APPLYING TRANSLATION THEORY TO A CORPUS-BASED ANALYSIS OF
The Count of Monte Cristo

by

LISA Y. BAO

Professor Nathan Sanders, Advisor

A thesis submitted in partial fulfillment
of the requirements for the
Degree of Bachelor of Arts with Honors
in Linguistics

SWARTHMORE COLLEGE

Swarthmore, Pennsylvania

December 2013
# Table of Contents

Table of Contents ........................................... i

Abstract .................................................. ii

Acknowledgements ......................................... iii

1 Introduction ............................................. 1

2 Related Literature ...................................... 3
   2.1 Translation theory .................................. 3
   2.2 Corpus-based translation studies ................. 7
   2.3 Impact and relevance ............................... 10

3 Quantitative Analysis .................................. 11
   3.1 Word frequencies ................................... 11
   3.2 Zipf distribution .................................... 17
   3.3 Measuring sentences and paragraphs ............. 19

4 Qualitative Analysis ................................... 26
   4.1 Domestication and foreignization ................. 27
   4.2 Fluency and readability ............................ 30
   4.3 Censorship and societal norms .................... 32

5 Conclusion ............................................... 35
   5.1 Results ............................................... 35
   5.2 Future work .......................................... 36

References ................................................ 37
Abstract

Alexandre Dumas père’s well-known 19th-century novel, *Le comte de Monte-Cristo [The Count of Monte Cristo]*, has been twice translated from the original French into English, in 1846 and again in 1996. In this thesis, I analyze corpora composed of French and English editions of Dumas’s novel through the lens of modern translation theory. I use quantitative and qualitative methods drawn from corpus-based translation studies to conduct both a word frequency analysis and a theoretically grounded close reading of selected passages, ultimately finding evidence for several hypotheses about translational strategies in English.
Acknowledgements

I thank my advisor, Professor Nathan Sanders, for his excellent guidance throughout the process of bringing this project to fruition. I owe an equal debt to Professor Sibelan Forrester for introducing me three years ago to the study of translation and for continuing as my mentor in all walks of life. Further thanks are due to Sarah Bristow (Bryn Mawr College) and Professor David J. Birnbaum (University of Pittsburgh) for their feedback on earlier drafts of this thesis.
Chapter 1
Introduction

The quality of a literary translation today can be judged on the basis of modern standards set by translation theorists. However, a translator from an earlier era might produce a translation that matches period cultural sensibilities—perhaps quite different from today’s expectations—and is informed by the predominant (even if implicit) theories of the time. For example, Victorian-era English translators could choose to reinterpret passages of the original source text that defied societal expectations or their sense of morality, whereas modern Anglo-American translators often favor reading fluency over a more literal translation of foreign terminology.

This thesis examines two English translations from markedly different time periods of *Le comte de Monte-Cristo [The Count of Monte Cristo]* by Alexandre Dumas père, as compared to each other and to the original French edition. The overarching goal is to discover how respective contemporary translation practices and theories may have influenced choices in translation. First, I employ a quantitative approach to gathering meaningful statistical data from the corpora. For example, word frequency counts and distributional analyses are performed separately on each of the three corpora. Dumas French often employs lengthy sentences punctuated in ways which would be ungrammatical in English, a topic that can be examined by computing sentence and paragraph lengths. Other translation issues, such as the translator’s enactment of foreignization or domestication, are examined through a qualitative close reading of selected manually aligned passages. Ultimately, I seek to answer the question of how translation theory has potentially influenced differing choices in translation between 1846 and 1996.

*Le comte de Monte-Cristo*, originally published in France in 1844, follows the unjust downfall and lengthy revenge of a sailor named Edmond Dantès. Betrayed by apparent friends just before his wedding, Dantès is imprisoned in the Chateau d’If for six years until
he meets fellow prisoner Abbé Faria, who tells him of a treasure hidden on the island of Monte Cristo and eventually provides Dantès a means of escape through his death. Now independently wealthy, Dantès returns to Marseille to track down everyone from his former life and then travels to Paris to plot an elaborate revenge on the three men who had betrayed him.

Dumas’s now-famous novel has been twice translated from French into English. The first major translation, which has entered the public domain and is still reprinted today, was published in 1846 with no translator attribution. However, the 1846 English translation has been criticized for clunky prose and for glossing over certain aspects of the novel that its Victorian audience would have decried, such as the lesbian relationship and implicit elopement of Eugénie Danglars and her music teacher Louise d’Armillly. In 1996, Penguin commissioned Robin Buss to complete a modern English translation of the novel, an undertaking which was distinct from previous and later attempts to simply modernize the language of the 1846 translation.

In the following chapters, I first present an overview of the interdisciplinary related work which foregrounds my analysis: translation theory, corpus linguistics, and corpus-based translation studies. Hence, I contextualize this current work in the historical foundation and modern developments of translation studies. Textual corpora have already been used to great effect in both corpus-based translation studies and the computational literary stylistics or other subfields of digital humanities. In Chapter 3, I present the results of my quantitative analysis using computational and statistical methods. In Chapter 4, I select certain passages, manually aligned across corpora, on which to conduct a qualitative analysis of specific cultural and stylistic issues in translation. Finally, I conclude by summarizing my results and describing possible future work.
Chapter 2

Related Literature

Much has been written from classical antiquity to the present day on the topic of translation: conveying in a target language an idea first expressed in the source language. Critical theories of translation have seen enormous shifts among varying foci, including the recognition of a translated text as equal in value to an original source text. Particularly in the field of corpus-based translation studies, the theoretical departure from equivalence has led to a flourishing of research into systematic features of translational corpora and a general awareness of translations as texts worthy of analysis.

2.1 Translation theory

Lawrence Venuti, a literary scholar preeminent in translation studies, describes the landscape of translation theory as a spectrum which must acknowledge many contributing disciplines: philosophical rhetoric and literary theory, modern linguistic theory based in the scientific method, and practical translator training (2012). A “complete” translation theory is defined by Kelly (1979) as having three parts: specifying a goal and function, describing and analyzing operations, and commenting critically on the relationship between the prior two parts.

While not disagreeing with Kelly’s definition, Venuti singles out the autonomy of a translation, or “factors that distinguish it from the source text and from texts originally written in the translating language,” as a key common characteristic of current translation research (2012: 4). However, throughout the history of translation studies, critics have focused on either equivalence or function as predominant theoretical categories. Equivalence, often synonymous with terms such as “accuracy,” “correspondence,” or “fidelity,” describes the relationship from a translation back to the source text (Venuti 2012). Fundamentally,
equivalence concerns itself with the best representation of the meaning of the original source, with varying definitions of how such a superlative is defined or attained. In contrast, function is the notion of how a translation relates to the target language and receiving culture, concerned with the social effects of translation, as in Brisset’s (1990) study of Québécois translation as a political act.

Furthermore, theorists like Gideon Toury establish translations as independent texts by critiquing or blurring the dichotomy between equivalence and function. Toury (1978/1995, 2012) defines the concept of translational norms and is considered the father of Descriptive Translation Studies, a discourse which has greatly influenced corpus-based translation studies.

2.1.1 Early notions of equivalence

In his 1680 preface to a translation of Ovid’s *Epistles*, John Dryden presents an early formalization of equivalence as a major concern in translation practice. Dryden divides all approaches to translation into three methods:

a) Metaphrase, a literal word-by-word translation. In an extreme example, the French word *machine à laver* ‘washing machine’ might be translated as *machine for washing*. More commonly, the method of metaphrase can lead to awkward calques (foreign borrowings) or the obvious syntactic clunkiness called “translationese.”

b) Paraphrase, a sense-by-sense translation permitting some measure of latitude. In contemporary practice, most translators¹ will translate by “sense” or meaning, deviating from an exact syntactic or semantic correspondence when necessary to avoid calques or translationese.

c) Imitation, a free translation which allows the translator to “forsake [both word and sense] as he sees occasion” (Dryden 1680: 38). The method of free translation has largely fallen out of favor but has seen historical popularity, such as among early English translations of classical Chinese poetry.

