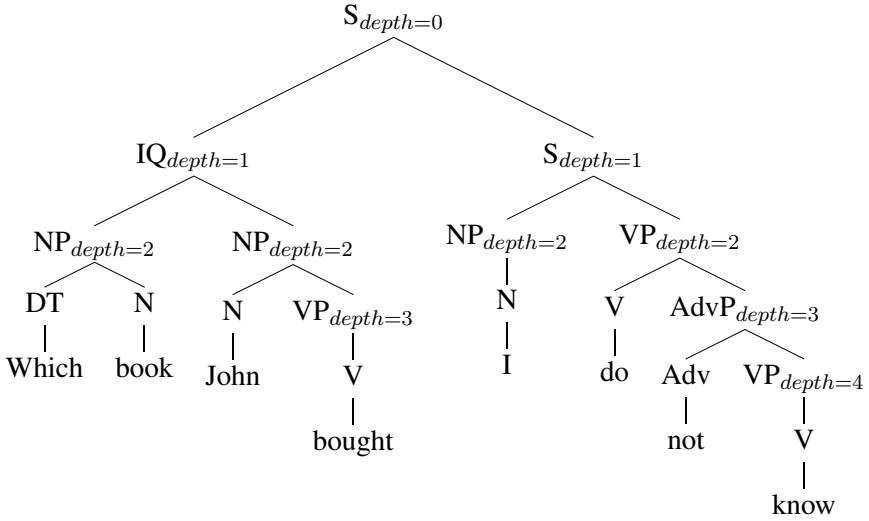


Figure 6-9: Binary tree structure for indirect question constructs.

| Sentence | Depth of top-level Subject & Predicate | |
|---|---|---|
| | Phrase | Clauses |
| $[\text{I}]_a$ [would not think that $[\text{this}]_b$ [was possible]$_b$]$_a$. | 1, 7 | 0, 2 |
| $[\text{I}]_a$ [have found $[\text{it}]_b$ [difficult to say that $[\text{I}]_c$ [like it]$_c$]$_b$]$_a$. | 1, 10 | 2, 2 |
| [That $[\text{she}]_b$ [would give such a violent reaction]$_b$]$_a$ [was unexpected]$_a$. | 8, 2 | 1, 1 |
| [For $[\text{her}]_b$ [to see this note]$_b$]$_a$ [is impossible]$_a$. | 5, 2 | 1, 1 |
| [Wearing the blue shirt]$_a$ [was a good idea]$_a$. | 4, 4 | 1, 1 |
| $[\text{It}]_a$ [is not known whether $[\text{he}]_b$ [actually libelled the queen]$_b$]$_a$. | 1, 9 | 0, 2 |
| $[\text{He}]_a$ [was shown that [the plan]$_b$ [was impractical]$_b$]$_a$. | 1, 6 | 0, 2 |
| $[\text{They}]_a$ [believed $[\text{him}]_b$ [to be their only hope]$_b$]$_a$. | 1, 7 | 0, 2 |
| $[\text{I}]_a$ [suggest $[\text{he}]_b$ [go alone]$_b$]$_a$. | 1, 4 | 0, 2 |
| $[\text{I}]_a$ [waited for [John]$_b$ [to come]$_b$]$_a$. | 1, 5 | 0, 2 |

Table 6.9: Sample sentences broken down into their clauses and the depth of the top-level subject and predicate, measured in terms of the depth of the lowest phrase and in terms of the number of clauses each branch contains. Top-level clause is always credited to the right branch.

> And as to a wheelbarrow, I fancied I could make all but the wheel, but that I had no notion of, neither did I know how to go about it; besides, I had no possible way to make the iron gudgeons for the spindle or axis of the wheel to run in, so I gave it over; and so for carrying away the earth which I dug out of the cave, I made me a thing like a hod which the laborers carry mortar in, when they serve the bricklayers.

Figure 6-10: Sentence from *Robinson Crusoe*.

For sentences that do not contain complicated use of punctuation, our algorithm produces the appropriate output. The language of the novels however, is arbitrarily complex; many phrases and clauses are appended together and separated by punctuation, e.g., Figure 6-10. In these cases, it is very difficult even for a full syntactic parser to find the correct syntactic structure. Naturally, our algorithm does not produce the correct depth in these cases; however, the final output is still a reasonable indication of the complexity of the sentences in terms of the structure of the left and right branches, subject–predicate pairs, and depths of these constituents.

In order to run a classification task for distinguishing between authors, we use the complexity information mostly in aggregate, representing documents with complexity information averaged over sentences in a document. This representation includes the average depths of the left and right branches of the trees, and the standard deviation of these depths, as well as the average number of clauses in each of the branches of the sentences and standard deviations of these values, average

number of prepositional phrases in each sentence and the depths of these phrases, the presence or absence of sentence-initial subordinate clauses and their depths, the percentage of left-heavy clauses (i.e., clauses where the left branch is deeper than the right branch), right-heavy clauses, right-embedded clauses, left-embedded clauses, and clauses where the two branches are equally heavy and equally embedded.

We claim that the measurements related to level of linguistic complexity can help differentiate between books and between authors. To validate this claim, we performed chi-square tests on a subset of these features.

Our chi-square tests on the frequencies (see Table 6.10(c)) of left-heavy and right-heavy clause structures, for example, indicated that these features were significantly different between the three translations of Madame Bovary with 99.9% confidence. However, for the two translations of *20000 Leagues*, the difference in the distributions could still be explained by chance. We interpret these results to mean that the differences in the left-heavy and right-heavy phrases can potentially be a manifestation of the expressive differences of some authors.

| | **20000 Leagues v.1** | | **20000 Leagues v.2** | |
|---|---|---|---|---|
| **Feature** | frequency | % | frequency | % |
| # left-heavy | 859 | 7.8% | 1335 | 8.3% |
| # right-heavy | 9706 | 87.9% | 13974 | 87.3% |
| equal-weight | 472 | 4.3% | 699 | 4.4% |
| Total | 11037 | 100% | 16008 | 100% |

(a) Raw counts and percentages of left-heavy, right-heavy, and equal-weight clauses in two translations of *20000 Leagues under the Sea*.

| | **M. Bovary v.1** | | **M. Bovary v.2** | | **M. Bovary v.3** | |
|---|---|---|---|---|---|---|
| **Feature** | frequency | % | frequency | % | frequency | % |
| # left-heavy | 865 | 6.2% | 789 | 6% | 803 | 5.4% |
| # right-heavy | 12484 | 89.1% | 11885 | 90.1% | 13459 | 91.1% |
| # equal-weight | 665 | 4.7% | 513 | 3.9% | 512 | 3.5% |
| Total | 14014 | 100% | 13187 | 100% | 14774 | 100% |

(b) Raw counts and percentages of left-heavy, right-heavy and equal-weight clauses in three translations of *Madame Bovary*.

| **Book** | **Degrees of Freedom** | **Chi-Square** | **p-value** |
|---|---|---|---|
| Two translations of *20000 Leagues* | 2 | 3.1 | $= 0.21$ (not significant) |
| Three translations of *M. Bovary* | 4 | 40.6 | $\leq 0.001$ |

(c) Results of chi-square.

Table 6.10: Raw counts, percentages, and chi-square test results for left- and right-heavy clauses as measured by the depth of subjects and predicates for *20000 Leagues* and *Madame Bovary*.

| | Anna Karenina part 1 | | Anna Karenina part 2 | |
|---|---|---|---|---|
| **Feature** | frequency | % | frequency | % |
| # left-heavy | 1082 | 4.9% | 1058 | 4.8% |
| # right-heavy | 20049 | 91.5% | 20293 | 91.8% |
| equal-weight | 791 | 3.6% | 756 | 3.4% |
| Total | 21922 | 100% | 22107 | 100% |

(a) Counts and percentages of left- and right-heavy and equal-weight clauses in odd and even chapters of *Anna Karenina*.

| Book | Degrees of Freedom | Chi-Square | p-value |
|---|---|---|---|
| Odd- and even chapters of*Anna Karenina* | 2 | 1.76 | $= 0.42$ |

(b) Results of chi-square.

Table 6.11: Counts, percentages, and chi-square results for left- and right-heavy clauses of *Anna Karenina*.

In contrast, when we compare the odd-numbered chapters of *Anna Karenina* to the even-numbered chapters, we find that we cannot reject the null hypothesis (see Table 6.11). More precisely, we find that there is 42% chance that the distributions of left- and right-heavy clauses in the two groups of chapters is the same and the differences between them can be explained by chance.

The ability of the complexity measurements to recognize different translations of the same work and their inability to recognize different portions of the same book indicate that these features vary between authors, and they do not in fact depend on content. We expect that when authors write about different content, these features will be able to identify expressive differences easily. Yet, the question remains as to whether these features capture the style of an author or her expression. Looking at *Emma* and *Sense and Sensibility*, and also *Emma* and *Robinson Crusoe*, we find that these features capture expression, however they also contain elements of style. While we can distinguish between both pairs of works easily and with more than 99% confidence, the chi-square value for *Emma* and *Robinson Crusoe* is much higher than that of *Emma* and *Sense and Sensibility*, indicating that the tested features vary more between authors and less between the works of the same author.

| Book | Degrees of Freedom | Chi-Square | p-value |
|---|---|---|---|
| *Emma* v. *Robinson Crusoe* | 2 | 99.49 | $p \leq 0.001$ |
| *Emma* v. *Sense and Sensibility* | 2 | 10.87 | $p = 0.004$ |

Table 6.12: Chi-square results for left- and right-heavy clauses for pairs of various books.

## 6.7 Expression as a Function of Verb Phrase Structure

We have so far discussed sentence-initial and -final phrase structures and sentence complexity as elements of an author's expression. In this section, we study the structure of phrases in more detail and identify characteristics of internal phrase structures that can be attributed to authors/translators.

Consider the following sentences from [8].

1. Someone is frying the fish.

2. The fish is frying.

3. The fish is being fried.

Each of these sentences convey the same basic information, but they vary in their syntactic structure and expression. The first sentence uses an active construct with a generic subject, the second sentence is an active construct where the object of the verb *fry* is used in the subject position, and the third sentence is a passive sentence. While the facts conveyed are the same, the syntactic differences produce different expression.

We analyze the internal phrase structure to capture additional elements of an author's expression. Verb phrases are diverse in their internal structures, which makes them a good target for internal phrase structure analysis. In particular, inspired by the notion of "richness of vocabulary", we define part of an author's "syntactic repertoire" and "linguistic richness" through an analysis of the classes of verbs used and the differences among these verbs with respect to their semantic classes and the syntactic structures underlying their alternations.

### 6.7.1 Linguistic Richness

The notion of "richness of vocabulary" has been used in the literature for distinguishing between authors' writing style. Most research directed at recognizing authors has analyzed vocabulary diversity by calculating the type–token ratio, i.e., the number of unique lexical items in the text divided by the total number of tokens in the text [117]. The basic assumption behind this approach to estimation of vocabulary richness is that authors have a repertoire of lexical items which they use at different frequencies [53]. Therefore, sampling the lexicon of a text provides us a sample of the author's lexical repertoire and her biases within that repertoire, giving us a measure of the "richness" or "diversity" of her vocabulary.

We extend the idea of vocabulary richness to elements of syntax. We define a notion of "linguistic richness" based on the syntactic constructs prevalent in an author's writing. In particular, we focus on the verb-driven aspects of syntax and describe an author's syntactic repertoire, and the level of her linguistic richness, in terms of the kinds of verb constructs she uses. Verb phrases have significant semantic and syntactic content, which can be used for identifying the linguistic repertoire of the authors.

### 6.7.2 Semantics of Verbs

Semantics of verbs have been studied in depth in the seminal work of Levin [72]. In this work, Levin exploited the observation that the syntax and the semantics of verbs are related, and showed that verbs that exhibit similar syntactic behavior are also related semantically (and vice versa).

Levin's study grouped 3024 verbs into 49 high-level semantic classes, and further classified them into 200 lower-level semantic groups. Verbs of "putting", such as `put`, `hang`, `scoop`, `drop`, `dribble`, `roll`, `drape`, `adorn`, `blanket`, and `bag`, for example, are collected under this high-level semantic class and can be further broken down into 10 semantically coherent lower-level classes which include "put verbs", "verbs of putting in a spatial configuration", "funnel verbs", "verbs of putting with a specified direction", "pour verbs", "coil verbs", "spray/load verbs", "fill verbs", "butter verbs", and "pocket verbs". Each of these lower-level classes represents a group of verbs that have similarities both in semantics and in syntactic behavior, i.e., they can grammatically undergo similar syntactic alternations.[4] For example, the verbs `coil`, `curl`, `loop`, `roll`, `spin`, `twirl`, `twist`, `whirl`, and `wind` belong to the semantic class of "coil verbs". These verbs can undergo "causative alternation" and "middle alternation" as shown by the examples from Levin [72] below:

1. Base Form
    - Cora coiled the rope around the post.
2. **Alternation 1**: Causative Alternation
    - The rope coiled around the post.

[4] Syntactic alternations are the alternate syntactic constructs that a verb can be involved in.

3. **Alternation 2**: Middle Alternation

    - That kind of rope coils easily around the post.

The semantic content of the verbs in general, and Levin's verb classes in particular, have previously been used for various natural language processing applications [59] and for evaluating document similarity [50, 64]. START was the first natural language processing system to use such classes in question answering [59]. For evaluating text similarity, given a set of news articles from the Wall Street Journal and semantically coherent verb classes such as those described in Levin [72], Klavans et al. found that different genres of news articles show biases towards different semantic verb classes. In addition, on a set of 50 news articles from the Wall Street Journal and using a set of verbs, from Levin, that belong to a relatively small number of semantic classes, Klavans et al. showed that verbs can be used to recognize common themes between articles [64].

Despite studies of verb classes for content and genre detection, verbs have been underutilized in studies of style and expression. We seek to use information from the semantic verb classes of Levin to describe the expression of an author in a particular work. In particular, we would like to exploit the fact that semantically similar verbs are often used in semantically similar syntactic alternations.

We consider the grammatically acceptable syntactic alternations involving semantically similar verbs to be also semantically similar. Then, given a particular content, the authors are free to choose among the available semantically similar verbs and their semantically similar syntactic constructs. This gives the authors flexibility in their expression. Therefore, to capture an author's syntactic repertoire and expression, we need to recognize the semantic classes of the verbs she uses and the particular alternations she prefers for them.

However, Levin's semantic classes cannot be directly used for modelling expression. More precisely, the lower-level semantic classes identified by Levin are too fine-grained and subdivide semantically related verbs based on their syntactic alternations, often resulting in synonyms being assigned to different low-level semantic classes. For example, the verbs `give` and `donate` belong to the high-level semantic class of "change of possession" however they are in two different low-level syntacto-semantic classes, namely "give" and "contribute" because of the differences in their syntactic behavior.

To avoid loss of synonymy information due to syntactic behavior, we use only Levin's higher-level semantic classes. The distribution of these semantic classes over works, without any consideration to syntax, should help us recognize titles. And, using this kind of semantic information in

| | | Chi-Square | | Likelihood Ratio | |
|---|---|---|---|---|---|
| **Book** | **#df** | $\chi^2$ | p-value | LR | p-value |
| Two translations of *20000 Leagues* | 48 | 174.6 | $p \leq 0.001$ | 138.7 | $p \leq 0.001$ |
| Three translations of *M. Bovary* | 92 | 247.18 | $p \leq 0.001$ | 263.9 | $p \leq 0.001$ |

Table 6.13: Chi-square and likelihood ratio test results for semantic classes in parallel translations.

coordination with syntactic structure of the alternations will help capture different expressions of semantically equivalent content.

Unfortunately, even when the study is limited to higher-level semantic classes, many verbs belong to multiple semantic classes. Therefore, successfully using semantic and syntactic verb information for any task requires a certain amount of sense disambiguation, i.e., selecting the most appropriate sense of a word in a given context. As word sense disambiguation is outside the scope of this thesis, to capture the semantics of the verbs used in a document, without identifying the most appropriate sense in a given context, we credit all semantic classes of all of the verbs inversely proportionally to the number of semantic classes that the verb belongs to, e.g., for each appearance of a verb that belongs to five different classes, each of its five semantic classes is credited with 0.20.

Contrary to our intuition, the results of significance tests indicate that semantic verb classes contribute to recognizing different translations of the same content. As shown in Table 6.13, both chi-square and likelihood ratio tests indicate that the differences in the distribution of semantic verb classes in parallel translations of *20000 Leagues* and *Madame Bovary* are significant with $p \leq 0.001$. This is partly because these syntacto-semantic classes are in fact limited in their assessment of semantics, because they represent a very small set of verbs, and because the semantic representation of documents that credits all semantic classes of verbs without selecting the most appropriate one introduces some noise to the analysis. Finally, the idiomatic verbs are omitted from this study, further constraining the ability to capture equivalent semantics. For example, the verb `remove`, from semantic class of "removing", appears 14 times in one of the translations of *20000 Leagues* and not at all in the other translation. Depending on the context, the instances of the verb `remove` in the first translation are replaced by `take off`, `make...fall`, `carry...away`, `clean...away`, `withdraw`, and other similar verbs (see Table 6.14). Among these, the idiomatic verbs do not appear in Levin's classes, and are therefore lost; and, the verb `make` belongs to the class of "verbs with predicative complements". As a result, we generate different representations for semantically equivalent sentences.

| translation 1 | translation 2 |
|---|---|
| ... diving suits were removed... | ...there our diving-dress was taken off... |
| ...second well-polished stone removed a tasty ringdove leg from conseil's hand... | ...a second stone, carefully aimed, that made a savory pigeon 's leg fall from conseil's hand... |
| ...a new one-meter slice was removed from this immense socket... | ...a new block a yard square was carried away... |
| ...I returned to my miner's trade, working to remove the fifth meter... | I resumed my miner's work in beginning the fifth yard. |
| ...four meters were left to be removed... | ...six yards of ice had been cleared, twelve feet only remaining to be cleared away. |
| Over the whole surface area, only two meters were left to be removed... | ...at the moment the manometer indicated that we were not more than twenty feet from the surface. |
| But the captain stopped me, signaled no, removed his dagger in one swift motion, and let the two valves snap shut. | But the captain stopped me, made a sign of refusal , and quickly withdrew his dagger, and the two shells closed suddenly. |

Table 6.14: Examples of paraphrases from two translations of *20000 Leagues*.

Testing the same features on two populations that are similar in content and expression, namely, two groups of chapters from *Anna Karenina*, gives a p-value of approximately $0.15$, indicating that the distribution of semantic classes among populations of chapters taken from the same book are too similar for us to reject the null hypothesis.

Finally, significance tests on works on different content by different authors and works on different content by the same author indicate that these features can tell authors apart but that their distributions are too similar among the works of an author for us to reject the null hypothesis. In other words, we can reject the null hypothesis when comparing books by different authors (with p-value $\leq 0.001$) but not when comparing works by the same author (see likelihood ratio test results for *Emma* and *Sense and Sensibility* in Table 6.15).[5]

The contribution of these semantic features to recognition of expression can be improved by coupling them with syntactic information. In particular, the grammatical syntactic alternations of semantically similar verbs and the differences in their use by different authors help capture more expression. In the next section, we discuss the syntax of verbs; in particular we consider embeddings, alternations, and the syntactic structure of the alternations of verbs.