For Dryden, the art of translation is comparable to the art of painting from life: neither has the right to alter existing features, and the goal should be to resemble the original without

¹ Vladimir Nabokov is a notable exception among modern translators to the general preference for paraphrase, as evidenced by his famously exact and heavily-footnoted translation of Pushkin’s *Onegin* from Russian to English (see discussion in Nabokov 1955).
slavishly rendering every word. Dryden thus chooses the method of paraphrase (albeit with
apology for greater than intended latitude) to render most of his own translations, although
he notes a few which were done by imitation.

Despite the subsequent Victorian period in England being a prolific time for translations
into English, translation practices and theories advanced relatively little beyond Dryden.
Throughout a large breadth of translation criticism, Victorian translators indicate only the
degree of fidelity to the source text in their prefaces to a translation (Young 1964). The lack
of theoretical sophistication, or indeed of a formalized translation theory emerging from this
secondary criticism, may be attributed to a perception of translation as “the poor step-sister
of the arts” or of English itself as a “barbaric tongue” unequal to Greek or Latin in artistic
expression, among other speculated reasons (Young 1964: 7–8). As such, nineteenth-century
translators can be divided approximately into three schools of thought—the free school
which looked up to Dryden’s authority, the literalists who made little claim to artistry, and
an intermediate group which borrowed features from both. This division, although not
adhering exactly to Dryden’s distinctions, nevertheless ignores any question outside of the
equivalence relationship between a translation and the original.

2.1.2 Social and cultural functions of translation

Writing from Germany in 1813, the philosopher Friedrich Schleiermacher essentially agrees
with Dryden’s dominant theorization but shifts toward an understanding of the translation in
cultural context, albeit still vis-à-vis its source text. Schleiermacher describes paraphrase as
an attempt to overcome the “irrationality of languages” and imitation as a simple surrender
to it, but he dismisses both methods as inadequate (1813: 48). Instead, he calls for the
translator to move either the reader toward the original author or the author toward the
reader, one or the other as strictly as possible without intermixing. Schleiermacher’s goal, in
accordance with his preferred method of moving reader toward author, is to give readers
from the target culture the same impression that they would receive if they were reading
natively in the source language and immersed in the source culture.

Venuti (1993) is clearly influenced by Schleiermacher when he coins the terms domest-
tication and foreignization to describe the orientation of a translation with respect to the
target culture. Domestication prioritizes the target audience’s familiarity with the text by
preferring the closest collocation in the target language, whereas foreignization deliberately
maintains the strangeness of the original text in translation as it might be experienced by
a foreign audience. The choice between foreignization and domestication can be broad in
scope, affecting characterization and personality, or as subtle as the spelling of a character’s
name. Glossing over any linguistic and cultural differences in the foreign text was common
practice during Dryden’s 17th century and established as canonical by the 19th century in
English-language translation; this domestication advocated for fluency of language and the
illusion of transparency to the original writer’s intent (Venuti 1993). In contrast, Venuti
extends Schleiermacher’s argument for a foreignized translation that emphasizes linguistic
and cultural differences, suggesting further that the “violence” of such a translation can act
as a strategic intervention in the English-language cultural hegemony.

Brisset (1990) discusses another example of translation as a strategic social influencer,
especially with respect to marginalized or minority languages. Translation has historically
been able to turn a local vernacular language into a referential language tied to oral and
written tradition, as with Luther’s German translation of the Bible and the edict of Villers-
Cotterêts requiring government acts to be written in French rather than Latin. Oftentimes
a target minority language lacks a written representation in the source language; for example,
French argot has no direct translation or cultural equivalent in English, so the translator may
design his or her own ad hoc equivalent or else revert to standard English. Furthermore,
Brisset considers the question of what happens when translating into a target language, such
as Québécois, which is closely linked to a higher-prestige language like Parisian French. For
example, Michel Garneau’s annotation “traduit en Québécois” [translated into Québécois]
on his translation of Macbeth signifies to a reader of the target language what that target
language actually is, to distinguish —and perhaps ideologically construct—a language
from its more powerful sibling with the same writing system. In the case of Québécois, a
translation is used neither to domesticate nor to foreignize the source text, but to appropriate
the source as a validating marker of cultural identity for an embattled language. Such a
translation, when perceived as a significant subject of study, contributes to the validation of
translation studies as a discipline.

On the theoretical front, Gideon Toury (2012) similarly contributes to the rising promi-
nence of translation criticism through his seminal proposal of Descriptive Translation Studies
(DTS). Toury asserts that translation studies cannot achieve autonomy without a descriptive
branch to complement the theoretical branch and the practical applications branch. Ad-
vancing forward from the initial concept of equivalence, Toury describes translation as a
“norm-governed activity” which involves at least two cultural norm-systems, since a text in
one language occupies some position in its culture and a translation represents the text in an-
other language/culture (Toury 1978/1995: 170). He distinguishes two types of translational
norms: preliminary norms relating to policy and directness of translation, and operational norms which affect low-level textual decisions in the act of translating. This theory of norms, using a functional awareness of cultural positioning, shapes the manifestation of equivalence in a given translation. DTS, and the descriptive branch of translation studies as a whole, has been particularly influential to corpus-based translation studies. Both DTS and corpus linguistics share an empirical perspective on language and investigate it through direct observation of data, contrasting with theories derived from intuitive speculation on the data (Laviosa 2002).

2.2 Corpus-based translation studies

Mona Baker is widely acknowledged as the first pioneering scholar in the effort to apply corpus linguistics to the field of translation studies. Moving past the assumption of primacy for the source text and an equivalence-based assessment of the translation, Baker argues for the study of translation itself as a phenomenon through corpus-based research. For example, corpus-based analysis can identify “features of translation” that recur uniquely in translated corpora versus non-translated corpora; translations may habitually simplify, add explicitation to, or normalize certain patterns of text production (Baker 2004a). Moreover, this technique has been successfully used to investigate non-Indo-European languages as well: see Xiao 2010 for a corpus-based comparison of native and non-translated Chinese and Hou 2013 on nominalization in English-Chinese literary parallel corpora.

Foundationally, Baker challenges the assumption of equivalence as central to translation research and theory. The traditional semantic understanding of equivalence assumes an outdated representational theory of meaning, while an alternative interpretation as “equivalent effect” relies on the impossibility of consistently predicting reader response (Baker 2004b: 20). As Maeve Olohan notes in evaluating the status of translational corpora vis-à-vis their non-translated counterparts, a set of bilingual parallel corpora is sometimes defined as “translations of one another,” implying that source texts and translations are the same unless “discrepancies” exist otherwise (2002: 420).

By discarding the traditional conceptions of equivalence and a sacred relationship to the source text, Baker (2000) delves into new areas such as a methodology for analyzing the style proper of a literary translator. Moreover, her application of narrative theory to translation studies is a direct reaction to existing frameworks such as Toury’s norms and Venuti’s domestication/foreignization dichotomy (Baker 2007). In other work, Baker has
attempted to identify what she calls (universal) “features of translation” (Baker 1993). Some features which could be investigated using a translational corpus include:

a) increased explicitation of concepts and ideas, which might be measured by sentence or paragraph length;

b) disambiguation and simplification of terminology;

c) omitting or rewording repetitions;

d) overrepresentation of certain stylistic features in the target text.

For example, Baker (2004a) looks for the possible overrepresentation of stylistic features by analyzing the distribution of fixed lexical phrases in the Translational English Corpus, a corpus of translated text which she created and continues to manage (Baker 2013). If such phrases occurred more frequently in translated than non-translated text, they would provide evidence for fluency as a dominant strategy in English-language translations. Moreover, as an exploratory study, Baker elaborates on the different possible avenues of analysis enabled by the corpus methodology rather than asserting a critical interpretation of the presented data.