[5]The similarity in the used semantic classes is partly due to the similarities and overlaps in the contents of *Emma* and *Sense and Sensibility*.

| | | Chi-Square | | Likelihood Ratio | |
|---|---|---|---|---|---|
| **Book** | **#df** | $\chi^2$ | p-value | LR | p-value |
| Odd & even chapters of *Anna Karenina* | 47 | 53.7 | $p = 0.23$ | 56.84 | $p = 0.154$ |
| *Emma* v. *Robinson Crusoe* | 45 | 1382.85 | $p \leq 0.001$ | 1463.3 | $p \leq 0.001$ |
| *Emma* v. *Sense and Sensibility* | 40 | 98.56 | $p \leq 0.001$ | 100.7 | $p = 0.154$ |

Table 6.15: Chi-square and likelihood ratio test results for semantic classes in pairs of various books chapters.

### 6.7.3 Syntax of Verbs

The syntax of verbs can be analyzed in many dimensions that range from the tense and aspect to voice and argument structure. Tense, aspect, and the voice of verbs have previously been studied in genre and authorship studies, e.g., [12]. Chapter 5 showed that writers can be distinguished by the distributions of active and passive voice sentences, and the particular differences in the distributions of *be-* and *get-passives* of works. In this section, we study the structure of sentences and the syntactic characteristics of verbs in more detail to get an insight into the contribution of these features to the syntactic repertoire and linguistic diversity of authors. Building on the results shown in the previous section about the value of semantic verb classes for capturing expression, we study the connection between syntax and semantics, and the differences in the syntactic choices made by authors who write about semantically equivalent content.

Levin's verb classes provide a good starting point for analysis of expressive characteristics of semantically equivalent texts. However, Levin's verb classes are limited to those that do not take clausal or verb phrase embeddings. To enlarge our set of verbs, we supplement Levin's classes with verbs that take complex arguments such as sentences, clauses, and verb phrases. We call this class of verbs "embedding verbs". In contrast, the verb classes of Levin contain "non-embedding verbs".

We study the syntax of embedding and non-embedding verbs in two different ways. For non-embedding verbs, we find the semantic class and the syntactic structure of the alternations of the verbs themselves. For embedding verbs, we identify their syntactic class by analyzing the syntax of their arguments and add to this the information we have about the semantics of the arguments.

#### Non-Embedding Verbs

To study the syntactic alternations of semantically related verbs gathered in the high-level classes of Levin, we used part-of-speech tags and shallow parsing, and broke sentences into their phrase-level constituents. For example, "spray/load verbs" can be seen in the following alternations [72]:

1. Base Form
    - Jessica sprayed paint on the wall.
    - NP + V + NP + PP.
2. **Alternation 1**: Locative Alternation
    - Jessica sprayed the wall with paint.
    - NP + V + NP + PP.
3. **Alternation 2**: Causative Alternation
    - Paint sprayed on the wall.
    - NP + V + PP.

We classified the alternations into 7 verb syntactic structure classes, gathering the remaining syntactic structures under the "other" category:

1. Verb followed by two consecutive noun phrases.
    (a) Mary gave John the book.
    (b) NP + V + NP1 + NP2.
2. Verb followed by a noun phrase and a prepositional phrase.
    (a) Mary gave the book to John.
    (b) NP + V + NP + PP.
3. Verb followed by a prepositional phrase.
    (a) Mary went to the store.
    (b) NP + V + PP.
4. Verb followed by a noun phrase.
    (a) Mary bought flowers.
    (b) NP + V + NP.

| | | Chi-Square | | Likelihood Ratio | |
|---|---|---|---|---|---|
| **Book** | **#df** | $\chi^2$ | p-value | LR | p-value |
| Two translations of *20000 Leagues* | 292 | 603 | $p \leq 0.001$ | 660.7 | $p \leq 0.001$ |
| Three translations of *M. Bovary* | 602 | 860.29 | $p \leq 0.001$ | 935.9 | $p \leq 0.001$ |

Table 6.16: Chi-square and likelihood ratio test results for semantic verb classes paired with their syntactic alternations in parallel translations.

| | | Chi-Square | | Likelihood Ratio | |
|---|---|---|---|---|---|
| **Book** | **#df** | $\chi^2$ | p-value | LR | p-value |
| Odd and even chapters of *Anna Karenina* | 283 | 287.1 | $p = 0.42$ | 316.7 | $p = 0.08$ |
| *Emma* v. *Robinson Crusoe* | 271 | 1852.15 | $p \leq 0.001$ | 2030.5 | $p \leq 0.001$ |
| *Emma* v. *Sense and Sensibility* | 238 | 440.65 | $p \leq 0.001$ | 471.4 | $p \leq 0.001$ |

Table 6.17: Chi-square and likelihood ratio test results for semantic verb classes paired with their syntactic alternations in various books.

5. Verb followed by a noun phrase and a prepositional phrase from which the noun phrase has moved.

   (a) Who is the book for —?

   (b) NP + V + NP + prep.

6. Verb followed by a prepositional phrase from which the noun phrase moved.

   (a) I don't know who she is waiting for —.

   (b) NP + V + prep.

7. Verb alone at either the sentence or the clause boundary.

   (a) She left.

   (b) NP + V

8. other

We studied the syntactic structure of the alternations of non-embedding verbs in parallel with their semantics, and tested the hypothesis that these pairs of semantic classes and their syntactic structures capture expression. The distributions of these features across translations of the same novel were found to be significantly different with $p \leq 0.001$ when tested either by chi-square and likelihood ratio (see Table 6.16).

As Table 6.16 shows, the distributions of these features across translations of the same novel are significantly different with $p \leq 0.001$ when tested either by chi-square or likelihood ratio.

In fact, the chi-square values for both sets of translations are very high, indicating that the syntacto-semantic preferences of translators are different when translating the same text. Similarly, the chi-square values are very high when we compare chapters from *Emma* with *Robinson Crusoe* and when we compare *Emma* with *Sense and Sensibility*. However, the difference between the distributions of these features over *Emma* and *Robinson Crusoe* is clearer, marked by higher chi-square value, than the difference in the distributions over *Emma* and *Sense and Sensibility*. Although they were written by the same author, *Emma* and *Sense and Sensibility* are still significantly different from each other, indicating presence of elements of expression unique to each work. On the other hand, testing the same features on chapters from *Anna Karenina*, we find that we are unable to reject the hypothesis that the two populations of chapters come from the same distribution.[6]

The results in Table 6.17 show that semantic verb classes capture the expressive differences between independent works by independent authors and independent works by the same author. As with previously evaluated features, the differences in the use of these features are more significant between independent works of independent authors when compared to independent works of the same author. This implies that these features capture some elements of style, as well as expression.

**Embedding Verbs**

Embedding verbs take clausal complements as either direct or indirect object. Syntactically, these complements range from infinitive verb clauses, to small clauses, indicative clauses, subjunctives, and indirect questions. Some of these complements require complementizers such as "that", others take wh-phrases that use "whether"or "if", while others do not take a complementizer at all.

For example, the verb "say" can take an indicative clause as its complement. This complement may or may not be preceded by the complementizer "that", as shown in Figure 6-11.

[6]The difference in the p-values of the chi-square and likelihood ratio tests (see Table 6.17) are due to the sparsity of the data, and therefore the likelihood ratio is considered more reliable.

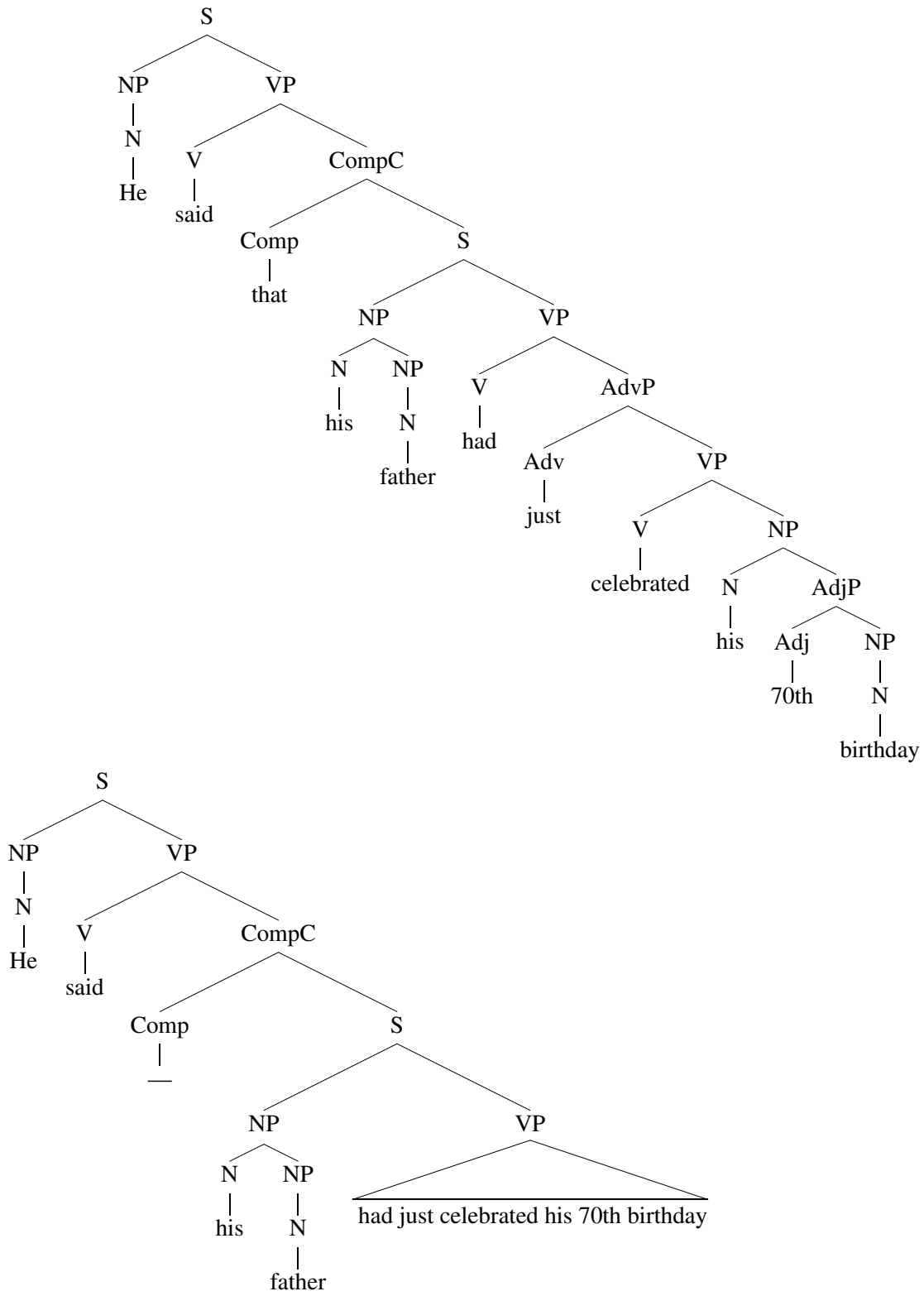


Figure 6-11: Structure for “say” with and without ‘that” complementizer.

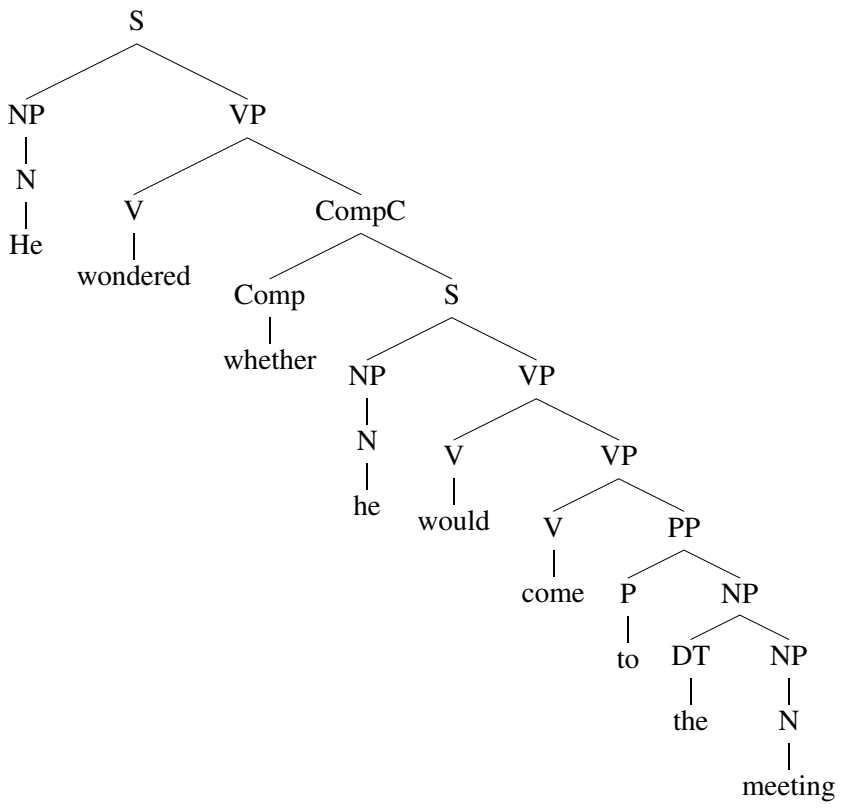


Figure 6-12: Structure for "wonder" with complementizer "whether".

On the other hand, the verb "wonder" embeds a complementizer clause headed with a wh-word (Figure 6-12), while the verb "want" embeds infinitive verb phrases (Figure 6-13).[7]

To find the embedding structure of these sentences, we study surface-level syntactic constructs and phrase-level constituents. For example, the structure of the sentence "He said that his father had just celebrated his 70th birthday" is analyzed in terms of its high-level constituents as shown below:

- **Original Sentence**: He said that his father had just celebrated his 70th birthday.
- **Basic Constituents**: He + said + that + his father had just celebrated his 70th birthday.
- **Formula**: Noun Phrase (NP) + Verb Phrase (VP) + that + Indicative Sentence (IS).

[7]Technically, what we call an "infinitive verb phrase" is in fact a clause with a covert subject, *pro*. However, surface methods for recognizing clauses cannot identify covert subjects. Therefore, we treat clauses with covert subjects as if they are phrases and call them phrases instead of clauses. In other words, for the sake of our analysis, "him to run" in "I want him to run" is a clause, while " to *pro* run" in "I want to run" is a phrase because *pro* is a covert subject and "him" is an overt subject.

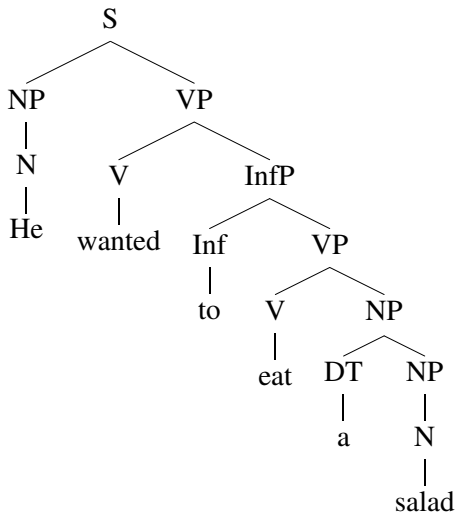


Figure 6-13: Structure for “want” + infinitive.

Similarly, sentences shown below for different syntactic alternations involving the verb "want" can be broken down into their high-level constituents through shallow-parsing using part-of-speech tags:

1. I want him destroyed. → NP + VP-head(Vh) + NP + Perfective VP

2. I want someone to destroy him. → NP + Vh + NP + to + VP

3. I want him to be destroyed. → NP + Vh + NP + to be + Perfective VP

These sentences convey almost the same meaning, but use very different embedding syntax, implying that authors have a choice in the verbs and alternations they choose to convey the content.[8]

The formulae for these sentences indicate that each of these sentences starts with a noun phrase, and continues with a verb, which we refer to as the "matrix verb" [60]. This matrix verb can be surrounded by an inflectional phrase, be preceded by auxiliaries, and/or be separated from the auxiliaries with adverbs. However, regardless of the presence or absence of auxiliaries, modals, or adverbs in the sentence, this verb determines the nature of the rest of the sentence—as sentences "I want him destroyed", "I want someone to destroy him", and "I want him to be destroyed" show, the verb "want" allows at least three different ways of conveying the same meaning. To signify its importance, we call this verb "verb phrase head" or "Vh". The part of the formulae that follow the Vh constitute the embedded argument of the Vh.

Kunz et al. have identified classes of embedding verbs and created comprehensive lists of verbs that follow the same syntactic formulae [1]. Bridgeman et al. [15] extended this work with more embedding verb classes, providing a great resource for studying classes of embedding verbs and the syntax of their embeddings. START was the first natural language processing system to use similar classes of embedded verbs in parsing and generation [60].

Kunz and Bridgeman's studies of classes of embedding verbs present the syntactic formulae that describe the structure of the embedded arguments of verbs in terms of their phrasal and clausal elements, such as verb phrases, participial phrases, infinitive phrases, indicative clauses, and small clauses. These formulae differentiate between embeddings of active and passive verbs, as well as

[8]Note that the sentence "I want his destruction" is yet another form of expressing the same meaning. In this case, the sentence structure is changed due to use of nominalization. Our primary experiments showed that nominalization does not add extra information to evaluation of expressive similarities when used in conjunction with syntactic repertoire and linguistic complexity features. This is because the resulting sentence structure is already captured by our analysis of sentence-initial and -final phrase structures. Therefore, we omit nominalization from this part of the analysis of expressive features.

the use of expletive `it` as the subject. For capturing expression, this distinction is useful, because, often, active and passive voice of even the same verbs take different kinds of clausal arguments. In particular, many of the embeddings used in active constructs are not grammatical when the Vh is in passive voice. Use of expletive `it`, e.g., formulae 26-29 in Table 6.18, further constrains the set of possible embeddings, resulting in only several constructs that are frequently encountered in texts.

For modelling expression, we used 29 of the more frequent verb embedding classes defined by Kunz and Bridgeman, shown in Table 6.18. We added to these classes information about the use of copular `be`.

In the formulae shown in Table 6.18 [1]:

- Noun Phrase (NP) refers to any phrase with a noun or pronoun as its head. e.g., John, she, the white house.
- Verb Phrase (VP) refers to any phrase with a verb as its head. e.g., go, cannot go.
- Participial Verb Phrase (Particip.) refers to a present participle of a verb followed by any NP, or adverb, etc., e.g., leaving, running fast, chewing gum.
- Infinitive Verb Phrase (Inf.) refers to the isolate form of a verb followed by any NP or adverb, e.g., sing, walk fast, eat chicken.
- Indicative Sentence (IS) refers to any subordinated declarative sentence which is not a subjunctive sentence, e.g., He ran, she does, I like fruit.
- Subjunctive (Subj.) refers to a subordinated declarative sentence where the verb is in the subjunctive form, e.g., she go alone, she run for president.
- A small clause (SC) looks like the subjunctive and can only be differentiated from it by studying the argument structure of the embedding verb.

| No. | Syntactic Formula | Example |
|---|---|---|
| 1 | NP + Vh + IS | I assume she attended the meeting. |
| 2 | NP + Vh + NP + IS | I will show him the problem is simple. |
| 3 | NP + Vh + SC | They saw John hide the money. |
| 4 | NP + Vh + NP + Partic. | I caught him stealing. |
| 5 | NP + Vh + NP + that + IS | She told him that she was flying at 7pm. |
| 6 | NP + Vh + NP + to + Inf. | They asked him to help. |
| 7 | NP + Vh + NP + wh + IS | He asked me if they were coming. |
| 8 | NP + Vh + NP + wh + to + Inf. | He told me how to act. |
| 9 | NP + Vh + Particip. | I began singing. |
| 10 | NP + Vh + Subj. | I request she go alone. |
| 11 | NP + Vh + for + NP + to + Inf. | We waited for John to come. |
| 12 | NP + Vh + from + Particip. | The seatbelt prevented him from getting hurt. |
| 13 | NP + Vh + that + IS | She admitted that she was not happy. |
| 14 | NP + Vh + that + Subj. | I request that she go alone. |
| 15 | NP + Vh + to + Inf. | My father wanted to see the world. |
| 16 | NP + Vh + to + NP + IS | She mentioned to him Smith's had a sale. |
| 17 | NP + Vh + to + NP + that + IS | She mentioned to him that Smith's had a sale. |
| 18 | NP + Vh + to + NP + wh + IS | He explained to me what methods were crucial. |
| 19 | NP + Vh + wh + IS | He asked if they were coming. |
| 20 | NP + Vh + wh + to (w/ covert subject) + Inf. | Mary knew how to keep them quiet. |
| 21 | NP + Vh + wh + to (w/ overt subject) + Inf. | They explained how to clean the sink. |
| 22 | NP + passive Vh (pass.) + IS | He was shown the plan was impractical. |
| 23 | NP + pass. + Particip. | He was seen stealing the keyboard. |
| 24 | NP + pass. + that + IS | He was shown that the plan was impractical. |
| 25 | NP + pass. + to + Inf. | He was advised to exercise. |
| 26 | it + Vh + NP + that + IS | It amazes me that she still has no idea. |
| 27 | it + Vh + that + IS | It seems that you were wrong. |
| 28 | it + pass. + that + IS | It can be deduced that he won the prize. |
| 29 | it + pass. + wh + IS | It is not known whether he passed the exam. |

Table 6.18: Syntactic formulae and examples of embedding verb classes based on Kunz and Bridgeman [1, 15].

| | | Chi-Square | | Likelihood Ratio | |
|---|---|---|---|---|---|
| **Book** | **#df** | $\chi^2$ | p-value | LR | p-value |
| Two translations of *20000 Leagues* | 25 | 57 | $p \leq 0.001$ | 64.3 | $p \leq 0.001$ |
| Three translations of *M. Bovary* | 58 | 136.8 | $p \leq 0.001$ | 144.2 | $p \leq 0.001$ |

Table 6.19: Chi-square and likelihood ratio test results for classes of embedding verbs in parallel translations.

| | | Chi-Square | | Likelihood Ratio | |
|---|---|---|---|---|---|
| **Book** | **#df** | $\chi^2$ | p-value | LR | p-value |
| Odd and even chapters from *Anna Karenina* | 29 | 26.9 | $p = 0.58$ | 29.7 | $p = 0.43$ |
| *Emma* v. *Robinson Crusoe* | 29 | 496.29 | $p \leq 0.001$ | 507.7 | $p \leq 0.001$ |
| *Emma* v. *Sense and Sensibility* | 29 | 95.14 | $p \leq 0.001$ | 99.6 | $p \leq 0.001$ |

Table 6.20: Chi-square and likelihood ratio test results for classes of embedding verbs in chapters from various books.

We used these syntactic verb classes to capture more of the syntactic repertoire of the authors. We captured the expressive use of embedding verbs by looking at the percentage of verbs in the document that belong to a class of embedding verbs and that are observed in the particular syntactic construct representing the embedding class. We used text tagged with the Brill tagger [16] to extract these verb structures.

Note that some kinds of embeddings, alternations, and even semantic classes appear very infrequently in some texts, making the data on these features sparse. For example, the embedding "NP + passive Vh + Participial VP" appears once in one translation of *20000 Leagues* and twice in the other.

For significance testing on this sparse data, we performed chi-square and likelihood ratio tests. The results shown in Table 6.19 indicate that for both sets of translations, the distributions of classes of embedding verbs are significantly different from each other and that the differences cannot be explained by chance. This implies that there is a dependency between the use of classes of embedding verbs and the translator, when given similar content.

The significance tests performed on the odd and even chapters of *Anna Karenina*, shown in Table 6.20, however, indicate that the distributions of classes of embedding verbs in these two populations are not different enough for us to reject the null hypothesis. These results indicate that classes of embedding verbs are useful for capturing expression.

In addition to capturing expressive differences between parallel translations and expressive consistency within a work, the distributions of embedding verbs also capture the expressive differences

between different authors when they write about different content and expressive differences in the works of the same author even when these works vary in content, as shown by the significance test in Table 6.20.

The results of the exploratory tests show that different expressions of the same content can be recognized by analyzing the syntactic structure. The choices of the authors regarding embedding or non-embedding verbs are particularly useful for capturing expressive differences and should be a part of the expressive elements describing a text.

In Chapter 8, we use the features discussed and tested in this section, to differentiate between books and writers.

## 6.8 Summary

In this chapter, we have described semantic and syntactic features of text documents centered around verbs and phrase structures for capturing the syntactic repertoire, linguistic diversity, and linguistic complexity of authors within a particular work. We identified sentence-initial and -final phrase structures, semantic verb classes and their alternations, and embedding verbs that take complex arguments, as elements of an author's syntactic repertoire. We studied the tendencies of authors to construct left- and right-heavy clauses, also as elements of syntactic repertoire, and the complexity of clause structures found in an author's sentences as a measure of her linguistic complexity. We evaluated these features with respect to their ability to differentiate between translations of the same title, between independent books by independent authors, between independent books by the same author, as well as their ability to measure expressive similarity throughout a book. Chi-square and likelihood ratio test results on these features showed that both syntactic and semantic elements of documents contribute to capturing expression. These features also capture elements of style.

# Chapter 7

# Semantic Categories and Content

Expression captures the language particular to the content presented in one work. To recognize expression, we find the link between syntactic elements of expression and the semantics surrounding them.

In Chapter 6, we described syntactic features that can capture how the content is presented. In this chapter, we model the semantic content of documents to create a high-level representation of the information that is conveyed.

## 7.1 Content in the Literature

Content refers to information and story contained in documents. Most approaches to modelling content rely predominantly on keywords. For example, search engines and information retrieval systems such as Google and SMART [18, 103] identify content similarity between a query and a set of documents using features based on keywords and verbatim phrases. Similarly, many text classification systems measure content similarity using these features. When the documents are long enough for keywords and phrases to capture the information and the story contained in the document, deeper analysis of syntax and semantics is usually omitted. However, for shorter documents, where the information is condensed into only a few paragraphs, evaluation of content similarity can take into consideration syntax and semantics, mainly because the documents are short enough for deeper analysis to be performed without much computational cost. For example, Hatzivassiloglou et al. [50] define similarity in terms of the focus of the text and find similarities between actors, objects, and actions.

An alternative to content modelling with keywords and phrases involves using a dictionary to create a higher-level representation of documents by mapping individual words to higher-level concepts, i.e., hypernyms. The most commonly used resource for hyponymy/hypernymy relations is WordNet [82]. But WordNet is not suitable for our purposes for two reasons: One problem is the fine-grained distinctions between the senses of the words in WordNet, i.e., many lexical entries in WordNet have senses that are very difficult even for humans to differentiate. The other problem is the irregularity of the lengths of the branches and the different levels of specificity of the concepts present at the same depth in WordNet. Many concepts that are at the same depth in the net differ in how specific or broad they are; e.g., "amnesia" and "psychological state" both appear at the same depth in WordNet, and it is very difficult to identify the depth at which a coherent set of broad concepts could be found.

These problems can be avoided by using a dictionary that contains a shallow hierarchy of semantically coherent and broad concepts, and that does not make as fine-grained distinctions between the senses of words. One such resource is the General Inquirer dictionary developed at Harvard University [116]. Using this dictionary, we disambiguate the words in a given context, map lexical entries to broader concepts, and create a semantic representation that can recognize conceptual similarities even between paraphrased versions of works.

In this chapter, we present the General Inquirer dictionary from which we gathered 62 broad, high-level semantic categories, and we test the ability of these GI categories to capture paraphrased content. Our results show that the GI categories provide a compact and effective representation of content. Although their representation is not as accurate as that of keywords, their small dimensionality and compactness make the GI categories suitable for studying the semantics of documents, and for identifying the link between syntactic elements of expression and the context in which they are used without significantly complicating the feature space of expressive elements.

## 7.2 General Inquirer (GI) Classes

The General Inquirer dictionary describes words in terms of general concepts, called "semantic marker categories" [35], that range from expressed emotions to actions to inanimate objects. This dictionary aggregates information from four different sources and contains more than 11,000 words that are marked with their senses, their parts of speech, and their categories.

The General Inquirer dictionary contains 26 high-level categories that are sub-divided into 182 lower-level ones. These categories include valence classes such as `Positiv` which contains "words of positive outlook" and `Negativ` which contains "words of negative outlook"; semantic dimensions such as `Ngtv` (which is "an earlier version of `Negativ`" and includes some elements that have also been tagged `Hostile` to indicate "an attitude or concern with hostility or aggressiveness") and `Pstv`; and semantic classes such as `Strong` and `Weak`, and `Active` and `Passive`; as well as semantic categories of "words of pleasure, pain, virtue and vice", "words indicating overstatement and understatement", and "words reflecting the language of a particular institution", etc. [35].

To capture high-level content using this dictionary, we preprocessed in three steps:

- We created parent nodes for groups of semantically coherent lower-level categories that were too fine-grained for our purposes. For example, we combined the categories that are presented in the General Inquirer as subcategories of "Motivation", namely `Need` ("words related to the expression of need or intent"), `Goal` ("names of end-states towards which muscular or mental striving is directed"), `Try` ("words indicating activities taken to reach a goal, but not including words indicating that the goals have been achieved"), `Means` ("words denoting objects, acts or methods utilized in attaining goals"), and `Persist` ("words indicating ... endurance") and formed the category `Motivation`.

- We merged lower-level categories with their common parent when the lower-level categories were too small and were semantically coherent with the existing parent node. For example, we combined the categories listed under "change process", i.e., `Begin`, `Vary` (words indicating change without connotation of increase, decrease, beginning or ending), `Increas`, `Decreas`, and `Finish` and created the `Change-Process` category.

- We eliminated redundant categories, e.g., kept `Negativ` but eliminated `Ngtv` ("an earlier version"), as well as categories that did not present a semantic class of content words, e.g., the

category `You`, which contains "9 pronouns indicating another person is being addressed directly."

We used the resulting 62 categories, shown in the appendix, for fingerprinting content. We refer to these categories as GI categories.

As with many dictionaries, each word in the General Inquirer dictionary can have multiple senses and can belong to multiple GI categories. For example, the word "make" has nine senses in this dictionary and each of its senses belongs to one or more GI categories (Table 7.1).

| Sense | GI Categories |
|---|---|
| Verb: create, produce, execute, construct; cause to be | Strong, Active, Work, IAV |
| Verb: coerce, force to, cause to | Negativ, Strong, PowTOT, Active, SocRel, IAV |
| Idiom-verb "make it": succeed | Strong, Active, Complet, IAV |
| Idiom-verb "make out": decipher; survive; complete | Active, Cognitive-Orientation, IAV |
| Idiom-verb "make up (one's) mind": decide | Active, Cognitive-Orientation, IAV |
| Verb "make up": compensate, make reparation | Positiv, AffTOT, Active, SocRel, IAV |
| Idiom-verb "make friends" | Positiv, AffTOT, SocRel, IAV |
| Noun "making": creation, production, execution | Strong, Active, Motivation |
| Noun "makings": ingredients | Object |

Table 7.1: Nine senses of "make" and their GI categories. See appendix for an explanation of the categories.

Polysemy of words presents a problem for choosing the correct sense and the relevant GI categories for a word in a given context. In the literature, this problem has been addressed in the context of word sense disambiguation, some approaches to which used online versions of regular dictionaries. Dictionaries contain descriptions of the different senses of words that can simplify the disambiguation process. For example, Wilks identified the correct sense of the word "bank" accurately 45% of the time by intersecting the keywords used in describing each sense of the word with the words that surrounded it in its context. In 85% of the cases, the correct sense was ranked in the top three [130].

For most of its entries, the General Inquirer dictionary lacks definition information. However, the syntactic and semantic information it contains can be used to represent the context of ambiguous words [61]. In addition, this dictionary provides a mapping between lexical entries and higher-level concepts which we used to represent the global context of documents, and to address the ambiguity problem.

Our representation of the global context of documents was based on the distribution of GI categories in a document. These distributions were obtained by counting all the GI categories associated with all of the senses of all of the words that appeared in the document. During this process, we considered the number of GI categories that each word sense belonged to, and allowed each category to contribute to the distribution of GI categories only in inverse proportion to the number of categories associated with that sense. For example, one of the noun senses of "make" belongs to the GI categories `Strong`, `Active`, and `Motivation`. For each appearance of the word "make", this particular sense of "make" would contribute 0.33 to the frequencies of each of these GI categories.

Given the distribution of GI categories representing the document, for each sense of each polysemous word in the document, we calculated a score based on the collective ability of its GI categories to represent the GI categories of the document. We selected the top scoring sense, i.e., the sense whose GI categories had the highest overlap with the GI categories of the document, for fingerprinting the content of the document. After disambiguation, we updated the distribution of the GI categories representing the document to reflect only the relevant sense of each ambiguous word (normalized by the document length).

As an alternative to content representation with GI categories, one could use a dictionary such as WordNet [82] to disambiguate the content words in a document and convert all synonymous words, their hyponyms, and hypernyms to the same canonical representation, highlighting semantic similarities. However, converting the documents to a canonical representation requires identification of the appropriate representation. If the canonical representation is based on the hypernyms of words, then the group of hypernyms that form a coherent class of broad concepts first have to be extracted from the dictionary. GI categories provide such a set of broad concepts, eliminating the need to generate such semantic categories from WordNet or any other dictionary.

In the next sections, we evaluate the contribution of GI categories to modelling content by comparing them with tfidf-weighted keywords for recognizing titles. The feature space of tfidf-weighted keywords represents the words that appear in a corpus, which in our case corresponds to a space of almost 11,000 dimensions. When the dimensionality is so high, many features are redundant; others are noisy, and can potentially cause overfitting. Therefore, before training classifiers on this very high dimensional space, we experiment with feature selection methods and boosting to estimate the parameters of the classifiers that can serve our purposes without overfitting the data.

## 7.3 Parameter Tuning

Models used in this thesis use the vector space model where each feature is a dimension and each document is a data point. This representation typically results in a very high dimensional feature space for classification. In the case of our largest data set of 45 novels (see Table 8.2 on page 171 for the details of this corpus), using keywords to represent documents yields more than 11,000 dimensions. Even on our smallest corpus of seven books, which contains translations of three titles, the number of keywords, and hence dimensions, is more than 4200. In both corpora, the number of dimensions reflects the number of stemmed [7] content words in the corpus that are not proper nouns or foreign words, and that appear in the corpus more than 5 times and in at least 3 documents. These limitations on the keyword-space eliminate the character names that would reveal titles without capturing content and also get rid of misspelled words that introduce noise to the representation.

In pattern recognition, in order to model the data accurately without overfitting, the training data is expected to include 5 to 10 samples from each class per dimension, i.e., the ratio of the number of training samples from each class to the number of dimensions in the feature space should be between 5:1 and 10:1 [57]. As classifiers get more complex, this heuristic ratio also increases [58], and using fewer than the prescribed number of samples per dimension usually results in a "peaking effect", i.e., the performance of classifiers reaches a maximum value after which adding more features hurts performance. On large dimensional spaces such as the vector space model with keywords, there is rarely a data set large enough to satisfy the heuristic ratio of ten samples per class per dimension without "peaking".

When the data is separable in low-dimensional spaces, the absence of a large enough data set can be compensated for by selecting and using only the most informative features for classification. For example, information gain can provide a measure of informativeness for the features in the training set, and this measure can be used to limit the feature space to only the more informative features; if the goal was to limit the feature space to ten features, information gain would provide a way of selecting those ten features. However, most data cannot be optimally separated in low-dimensional spaces. As a result, limiting the feature space to a small number of features results in loss of useful information. To find the correct feature space size for our data, we select features by cross-validating on the training set.

When the feature space is too complex, boosting and limiting the complexity of the decision tree can also regularize the feature space, i.e., focus the models on only the features that are most useful for more generalizable models. Therefore, to further minimize the risk of overfitting, without directly eliminating features and losing information, we also determined:

1. The number of rounds of boosting, and
2. The number of samples in each leaf of the decision tree,

which minimized the cross-validation error on the training set.

**Feature Selection**

Limiting the feature space can help classifier performance [57]. However, the heuristic of limiting the space to one tenth as many features as samples present in a class severely limits the feature space, resulting in significant information loss. Therefore, to constrain the feature space without eliminating useful features, and without imposing hard limitations on the classifiers, we used information gain.

Information gain has frequently been used for feature selection in the context of text classification. Other methods of feature selection, such as document frequency of words, mutual information, $\chi^2$ test, and term strength have been compared to information gain by Yang and Pedersen [124], who have shown that information gain was more successful than the rest of the approaches at reducing the feature space while improving classification performance. Similar results were shown by Forman [40].

We used information gain to select the features whose individual contribution to classification is above a threshold, with the objective of eliminating the features that are least useful for separating the data. Our experiments on the training set showed that setting the threshold information gain value to zero significantly reduced the dimensionality.[1]

On our parallel translation corpus (chapters from translations of the titles *Madame Bovary*, *20000 Leagues under the Sea*, and *The Kreutzer Sonata*), selecting only the features with non-zero information gain eliminated approximately 3000 features, resulted in a space of 1096 dimensions,

[1]Eliminating the features that have zero information gain value on their own eliminates some features that could potentially contribute to classification when combined with other features. However, keeping the features with zero information gain value in the feature space would require keeping all of the features or setting an arbitrary threshold for the number of features to be used. In the case that we choose to keep all of the features, there would be a significant increase in the computational cost. Decision trees scan all values of each of the dimensions of the feature space at every iteration and therefore the computation time is exponential in the number of features.

and gave a cross-validation accuracy of 97% on the training set using a decision tree boosted with ten rounds of AdaBoost. Similarly, limiting the feature space to GI categories that had non-zero information gain eliminated 18 of the 62 GI categories, resulting in a feature space of 44 dimensions, and gave cross-validation accuracy of 91% on the training set.

**Boosting**

Boosting is known to improve the performance of classifiers by iteratively identifying the more difficult sample points, training classifiers that can give better results on these difficult sample points, and combining the votes of classifiers. Schapire argues that increasing the number of rounds of boosting does not usually overfit [108]. However, boosting is expensive. Therefore, we identified the number of rounds of boosting beyond which further boosting did not improve the performance. Table 7.2 shows that, in our experiments on the training corpus of three titles, while modelling titles (i.e., three-way classification), the tfidf-weighted keywords benefitted from boosting up to 30 rounds; the GI categories benefitted from boosting up to 50 rounds. After 50 rounds, further boosting did not improve performance of classifiers built with either feature set.

| | **tfidf-weighted keywords** | **GI categories** |
|---|---|---|
| 10 rounds of boosting | 97% | 98% |
| 30 rounds of boosting | 98% | 98% |
| 50 rounds of boosting | 98% | 99% |
| 70 rounds of boosting | 98% | 99% |
| 90 rounds of boosting | 98% | 99% |

Table 7.2: Confusion matrix for tenfold cross-validation performance on the training set of three titles using boosted decision trees.

Similarly, the classifiers performed best when the decision trees were required to have at least a threshold number of samples in each leaf. This threshold was found to be equivalent to 15% of the number of samples in the largest class in the data set.

In the following sections, we use the parameter values set in this section to report on the performance of tuned classifiers on the test set. Thereafter, we run classification experiments on a corpus that contains 45 titles, using boosted decision trees for which the parameter values are set similarly on the training set.