Another prominent scholar in corpus-based translation studies, Maeve Olohan, maintains that translations must be assessed within both a cultural context and through a theoretical lens such as descriptive translational norms (Toury 1978/1995) or prototype theory (Olohan 2002). Olohan describes prototype theory as a way to consider the “best examples” or prototypical translations in a translational corpora without invalidating the less prototypical examples which may not display a particular investigated feature. Although Olohan acknowledges the concerns of some translation scholars that quantitative corpus-based analysis cannot speak to the social status of a translation or the rhetorical purposes governing language production, she argues for the integration of corpus-linguistic methodologies with a theoretical framework as well as additional qualitative studies.

2.2.1 Corpus linguistics

Corpus linguistics is a methodology for analyzing language use and characteristics using data drawn from a “body of linguistic evidence typically composed of attested language use” called a corpus (McEnery 2003: 449). According to Biber et al. (1998), a corpus-based analysis focuses on “association patterns” measuring the relationship between features or
variants and contextual factors, combining this empirical data with a qualitative interpretation. In almost all cases, the analysis is conducted using computational methods, whether a concordancing software or a customized computer program. Elena Tognini-Bonelli (2001) distinguishes between a corpus-based approach, which uses the corpus to test or exemplify an external theory or description, and a corpus-driven approach using the corpus data in order to formulate a theoretical statement. However, Laviosa (2011) uses the term “corpus-based” to encompass both of Tognini-Bonelli’s “corpus-based” and “corpus-driven” terminology.

As Olohan notes, parallel corpora—meaning a set of texts in one language and their translations in a different language—have been used by corpus linguists to facilitate contrastive linguistic analysis, but the research frequently “fails to recognize that translators’ choices may be motivated by something other than language systemic conventions” (2004: 28). Alternatively, Baker has suggested the use of comparable corpora—two collections of translated and non-translated texts in the same language—among which her own Translational English Corpus is most well-known. Comparable corpora are used to study recurring patterns of syntactic or lexical features in translations which may be evidence for translators’ use of explicitation, simplification, or normalization. However, Olohan acknowledges that this type of corpus-based analysis privileges the product of translation over the translation process itself.

2.2.2 Literary stylistics

Corpus linguistics has long been used in the area of literary stylistics, or the linguistic analysis of literary texts. Hence, “corpus stylistics” is the study of encoded meaning—and how to decode that meaning—in a literary text (Fischer-Starcke 2010). According to Fischer-Starcke, the corpus-based approach permits a systematic analysis of lexical and grammatical patterns (representing an author’s unique style) in a large corpus of language data, so that the meanings behind those patterns may be decoded. By the principles of corpus linguistics, the more frequently a pattern or feature occurs, the more significance it holds for the data. However, some scholars dispute this assumed correlation between frequency and significance, arguing that infrequent, yet prototypical features may nonetheless be significant.

Fischer-Starcke asserts that the structuralist assumption of a correlation between form and meaning in language is the foundation of stylistics. The study of literary stylistics can cover any number of levels of linguistic features in a text or corpus, principally syntax, semantics, phonology, or pragmatics. For instance, Fischer-Starcke applies corpus stylistics
to the complete oeuvre of Jane Austen’s novels with a special focus on the single text of her less commonly studied novel *Northanger Abbey*, using a quantitative keyword analysis as a starting point for a qualitative analysis of literary meanings from the keywords’ concordance lines. Mahlberg (2013) takes a similar approach to a corpus of Charles Dickens’s novels, focusing on syntactic and semantic data to drive a literary analysis.

Phonologically, character and author idiolects have been frequently studied since Burrows (1987) first analyzed Jane Austen’s character idiolects. Burrows (2002) later proposes a measure of identifying authorship based on multivariate statistical analysis of a text’s style. His “delta measure” represents the “pure difference,” both positive and negative, in divergence from the norm for a given corpus (Burrows 2002: 680). Burrows’s techniques have been widely adopted in the field of computational stylistics, as Rybicki (2006) attests.

Moreover, corpus stylistics has been recently applied to translation studies in several case studies. Bosseaux (2004) examines the translator’s treatment of free indirect discourse in Virginia Woolf’s novel, *To the Lighthouse*, while van Dalen-Oskam (2013) investigates the translation of proper names from the perspective of computational stylistics using a parallel corpus of Dutch and English novels and their respective translations. Furthermore, Li et al. (2011) conduct a corpus analysis on two English translations of the classical Chinese novel *Hongloumeng* in the tradition of Baker (2000)’s methodology for analyzing translator style. Clearly, the intersection of literary stylistics and translation studies includes both analyses on the level of the translator and analyses comparing literary texts and their translations in a parallel corpus.

### 2.3 Impact and relevance

In summary, the primary research of this thesis, to be discussed in Chapters 3 and 4, integrates two distinct fields of related work. First, literary translation theory provides an important historical and theoretical foundation for my analysis, especially in support of Venuti’s foreignization/domestication paradigm. Moreover, Toury’s theory of DTS lays the groundwork for my second, no less significant area of related literature in corpus-based translation studies. Researchers such as Baker, Olohan, and Laviosa have studied translations using formal corpora and quantitative methods similar to their counterparts in corpus-based literary stylistics. However, no published literature has combined a corpus-based approach to translation studies with a rigorous background and parallel analysis originating from contemporary translation theory.
Chapter 3
Quantitative Analysis

In this chapter, I describe the methodology and computational techniques used to analyze the corpora. Moreover, I present the results of my analyses and interpret the data as evidence toward an empirically-based theory of translation. All statistical analyses were conducted using scripts written in the Python programming language. I refer to the anonymous 1846 English translation of *Le comte de Monte-Cristo* as the GUTENBERG corpus, Robin Buss’s 1996 English translation as BUSS, and the original 1844 French text as DUMAS.

The overarching goal of the analysis is to test the hypothesis that distinct translator’s styles exist in the two English translational corpora, as compared to each other and to the French source corpus. I produce quantitative results to distinguish differences of possible significance using techniques such as basic word frequencies, Zipf distributional estimations, and sentence and paragraph lengths.

### 3.1 Word frequencies

For a preliminary investigation of the corpora, several word frequency analyses were run on the two English translations, GUTENBERG and BUSS. In those cases where meaningful comparisons or conclusions stemmed from a language-independent computational technique, the French source text DUMAS was also analyzed. Underpinning all of the following results is a basic word count data structure, coincidentally called a *dictionary* in Python, created by splitting the text on whitespace, lowercasing each word, and removing punctuation.

#### 3.1.1 Relative frequency

The absolute frequency, or raw count, of a given word in a text is equal to its number of occurrences; hence, the relative frequency is its percentage occurrence in the text, calculated
by dividing the absolute frequency by the total number of all words in the corpus. Using a frequency measure, I can compute the most frequent words in each corpus. After observing the initial list of most frequent words, which is composed primarily of common grammatical function words such as the or of, I also conduct the same analysis after filtering out basic stopwords, or universally common words that may obscure an analysis of less common words. The list of stopwords chosen is derived from the default stopwords filter used in Apache Lucene, a free and open-source information retrieval system. Although the resulting filtered list of most frequent words still contains many function words, I am wary that continuing to expand the stopwords filter will obscure interesting differences in the frequency of function words which may indicate differences in sentence structure.