## 7.4 Evaluation of GI Categories for Capturing Content Similarity

Representation of documents based on semantic categories, such as those found in the General Inquirer dictionary, abstracts away from keywords and has the potential to capture content similarity between paraphrased versions of works. For example, when used for modelling the content of parallel translations of three titles,[2] the boosted decision trees using GI categories correctly identified chapters from these titles 91% of the time, using 44 GI categories. Table 7.3 shows that the classifier made six mistakes.

| **Real Label** | **Predicted Label** | | |
|---|---|---|---|
| | **Bovary** | **20000 Leagues** | **Kreutzer** |
| Bovary | 18 | 4 | 0 |
| 20000 Leagues | 0 | 22 | 0 |
| Kreutzer | 2 | 0 | 20 |

Table 7.3: Confusion matrix for test performance of boosted decision trees (30 rounds of boosting with all GI categories whose information gain value was greater than zero) using semantic category information for content modelling with accuracy of 91%. The complete corpus contains 56 chapters from each title, 32 of which were used for training and parameter tuning. Results are reported on the 22 test samples from each title.

Ranking the GI categories with respect to information gain identified `PLACE`, `PowTot`, `DIST`, `AffTot`, and `Strong` as the most informative categories. The contents of each of the top five GI categories is described in Table 7.4.

Although 44 features had non-zero information gain value for classifying this data, separation of the chapters into different titles was quite successful even on two-dimensional feature spaces of GI categories. Figures 7-1, 7-2, and 7-3 show clear separation of the titles from each other, and clustering of chapters that belong to the same title on three different two-dimensional spaces, even when the chapters come from different translations.

[2]We used a total of 28 chapters from each of the two translations of *20000 Leagues under the Sea*, 18 or 19 chapters from each of the three translations of *Madame Bovary*, and 28 chapters from each of the two translations of *The Kreutzer Sonata*. We separated this set into training and test sets—a total of 24 chapters from each title was used for testing and the rest for training. Results are reported on the test set.

| GI Category | Contents |
|---|---|
| PLACE | words referring to places of social gatherings, regions, and routes, as well as places in nature, e.g., land, and sky |
| PowTot | words referring to a valuing of having the influence to affect the policies of others, power gain, power loss, the goals of the power process, political places, and environments except nation-states, power conflicts, power cooperation, individuals and collective actors in power process, non-authoritative actors in the power process, ideas about power relations and practices, and tools or forms of invoking formal power |
| DIST | words referring to distance and its measures |
| AffTot | words referring to valuing of love and friendship, reaping affect, affect loss and indifference, affect participants such as friends and family |
| Strong | words implying strength, including those that indicate a concern with power, control or authority |

Table 7.4: Top five GI categories, their descriptions from the Inquirer dictionary.

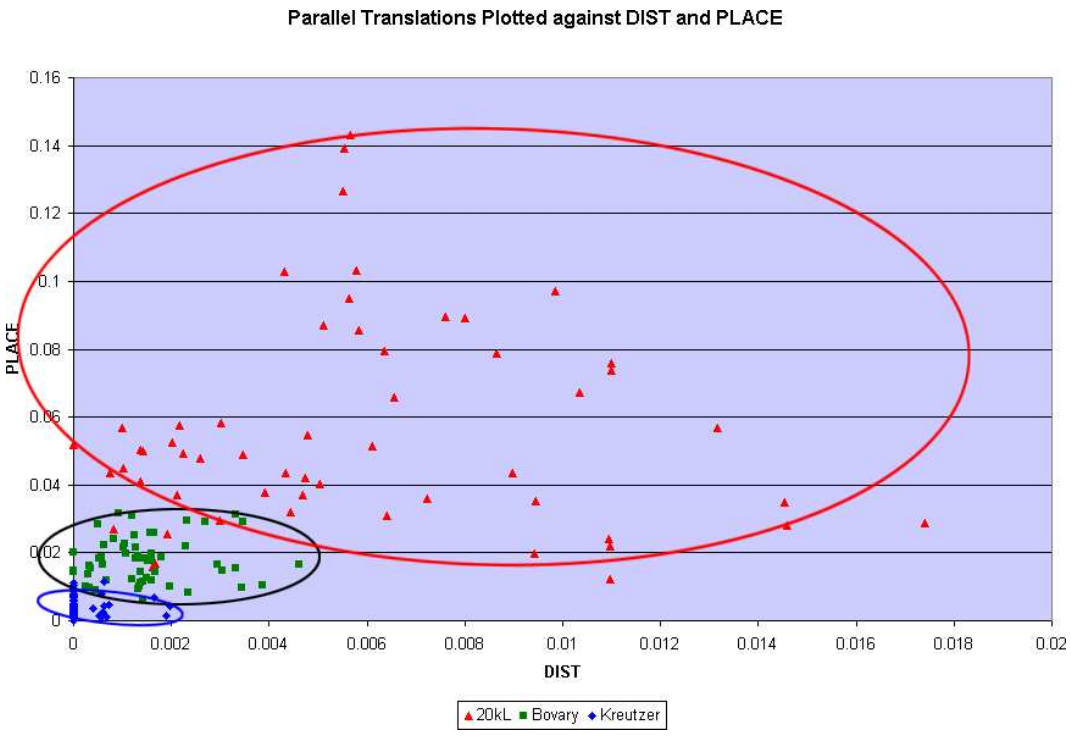


Figure 7-1: Translations of titles plotted against the GI categories “DIST” and “PLACE”.

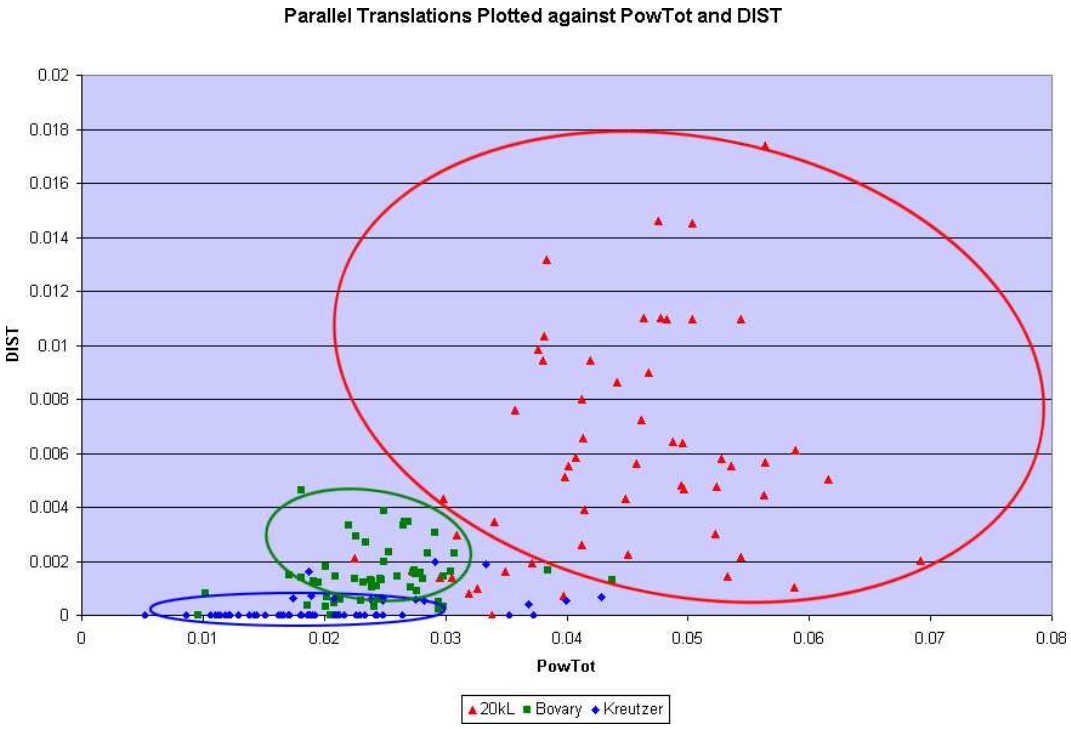


Figure 7-2: Translations of titles plotted against the GI categories “PowTot” and “DIST”.

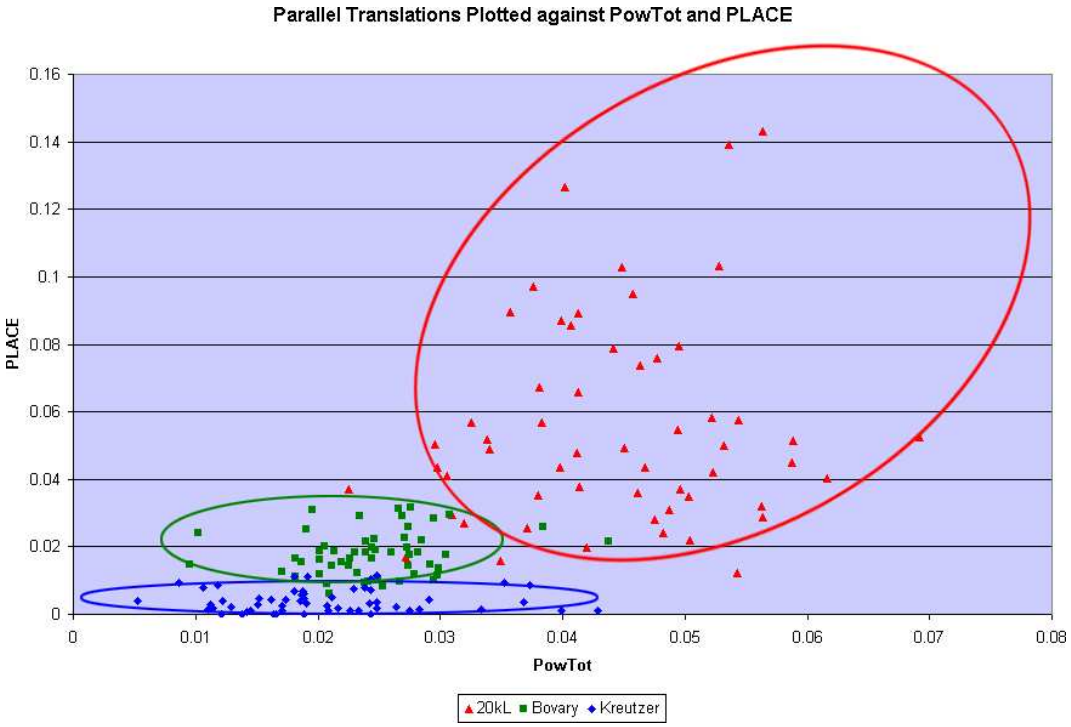


Figure 7-3: Translations of titles plotted against the GI categories “PowTot” and “PLACE”.

In comparison, models based on keywords recognized titles 97% of the time when we weighted the keywords with tfidf, normalized for document (i.e., chapter) length, and tuned the parameters for the classifier on the training set.

These two approaches to representing content differ in their strengths. Keywords take advantage of the unambiguous words that represent the disjoint contents of these titles. For example, the words "doctor", "ship", "west", "sea", and "lady" are highly predictive on this corpus. The GI categories, on the other hand, provide a higher-level representation of content, grouping semantically similar words under the same category. Keywords recognize the content similarity between paraphrases by taking advantage of words that are common to the paraphrases; GI categories recognize the content similarity by recognizing that the paraphrased words belong to the same semantic class. The strengths and weaknesses of the two approaches are complementary: when there exist monosemous words that are unique to a class, keywords are very informative, and when the content is paraphrased with synonymous words, GI categories are very informative.

The ability of GI categories to map content to a small number of semantic classes makes them particularly appropriate for obtaining a general overview of content. However, because these semantic classes are high level, and because they shrink a feature space of thousands of words to only tens of dimensions (at the expense of eliminating also some dimensions that do not exist in the GI dictionary), we cannot expect the GI categories to capture content as accurately as keywords. For example, on our corpus of 45 titles, the GI categories recognized titles 44% of the time, while keywords recognized titles 64% of the time. This difference is statistically significant at $\alpha = 0.05$ as measured by the z-test.[3] The parameters for these classifiers were tuned by cross-validating on the training corpus. Information gain value threshold was set to zero. The number of rounds of boosting was 210, after which further boosting did not improve the performance of either feature sets. The number of samples in each leaf was set to 15% of the number of samples in the largest class.

| **Corpus Size** | **Accuracy of GI Categories** | **Accuracy of tfidf-weighted keywords** |
|---|---|---|
| 45 titles | 44% | 64% |

Table 7.5: Test performance of GI categories and keywords on the corpus of 45 titles.

As more titles are added to the corpus, both the keywords and the GI categories lose some predictive capability, and confuse titles for each other. The ability of GI categories to recognize independent titles deteriorates more drastically than the ability of keywords to recognize independent

[3] On this corpus, a difference of 4% is statistically significant at $\alpha = 0.05$.

titles. However, the ability of keywords to recognize paraphrases decreases more drastically than the ability of GI categories to recognize paraphrases. This is because GI categories are broad semantic classes and a very simple representation that can reduce independent works on the overlapping content to the same representation. As a result, they may misclassify independent titles. However, for the same reasons, they are more accurate at recognizing paraphrases of titles. As Table 7.6 shows, GI cateogories correctly recognize paraphrased titles 83% of the time (45-way classification experiment), whereas tfidf-weighted keywords recognize paraphrases of titles only 71% of the time.

| **Corpus Size** | **Accuracy of GI Categories** | **Accuracy of weighted keywords** |
|---|---|---|
| 45 books<br>(3 parallel translations) | 83% | 71% |

Table 7.6: Test performance of GI categories and keywords on only the paraphrased titles (parallel translations) contained in the corpus of 45 titles.

Despite not performing as well as keywords on recognizing individual titles, the GI categories capture significant content—their accuracy is significantly better than the baseline of picking the class randomly, i.e., 2.2%. In addition, these features recognize paraphrases of titles more accurately than tfidf-weighted keywords and the compactness of the feature space generated by these semantic categories makes them more appropriate for capturing the semantic elements of expression than keywords.

## 7.5 Conclusion

The results reported in this chapter indicate that high-level semantic categories, such as those presented in the General Inquirer dictionary, capture content and recognize paraphrases of works more accurately than tfidf-weighted keywords on our largest corpus. What is more, the information captured is independent of keywords. The simplicity of the feature space provided by the GI categories and the small number of dimensions they encompass, in addition to their ability to recognize paraphrases, make GI categories especially appropriate for capturing content.

We use the GI categories to obtain a low-dimensional and high-level content fingerprints of documents for identifying the context which ties syntactic linguistic elements of expression, described in Chapter 6, to content. In Chapter 8, we combine the expressive linguistic elements with semantic categories that capture content to generate fingerprints that evaluate text similarity with respect to both syntax and semantics, and to capture the unique expression of an author in a particular content. If two works include the same general concepts and use similar expressions, their expression–content fingerprints should be similar. If, however, they differ in either content or expression, their fingerprints should reflect this difference. We capture these fingerprints using syntactic features described in Chapter 6 and GI categories discussed in this chapter.

## 7.6 Summary

In this chapter, we presented the General Inquirer dictionary which we used to recognize titles even when they are paraphrased. We showed that the semantic classes of the General Inquirer dictionary can recognize titles significantly better than chance, without requiring a high dimensional, complex feature space. These semantic classes can recognize paraphrased titles more accurately than tfidf-weighted keywords—on a corpus of 45 titles which contains seven translations of three titles, they recognized paraphrased titles accurately 83% of the time, whereas keywords recognized paraphrased titles accurately 71% of the time.

# Chapter 8

# Evaluation

In Chapters 6 and 7, we described syntactic and semantic elements of expression. We used the semantic elements of expression to represent content and we used the syntactic elements of expression to represent linguistic choices that affect how the content is conveyed.

In this chapter, we combine the semantic and syntactic elements of expression to form linguistic elements of expression. Using these linguistic elements, we generate language models that identify:

- Paraphrases of titles (irrespective of paraphrases, all versions of *The Kreutzer Sonata* is identified as the same title),
- Expression in individual books (all works, including different paraphrases of the same title, are considered different books), and
- Style of authors.

For identification of titles, books, and authors, we use a corpus of 45 titles. 3 of the titles included in this corpus contain multiple translations (*Madame Bovary* has three translations, *20000 Leagues* has two translations, and *The Kreutzer Sonata* has two translations). Throughout our experiments, we treat translations of the same original as paraphrases of the same content and we consider them to be different books (derived from the same title). The corpus therefore contains 49 books.

We use boosted decision trees [133] for modelling titles, books, and authors. For each of these experiments, we compare the performance of linguistic elements of expression with different sets of baseline features, such as function words, distribution of word lengths, distribution of sentence lengths, and tfidf-weighted keywords.

Our results show that the linguistic elements of expression outperform all baseline features in identifying books (even when some books are derived from the same original and therefore share content) and titles (even when some titles appear in the form of different paraphrases). In both of these cases, the linguistic elements of expression give an accuracy of more than 80%. Our results also show that the linguistic elements of expression are more relevant for capturing expression of content than capturing the style of authors.

## 8.1 Data

Authorship attribution literature uses corpora consisting of literary works that are written by native speakers of English, that are in the same genre, and that are written around the same time periods [65, 80, 85, 91, 128]. By controlling time period and genre, these corpora help expose the linguistic differences that are due to authors.

For the experiments in this section, we used a similarly controlled corpus. For studies on recognizing books and titles, we used titles from authors who lived in approximately the same time periods. For experiments on recognizing authors, we also limited the corpus to works written by native speakers of English.

The books in our corpus are:

- Jane Austen (1775-1817): *Northanger Abbey*, *Emma*, *Sense and Sensibility*, *Mansfield Park*, *Lady Susan*, *Persuasion*, *Pride and Prejudice*.
- Fyodor Dostoyevski (1821-1881): *The Idiot*.
- Charles Dickens (1812-1870): *A Tale of Two Cities*, *David Copperfield*, *Old Curiosity Shop*, *Oliver Twist*, *Pickwick Papers*, *The Life and Adventures of Nicholas Nickleby*.
- Arthur Conan Doyle (1859-1887): *The Firm of Girdlestone*.
- George Eliot (1819-1880): *Adam Bede*, *Middlemarch*, *Daniel Deronda*, *The Mill on the Floss*.
- Gustav Flaubert (1821-1880): 3 translations of *Madame Bovary*.
- Thomas Hardy (1840-1928): *The Mayor of Casterbridge*, *A Laodicean: A Story of To-Day*, *The Hand of Ethelberta: A Comedy in Chapters*, *Far from the Madding Crowd*, *Jude the Obscure*, *Tess of the d'Urbervilles: A Pure Woman*.

- Ivan Turgenev (1818-1883): *The Torrents of Spring*, *A House of Gentlefolk*.
- Victor Hugo (1802-1885): *The Hunchback of Notre Dame*, *The History of a Crime*, *The Man Who Laughs*.
- Washington Irving (1789-1859): *Life and Voyages of Christopher Columbus Vol. II*, *Chronicle of the Conquest of Granada*, *Knickerbockers History of New York*.
- Jack London (1876-1916): *The People of the Abyss*, *Adventure*, *The Little Lady of the Big House*, *The Sea Wolf*, *The Cruise of the Snark*, *Michael, Brother of Jerry*, *Burning Daylight*, *The Iron Heel*, *The Mutiny of the Elsinore*.
- William Makepeace Thackeray (1811-1863): *Catherine: A Story*, *The Memoirs of Barry Lyndon, Esq.*, *The Great Hoggarty Diamond*, *The Newcomes: Memoirs of a Most Respectable Family*, *The Tremendous Adventures of Major Gahagan*, *The History of Henry Esmond, esq: A Colonel in the Service of Her Majesty Queen Anne*, *The Virginians: A Tale of the Eighteenth Century*, *The History of Pendennis*, *The Book of Snobs*.
- Leo Tolstoy (1828-1910): *Anna Karenina*, 2 translations of *The Kreutzer Sonata*, *Resurrection*.
- Mark Twain (1835-1910): *The Mysterious Stranger*, *A Connecticut Yankee in King Arthur's Court*, *The Adventures of Huckleberry Finn*, *Following the Equator: A Journey Around the World*, *The Gilded Age: A Tale of Today*, *Those Extraordinary Twins*, *Christian Science*, *The Adventures of Tom Sawyer*.
- Jules Verne (1828-1905): *Journey into the Interior of the Earth*, 2 translations of *20000 Leagues under the Sea*, *In Search of the Castaway*, *The Mysterious Island*.

For each of our classification experiments, we selected a subset of titles from this corpus.

- To recognize titles, even when they are paraphrased, we used a corpus of 45 titles (i.e., all the titles in our corpus that contained more than 40 chapters) that included translations of some titles.[1] This corpus of 45 titles was split into training and test sets; 60% of the chapters of each title were used for training and 40% were used for testing. For titles with multiple translations,

[1]Translations are derived from the same original and convey the same content with different expression. Although they are not paraphrases of each other, they are paraphrases of the same original (in another language). Therefore, for this experiment, translations represent paraphrases.

chapters from one of the translations were used for training and the rest were used for testing. This experiment was repeated three times; at each round, a different translation was trained on.

- To capture the expression in individual books, we used a corpus of 47 books which included books that were translations of *Madame Bovary* and *20000 Leagues*. For this experiment, we split the corpus into training and test sets; 60% of the chapters of each book were used for training and 40% were used for testing.

- To capture the style of authors, we used a corpus that contained multiple books by each of the native English-speaker authors in our corpus. For this experiment, the classifiers were trained by studying several books by 8 authors and tested on a different set of books from the same authors. The books were split into training and test sets randomly. This experiment was repeated with 5 different sets of training and test sets.

To eliminate biases, the corpora designed for each of the experiments contained the same number of samples from each class. Parameters were tuned on the training set through cross-validation and the results were reported on the test set.

## 8.2 Feature Set

The main goal of this chapter is to evaluate the linguistic elements of expression, described in Chapter 6 and Chapter 7, with respect to their ability to recognize titles from their expression of content. We also use these features to identify books from the expression of their content. For this evaluation, we ran experiments with the linguistic elements of expression and compared the performance of these features with the performance of baseline features. In all, we ran experiments with six sets of features:

1. Phrase structure features:

    - The syntactic repertoire of the authors as captured by:
        - Sentence-initial and -final phrase structure;
        - Syntactic classes of embedding verbs;
        - Semantic classes of non-embedding verbs;
        - The context provided by the syntactic classes of embedding verbs for particular classes of semantic non-embedding verbs, i.e., which semantic verb classes tend to be embedded by which syntactic embedding classes;
        - The context provided by semantic non-embedding verb classes for particular syntactic structures of the alternations of verbs, i.e., which semantic verb classes tend to appear in which verb phrase structure;
        - The context provided by semantic classes of embedding verbs for particular verb phrase structures, i.e., which verb phrase structures tend to be embedded by which syntactic embedding classes;
    - Linguistic complexity as measured by:
        - Occurrences of left-, right-, and equally-heavy clauses;
        - Occurrences of left-, right-, and equally-embedded clauses;
        - The mean and the standard deviation of the depths of the left branches of sentences;
        - The mean and the standard deviation of the depths of the right branches of sentences;
        - The mean and the standard deviation of the number of clauses embedded in the left branches of sentences;

- The mean and the standard deviation of the number of clauses embedded in the right branches of sentences;
- Number of prepositional phrases in sentences;
- The mean and the standard deviation of the depths of the heaviest prepositional phrases in sentences;
- The mean and the standard deviation of the depths of (if any) sentence-initial subordinating clauses in sentences;

- GI Categories;

2. Function words;
3. Word lengths;
4. Sentence lengths;
5. Surface, syntactic, and semantic features used for the preliminary experiments presented in Chapter 5; and
6. Tfidf-weighted keywords.

Our corpus includes books of various lengths, measured in terms of the number of words, sentences, pages, etc. Each of the natural discourse units of these texts, e.g., paragraphs and chapters, also differ in their length. For example, the lengths of the chapters differ among books as well as within a book. Therefore, the raw frequencies of features in each of the chapters, or any other natural discourse unit, do not provide a good representation of this data—units that are longer are likely to contain more examples of features simply because they are longer.

In text classification literature, this problem is traditionally handled in one of two ways: breaking the texts into same size chunks [91], or normalizing the extracted features by the length of the documents [50].

Controlling for length does not usually correlate text units with natural discourse units such as sentences, paragraphs, and chapters. This method loses important discourse information. Therefore, we prefer to gather statistics at the chapter level and to normalize the obtained values for the length of each chapter. For normalization purposes, the length of each chapter is measured by the total number of structures present in the chapter; i.e., clause-level features are normalized by the number of clauses in the chapter, and sentence-level features are normalized by the number of sentences in the chapter.

## 8.3 Baseline Features

Books and titles can be distinguished (from other books and other titles respectively) based on either their content or the way they are written. To evaluate the linguistic elements of expression on book and title recognition, we used as baselines, features that capture content (i.e., distributions of tfidf-weighted keywords), as well as features that capture the way these works are written (i.e., distributions of function words, distributions of word lengths, distributions of sentence lengths, and the preliminary set of linguistic features described in Chapter 5).

Authors can be distinguished from other authors based on the way they write, independently of content. Therefore, for authorship attribution, we use as baselines only the features that capture the way authors write, i.e., distributions of function words, distributions of word lengths, distributions of sentence lengths, and the preliminary set of linguistic features described in Chapter 5.

### 8.3.1 Baseline Features

#### Tfidf-weighted Keywords

Most content-based text classification tasks use unordered sets of keywords [101] to classify documents. Keywords, for the purposes of most classification applications, are stemmed [7] nouns, verbs, adjectives, and adverbs; these words are informative of the information and the story contained in documents. Despite polysemy and ambiguity of words, i.e., the fact that some words have multiple meanings, when considered along with other words that appear in the same document, keywords capture content accurately.

The standard approach to using keywords for document classification represents documents as vectors [101]. In this representation, each document vector contains slots for all of the words in the corpus and encodes information about the words in the document by setting the values of these slots. In one example of this kind of representation, the vector encodes information about the presence/absence of each word in the document by filling the document vector with a positive bit, i.e., value 1, for each word that is present in the document and negative bit, i.e., value 0, for each word that is absent. In another example of this representation, the document vector encodes information about the frequency of words, and sets the slot value for each word to the frequency of that word in the document [78].

Studies using frequency-based vector representation of documents usually weight each of the vector entries for a document (where each entry corresponds to a word) proportionally to the fre-

quency of the word in the document and inversely proportionally to the number of documents that contain this word in the complete corpus, i.e., tfidf-weighting [102]. This weighting is based on the intuition that a word is more indicative of the content of a document if it appears frequently in that document but is rare in the corpus in general. We apply this weighting to the keyword-based vector representation of documents in our corpus.

Most keyword-based classification approaches treat proper nouns as regular keywords. In order to recognize copies of works for copyright infringement detection purposes, use of proper nouns in the document representation is inappropriate because although they may facilitate recognition of documents, proper nouns achieve this goal without capturing the semantic contents of documents—changing proper nouns in a document does not change the semantics or the story, but can fool a copy recognition system that relies on lexical overlap to identify copies. We are interested in evaluating the ability of words to capture the actual information and the story contained in documents. For this reason, we eliminate proper nouns from the vector representation of documents.

Finally, words that are very infrequent in the corpus are omitted from the vector representation, because such words are very likely misspellings and would introduce noise [83].

#### Baseline Linguistic Features

Experiments in Chapter 5 showed that features related to sentence structure such as the frequency of use of "get-passives", "be-passives", "of-genitives", and "'s-genitives" are useful for capturing the differences in the linguistic choices of translators of the same content. This observation inspired us to extract a novel set of features based on sentence structure in order to capture expression.

In the evaluation of this novel set of linguistic expression features, we go back to the features from which the inspiration was drawn and compare the abilities of the two sets of features to capture expression. The baseline features used in these comparisons include: number of words in the document, type–token ratio, the mean and standard deviation of the lengths of words in the document, the mean and standard deviation of the lengths of sentences in the document, number of sentences in the document, frequency of different sentence types (e.g., declarative, interrogative, etc.), frequency of use of alternate passives, frequency of use of active voice, frequency of use of alternate genitives, and frequencies of use of negations. All frequencies are normalized for document length, measured in terms of the number of sentences, clauses, or words, as appropriate for different features.

### Function Words

Function words include pronouns, prepositions, determiners, and auxiliary verbs that are mostly independent of content and can capture differences in the way authors write.

In studies of authorship attribution, many researchers have taken advantage of the differences in the way authors use function words. Mosteller and Wallace [85] studied the use of 70 function words for discriminating between Madison and Hamilton in order to identify the authors of 12 Federalist papers with disputed authorship. Other statisticians followed the example of Mosteller and Wallace, and many used the same 70 function words for studying stylistic differences among authors, e.g., Peng [91]. In our studies, we used the set of 363 function words from which Mosteller and Wallace's 70 function words were selected. These 363 function words were gathered by Miller, Newman and Friedman [81] and came from the King James Bible, from William James, and from The Atlantic (1957). We augmented this list with 143 function words, for a total of 506, that are frequently used in modern English, e.g., `until`, and were absent from the list provided by Miller, et al.

In our experiments, as with the rest of the frequency features, the frequencies of function words were normalized by document length.

### Word Length Distribution

Distributions of word lengths have been used in the literature for distinguishing authors from each other. T. C. Mendenhall [80] studied the distribution of word lengths for words that range from 1-letter words to 13-letter words (with a separate category for 13-or-more-letter words) in the works of Shakespeare, Marlowe, and Bacon. He found that the distribution of word lengths in the works of Shakespeare were different from the distribution representing the works of Bacon, but these distributions did not help separate the works of Shakespeare from the works of Marlowe. Later, Williams showed that Mendenhall did not take genre into account in this study [128]. Using Sir Philip Sidney's works in addition to those of Shakespeare and Bacon, he showed that the distribution of word lengths in Sidney's prose looked more similar to the distribution of word lengths in Bacon's prose than Sidney's own verse. Similarly, the distribution of word lengths in Sidney's verse resembled more closely the distribution of word lengths in Shakespeare's verse than Sidney's own prose. "On the other hand, the pattern of difference between Shakespeare's verse and Bacon's prose is almost exactly comparable with the difference between Sidney's prose and his own verse" [128].

Word length distributions provide an easy and straightforward way of modeling the text. However, their ability to capture an author's style is questionable; they may capture genre rather than style. However, for completeness, we evaluate these features, by calculating the distributions of 1-letter to 15-or-more-letter words (with a separate category for 15-or-more-letter words) in the works of each of the authors in our corpus.

### Distribution of Sentence Lengths

Chapter 5 showed that sentence length means and standard deviations played a role in identifying translators even when they translated the same original. Work in the literature showed that distributions of sentence lengths are usually more useful for identifying authors than aggregate measures such as means and standard deviations [53]. As a baseline, we use both sets of information: the sentence length distribution statistics and their means and standard deviations.

## 8.4 Linguistic Features

The linguistic elements of expression used in this section include all of the features described in Chapters 6 and 7, normalized by the lengths of chapters, shown in Table 8.1.

| No. | Feature |
|---|---|
| 1 | Percentage of sentences with each sentence-initial phrase structure |
| 2 | Percentage of sentences with each sentence-final phrase structure |
| 3 | Mean and standard deviations of linguistic complexity (in terms of depth and embedding features) of sentences |
| 4 | Percentage of clauses that are left-embedded |
| 5 | Percentage of clauses that are right-embedded |
| 6 | Percentage of clauses that are left-heavy |
| 7 | Percentage of clauses that are right-heavy |
| 8 | Percentage of sentences that contain each semantic class |
| 9 | Percentage of sentences that contain each semantic class–verb phrase structure pair |
| 10 | Percentage of sentences that contain each embedding class |
| 11 | Percentage of sentences that contain each embedding class–semantic class pair |
| 12 | Percentage of sentences that contain each embedding class–verb phrase structure pair |
| 13 | Percentage of each of the GI categories in the text |

Table 8.1: Linguistic elements of expression, normalized for chapter length.

The instances of these features that appear in our corpus, e.g., 62 GI categories, 29 syntactic formulae, combinations of syntactic formulae with semantic verb classes, etc., generate more than 1,400 features.

As with our previous experiments, we used decision trees to model our data. The computational cost associated with building decision trees increases with the number of features used. Therefore, to control the computational complexity, similar to the methodology used in Chapter 7, we limited our models only to those features that had non-zero information gain value on the training corpora.

## 8.5 Classification Experiments

To evaluate our features, we compared the linguistic elements of expression with the baseline features on three separate tasks: recognizing titles even when they are paraphrased, recognizing books even when some of them are derived from the same original, and recognizing authors.

### 8.5.1 Recognizing Paraphrases (Recognizing Titles)

To test the linguistic elements of expression on recognizing paraphrases of works, we used corpora that consisted of all the titles in our corpus that contained at least approximately 40 chapters (shown in Table 8.2). In order to create a balanced data set, we randomly selected approximately 40 chapters from each of these titles. For titles without paraphrases, we set aside 40% of the chapters (around 20 chapters) for testing. We used the remaining 60% of the chapters from each title (around 25 chapters) for training. For paraphrased titles, we used chapters from one of the paraphrases for training and chapters from the remaining paraphrases for testing.

Parameter tuning on the training corpus before the classification experiments showed that the performance of all models stabilized at around 200 rounds of boosting, after which further boosting did not improve performance. In addition, in order to control the computational complexity of the models and to eliminate noisy features, after parameter tuning we limited our models to features that had non-zero information gain on the training set.

| Title | Author |
|---|---|
| *20000 Leagues under the Sea*<br>*In Search of the Castaway*<br>*Journey into the Interior of the Earth*<br>*The Mysterious Island* | Verne |
| *A Connecticut Yankee in King Arthur's Court*<br>*Following the Equator: A Journey Around the World*<br>*The Gilded Age: A Tale of Today* | Twain |
| *A House of Gentlefolk*<br>*The Torrents of Spring* | Turgenev |
| *A Laodicean: A Story of To-Day*<br>*Far from the Madding Crowd*<br>*Jude the Obscure*<br>*Tess of the d'Urbervilles: A Pure Woman*<br>*The Mayor of Casterbridge*<br>*The Hand of Ethelberta: A Comedy in Chapters* | Hardy |
| *A Tale of Two Cities*<br>*David Copperfield*<br>*Oliver Twist*<br>*The Life and Adventures of Nicholas Nickleby*<br>*The Old Curiosity Shop*<br>*The Pickwick Papers* | Dickens |
| *Adam Bede*<br>*Daniel Deronda*<br>*Middlemarch*<br>*The Mill on the Floss* | Eliot |
| *Anna Karenina*<br>*Resurrection*<br>*The Kreutzer Sonata* | Tolstoy |
| *Chronicle of the Conquest of Granada*<br>*Knickerbockers History of New York*<br>*Life and Voyages of Christopher Columbus Vol. II* | Irving |
| *The History of a Crime*<br>*The Hunchback of Notre Dame*<br>*The Man Who Laughs* | Hugo |
| *Madame Bovary* | Flaubert |
| *Emma*<br>*Mansfield Park*<br>*Pride and Prejudice*<br>*Sense and Sensibility* | Austen |
| *The Firm of Girdlestone* | Doyle |
| *The History of Pendennis*<br>*The Newcomes: Memoirs of a Most Respectable Family*<br>*The Virginians: A Tale of the Eighteenth Century* | Thackeray |
| *The Idiot* | Dostoyevski |
| *The Mutiny of the Elsinore* | London |

Table 8.2: Corpus for experiments on recognizing titles (even when they are paraphrased) and books (expression).

Models built with linguistic expression features accurately recognized titles (even when some titles were paraphrased) 81% of the time and significantly outperformed the baseline models built with function words, tfidf-weighted keywords, word length distributions, sentence length distributions, and the baseline linguistic features (see Table 8.3).[2]

| **Feature Set** | **Run 1** | **Run 2** | **Run 3** | **Avg. Accuracy** |
|---|---|---|---|---|
| Linguistic Features w/ GI | 81% | 81% | 80% | 81% |
| Linguistic Features w/o GI | 73% | 74% | 73% | 73% |
| Function Words | 52% | 52% | 54% | 53% |
| Tfidf-weighted Keywords | 49% | 48% | 45% | 47% |
| Baseline Linguistic | 40% | 40% | 40% | 40% |
| Distribution of Word Length | 16% | 21% | 17% | 18% |
| Distribution of Sentence Length | 13% | 13% | 11% | 12% |

Table 8.3: Classification results on the complete test set for recognizing titles even when some titles are paraphrased. Train on 32 chapters from each title and test on 22 chapters.

Some of the most useful linguistic features for recognizing titles related to the standard deviations of the top-level constituents of sentences and GI categories. Eliminating the GI categories, i.e., the semantic elements of expression, from the feature set reduced the accuracy; when we reran the classification experiment without GI categories, we lost 8% from the performance. However, the syntactic elements of expression alone identified titles accurately 73% of the time and still significantly outperformed the models built with all baseline features. In this case, sentence-initial and -final phrase structures, as well as embedding constructs (e.g., sentential) and verb argument structures (0-V-NP) were among the most useful features identified by information gain.

[2]For this corpus, a difference of 4% or more is statistically significant with $\alpha = 0.05$.

| Weighted keywords | Function Words | Linguistic Features w/ GI Categories |
|---|---|---|
| sister | the | Std. dev. of the depths of the top-level left branches (measured in phrase depth) |
| reply | of | Std. dev. of the depths of the top-level right branches (measured in phrase depth) |
| captain | 'll | Std. dev. of the depths of the deepest prepositional phrases of sentences (measured in phrase depth) |
| lady | she | Percentage of words in a chapter that belong to GI category "cognitive orientation" |
| dear | her | Percentage of words that belong to GI category "adjectives referring to people" |
| uncle | and | Percentage of words that belong to GI category "Emotion" |
| miss | 's | Percentage of sentences that contain unembedded verb phrases |
| gentleman | upon | Percentage of words that belong to GI category "Role" |
| aunt | 've | Percentage of words that belong to GI category "human" |
| till | 'm | Percentage of sentences that contain an unembedded verb with noun phrase direct object (0-V-NP) |

Table 8.4: Some of the most useful features for recognizing titles even when some titles are paraphrased.

| **Linguistic Features w/o GI Categories** |
|---|
| Standard deviation of the depths of the top-level left branches (measured in phrase depth) |
| Standard deviation of the depths of the top-level right branches (measured in phrase depth) |
| Standard deviation of the depths of the deepest prepositional phrases in sentences (measured in phrase depth) |
| Percentage of sentences that contain an unembedded verb with noun phrase object (0-V-NP) |
| Percentage of sentences that contain a sentence-final adverb phrase |
| Average depth of the subordinating clauses at the beginning of sentences (measured in phrase depth) |
| Percentage of sentences that contain unembedded verbs |
| Percentage of sentences that contain unembedded intransitive verb phrases (0-V) |
| Standard deviation of the depths of sentence-initial subordinating clauses |
| Average number of embedded clauses in the top-level right branch |

Table 8.5: Top ten syntactic elements of expression that recognize titles—in absence of GI categories.

| Feature Set | Run 1 | Run 2 | Run 3 | Avg. Accuracy |
|---|---|---|---|---|
| Linguistic Features w/ GI | 94% | 95% | 95% | 95% |
| Linguistic Features w/o GI | 95% | 95% | 95% | 95% |
| Baseline Linguistic | 67% | 68% | 65% | 67% |
| Distribution of Word Length | 45% | 62% | 54% | 54% |
| Tfidf-weighted Keywords | 45% | 36% | 32% | 38% |
| Function Words | 29% | 38% | 35% | 34% |
| Distribution of Sentence Length | 23% | 17% | 10% | 17% |

Table 8.6: Classification results only on the paraphrased titles included in the 45-title corpus. Random chance would recognize a paraphrased title 2% of the time.