Table 3.1 displays both the unfiltered and filtered most frequent words in the GUTENBERG and BUSS corpora. A brief observation of the two tables indicates that GUTENBERG and BUSS share almost all of the same most common words in both the unfiltered and the filtered lists. The relative rank of some words does change: for instance, he is at rank 8 in the unfiltered GUTENBERG list and at rank 10 in the unfiltered BUSS list, while which shows up in the filtered GUTENBERG list but not in the filtered BUSS list. These frequency results form the basis of more sophisticated analyses. However, a test of statistical significance would need to precede quantitative confirmation of any particular hypothesis.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Unfiltered</th>
<th>Filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G</td>
<td>B</td>
</tr>
<tr>
<td>1</td>
<td>the</td>
<td>the</td>
</tr>
<tr>
<td>2</td>
<td>to</td>
<td>to</td>
</tr>
<tr>
<td>3</td>
<td>of</td>
<td>and</td>
</tr>
<tr>
<td>4</td>
<td>and</td>
<td>of</td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>6</td>
<td>i</td>
<td>you</td>
</tr>
<tr>
<td>7</td>
<td>you</td>
<td>i</td>
</tr>
<tr>
<td>8</td>
<td>he</td>
<td>that</td>
</tr>
<tr>
<td>9</td>
<td>in</td>
<td>in</td>
</tr>
<tr>
<td>10</td>
<td>his</td>
<td>he</td>
</tr>
</tbody>
</table>

Table 3.1: Top 10 most common words by relative frequency, with and without stopwords filtering, for the GUTENBERG (G) and BUSS (B) corpora.
3.1.2 Frequency ranks and differences

To narrow in on differing word frequencies between GUTENBERG and BUSS, I first compute the difference between relative frequencies for each word common to both corpora. Since both English translations draw from the same source text, their relative frequencies should be roughly comparable. If a word’s relative frequency in BUSS is subtracted from the same word’s relative frequency in GUTENBERG, then the resulting list of words—each associated with its change in relative frequency across corpora—can be sorted to produce the words with the largest change in relative frequency from GUTENBERG to BUSS. Table 3.2 demonstrates the process of moving from a given word’s relative frequency in each corpus to a composite difference value. The same procedure is performed for both raw frequency values and relative frequency ranks.

Partial results for both unfiltered and filtered word sets are shown in Table 3.3. A high G–B value represents words which occur more frequently in GUTENBERG compared to BUSS, based on their relative frequencies in each corpus. For instance, the word which has a relative frequency of about 0.60% in GUTENBERG and about 0.44% in BUSS, so its G–B score is approximately 0.6% − 0.44% = 0.16%. After calculating these scores for each word, the words are ranked by the greatest difference in relative frequency.

One interesting observation from the data is a choice between m and monsieur, mme and madame. It appears that the abbreviation M. for monsieur is preferred in GUTENBERG, while the honorific is more frequently written out in its entirety in BUSS. However, the opposite relationship holds for madame on the GUTENBERG-positive list (G–B) and its abbreviated version, Mme, on the BUSS-positive list (B–G). For example, since madame is ranked among the top 10 words by G–B value, it must have a highly negative B–G value, especially compared to the highly positive B–G value of mme.

<table>
<thead>
<tr>
<th>Word</th>
<th>Rel Freq G</th>
<th>Rel Freq B</th>
<th>G–B</th>
<th>B–G</th>
</tr>
</thead>
<tbody>
<tr>
<td>of</td>
<td>0.0277</td>
<td>0.0233</td>
<td>0.0044</td>
<td>-0.0044</td>
</tr>
<tr>
<td>the</td>
<td>0.0614</td>
<td>0.0574</td>
<td>0.0041</td>
<td>-0.0041</td>
</tr>
<tr>
<td>that</td>
<td>0.0118</td>
<td>0.0161</td>
<td>-0.0043</td>
<td>0.0043</td>
</tr>
<tr>
<td>because</td>
<td>0.0005</td>
<td>0.0015</td>
<td>-0.0010</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

Table 3.2: A sample table demonstrating how G–B and B–G values are calculated for unfiltered relative frequencies (numbers have been truncated for display).
<table>
<thead>
<tr>
<th>Rank</th>
<th>Unfiltered</th>
<th>Filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G–B</td>
<td>B–G</td>
</tr>
<tr>
<td>1</td>
<td>of</td>
<td>that</td>
</tr>
<tr>
<td>2</td>
<td>the</td>
<td>monsieur</td>
</tr>
<tr>
<td>3</td>
<td>and</td>
<td>in</td>
</tr>
<tr>
<td>4</td>
<td>m</td>
<td>a</td>
</tr>
<tr>
<td>5</td>
<td>not</td>
<td>was</td>
</tr>
<tr>
<td>6</td>
<td>dantes</td>
<td>on</td>
</tr>
<tr>
<td>7</td>
<td>which</td>
<td>out</td>
</tr>
<tr>
<td>8</td>
<td>will</td>
<td>up</td>
</tr>
<tr>
<td>9</td>
<td>sir</td>
<td>dont</td>
</tr>
<tr>
<td>10</td>
<td>then</td>
<td>because</td>
</tr>
</tbody>
</table>

Table 3.3: Top 10 words, with and without stopwords filtering, showing the highest positive difference in relative frequency for GUTENBERG–BUSS (G–B) and BUSS–GUTENBERG (B–G).

Word frequency distributions can also be compared between two corpora by subtracting each word’s relative ranking in one corpus from its ranking in the other corpus, finding the words with the largest change in frequency rank. After sorting all of the words in a corpus in descending order of absolute frequency (counts), I assign a monotonically increasing rank number to each successive word, going down the list. For instance, the word *which* is ranked as the 26th most frequent word in GUTENBERG and the 31st most frequent word in BUSS, so its G–B score is \(26 - 31 = -5\) and its B–G score is \(31 - 26 = 5\). If a tie occurs between successive words with the same word count, then those words are all assigned the same rank, and the rank counter continues to increase monotonically so that the next-lowest-frequency word will be assigned a lower rank that is properly representative of its relative location in the list.

Tables 3.4 and 3.5 show the results of a frequency rank difference computed for each word common to both GUTENBERG and BUSS. The presence of *mme* as the highest-ranking word by G–B value (this time for rank difference) confirms the prior hypothesis drawn from relative frequency differences. Moreover, the two sets of words occurring most frequently in each corpus are distinct from each other. Specific nouns like *stairway* and *commissioner* rank highly on this measure of (dis)similarity in the GUTENBERG corpus; perhaps BUSS translates these terms to different English words, depending on context. Further, the words occurring more frequently in BUSS than in GUTENBERG display a pattern of formality
in words like *arose*, *recollect*, and *whence*. This tone and register suggest that the Buss translation may be striving for a more classical literary style, since *whence* is a rare word to find in late-twentieth-century vocabulary.

Relative frequencies and frequency ranks can act as a coarse measure of similarity or difference. However, the preliminary interpretation supported by these data of a classical style in Buss is directly contradicted by its translator’s stated intent (see §3.1.3). Comparing only a few most frequent words—even after a preliminary stopwords filtering—does not provide enough significant evidence to make a judgment regarding stylistic similarity or a distinguishable translator’s style.
### 3.1.3 Types and tokens

Another simple statistic which succinctly summarizes corpus word distribution is the type-token ratio. Each instance of any word, including repetitions, is a token; so the usual word count of a text is simply the total number of tokens. In contrast, the number of types is how many distinct words occur throughout the entire corpus, or the size of the vocabulary used. Counting the number of types and tokens in each corpus, then computing the type-token ratios, gives the results shown in Table 3.6.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Types</th>
<th>Tokens</th>
<th>Type-Token Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>GUTENBERG</td>
<td>19,026</td>
<td>459,127</td>
<td>0.041</td>
</tr>
<tr>
<td>Buss</td>
<td>16,880</td>
<td>481,541</td>
<td>0.035</td>
</tr>
<tr>
<td>DUMAS</td>
<td>26,729</td>
<td>497,957</td>
<td>0.054</td>
</tr>
</tbody>
</table>

**Table 3.6**: Type and token counts for the three corpora.

Just comparing the counts across all three corpora, DUMAS has the highest number of both types and tokens. However, between GUTENBERG and Buss the patterns are not so simple: while GUTENBERG has a higher type count, Buss has a higher token count. Hence, the more recent Buss translation is longer in absolute length, with more words, but the older 1846 translation employs greater variety of word choice and has a larger overall vocabulary size. This direct comparison is appropriate because both translations claim to represent the same source text in the target language; I am interested in exploring the differences between their representations.