We analyzed the results of this experiment in more detail in Table 8.6. These results indicated that:

- Linguistic expression features recognize on average 95% of the paraphrased titles accurately;
- Removing the GI categories from the feature set does not change this result;
- Baseline linguistic features identify on average 67% of the paraphrased titles accurately; and
- Distribution of word lengths provide an accuracy of 54% on this task and outperform tfidf-weighted keywords as well as function words (See Table 8.6).

That linguistic features recognize paraphrased titles (with or without GI categories) indicates that some of our linguistic elements of expression are common to books that are derived from the same source. While our features and experiments do not reveal the reason behind the expressive similarities of these paraphrases, these similarities could be due to their common content (which implies that semantics dictate syntax and works on the same content exhibit similarities in the expression of content also), or due to the underlying expression of the original author (which the translators reflect in their translations, or which bleeds through the expression of the translators).

Because of their content similarity, we expect paraphrases to exhibit certain level of similarity with respect to keywords. The results in Table 8.6 indicate that keywords can indeed capture some of the similarity between these books.

Function words are devoid of any content information; however, they also recognize paraphrased titles reasonably well. This may be an indication that some of the similarities between the paraphrases are due to the residual expression of the original author.

### 8.5.2 Recognizing Expression (Recognizing Books)

To test the linguistic elements of expression on recognizing expression in individual books, we used corpora that consisted of all the books in our corpus that contained at least approximately 40 chapters. This corpus differed from the corpus used in Section 8.5.1 in three ways: it did not contain translations of *The Kreutzer Sonata* because each of these translations did not have enough chapters to contribute both to training and test sets; the translations of *Madame Bovary* were treated as separate books on the same content; the translations of *20000 Leagues under the Sea* were treated as separate books on the same content.

As before, in order to create a balanced data set, we randomly selected approximately 40 chapters from each of these books. For each book, we set aside 40% of the chapters (around 20 chapters) for testing. We used the remaining 60% of the chapters from each book (around 25 chapters) for training.

Parameter tuning on the training corpus before the classification experiments showed that the performance of all models stabilized at around 200 rounds of boosting, after which further boosting did not improve performance. In addition, in order to control the computational complexity of the models and to eliminate noisy features, as before, after parameter tuning we limited our models to features that had non-zero information gain on the training set.

Table 8.7 shows that, in this experiment, the linguistic expression features gave an accuracy of 82% when recognizing the expression unique to each book, even when some books are based on the same content. For this experiment also, GI categories dominated the top ten features—five out of the top ten most useful features identified by information gain were GI categories. Removing the GI categories from the feature set resulted in a loss of 6% from the accuracy. In either case, the expression features significantly outperformed all baseline features. On this corpus, a performance difference of 4% or more is statistically significant at $\alpha = 0.05$.

| Feature Set | Accuracy |
|---|---|
| Linguistic Features w/ GI | 82% |
| Linguistic Features w/o GI | 76% |
| Tfidf-weighted Keywords | 66% |
| Function Words | 61% |
| Baseline Linguistic | 42% |
| Word Length | 29% |
| Sentence Length | 13% |

Table 8.7: Classification results on the test set for expression recognition even when some books contain similar content.

| Linguistic Features with GI | Linguistic Features without GI |
|---|---|
| Standard deviation of the depths of the top-level left branches (measured in phrase depth) | Standard deviation of the depths of the top-level left branches (measured in phrase depth) |
| Standard deviation of the depths of the top-level right branches (measured in phrase depth) | Standard deviation of the depths of the top-level right branches (measured in phrase depth) |
| Percentage of words that belong to GI category "human" | Standard deviation of the depths of the deepest prepositional phrases of sentences (measured in phrase depth) |
| Percentage of words that belong to GI category "adjectives referring to people" | Percentage of sentences that contain unembedded verbs |
| Standard deviation of the depths of the deepest prepositional phrases (measured in phrase depth) in each sentence | Percentage of sentences that contain an unembedded verb with noun phrase object (0-V-NP) |
| Percentage of words that belong to GI category "Role" | Average depth of the subordinating clauses at the beginning of sentences (measured in phrase depth) |
| Percentage of words that belong to GI category "Emotion" | Percentage of sentences that contain equal numbers of clauses in left and right branch |
| Percentage of sentences that contain unembedded verbs | Percentage of sentences that contain a sentence-final adverb phrase |
| Percentage of words that belong to GI category "cognitive orientation" | Percentage of sentences that contain unembedded intransitive verb phrase (0-V) |
| Percentage of sentences that contain an unembedded verb with noun phrase object (0-V-NP) | Percentage of sentences that contain more clauses on the right branch than on the left |

Table 8.8: Top ten linguistic expression features that recognize books even when some books share content.

The GI categories appear among the top ten features in the experiments related to recognition of books and titles. Intuitively, we expect that GI categories capture content more than expression—they only capture high-level semantics. The ability of these features to recognize titles is expected: if the content of the titles are disjoint, then the GI categories will recognize the titles based on this difference.

Our results show that, without any contribution from GI categories, the syntactic elements of expression give approximately 76% accuracy when recognizing books and gain only 7% from the addition of GI categories to the feature set. These results indicate that, although GI categories appear among the top ten most useful features (as identified by information gain), the syntactic elements of expression also contribute significantly to recognition of books.

The significant power of syntactic elements of expression in recognizing books implies that we can recognize books from the way they are written. In other words, the high-level observations about syntactic and structural characteristics of sentences help differentiate books from each other.[3]

| **Feature Set** | **Accuracy** |
|---|---|
| Linguistic Features w/ GI | 92% |
| Linguistic Features w/o GI | 89% |
| Tfidf-weighted Keywords | 88% |
| Function Words | 81% |
| Word Length | 72% |
| Baseline Linguistic | 53% |
| Sentence length | 14% |

Table 8.9: Expression recognition results on the test set for paraphrased books only.

Studying the results of this experiment in more detail, with particular focus on the recognition results on the test set of only paraphrased books, we found that (see Table 8.9).

- Liguistic elements of expression recognize translations of the same content accurately 92% of the time and the remaining 8% of the time, they confuse the translations of the same title with each other, i.e., chapters from one translation of Madame Bovary get misclassified as chapters from another translation of Madame Bovary;

[3]The only semantic information included in our feature set, in addition to GI categories, is the semantic classes of verbs which are more accurately described as syntacto-semantic classes of verbs because they represent verbs with similar semantics that exhibit similar syntactic behavior.

- Removing GI categories from the feature does not change the results significantly and all of the classification errors on translations still occur due to confusions between translations of the same title;
- Tfidf-weighted keywords recognize translations accurately 88% of the time; only 10% of the mistakes are due to confusion between different translations of the same title;
- Function words recognize translations accurately 81% of the time. 17% of the time, the classification errors are due to confusions between translations of the same title; and
- The performance of baseline linguistic features on this task is much worse than their performance on the task of recognizing titles (even when some titles are paraphrased). While these features are more useful than tfidf-weighted keywords and function words in recognizing the similarities between paraphrases of titles, they fail to capture the expressive differences between individual paraphrases.

That linguistic features can differentiate between translations of the same title indicates that translators add their own expression to works, even when they write about the same content, and even when their works are derived from the same original. However, that these translations cannot be separated from each other 100% of the time indicates that some linguistic similarities remain in these translations, despite the addition of translators' expression. While some other features may be able to capture more of the linguistic difference between translations and improve performance, the observed linguistic similarities could also be due to the residual expression (as defined in terms of the features described in this thesis) of the original author of the title, or the limitations in the syntactic choices available for expression of semantically equivalent content.

Despite not capturing any linguistic information, keywords and function words recognize translations of the same title reasonably well. However, because models built with these features are based simply on the word frequency and overlap, these features are more likely to misclassify some chapters as parts of an unrelated title simply because the same words are distributed similarly (even when used in different senses) in these works; whereas the underlying syntactic structure would have revealed the original title that the translation is based on.

### 8.5.3 Recognizing Authors

In Chapter 4, we described the differences in style and expression. These concepts, though different, both relate to the way people write. Then, an interesting question to answer is: Can the same set of linguistic features help recognize both expression and style?

In order to answer this question, we ran classification experiments on all of the titles in our corpus that were written by native speakers of English. To test the ability of different sets of features to capture style, we trained models on a subset of the titles by eight authors and tested on a different subset of titles by the same authors. We repeated this experiment five times so that several different sets of titles were trained and tested on. At each iteration, we used 150 chapters from each of the authors for training and 40 chapters from each of the authors for testing. The titles used in this experiment are shown in Tables A-1 and A-2. The column labelled "Status" indicates whether a title was used for training or testing at each iteration.

Parameter tuning, as before, was performed on the training set. The threshold value of zero information gain was selected for eliminating excess features from the feature space, and boosting was limited to 200 rounds. Keeping these threshold values constant, across different feature sets, we obtained the following classification results on the test corpora:

| Feature Set | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 | Avg. Accuracy |
|---|---|---|---|---|---|---|
| Function Words | 86% | 89% | 87% | 90% | 81% | 87% |
| Linguistic Features w/ GI | 73% | 77% | 62% | 76% | 65% | 71% |
| Linguistic Features w/o GI | 64% | 63% | 64% | 55% | 62% | 62% |
| Distribution of Word Length | 33% | 37% | 44% | 53% | 35% | 40% |
| Baseline Linguistic | 39% | 39% | 41% | 48% | 28% | 39% |
| Distribution of Sentence Length | 33% | 41% | 31% | 41% | 25% | 34% |

Table 8.10: Results for authorship attribution. Classifier is trained on 150 chapters from each author, and tested on 40 chapters from each author. The chapters in the training and test sets come from different titles.

The results in Table 8.10 show that high level features such as function words capture the style of authors better than any other features. Top ten most predictive function words are: `the`, `not`, `of`, `she`, `very`, `be`, `her`, `'s`, `and`, and `it`.

Among the linguistic expression features, GI categories again play a role in determining authorship: authors seem to repeat concepts and themes which make these features useful for authorship attribution. Removing the GI categories and rerunning the experiments reduces the performance of the linguistic features on average by 9%.

The results indicate that linguistic expression features are not as effective as function words in capturing the style of authors. This finding is consistent with our intuition: we selected the linguistic elements of expression for their ability to differentiate between different books and titles, even when some titles are written by the same author. Recognizing the style of an author requires focus on the

elements that are similar in the titles written by the same author, instead of focus on elements that differentiate these titles.

However, the linguistic elements of expression are not completely devoid of any style information. In Chapter 6, we indicated that some elements of expression, such as distributions of left- and right-heavy sentences, despite successfully differentiating between translations of the same title by recognizing the expressive differences between them, were not always successful in differentiating between the different works of an author, i.e., that they captured style also. Presence of such features among the linguistic elements of expression enabled authorship attribution with 71% accuracy despite the focus of the general feature set on expression rather than style.

## 8.6 Conclusion

Linguistic elements of expression, described in Chapters 6 and 7, recognize books and titles more accurately than any of the baseline features. These features can recognize titles even when they are paraphrased, enabling recognition of derivatives. They also recognize different expressions of the same content, and enable recognition of independently copyrighted derivatives of the same title, e.g., different derivatives of a public domain work. Because of their focus on identifying individual books, even when they share content, and even when they are written by the same person, these features are not as successful in capturing style of authors—style can be better captured by features that are used similarly in different works of an author and that would not be able to differentiate between the author's works. Function words provide one such feature set; they recognize an author's style more accurately than any of the other feature sets.

## 8.7 Summary

In this chapter, we compared the ability of different sets of features to classify books, titles, and authors. Through classification experiments on a corpus of novels, we showed that linguistic expression features significantly outperform all baseline features in recognition of individual books and titles. However, function words give the best performance when recognizing authors.

# Chapter 9

# Conclusion

Technological developments that led to development of digital media changed the way publishing and entertainment companies work by enabling these companies to reduce some of the costs associated with production and distribution of digital works. This cost reduction promises to promote progress by enabling copyright holders to reach wider audiences, and by enabling users to gain acess to more and varied works at low cost. However, this cost reduction also enables users to make and distribute unauthorized copies of digital works, at the expense of copyright holders: copyright holders rely on control over distribution of works in order to raise revenues; ease of distribution of digital works by anyone eliminates this control mechanism. The result threatens the revenues and the existing (royalty-based) businesses of copyright holders. To promote progress in the digital world, copyright holders should have incentives to produce and disseminiate their works digitally, and the users should have access to these works for fair use, for free speech, and without threats to their privacy.

The legal and technological protections adopted in response to the digital world have failed to re-establish a balance between the rights of copyright holders and the rights (and privileges) of the public. These protections expanded the rights of copyright holders at the expense of interests of the public, and have raised issues regarding fair use, privacy, first sale, and freedom of speech.

As a result of acts of civil disobedience against overreaching uses of legal and technological protections, many technical protection mechanisms are circumvented, and thus fail to achieve the main goal for which they were designed—promotion of progress through protection of copyrights to encourage authors to produce and disseminate works for public to use.

To achieve the goal of copyrights, to promote progress, and to encourage learning, the incentives of copyright holders need to be balanced with the ability of the public to access and use works. Some proposals that can achieve such a balance eliminate infringement claims related to use, modification, distribution, and many other personal uses of digital works, arguing for free flow of digital works within society in return for fair compensation of copyright holders proportional to the value of their works for society. Fisher and Netanel argued that the compensation can come from taxes imposed on digital equipment and media that facilitate copying, use, and distribution of digital works [39, 87].

These solutions are promising for promotion of progress because they do not limit use of works, and they also succeed in protecting the revenues of copyright holders in a world where control over distribution is no longer pragmatically possible. However, successful implementation of these solutions requires mechanisms for measuring the value of works for society. Netanel argues that popularity of works, as measured by digital tracking mechanisms that can identify even modified copies, and that can meter use and distribution, could provide a good proxy for this value [87].

In this thesis, we described a set of syntactic and semantic features that can capture expressive fingerprints of works. These fingerprints recognize both verbatim and paraphrased copies of works, without relying on labeling information such as watermarks [92]. These fingerprints cannot be removed or altered because they are generated from the works themselves on demand. Thus, they make it possible to implement proposals that rely on accurate metering of use of works, even when works are paraphrased and even when watermarks and other identifying information are removed.

Our approach to fingerprinting literary works relies on linguistically grounded features in order to capture expression, and uses natural language processing techniques to recognize novels and their paraphrases. This method analyzes expression in terms of the syntactic choices of authors when writing about a particular content, and high-level semantic categories that capture the global context in which these syntactic choices are made.

Evaluation of these features and comparisons with standard baselines from the text classification literature showed that, even in the presence of paraphrases, linguistic elements of expression identify titles correctly more than 80% of the time, significantly outperforming all standard baselines, including keywords, function words, sentence lengths, and word lengths. These results confirm our initial findings in Chapter 5 that indicate that expression of content in literary works can be captured through analysis of syntax. Indeed, in particular, verb phrase structure and observations about high-level syntactic patterns that reveal deeper syntactic phenomena can provide much information about the manner of expression of content. In addition, a small number of high-level semantic classes are

adequate for capturing the content surrounding these syntactic elements of expression.

Given a mechanism that can recognize literary works from their linguistic elements of expression, these works can be identified accurately even when they are paraphrased. Digital tracking based on such a technology provides an accurate measurement of the value of works for society, by also identifying the modified versions that are used and distributed digitally. Compensation mechanisms that build on the information gathered by these tracking mechanisms may enable implementation of various policy solutions, e.g., taxation-based solutions.

# Chapter 10

# Contributions

The main contribution of this thesis is a novel set of linguistic features that capture the expression of content in literary works, particularly novels, and that enable recognition of these works and their copies even when they are modified. The approach taken to achieve this goal is linguistically informed.

The linguistic features described in this thesis have been captured through low cost syntactic analysis. Only part of speech tagged corpora have been used, and all features have been extracted based on analysis of the patterns of these tags. Despite the tradeoff between accuracy and computational cost, the extracted features were very useful for our purposes. The results demonstrate that extraction of linguistic information for text classification does not need to be computationally prohibitively expensive.

The fingerprinting system based on these linguisitc features contributes to the solution of the digital copyright problem by enabling recognition of works from the expression of their content. Some of the promising proposals to the digital copyright problem can benefit from (and others rely on) digital tracking systems that would recognize works even when they are paraphrased. The technology described here is a proof of concept that facilitates implementation of these policy proposals. In Chapter 8, we have shown that we can recognize paraphrases of works accurately by considering linguistic elements of expression and that this approach outperforms all baseline approaches. In addition to recognizing paraphrases, the fingerprints described in this thesis capture expression of content and enable digital tracking based on the copyrightable aspects, i.e., expression, of works.

Syntactic elements of expression based on the structure of sentences and phrases play a significant role in identifying titles and their paraphrases. These features are found in similar distributional

patterns throughout a work and in different distributional patterns in different works. Some of these syntactic elements of expression remain common to translations derived from the same original, even when translators add their own expression. However, enough of the expressive elements added by translators can be recognized and enable identification of each of the translations.

Semantic elements of expression, defined by a small number of broad, semantic categories capture the content surrounding the syntactic elements of expression adequately. A set of 62 broad semantic categories is adequate to capture the content for our purposes and does not require the 11000-dimensional feature space that would be provided by keywords. These high level semantic categories recognize paraphrases of the same content more accurately than keywords.

# Chapter 11

# Future Work

There are two broad classes of future work that extend this thesis: implementation of a digital tracking system based on the expression fingerprints and use of linguistic information for other applications.

## 11.1 Implementation of Digital Tracking with Expression Fingerprints

Expression fingerprints enable implementation of policy proposals that rely on accurate identification of intellectual property. In this thesis, we focused on developing these expression fingerprints and did not attempt to address questions related to their use in implementation of a digital tracking system. In order to go from an expression fingerprinting system to a digital tracking system, some questions need to be answered. For example, in order to be able to track use, would the fingerprinting system have to reside on each individual machine? Would that imply that individual machines have to be trusted? Could the fingerprinting system be an integral part of peer-to-peer distribution systems? How, when, and where should the fingerprinting be done in order to track files which are transmitted and stored encrypted? What policy changes are necessary for the fingerprinting system to be adopted widely? Answers to these and similar questions can be determined by deploying the fingerprinting system under real life situations at least within a small sample population.

Our expressive fingerprints highlight the expressive differences between novels. In order to develop these fingerprints, we collected and studied a corpus (a repository) of documents. The fingerprints and the discriminative models generated based on the expression of these documents can be used to determine whether any document (within this repository, or later presented to the system) is a copy of any of the documents in the repository. The discriminative models that make

classification decisions do not have to be generated on the fly; for any repository, the models have to be created once. The models have to be re-created only when new documents are added to the repository.

The repository and the discriminative models require disk space that holds all the original works, and identification of the linguistic elements of documents requires some computation time. In order to minimize the computational cost, in this thesis, we chose to study linguistic features captured through high-level, low-cost analysis, sometimes at the expense of accuracy of identification of the features, e.g., features related to sentence depth. Experiments with improved features and the changes in the computation time they require would reveal whether higher accuracy in identification of features translates into higher accuracy in identification of copyrighted works, and whether the changes in computational cost justify the gain.