The type-token ratio for each corpus further refines these distinctions. Although DUMAS has the highest ratio of the three corpora, as the only French-language corpus studied here, a direct cross-lingual comparison may be inaccurate. GUTENBERG has a higher type-token ratio than Buss, supporting the conclusion that it uses more distinct words after adjusting for relative document length. The larger vocabulary of GUTENBERG could potentially restrict the range of capable readers and potential audience, given the potential for more obscure or uncommon words. This accessibility concern is more likely to have impacted translational choices in Buss because, although GUTENBERG was produced for a bourgeois audience, the expected education of literate 1846 English readers remained relatively uniform.

Hence, I hypothesize that GUTENBERG may have been written for an audience which expected a more formal register, whereas the author of Buss—translator Robin Buss—states
that his goal was to “produce a version that is accurate and readable” (Buss 1996: xxvii). Although Buss takes issue with academic theories of translation, alluding in particular to Venuti’s technique of foreignization, his translational intent clearly supports “fluency as the overriding strategy in Anglo-American translations,” a claim that Baker has tried to evaluate through corpus-based analysis of translated text (Baker 2004a: 173).

3.2 Zipf distribution

Given a list of each type and its number of tokens for a particular corpus, I attempt to find the best-fit Zipf distribution curve. Zipf’s law states that the rank-proportional frequency (restated here as Zipf frequency) of a word is inversely proportional to its frequency rank among all words in a given natural language corpus. Thus,

\[ \text{zipfFreq}(\text{word}) = \frac{1}{\text{rank}^c} \]

where \( c \) represents some constant that is computed for a given corpus. Furthermore, this rank-proportional frequency can be normalized over the sum of all \( n \) word frequencies to give the predicted relative frequency (PRF):

\[ \text{PRF}(w) = \frac{\text{zipfFreq}(w)}{\sum_{i=1}^{n} \text{zipfFreq}(w_i)}. \]

By systematically testing different values of \( c \) for the Zipf frequency, I estimate the best values for each corpus by choosing the minimum sum squared error, or the total of the difference between the PRF and the actual relative frequency (number of tokens divided by total tokens) for each word in a given corpus. These values, along with their calculated total errors, are shown in Table 3.7. The best Zipf constant for GUTENBERG is the highest of the three corpora at \( c = 0.93 \), and it also produced the lowest total error. BUSS follows very closely at \( c = 0.92 \), while DUMAS is the lowest at \( c = 0.90 \). Overall, the best \( c \) values are distinct but close to one another, and a test of statistical significance would be needed to measure the actual difference.

To illustrate these Zipf constants, Figure 3.1 displays the empirically discovered best constant for DUMAS alongside the same normalized rank-proportional word frequencies calculated with smaller and greater constants. For the most frequent words shown in the upper left corner of the graph, the smallest Zipf constant produces the highest Zipf predicted frequency. Conversely, as absolute word frequency decreases to the right of the graph—the values being ordered by frequency rank—the largest Zipf constant gives the highest Zipf
Table 3.7: Constant values for the Zipf distribution which minimize total squared error across all words in a given corpus.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Zipf Constant</th>
<th>Sum Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>GUTENBERG</td>
<td>0.93</td>
<td>0.00063</td>
</tr>
<tr>
<td>BUSS</td>
<td>0.92</td>
<td>0.00066</td>
</tr>
<tr>
<td>DUMAS</td>
<td>0.90</td>
<td>0.00116</td>
</tr>
</tbody>
</table>

predicted frequency. The underlying reason for this progressive relationship remains unclear, but importantly, the best Zipf constant produces a Zipf predicted frequency which is always bounded by those computed using the smaller and greater Zipf constants.

![Figure 3.1: A graph of normalized Zipf rank-proportional frequencies using the best-fit Zipf constant for DUMAS compared to smaller and larger constant values, shown on a logarithmic scale.](image)

Since an inspection of my absolute word frequency data shows a disproportionately high number of words occurring only once in a given corpus, I also perform the same analysis after filtering out these singleton words. The results are shown in Table 3.8. This case shows slightly more differentiation among the three corpora’s best-fit Zipf constants, which range between $c = 0.88$ and $c = 0.92$. With respect to the sum squared errors, the data maintain a trend of decreased total error as the Zipf constant value increases, but the statistical significance of this pattern is unclear.
Table 3.8: Constant values for the Zipf distribution that minimizes total squared error across all words in a given corpus, after filtering out singleton words.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Zipf Constant</th>
<th>Sum Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>GUTENBERG</td>
<td>0.92</td>
<td>0.00058</td>
</tr>
<tr>
<td>Buss</td>
<td>0.90</td>
<td>0.00061</td>
</tr>
<tr>
<td>DUMAS</td>
<td>0.88</td>
<td>0.00112</td>
</tr>
</tbody>
</table>

As filtering out single-count words produces only very slightly lower error values and Zipf constants. I also perform the same analysis filtering out duplicate ranks, which artificially truncates the tail end of the distribution. This procedure gave lower best-fit Zipf constants compared to the previous iterations, as expected, and the data are shown in Table 3.9. The calculated Zipf constants now vary from $c = 0.77$ to $c = 0.84$, a greater range, with DUMAS still having the lowest constant value and GUTENBERG having the highest constant value. Similar to previous iterations, the same inverse relationship holds between the constant and the sum error. However, increased filtering also leads to higher error—compare, for instance, the sum squared error values for Table 3.7 and Table 3.9—and strays from a strict interpretation of Zipf’s law, which should ostensibly apply to any natural language corpus.

Table 3.9: Constant values for the Zipf distribution that minimizes total squared error across all words in a given corpus, after filtering out words of duplicate rank.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Zipf Constant</th>
<th>Sum Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>GUTENBERG</td>
<td>0.84</td>
<td>0.00073</td>
</tr>
<tr>
<td>Buss</td>
<td>0.81</td>
<td>0.00067</td>
</tr>
<tr>
<td>DUMAS</td>
<td>0.77</td>
<td>0.00150</td>
</tr>
</tbody>
</table>

3.3 Measuring sentences and paragraphs

3.3.1 Counts

As a preliminary measure of length, I count the number of sentences and paragraphs per chapter of each corpus, enumerated in Table 3.10. The end of a sentence was marked by
a final punctuation mark (period, question, or exclamation) followed by whitespace and excluding the common abbreviations *M.*, *Mme.*, and *Mlle.*. Similarly, the end of a paragraph was marked by a line break at the end of a line containing at least one non-whitespace character and followed by a blank line.

**Table 3.10:** Sentence and paragraph counts (rounded to the nearest whole number) for each corpus, overall and on average per chapter.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sentences</th>
<th>Sent / Chap</th>
<th>Paragraphs</th>
<th>Par / Chap</th>
</tr>
</thead>
<tbody>
<tr>
<td>GUTENBERG</td>
<td>26,415</td>
<td>226</td>
<td>11,328</td>
<td>97</td>
</tr>
<tr>
<td>Buss</td>
<td>33,666</td>
<td>288</td>
<td>15,951</td>
<td>136</td>
</tr>
<tr>
<td>DUMAS</td>
<td>28,567</td>
<td>244</td>
<td>10,502</td>
<td>90</td>
</tr>
</tbody>
</table>

**BUSS,** with more words (tokens) but fewer unique types than **GUTENBERG,** has the highest number of sentences overall and sentences per chapter, even compared to **DUMAS** which has the highest number of both types and tokens (cf. Table 3.6). A similar relationship holds for paragraph counts, although here **GUTENBERG** slightly beats **DUMAS.** Thus, under the assumption that each chapter of both **GUTENBERG** and **BUSS** corresponds to the same chapter exactly of **DUMAS,** with no divergence in chapter splits, **BUSS** appears to employ comparatively longer sentences and paragraphs.