Our fingerprinting mechanism may be improved in the following ways:

- The fingerprints presented in this thesis are useful for differentiating between novels and have not been applied to other genres. Study of a variety of genres would reveal a richer set of features that capture expressive differences in other genres.

- Measurement of complexity of sentences can be modified to consider relative clauses and movement of noun phrases in more detail.

- A low-cost constituent parser can be developed to improve the parse of arbitrarily long sentences in novels, and to improve measurement of depth and embeddedness in these sentences even in the presence of complex punctuation, embedded quotes, etc., which we currently do not attempt to recognize.

- Disambiguation of syntacto-semantic categories of verbs can be improved by resolving ambiguity in the case of verbs that belong to multiple semantic classes and that appear in multiple syntactic alternations. In this thesis, we disambiguated broad semantic categories used for capturing content; however, we did not attempt to disambiguate the syntacto-semantic classes of verbs. Syntacto-semantic classes do not imply synonymy and need to be treated differently from semantic categories of words. Many verbs appear in several syntacto-semantic classes, and they may be observed in different argument structures and alternations in each case. This many-to-many matching complicates the identification of the correct syntacto-semantic classes of verbs. A corpus study on the frequency of use of particular syntacto-semantic

classes with particular syntactic alternations and argument structures provide a good starting point for addressing this problem.

- Our disambiguation method for broad semantic categories of words can be evaluated more thoroughly on a sense-tagged corpus, and result may indicate improvements that can be made to the fingerprinting algorithm.

These improvements could increase computation time, and the gain obtained from them would have to be evaluated with this computational cost in mind. In addition, before full-scale deployment, the scalability of the system, in terms of both computation time and classification accuracy, should be studied.

## 11.2 Study of Linguistic Features for Other Applications

In addition to author and expression recognition, there is a rich set of problems that can benefit from statistical language modelling and natural language processing using low-cost syntactic analysis. Examples include text similarity evaluation in the context of tasks such as information retrieval, opinion classification, and a variety of educational applications such as essay evaluation.

Information retrieval lies at the core of search engines, most of which rely on verbatim keyword and phrase matches to identify and rank documents that are relevant to a query. There are many dimensions along which the returned documents can be organized to make information access much easier. For example, each document can be linked to a group of documents that are similar to it in content and/or expression, i.e., paraphrased documents, plagiarized documents, or documents that make the same point in similar ways. Similarly, documents that are at the same level of linguistic complexity can be grouped together, making it easier for less sophisticated readers, such as elementary school students, to get easy access to appropriate content. These problems are natural extensions of the problem addressed in this thesis: the semantic and syntactic elements of language can be extended to group documents based on their similarity in content and expression, addressing the issues of recognition of paraphrases, plagiarized documents, and documents that express the same content in similar ways; similarly, the syntactic elements of language devised to measure sentence complexity can be used to go beyond lexical cues in classifying documents based on their level of linguistic complexity.

Another natural extension of the work presented in this thesis is classification of documents with respect to opinions. Since part of the meaning in a document lies in how things are said, in addition to what is being said, the analysis of syntactic and semantic elements of language can be extended to improve opinion classification beyond what is possible through analysis of keywords. Improved accuracy on this ask can benefit many applications such as "computational politics" (a new research area emerging from the collaboration of computer scientists with political scientists) and electronic rulemaking for e-goverment (the use of the Web to increase public access to government agencies) that rely on quick and correct evaluation of the polarity of opinions and of reactions to news and proposed e-rules.

Statistical language modelling using different modes of linguistic information also has a diverse set of pedagogical applications such as essay evaluation. Findings and methodologies of this thesis can be applied to essay evaluation and would improve the state of the art in these domains by

considering expressive aspects of documents in addition to content: we can model the language behind good writing practices by analyzing syntax and measuring the coherence of text through syntactic and semantic language models, and we can provide students with feedback on different linguistic aspects of their writing.

# Bibliography


[1] D. Alexander and W. J. Kunz. *Some Classes of Verbs in English*. Linguistics Research Project. Indiana University. June 1964.

[2] American Association of Publishers. Anti-Piracy Program. [web page] 2004. URL: http://www.publishers.org/antipiracy/index.cfm. [Accessed 31 January, 2005].

[3] American Library Association. Copyright Issues. [web page] 2004. URL: http://www.ala.org/ala/washoff/WOissues/copyrightb/copyright.htm. [Accessed 31 January, 2005].

[4] American Library Association. DRM: Statement of Library and Higher Education Concerns. [web page] 2003. URL: http://www.ala.org/ala/washoff/WOissues/copyrightb/digitalrights/digitalrightsmanagement.htm.

[5] R. Anderson. "Trusted Computing" Frequently Asked Questions. [web page] August 2003. URL: http://www.cl.cam.ac.uk/users/rja14/tcpa-faq.html. [Accessed 31 January, 2005].

[6] Association of Research Libraries. Fair Use In The Electronic Age: Serving The Public Interest. [web page] 2001. URL: http://www.arl.org/info/frn/copy/fairuse.html. [Accessed 31 January, 2005].

[7] R. H. Baayen, R. Piepenbrock, and H. van Rijn. *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1993.

[8] C. L. Baker. *English Syntax*. Second Edition. ISBN 0-262-52198-9. Cambridge, MA: MIT Press. 1995.

[9] J. C. Baker. A Test of Authorship Based on the Rate at which New Words Enter an Author's Text. *Journal of the Association for Literary and Linguistic Computing*, 3(1), 36–39, 1988.

[10] S. Bechtold. The Present and Future of Digital Rights Management—Musings on Emerging Legal Problems. In *Digital Rights Management*, E. Becker et al. (eds.), LNCS 2770, 597–654, 2003. Springer-Verlag Berlin Heidelberg.

[11] Benefit Authors without Limiting Advancement or Net Consumer Expectations Act of 2003. H.R. 1066. 108th Congress. 2003.

[12] D. Biber. A Typology of English Texts. *Language*, 27, 3–43. 1989.

[13] D. Biber, S. Conrad, and R. Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press. 1998.

[14] P. Biddle, P. England, M. Peinado, and B. Willman. The Darknet and the Future of Content Protection. In *Digital Rights Management*, E. Becker et al. (eds.) LNCS 2770, 344–363, 2003. Springer-Verlag Berlin Heidelberg.

[15] L. I. Bridgeman, D. Dillinger, C. Higgins, P. D. Seaman, and F. A. Shank. *More Verb Classes in English*. Linguistics Research Project. Indiana Univeristy. August 1965.

[16] E. Brill. A Simple Rule-Based Part of Speech Tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, 1992.

[17] C. S. Brinegar. Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship. *Journal of the American Statistical Association*, 58, 85–96. 1963.

[18] C. Buckley, A. Singhal, M. Mitra, and G. Salton. New retrieval approaches using SMART: TREC 4. *TREC*. 1995.

[19] D. Burk and J. Cohen. Fair Use Infrastructure for Copyright Management Systems. *15 Harvard Journal of Law and Technology*, 1, 41, 2001.

[20] D. Chaum. Untraceable electronic mail, return address and digital pseudonyms. *Communications of the ACM*, 24(2), 84–88, 1981.

[21] Consumer Broadband and Digital Television Promotion Act. S. 2048. 2002.

[22] Creative Commons. [web page]. URL: http://creativecommons.org. [Accessed 31 January, 2005].

[23] J. Cohen. A Right to Read Anonymously. A Closer Look at "Copyright Management" in Cyberspace. *28 Connecticut Law Review*, 981, 1996.

[24] J. Cohen. Copyright and the Jurisprudence of Self-Help. *13 Berkeley Technology Law Journal*, 1089, 1998.

[25] J. Cohen. Some Reflections on Copyright Management Systems and Laws Designed to Protect Them. *12 Berkeley Techology Law Journal*, 161. 1997.

[26] J. Cohen. WIPO Treaty Implementation in the United States: Will Fair Use Survive? *21 European Intellectual Property Review*, 236. 1999.

[27] J. Cohen. Overcoming Property: (Does Copyright Trump Privacy?) *Journal of Law, Technology and Policy*, 375. 2002.

[28] J. Cohen. DRM and Privacy. *18 Berkeley Techology Law Journal*. 2003.

[29] M. Collins. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. Dissertation, University of Pennsylvania. 1999.

[30] Consumers, Schools and Libraries Digital Rights Management Awareness Act of 2003. S. 1621, 108th Congress. 2003.

[31] M. Diab, J. Schuster, and P. Bock. A Preliminary Statistical Investigation into the Impact of an N-Gram Analysis Approach based on Word Syntactic Categories toward Text Author Classification. In *Proceedings of Sixth International Conference on Artificial Intelligence Applications*, 1998.

[32] Digital Millennium Copyright Act. 17 U.S.C. Chapter 12. 1998.

[33] Digital Media Consumers' Rights Act of 2003. H.R. 107. 108th Congress. 2003.

[34] Digital Choice and Freedom Act of 2002. H.R. 5522. 107th Congress. 2002.

[35] Descriptions of Inquirer Categories and Use of Inquirer Dictionaries. [web page] 2002. URL: http://www.wjh.harvard.edu/ inquirer/homecat.htm. [Accessed 31 January, 2005].

[36] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Second Edition. ISBN 0-471-05669-3. Wiley-Interscience. 2001.

[37] E. Ejerhed. Finding Clauses in Unrestricted Text by Finitary and Stochastic Methods. In *Proceedings of the 2nd Conference on Applied Natural Language Processing*, 219–227. 1988.

[38] Electronic Future Foundation. [web page]. URL: http://www.eff.org/. [Accessed 31 January, 2005].

[39] W. Fisher. *Promises to Keep. Technology, Law, and the Future of Entertainment.* Stanford University Press, 2004.

[40] G. Forman. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, 3, 1289–1305. 2003.

[41] M. J. Freedman and R. Morris. Tarzan: A Peer-to-Peer Anonymizing Network Layer. In *Proceedings of the CSS'02*. 2002.

[42] U. Gasser. iTunes: How Copyright, Contract, and Technology Shape the Business of Digital Media. [web page] June 2004. URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=556802. [Accessed 31 January, 2005].

[43] B. Gerovac and R. J. Solomon. Protect Revenues, Not Bits: Identify Your Intellectual Property. In *Proceedings of Technological Strategies for Protecting Intellectual Property in the Networked Multimedia Environment.* Interactive Multimedia Association.

[44] Giant Steps. DRM Features. DRM Concepts/Standards. [web page] 2003. URL: http://www.giantstepsmts.com/drmarticle.htm. [Accessed 31 January, 2005].

[45] A. Glover and G. Hirst. Detecting Stylistic Inconsistencies in Collaborative Writing. In *Sharples, Mike and van der Geest, Thea (eds.), The new writing environment: Writers at work in a world of technology*. London: Springer-Verlag. 1996.

[46] D. M. Goldschlag, M. G. Reed, and P. F. Syverson. Hiding Routing Information. In *Proceedings of Workshop on Information Hiding*, 1996.

[47] M. H. deGroot. *Probability and Statistics*. Second Edition. Addison-Wesley Publishing Company. ISBN 0-201-11366-X. 1989.

[48] M. Halliday and R. Hasan. *Cohesion in English*. London: Longman. 1976.

[49] M. Halliday. *An Introduction to Functional Grammar*. London; Baltimore, Md., USA : Edward Arnold, 1985.

[50] V. Hatzivassiloglou, J. Klavans, and E. Eskin. Detecting Similarity by Applying Learning over Indicators. *37th Annual Meeting of the ACL*, 1999.

[51] V. Hatzivassiloglou, J. Klavans, M. Holcombe, R. Barzilay, M.Y. Kan, and K.R. McKeown. SimFinder: A Flexible Clustering Tool for Summarization. *NAACL'01 Automatic Summarization Workshop*, 2001.

[52] L. Hidalgo-Downing. Negation in Discourse: A Text World Approach to Joseph Heller's Catch-22. *Language and Literature*, 9(3). 2000.

[53] D. I. Holmes. Authorship Attribution. *Computers and the Humanities*, 28, 87–106. Kluwer Academic Publishers. Netherlands. 1994.

[54] D. Hughes and V. Shmatikov. Information Hiding, Anonymity and Privacy: A Modular Approach. In *Workshop on Issues in the Theory of Security (WITS '02)*, 2002.

[55] Inducing Infringement of Copyright Act. S. 2560.

[56] Information Infrastructure Task Force. *Intellectual Property and the National Information Infrastructure: The Report of the Working Group on Intellectual Property Rights*. 1995.

[57] A. K. Jain and B. Chandrasekaran. Dimensionality and Sample Size Considerations in Pattern Recognition Practice. *Handbook of Statistics*, P.R. Krishnaiah and L.N. Kanal (eds.), 2, 835–855, North Holland, 1982.

[58] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical Pattern Recognition: A Review. *IEEE Transations on Pattern Analysis and Machine Intelligence*, 22(1), 4–37, 2000.

[59] B. Katz and B. Levin. Exploiting Lexical Regularities in Designing Natural Language Systems. In *Proceedings of the 12th International Conference on Computational Linguistics*, COLING '88. 1988.

[60] B. Katz. Using English for Indexing and Retrieving. *In Artificial Intelligence at MIT: Expanding Frontiers*. P. H. Winston and S. A. Shellard, Eds. ISBN: 0-262-23150-6. MIT Press. Cambridge, MA. 1990.

[61] E. Kelly and P. Stone. *Computer Recognition of English Word Senses*. North-Holland Publishing, 1975.

[62] D. Khmelev and F. Tweedie. Using Markov Chains for Identification of Writers. *Literary and Linguistic Computing*, 16(4), 299–307. 2001.

[63] G. Kjetsaa. *The Authorship of the Quiet Don*. ISBN 0391029487. International Specialized Book Service Inc. 1984.

[64] J. Klavans and M. Kan. Role of Verbs in Document Analysis. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, 680–688. 1998.

[65] M. Koppel, N. Akiva, and I. Dagan. A Corpus-Independent Feature Set for Style-Based Text Categorization. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[66] D. Kugler An Analysis of GNUnet and the Implications for Anonymous, Censhorship-resistant Networks. In *Proceedings of Privacy Enhancing Technologies workshop (PET 2003)*, 2003.

[67] O. V. Kukushkina, A. A. Polikarpov, and D. V. Khemelev. Using Literal and Grammatical Statistics for Authorship Attribution. Published in *Problemy Peredachi Informatsii*,37(2), April-June 2000, 96–108. Translated in "Problems of Information Transmission", 172–184.

[68] M. A. Lemley and R. A. Reese. Reducing Digital Copyright Infringement without Restricting Innovation. *56 Stanford Law Review*, June 2004.

[69] M. A. Lemley. Property, Intellectual Property, and Free Riding. *Stanford Law and Economics Olin Working Paper No. 291*. URL: http://ssrn.com/abstract=582602. [Accessed 31 January, 2005].

[70] L. Lessig. *Code and Other Laws of Cyberspace*. Basic Books, NY, 2000.

[71] L. Lessig. The Limits of Copyright. *The Industry Standard*. 2000.

[72] B. Levin. *English Verb Classes and Alternations. A Preliminary Investigation*. ISBN 0-226-47533-6. University of Chicago Press. Chicago. 1993.

[73] Limiting the Liability of Copyright Owners for Protecting their Works on Peer-to-Peer Networks. H.R. 5211. 109th Congress. 2004.

[74] J. Litman. New Copyright Paradigms. In *Growing Pains: Adapting Copyright to Libraries, Education and Society*, Laura Gassaway, ed., 63, 1997.

[75] J. Litman. Copyright Legislation and Technological Change. *68 Oregon Law Review*, 275. 1989.

[76] J. Litman. *Digital Copyright.* Prometheus Books, 2001. ISBN: 1-57392-889-5.

[77] C. Love. Courtney Love Does the Math. *Salon.com*. [web page] June 14, 2000. URL: http://dir.salon.com/tech/feature/2000/06/14/love/index.html. [Accessed 31 January, 2005].

[78] A. McCallum and K. Nigam. A Comparison of Event Models for Naive Bayes Text Classification. In *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[79] L. W. McKnight, D. Anius, and Ö. Uzuner. Virtual Markets in Wireless Grids: Peering Policy Prospects. In *Internet Services. Quality of Service in Grids, Networks and Markets.* MIT Press, forthcoming.

[80] T. C. Mendenhall. Characteristic Curves of Composition. *Science*, 11, 237–249. 1887.

[81] G. A. Miller, E. B. Newman, and E. A. Friedman. Length-Frequency Statistics for Written English. *Information and Control*,1(4), 370–389. 1958.

[82] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to Wordnet: An On-Line Lexical Database. *International Journal of Lexicography*, 3(4), 235–312, 1990.

[83] D. Mladenic and M. Grobelnik. Feature Selection for Unbalanced Class Distribution and Naive Bayes. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 1999.

[84] A. Q. Morton. The Authorship of Greek Prose. *Journal of the Royal Statistical Society (A)*, 128, 169–233. 1965.

[85] F. Mosteller and D. L. Wallace. Inference in an Authorship Problem. *Journal of the American Statistical Association,* 58(302), 275–309. 1963.

[86] National Research Council. R. Davis, committee chair. *The Digital Dilemma: Intellectual Property in the Information Age Committee on Intellectual Property Rights in the Emerging Information Infrastructure*. ISBN: 0309064996. National Academies Press. 2000.

[87] N. W. Netanel. Impose Noncommercial Use Levy to Allow Free P2P File-Swapping and Remixing. *17 Harvard Journal of Law & Technology*, 2003.

[88] N. W. Netanel. Copyright and the First Amendment; What Eldred Misses – and Portends. *Copyright and Free Speech: Comparative and International Analyses*. Oxford University Press, forthcoming 2005.

[89] Network Fusion. DRM (Digital Rights Management). [web page]. URL: http://www.nwfusion.com/details/699.html. [Accessed 31 January, 2005].

[90] A. Odlyzko. Internet Pricing and the History of Communications. 36 *Computer Networks*, 493. 2001.

[91] R. D. Peng and H. Hengartner. Quantitative Analysis of Literary Styles. *The American Statistician*, 56(3), 175–185. 2002.

[92] F. Petitcolas. Digital Watermarking. In *Digital Rights Management*, E. Becker et al. (eds.) LNCS 2770, 81–92, 2003. Springer-Verlag Berlin Heidelberg.

[93] F. Petitcolas, R. J. Anderson, and M. G. Kuhn. Attacks on Copyright Marking Systems. In *Workshop on Information Hiding*. 1998.

[94] R. Quirk, S. GreenBaum, G. Leech, and J. Svartvik. *A Comprehensive Grammar of the English Language*. Longman. ISBN: 0-582-51734-6. 1985.

[95] Recording Industry Association of America. [web page] 2003. Issues. URL: http://www.riaa.com/issues/default.asp. [Accessed 31 January, 2005].

[96] M. Rennhard and B. Plattner. Introducing MorphMix: Peer-to-Peer based Anonymous Internet Usage with Collusion Detection. In *Proceedings of the Workshop on Privacy in the Electronic Society (WPES)*, in association with the *9th ACM Conference on Computer and Communications Security (CCS 2002)*, 91–102, 2002.

[97] A. De Roeck, A. Sarkar, and P. H. Garthwaite. Defeating the Hoogeneity Assumption. *Technical Report Number 2004/07*. The Open University.

[98] M. J. Rose. Stephen King's "Plant" Unprooted. *Wired News*, [web page] November 28, 2000. URL: http://www.wired.com/news/culture/0,1284,40356,00.html. [Accessed 31 January, 2005].