The distribution of chapter length varies greatly, ranging from 72-92 sentences in Chapter 49 to 588-760 sentences in Chapter 73. The number of sentences does not always correlate with the number of paragraphs; for instance, Chapter 25 has just 19 paragraphs but 91 sentences in **GUTENBERG.** This high variance across chapters can be seen in Figures 3.2 and 3.3. **BUSS** generally has the highest number of sentences for a given chapter, whereas **GUTENBERG** and **DUMAS** track each other more closely. A similar trend holds for the average number of paragraphs for a given chapter with **BUSS** topping the graph, although in this case **GUTENBERG** and **DUMAS** alternate having the comparatively fewest paragraphs per chapter.
Figure 3.2: The average number of sentences for each chapter of the three corpora.

Figure 3.3: The average number of paragraphs for each chapter of the three corpora.
3.3.2 Ratios and ranges

Examining the length of paragraphs, as measured in sentences, provides another angle on quantifying “wordiness” or word density. Table 3.11 displays the ten highest average sentences per paragraph and their corresponding chapters for each corpus. By this measure, the top 6 chapters with the highest sentence/paragraph ratio are all in GUTENBERG, which also includes the only two chapters with more than 5 sentences per paragraph on average. On the other end of the spectrum, both GUTENBERG and BUSS have a lower bound of between 1 and 2 average sentences per paragraph in any one chapter, but DUMAS’s lowest average is about 2 sentences per paragraph in Chapter 24.

<table>
<thead>
<tr>
<th>Rank</th>
<th>GUTENBERG</th>
<th></th>
<th>BUSS</th>
<th></th>
<th>DUMAS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sent / Par</td>
<td>Chapter</td>
<td>Sent / Par</td>
<td>Chapter</td>
<td>Sent / Par</td>
<td>Chapter</td>
</tr>
<tr>
<td>1</td>
<td>6.27</td>
<td>24</td>
<td>3.12</td>
<td>67</td>
<td>4.31</td>
<td>51</td>
</tr>
<tr>
<td>2</td>
<td>5.22</td>
<td>45</td>
<td>3.05</td>
<td>51</td>
<td>4.06</td>
<td>67</td>
</tr>
<tr>
<td>3</td>
<td>4.95</td>
<td>22</td>
<td>2.80</td>
<td>48</td>
<td>3.84</td>
<td>80</td>
</tr>
<tr>
<td>4</td>
<td>4.79</td>
<td>25</td>
<td>2.79</td>
<td>45</td>
<td>3.83</td>
<td>89</td>
</tr>
<tr>
<td>5</td>
<td>4.56</td>
<td>111</td>
<td>2.65</td>
<td>91</td>
<td>3.69</td>
<td>57</td>
</tr>
<tr>
<td>6</td>
<td>4.37</td>
<td>44</td>
<td>2.63</td>
<td>35</td>
<td>3.65</td>
<td>83</td>
</tr>
<tr>
<td>7</td>
<td>4.30</td>
<td>102</td>
<td>2.61</td>
<td>112</td>
<td>3.64</td>
<td>95</td>
</tr>
<tr>
<td>8</td>
<td>4.21</td>
<td>86</td>
<td>2.61</td>
<td>89</td>
<td>3.47</td>
<td>48</td>
</tr>
<tr>
<td>9</td>
<td>3.93</td>
<td>114</td>
<td>2.60</td>
<td>80</td>
<td>3.45</td>
<td>107</td>
</tr>
<tr>
<td>10</td>
<td>3.92</td>
<td>23</td>
<td>2.56</td>
<td>111</td>
<td>3.41</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 3.11: Top 10 words, with and without stopwords filtering, showing the highest positive difference in relative frequency for GUTENBERG–BUSS (G–B) and BUSS–GUTENBERG (B–G).

Laviosa (2002) points out that measuring sentence length is a common quantitative method in studies of authorship and literary stylistics. Using the average sentence length of a corpus to approximate one feature (among many) of written style, Laviosa considers shorter sentences to be characteristic of stylistic simplification, and she hypothesizes that translated texts will have lower average sentence length than non-translated texts (including a number of different source languages) in her English Comparable Corpus (ECC). Via a computational analysis of the ECC, Laviosa confirms this hypothesis for a subcorpus drawn from newspaper articles, although not for the narrative subcorpus. In the case of the GUTENBERG and BUSS corpora, shorter sentence lengths could correlate with a more fluent translation.
Furthermore, Table 3.12 shows the average sentence length, or the total number of words divided by the total number of sentences, as well as the average number of sentences per paragraph for each corpus. GUTENBERG and DUMAS both have about 17 words per sentence on average, compared to just 14 words per sentence in BUSS. When counting sentences in paragraphs, the differences are more subtle; BUSS has the lowest number at 2.11 sentences per paragraph, but GUTENBERG follows closely with 2.33 sentences per paragraph and DUMAS, which has the highest number, still only averages 2.72 sentences per paragraph. Based on both whole-sentence and paragraph counts, BUSS is stylistically the longest text. This result is surprising because French as a language tends towards longer sentences than English.

<table>
<thead>
<tr>
<th>Metric</th>
<th>GUTENBERG</th>
<th>BUSS</th>
<th>DUMAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words / Sent</td>
<td>17.38</td>
<td>14.30</td>
<td>17.43</td>
</tr>
<tr>
<td>Sent / Par</td>
<td>2.33</td>
<td>2.11</td>
<td>2.72</td>
</tr>
</tbody>
</table>

Table 3.12: The average sentence length and paragraph length (in sentences) for all three corpora.

But to fully support such a claim, it is also necessary to investigate the lengths of sentences themselves, as well as the number of sentences which form a given paragraph. For example, Figure 3.4 illustrates the general trend in sentence lengths across all three corpora. Overall, BUSS tends to have a lower sentence length than the other two corpora, whereas GUTENBERG and DUMAS vary more frequently across chapters in their sentence-length relationship to each other.

Similarly, Figure 3.5 illustrates paragraph lengths across each chapter in the three corpora. Here GUTENBERG tends to have the highest paragraph length, although DUMAS sometimes reverses this trend; BUSS more consistently has the lowest average paragraph length of the three. From these data, I hypothesize that BUSS employs shorter sentences and fewer paragraphs than GUTENBERG in translating the same text, DUMAS, in order to achieve a more fluent and readable translation.

In combination with BUSS’s smaller vocabulary shown in Section 3.1.3, the data support an overall theory of producing an “easier,” more accessible text. However, the translator of GUTENBERG cannot be faulted for intentionally excluding some readers, since the makeup of its contemporary audience would have been significantly more uniform (and smaller) than a possible English-literate audience today.
Figure 3.4: The average sentence length (in words) for each chapter of the three corpora.

Figure 3.5: The average paragraph length (in sentences) for each chapter of the three corpora.
Finally, calculating the range between the lowest and highest average values acts as a whole-corpus measure of variance. As seen in Table 3.13, the ranges themselves vary considerably. Even after accounting for differences in relative scale, since a paragraph should have many fewer sentences than a sentence will have words, the results remain inconclusive. For instance, GUTENBERG has the largest range of sentence and paragraph lengths at nearly 5 sentences per paragraph and around 23 words per sentence, respectively. However, BUSS has the largest range of 668 sentences per chapter, and DUMAS has the largest range of 329 paragraphs per chapter. It appears from the data that all three corpora vary greatly, albeit on different text metrics.

<table>
<thead>
<tr>
<th>Metric</th>
<th>GUTENBERG</th>
<th>BUSS</th>
<th>DUMAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences/Chapter</td>
<td>507</td>
<td>668</td>
<td>586</td>
</tr>
<tr>
<td>Paragraphs/Chapter</td>
<td>265</td>
<td>241</td>
<td>329</td>
</tr>
<tr>
<td>Sentences/Paragraph</td>
<td>4.90</td>
<td>1.51</td>
<td>2.28</td>
</tr>
<tr>
<td>Words/Sentence</td>
<td>23.48</td>
<td>17.17</td>
<td>21.65</td>
</tr>
</tbody>
</table>

Table 3.13: The range of average sentences per chapter, paragraphs per chapter, sentences per paragraph, and words per sentence in all three corpora.
Chapter 4
Qualitative Analysis

As a counterpart to empirical data collection and analysis, I select one significant excerpt from the novel to examine through manual concordance. Chapter 97, titled “La route de Belgique” in DUMAS and translated as either “The Departure for Belgium” (GUTENBERG) or “The Road for Belgium” (BUSS), narrates the secret departure—perhaps an elopement—of Eugénie Danglars, daughter of the wealthy banker who contributed to betraying Dantès, and her music teacher Louise d’Armilly. Eugénie has just discovered that her fiancé (whom she does not wish to marry) is accused of being an escaped convict; but even before this shocking revelation, she had planned to run away to Italy with Louise rather than submit to a political marriage.

By analyzing this passage through the lens of contemporary translation theories, I attempt to answer several key questions:

• Where do specific words rendered differently in GUTENBERG versus BUSS fall on the spectrum of domestication and foreignization?

• As compared to GUTENBERG, is BUSS a more “fluent” translation (Baker 2004a)? Does BUSS achieve its translator’s explicit goal of readability (Buss 1996)?

• Did the Victorian-era translator who produced GUTENBERG elect to censor the lesbian relationship hinted at between Eugénie and Louise in the original text? Conversely, does the modern BUSS translation portray a more “equivalent” (Baker 2004b) representation of their relationship?
4.1 Domestication and foreignization

Venuti (1993) uses the terms domestication and foreignization (see §2.1.2) to describe a decision which must be made by the translator with respect to fluency and ease, or lack thereof, experienced by the expected reader of a target language. The dichotomy between domestication and foreignization does not always hold, but it is especially notable when translating both connotation and denotation of culture-specific vocabulary. For instance, Li et al. (2011) examine two translations of the classical Chinese novel *Hongloumeng* in which one translation conveyed cultural references through paraphrase or amplification and the other translation gave a literal rendering with copious footnotes. Moreover, even a factor as subtle as the spelling and transliteration of proper names can be interpreted through the framework of domestication or foreignization.

4.1.1 Vocabulary

The translation of particular vocabulary is one area where translators often face a choice on the spectrum between domestication and foreignization. In the case of French-to-English translation, a translator may have access to French loans or other recent borrowings which exist but are less commonly used in English. However, some words are culturally specific to one language and would make no sense as a literal translation, leaving the translator to decide how foreign (or un-foreign) their rendering should appear.

In Chapter 97 of the corpora, the French term *la valise* ‘luggage’ appears in DUMAS (1a) and is translated differently in GUTENBERG (1b) and BUSS (1c). In GUTENBERG, the term is consistently translated as *portmanteau*, an English word that is uncommon today, of obviously French origin. This translation choice is a clear instance of foreignization, emphasizing to the English reader that the novel is set in France with French-speaking characters. Since French loanwords in English tend to have high-prestige connotations, the reader is also impressed with a sense of formality. In fact, the French word *valise* from DUMAS has neutral connotations and is still often used in contemporary colloquial French. Furthermore, DUMAS distinguishes the French *la malle* from *la valise*, a distinction which BUSS maintains by translating the former as *trunk* and the latter as *suitcase*, whereas GUTENBERG merges the two senses by translating both words as *portmanteau*. 
a. **Dumas:**

—Et pour autant au moins de perles, de diamants et bijoux, dit Eugénie. Nous sommes riches. Avec quarante-cinq mille francs, nous avons de quoi vivre en princesses pendant deux ans ou convenablement pendant quatre.

“Mais avant six mois, toi avec ta musique, moi avec ma voix, nous aurons doublé notre capital. Allons, charge-toi de l’argent, moi, je me charge du coffret aux pierreries ; de sorte que si l’une de nous avait le malheur de perdre son trésor, l’autre aurait toujours le sien. Maintenant, la valise : hâtons-nous, la valise !

b. **Gutenberg:**

“And as much, at least, in pearls, diamonds, and jewels,” said Eugenie. “We are rich. With forty-five thousand francs we can live like princesses for two years, and comfortably for four; but before six months—you with your music, and I with my voice—we shall double our capital. Come, you shall take charge of the money, I of the jewel-box; so that if one of us had the misfortune to lose her treasure, the other would still have hers left. Now, the portmanteau—let us make haste—the portmanteau!”

c. **Buss:**

“And at least as much again in pearls, diamonds and jewels,” said Eugénie. “We are rich. With forty-five thousand francs, we can live like princesses for two years, or more modestly for four. But in less than six months, you with your music and I with my voice, we shall have doubled our capital. Come, you take the money, I’ll look after the jewel box, so that if one of us is unlucky enough to lose her treasure, the other will still have hers. Now the suitcase! Quickly, the suitcase!”

Another point of divergence due to culture-specific vocabulary is the translation of *poltronne* ‘coward.’ From the original text in **Dumas** (2a), Eugénie clearly directs the word toward Louise as a term of affection or endearment. This connotation is the opposite of how *coward* is typically perceived in English. *Poltron* does exist in the English dictionary as a French borrowing, but its obscurity makes it a poor translation choice.

The two English translations handle this issue in very different ways. In **Gutenberg,** the translator renders *poltronne* as the English phrase *you little coward* (2b), which can carry a sense of playfulness but still invokes the negativity of *coward.* On the other hand, in **Buss** (??) the French word is rendered by coining the neologism *chicken-heart,* a partly literal translation which draws on the same animal association in both languages (*poltronne* being connected to the French *poulet* ‘chicken’). The disadvantage of this strategy is that the reader must correctly interpret the neologism.
(2) a. **DUMAS:**
—— N’aie donc pas la moindre inquiétude, *poltronne* ; tous nos gens sont occupés de la grande affaire. D’ailleurs, qu’y a-t-il d’étonnant, quand on songe au désespoir dans lequel je dois être, que je me sois enfermée, dis ?

b. **GUTENBERG:**
“Do not be uneasy, you little coward! All our servants are busy, discussing the grand affair. Besides, what is there astonishing, when you think of the grief I ought to be in, that I shut myself up?—tell me!”

c. **BUSS:**
“Don’t worry, chicken-heart. No one is thinking of anything except the great affair. And what is there so surprising about my shutting myself in my room, when you think of how desperate I must be?”

### 4.1.2 Names

In translating proper names from one language to another, a decision of domestication versus foreignization often arises. Where English is the target language, name translation can be as dramatic as a choice between transliteration or literal rendition, such as when translating a name like *lanhua* ‘orchid’ common in classical Chinese literature. It can also result in a loss of meaning, as in the case of a devoted and subservient wife named Amina, literally ‘faithful’ in Arabic (Al Rabadi 2012).

In the case of French-English translation illustrated here by the three corpora, the translation issue is more subtle but nonetheless decisive. First, BUSS preserves the original accented letters in characters’ names while GUTENBERG uses only the base non-diacritic letters. So Dantès becomes Dantes, Mercédès becomes Mercedes, Eugénie becomes Eugenie, and so on. In many cases, the resulting names have been borrowed into English with anglicized pronunciations, so this seemingly small difference can change how readers hear and remember characters’ names in their minds—a clear example of domestication.