[99] B. Rosenblatt. Moral Rights Basics. [web page] 1998. URL: http://cyber.law.harvard.edu/property/library/moralprimer.html. [Accessed 31 January, 2005].

[100] B. Rosenblatt, B. Trippe, and S. Mooney. *Digital Rights Management: Business and Technology.* New York, Cleveland, OH, and Indianapolis, IN: M&T Books, 2002.

[101] G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11), 613–620. 1975.

[102] G. Salton and C. Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5), 513–523, 1998.

[103] G. Salton, A. Singhal, C. Buckley, and M. Mitra. Automatic Text Decomposition Using Text Segments and Text Themes. In *Proceedings of the Hypertext '96 Conference*. 1996.

[104] P. Samuelson. Technological Protection for Copyrighted Works. Draft Article as of Jan. 1996.

[105] P. Samuelson. Intellectual Property and the Digital Economy: Why the Anti-Circumvention Regulations Need to be Revised. *14 Berkeley Tech. L. J.*, 519, 1999.

[106] P. Samuelson. Digital Rights Management and, or, vs. the Law. *Communications of the ACM*, 46(4), 41–45. 2003.

[107] R. E. Schapire. A Brief Introduction to Boosting. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 1999.

[108] R. E. Schapire. The Boosting Approach to Machine Learning. In *MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

[109] R. K. Sharma and S. Decker Practical Challenges For Digital Watermarking Applications. *EURASIP Journal on Applied Signal Processing*, 2002(2), 133–139, February 2002.

[110] N. Shivakumar and H. Garcia-Molina. SCAM: A Copy Detection Mechanism for Digital Documents. In *Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries*, 1996.

[111] N. Shivakumar and H. Garcia-Molina. Building a Scalable and Accurate Copy Detection Mechanism. *Proceedings of the 3rd International Conference on Theory and Practice of Digital Libraries*, 1996.

[112] H. S. Sichel. On a Distribution Representing Sentence-Length in Written Prose. *Journal of the Royal Statistical Society (A)*, 137, 25–34. 1974.

[113] F. H. Slowinski. AAP/ALA White Paper: What Consumers Want in Digital Rights Management. American Association of Publishers. [web page]. URL: http://dx.doi.org/10.1003/whitepaper1. [Accessed 31 January, 2005].

[114] M. W. A. Smith. Recent Experience and New Developments of Methods for the Determination of Authorship. *Association for Litrary and Linguistic Computing Bulletin*, 11, 73–82. 1983.

[115] Statute of Anne. 8 Anne, c. 19. [web page]. 1710. URL: http://www.swarb.co.uk/acts/1710AnneStatute.html. [Accessed 31 January, 2005].

[116] P. Stone. Thematic Text Analysis: New Agendas for Analyzing Text Content. In *Text Analysis for the Social Sciences. Chapter* C. Roberts, ed. Lawrence Erlbaum Associates, Publishers, 1997.

[117] D. R. Tallentire. Towards an Archive of Lexical Norms - A Proposal. In *The Computer and Literary Studies*, eds. A. J. Aitken, R. W. Bailey and N. Hamilton-Smith. Edinburgh University Press. 1973.

[118] D. R. Tallentire. *An Appraisal of Methods and Models in Computational Stylistics, with Particular Reference to Author Attribution*. PhD Thesis. University of Cambridge. 1972.

[119] R. Thisted and B. Efron. Did Shakespeare Write a Newly-discovered Poem? *Biometrika*, 74, 445–455. 1987.

[120] Trusted Computing Group. [web page]. URL: https://www.trustedcomputinggroup.org/home. [Accessed 31 January, 2005].

[121] Ö. Uzuner, R. Davis, and B. Katz Using Empirical Methods for Evaluating Expression and Content Similarity. In the *Proceedings of the 37th Hawaiian International Conference on System Sciences (HICSS-37)*, 2004. IEEE Computer Society.

[122] Ö. Uzuner and R. Davis. Content and Expression-Based Copy Recognition for Intellectual Property Protection. In the *Proceedings of the 3rd ACM Workshop on Digital Rights Management (DRM'03)*, 2003.

[123] Uniform Computer Information Transactions Act. [web page] 2001. URL: http://www.law.upenn.edu/bll/ulc/ucita/ucita01.htm. [Accessed 31 January, 2005].

[124] Y. Yang and J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*. 412–420. 1997.

[125] G. U. Yule. On Sentence-Length as a Statistical Characteristic of Style in Prose, with Application to Two Cases of Disputed Authorship. *Biometrika*, 30, 363–390. 1938.

[126] E. Walterscheid. *The Nature of the Intellectual Property Clause: A Study in Historical Perspective*. William S. Hein & Co. ISBN 1-57588-709-6. 2002.

[127] E. Weisstein. Paired t-Test. *Mathworld*. A Wolfram Web Resource. [web page]. URL: http://mathworld.wolfram.com/Pairedt-Test.html. [Accessed 31 January, 2005].

[128] C. B. Williams. Mendenhall's Studies of Word-Length Distribution in the Works of Shakespeare and Bacon. *Biometrika*, 62(1), 207–212. 1975.

[129] C. B. Williams. A Note on the Statistical Analysis of Sentence-Length as a Criterion of Literary Style. *Biometrika*, 31, 356–361. 1940.

[130] Y. Wilks, D. Fass, C-M. Guo, J. McDonald, T. Plate, and B. Slator. A Tractable Machine Dictionary as a Resource for Computational Semantics. In *Computational Lexicography for Natural Language Processing*. Briscoe and Boguraev (eds.). Longman, 1989.

[131] P. H. Winston. Artificial Intelligence. Third Edition. ISBN: 0-201-53377-4. Addison-Wesley Publishing. 1992.

[132] WIPO Copyright Treaties Implementation Act: Hearing on H.R. 2281. Subcommittee on Telecommunications Trade and Consumer Protection of the House Committee on Communication, 105th Congress. 1998.

[133] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools with Java Implementations. Morgan Kaufmann, San Francisco, 2000.

[134] World Intellectual Property Organization. WIPO Copyright Treaty, CRNR/DC/94. Diplomatic Conference on Certain Copyright and Neighboring Rights Questions. 1996.

[135] World Intellectual Property Organization. WIPO Performance and Phonograms Treaty. 1996.

[136] R. B. Wolfgang, C. I. Podilchuck, and E. J. Delp. Perceptual Watermarks for Digital Images and Video. In *Proceedings of IEEE*, 87, 1108–1126, Jul. 1999.

[137] X. Wu. *Knowledge Acquisition from Databases*. ISBN: 1-56750-206-7. Ablex Publishing Corp., U.S.A., 1995.

[138] J. Zittrain. *Calling off the Copyright War. In battle of property vs. free speech, no one wins*. Boston.com. [web page] 2002. URL: http://cyber.law.harvard.edu/is03/Readings/Zittrain_1.pdf. [Accessed 31 January, 2005].

**APPENDIX**

## A-1 General Inquirer Semantic Categories

General Inquirer semantic categories used in this thesis as described (mostly verbatim) in the General Inquirer website[1].

- ABS: 276 words reflecting tendency to use abstract vocabulary. Includes category `Abs@`.
- ANI: 72 references to animals, fish, birds, and insects, including their collectivities.
- Academ: 153 words relating to academic, intellectual or educational matters, including the names of major fields of study.
- Active: 2045 words implying an active orientation.
- AffTot: 196 words in the affect domain. Includes 35 words for reaping affect; 11 words for affect loss and indifference; affect participant; 55 words for friends and family; and 96 affect words not in other categories.
- COLL: 191 words referring to all human collectivities (not animal). Used in disambiguation.
- COLOR: 21 words of color.
- COM: 412 communications words.
- Change-Process: Change process categories `Begin` (56 words), `Vary` (98 words indicating change without connotation of increase, decrease, beginning or ending), `Increas` (increase, 111 words), `Decreas` (decrease, 82 words) and `Finish` (87 words).
- Cognitive-Orientation: Words for knowing, assessment, and problem solving.
- Complet: 81 words indicating that goals have been achieved, apart from whether the action may continue.
- DAV: 540 straight descriptive verbs of an action or feature of an action, such as "run, walk, write, read".

[1] http://www.wjh.harvard.edu/ inquirer/homecat.htm

- DIST: 19 words referring to distance and its measures.

- Doctrin: 217 words referring to organized systems of belief or knowledge, including those of applied knowledge, mystical beliefs, and arts that academics study.

- ECON: Includes category `Econ@` with 510 words of an economic, commercial, industrial, or business orientation, including roles, collectivities, acts, abstract ideas, and symbols, including references to money. Includes names of common commodities in business. Also includes category `Exch` with 60 words concerned with buying, selling, and trading.

- EMOT: 311 words related to emotion that are used as a disambiguation category, but also available for general use.

- EnlTot: Total of about 835 words which include 146 words likely to reflect a gain in enlightenment through thought, education, etc., 27 words reflecting misunderstanding, being misguided, or oversimplified, 18 words "denoting pursuit of intrinsic enlightenment ideas.", 61 words referring to roles in the secular enlightenment sphere, and 585 other enlightenment words.

- Exprsv: 205 words associated with the arts, sports, and self-expression.

- FREQ: 46 words indicating an assessment of frequency or pattern of recurrences, as well as words indicating an assessment of nonoccurrence or low frequency. (Also used in disambiguation).

- Fail: 137 words indicating that goals have not been achieved.

- HU: 795 general references to humans, including roles.

- IAV: 1947 verbs giving an interpretative explanation of an action, such as "encourage, mislead, flatter".

- IPadj: 117 adjectives referring to relations between people, such as "unkind, aloof, supportive".

- IndAdj: 637 adjectives describing people apart from their relations to one another, such as "thrifty, restless".

- Intrj: 42 words and includes exclamations as well as casual and slang references, words categorized “yes” and “no” such as “amen” or “nope”, as well as other words like “damn” and “farewell”.

- Legal: 192 words relating to legal, judicial, or police matters.

- Milit: 88 words relating to military matters.

- Motivation: Includes category `Need` with 76 words related to the expression of need or intent, category `Goal` with 53 names of end-states towards which muscular or mental striving is directed, category `Try` with 70 words indicating activities taken to reach a goal, but not including words indicating that the goals have been achieved, category `Means` with 244 words denoting objects, acts or methods utilized in attaining goals, category `Persist` with 64 words indicating “stick to it” and endurance.

- Movement: Movement categories, including `Stay` (125 words), `Rise` (25 words), `Exert` (194 words), `Fetch` (79 words, includes carrying), `Travel` (209 words for all physical movement and travel from one place to another in a horizontal plane), and `Fall` (42 words).

- NatrPro: 217 words for processes found in nature, birth to death.

- Negate: has 217 words that refer to reversal or negation, including about 20 “dis” words, 40 “in” words, and 100 “un” words, as well as several senses of the word “no” itself; generally signals a downside view.

- Negativ: 2,291 words of negative outlook (not including the separate category no in the sense of refusal).

- No: is 7 words directly indicating disagreement, with the word “no” itself disambiguated to separately identify absence or negation.

- Object: category with 661 words subdivided into `Tool`, (318 words), `Food` (80 words), `Vehicle` (39 words), `BldgPt` (46 words for buildings, rooms in buildings, and other building parts), `ComnObj` (104 words for the tools of communication), and `NatObj` (61 words for natural objects including plants, minerals and other objects occurring in nature other than people or animals). Last, a list of 80 parts of the body (`BodyPt`)

- Ovrst: “Overstated”, 696 words indicating emphasis in realms of speed, frequency, causality, inclusiveness, quantity or quasi-quantity, accuracy, validity, scope, size, clarity, exceptionality, intensity, likelihood, certainty, and extremity.

- POLIT: includes category `Polit@` with 263 words having a clear political character, including political roles, collectivities, acts, ideas, ideologies, and symbols.

- Passive: 911 words indicating a passive orientation.

- Place: category with 318 words subdivided into `Social` (111 words for created locations that typically provide for social interaction and occupy limited space), `Region` (61 words), `Route`, (23 words), `Aquatic` (20 words), `Land` (63 words for places occurring in nature, such as desert or beach), and `Sky` (34 words for all aerial conditions, natural vapors and objects in outer space).

- Positiv: 1,915 words of positive outlook. (It does not contain words for yes, which has been made a separate category of 20 entries.)

- PowTot: includes category `PowGain` for with 65 words about power increasing, `PowLoss` with 109 words of power decreasing, `PowEnds` with 30 words about the goals of the power process, `PowAren` with 53 words referring to political places and environments except nation-states, `PowCon` with 228 words for ways of conflicting, `PowCoop` with 118 words for ways of cooperating, `PowAuPt` with 134 words for individual and collective actors in power process, `PowPt` with 81 words for non-authoritative actors (such as followers) in the power process, `PowDoct` with 42 words for recognized ideas about power relations and practices, `PowAuth` with 79 words concerned with a tools or forms of invoking formal power, `PowOth` with 332 power words not in other sub-categories.

- Quality: 344 words indicating qualities or degrees of qualities which can be detected or measured by the human senses. Virtues and vices are separate.

- Quan: 314 words indicating the assessment of quantity, including the use of numbers. Numbers are also identified by the `NUMB` category (51 words) which in turn divides into `ORD` of 15 ordinal words and `CARD` for 36 cardinal words.

- RcTot: includes `RcEthic` with 151 words of values concerning the social order, `RcRelig` 83 words that invoke transcendental, mystical or supernatural grounds for rectitude, `RcGain`

with 30 words such as worship and forgiveness, RcLoss with 12 words such as sin and denounce, RcEnds with 33 words including heaven and the high-frequency word "ought".

- Rel: 136 words indicating a consciousness of abstract relationships between people, places, objects and ideas, apart from relations in space and time.
- Relig: 103 words pertaining to religious, metaphysical, supernatural, or relevant philosophical matters.
- Ritual: 134 words for non-work social rituals.
- Role: 569 words referring to identifiable and standardized individual human behavior patterns, as used by sociologists.
- RspTot: includes RspGain with 26 words for the garnering of respect, such as congratulations, RspLoss with 38 words for the losing of respect, such as shame, RspOth with 182 words regarding respect that are neither gain nor loss.
- SV: 102 state verbs describing mental or emotional states. Usually detached from specific observable events, such as "love, trust, abhor".
- SklTot: includes SklAsth with 35 words mostly of the arts, SklPt with 64 words mainly about trades and professions, SklOth with 158 other skill-related words.
- SocRel: 577 words for socially-defined interpersonal processes (formerly called "IntRel", for interpersonal relations).
- Space: 302 words indicating a consciousness of location in space and spatial relationships.
- Strong: 1902 words implying strength.
- TIME: includes Time@ with 273 words indicating a time consciousness, including when events take place and time taken in an action. Includes velocity words as well.
- Undrst: "Understated", 319 words indicating de-emphasis and caution in these realms.
- Vice: 685 words indicating an assessment of moral disapproval or misfortune.
- Virtue: 719 words indicating an assessment of moral approval or good fortune, especially from the perspective of middle-class society.

- Weak: 755 words implying weakness.
- WlBtot: includes `WlbGain` with 37 various words related to a gain in well being, `WlbLoss` with 60 words related to a loss in a state of well being, including being upset, `WlbPhys` with 226 words connoting the physical aspects of well being, including its absence, `WlbPsyc` with 139 words connoting the psychological aspects of well being, including its absence, `WlbPt` with 27 roles that evoke a concern for well-being, including infants, doctors, and vacationers.
- WltTot: includes `WltPt` with 52 words for various roles in business and commerce, `WltTran` with 53 words for pursuit of wealth, such as buying and selling, `WltOth` with 271 wealth-related words not in the above, including economic domains and commodities
- Work: 261 words for socially defined ways for doing work.
- Yes: is 20 words directly indicating agreement, including word senses “of course”, “to say the least”, “all right”.

## A-2 Tables of Novels Used in Authorship Attribution

| Book | Author | Status |
|---|---|---|
| Pride and Prejudice | Austen | train, train, train, test, train |
| Mansfield Park | Austen | train, test, train, train, train |
| Sense and Sensibility | Austen | test, train, train, train, train |
| Lady Susan | Austen | train, train, test, train, train |
| Emma | Austen | train, train, train, train, test |
| David Copperfield | Dickens | train, train, test, train, train |
| Life and Adventures of Nicholas Nickleby | Dickens | train, test, train, train, train |
| Oliver Twist | Dickens | train, train, train, train, test |
| The Pickwick Papers | Dickens | train, train, train, test, train |
| The Old Curiosity Shop | Dickens | test, train, train, train, train |
| A Tale of Two Cities | Dickens | train, train, train, train, train |
| Adam Bede | Eliot | train, train, train, test, train |
| Middlemarch | Eliot | test, train, train, train, test |
| Daniel Deronda | Eliot | train, test, train, train, train |
| The Mill on the Floss | Eliot | train, train, test, train, train |
| The Hand of Ethelberta: A Comedy in Chapters | Hardy | train, train, train, train, test |
| Tess of the d'Urbervilles: A Pure Woman | Hardy | train, train, train, train, train |
| Jude the Obscure | Hardy | train, train, test, train, train |
| The Mayor of Casterbridge | Hardy | train, train, train, test, train |
| A Laodicean: A Story of To-Day | Hardy | test, train, train, train, train |
| Far from the Madding Crowd | Hardy | train, test, train, train, train |
| Chronicle of the Conquest of Granada | Irving | test, train, train, test, train |
| Knickerbocker's History of New York | Irving | train, test, train, train, test |
| Life and Voyages of Christopher Columbus Vol. II | Irving | train, train, test, train, train |
| The Sea Wolf | London | train, train, test, train, train |

Table A-1: Part 1 of data used for studying the style of authors.

| Book | Author | Status |
|---|---|---|
| Michael, Brother of Jerry | London | train, train, train, test, train |
| The Iron Heel | London | train, train, train, test, train |
| The Cruise of the Snark | London | train, train, test, train, train |
| Burning Daylight | London | train, test, train, train, train |
| The People of the Abyss | London | train, train, train, train, test |
| Adventure | London | train, train, train, train, train |
| The Little Lady of the Big House | London | train, train, train, train, test |
| The Mutiny of the Elsinore | London | test, train, train, train, train |
| The History of Pendennis | Thackeray | train, train, test, train, train |
| Catherine: A Story | Thackeray | train, train, train, train, train |
| The Memoirs of Barry Lyndon, Esq. | Thackeray | train, train, train, train, train |
| The Tremendous Adventures of Major Gahagan | Thackeray | train, train, train, train, train |
| The Book of Snobs | Thackeray | train, train, train, train, test |
| The Virginians: A Tale of the Eighteenth Century | Thackeray | test, train, train, train, train |
| The Newcomes: Memoirs of a Most Respectable Family | Thackeray | train, test, train, train, train |
| The Great Hoggarty Diamond | Thackeray | train, train, train, train, train |
| The History of Henry Esmond... | Thackeray | train, train, train, test, train |
| The Adventures of Tom Sawyer | Twain | train, train, train, train, test |
| The Gilded Age: A Tale of Today | Twain | train, test, train, train, train |
| Following the Equator: A Journey Around the World | Twain | test, train, train, train, train |
| The Adventures of Huckleberry Finn | Twain | train, train, train, test, train |
| A Connecticut Yankee in King Arthur's Court | Twain | train, train, test, train, train |
| Those Extraordinary Twins | Twain | train, train, train, train, test |
| The Mysterious Stranger | Twain | train, train, train, train, test |
| Christian Science | Twain | train, train, train, train, train |

Table A-2: Part 2 of data used of studying the style of authors.