With respect to titles, such as the extremely common French honorifics of *Monsieur* (*M.*), *Madame* (*Mme*), and *Mademoiselle* (*Mlle*) equivalent to *Mr.*, *Mrs.*, and *Miss*, both GUTENBERG and BUSS retain a version of the French title rather than domesticating to English honorifics. However, GUTENBERG spells out the abbreviated forms used in DUMAS and kept identically in BUSS, thus explaining the titles to an unfamiliar reader and paradoxically also emphasizing their foreignness. When *M. Danglars* becomes *Monsieur Danglars*, an address equivalent to *Mister Danglars*, the reader is repeatedly reminded that the characters are French, not English (or American). In BUSS, the reader must puzzle
through the standard French abbreviations at first, but as the novel progresses these titles are more likely to fade into the background. Both corpora illustrate the choice of foreignization over domestication, but their ultimate translational effects will differ.

4.2 Fluency and readability

In the preface to his translation, which may be read as a statement of the translator’s own theory, Robin Buss critiques the work of translation theorists and states plainly that the goal of his work is to be simultaneously “accurate and readable” (Buss 1996: xxvii). Baker has claimed that fluency is an overriding concern in English translation and supports this hypothesis through corpus-based research (Baker 2004a). I define a fluent translation as one which is readable and does not confront the reader with intrusive foreignization, thus preferring techniques of domestication at a certain level. The question of fluency can be partly answered with a combination of corpus-based and theoretical approaches: sentence and paragraph lengths may be manually computed to support generalizations made from quantitative data, whereas issues of register or formality are more easily examined through a close reading than through computation.

An examination of sentence and paragraph lengths may be used to either support or question an interpretation based on corpus data. The first paragraph of Chapter 97 in GUTENBERG is quite dense: only 5 sentences yet a total of 247 words, which works out to about 49 words per sentence. In BUSS, the same content consists of 6 sentences and 276 words—slightly more than GUTENBERG on both counts—but it is divided up among 4 paragraphs and has a slightly lower 46 words per sentence on average. Hence, BUSS has more paragraphs and shorter sentences than GUTENBERG, both traits which contribute to increased ease in reading. In illustration, the first paragraph of Chapter 97 from GUTENBERG (3) is marked up with /S/ representing additional sentence breaks and /P/ representing additional paragraph breaks employed in BUSS.

In comparison, the same text in DUMAS comprises 3 paragraphs with 293 words and just 3 sentences, almost 98 words per sentence (and per paragraph). Although DUMAS did not necessarily have higher sentence lengths in the corpus overall, this passage makes clear that at times the corpus does follow the French stylistic stereotype of very long sentences.

The readability of GUTENBERG is evidently impacted by the translator’s decision to combine three French paragraphs, which each contain only one sentence, into a single longer paragraph. Buss takes the opposite approach by adding an additional paragraph.
break. Both GUTENBERG and BUSS break up the three lengthy sentences in DUMAS, thus increasing readability to an English audience who would be unused to the fluidly connected, rambling style of Dumas’s French. Although in one sense this decision could be construed as domestication, in another sense it merely translates a typical French style into a typical English style and maintains a similar audience-text relationship.

(3) GUTENBERG:

A few minutes after the scene of confusion produced in the salons of M. Danglars by the unexpected appearance of the brigade of soldiers, and by the disclosure which had followed, the mansion was deserted with as much rapidity as if a case of plague or of cholera morbus had broken out among the guests. In a few minutes, through all the doors, down all the staircases, by every exit, everyone hastened to retire, or rather to fly; /S/ for it was a situation where the ordinary condolences,—which even the best friends are so eager to offer in great catastrophes,—were seen to be utterly futile. /P/ There remained in the banker’s house only Danglars, closeted in his study, and making his statement to the officer of gendarmes; Madame Danglars, terrified, in the boudoir with which we are acquainted; and Eugenie, who with haughty air and disdainful lip had retired to her room with her inseparable companion, Mademoiselle Louise d’Armilly. /P/ As for the numerous servants (more numerous that evening than usual, for their number was augmented by cooks and butlers from the Cafe de Paris), venting on their employers their anger at what they termed the insult to which they had been subjected, they collected in groups in the hall, in the kitchens, or in their rooms, thinking very little of their duty, which was thus naturally interrupted. /P/ Of all this household, only two persons deserve our notice; these are Mademoiselle Eugenie Danglars and Mademoiselle Louise d’Armilly.

One marker of informality in English is the use of contractions. GUTENBERG and BUSS take different approaches to this issue: as illustrated by a short excerpt in (4), GUTENBERG includes no contractions at all in Chapter 97 while BUSS has 37 instances—all occurring in character dialogue—in the same chapter. Contractions are perceived as more informal than uncontracted words (compare, for instance, don’t and do not). Furthermore, the use of contractions in dialogue more closely mimics everyday speech. The translator of BUSS may have used contractions to translate the informal pronoun tu ‘you’ in French, a distinction which is left untranslated in GUTENBERG. The informal register of contracted dialogue correlates to a more readable text for an audience used to similar speech patterns in original English-language novels.
a. GUTENBERG:

Louise pressed with all the strength of her little hands on the top of the portmanteau. “But I cannot,” said she; “I am not strong enough; do you shut it.”

“Ah, you do well to ask,” said Eugenie, laughing; “I forgot that I was Hercules, and you only the pale Omphale!”

b. BUSS:

Louise pressed with all the strength of her little hands on the lid of the trunk. “I can’t do it,” she said. “I’m not strong enough. You close it.”

“Of course,” Eugenie said with a laugh. “I was forgetting that I’m Hercules and you’re just a feeble Omphale.”

4.3 Censorship and societal norms

Chapter 97 of the novel is an ideal sample for close qualitative analysis in order to test the hypothesis that Victorian cultural norms led to a different translation of the relationship between Eugénie and Louise than its portrayal in the modern translation. Are they seen and interpreted as close friends driven by Eugénie’s reckless spirit, or does the timid Louise actually follow Eugénie out of romantic love?

One significant interaction between the two young women occurs near the beginning of the chapter when Eugénie reveals her plan to run away despite her impending marriage having been called off. Of particular note is Louise’s response to Eugénie’s plan; in BUSS (5b) her reply has a questioning tone, as if seeking confirmation, whereas in GUTENBERG (5a) the line begins with a surprised exclamation, “What?”, and continues in a doubting tone.

(5) a. GUTENBERG:

“What shall we do?” asked Louise.

[...]

“Why, the same we had intended doing three days since—set off.”

“What?—although you are not now going to be married, you intend still”—

b. BUSS:

“What can we do?” Louise asked.

[...]

“The very thing we should have done three days ago: leave.”

“So, as you are not getting married, you still want to?”
(6)  a. Gutenberg:

“You shall see,” said Eugenie. And with her left hand seizing the thick mass, which her long fingers could scarcely grasp, she took in her right hand a pair of long scissors, and soon the steel met through the rich and splendid hair, which fell in a cluster at her feet as she leaned back to keep it from her coat. Then she grasped the front hair, which she also cut off, without expressing the least regret; on the contrary, her eyes sparkled with greater pleasure than usual under her ebony eyebrows. “Oh, the magnificent hair!” said Louise, with regret.

“And am I not a hundred times better thus?” cried Eugenie, smoothing the scattered curls of her hair, which had now quite a masculine appearance; “and do you not think me handsomer so?”

“Oh, you are beautiful—always beautiful!” cried Louise. ”Now, where are you going?”

[...]

“What are you looking at?”

“I am looking at you; indeed you are adorable like that! One would say you were carrying me off.”

b. Buss:

“You'll see,” said Eugnie. And with her left hand she grasped the thick plait of hair which her slender fingers could barely reach around, while with the right she took a pair of long scissors. Very soon the steel blades were squeaking in the midst of the magnificent and luxuriant head of hair, which fell in tresses around the young woman’s feet as she bent backwards to prevent it covering her coat.

Then, when the hair on the crown of her head was cut, she turned to the sides, shearing them without the slightest sign of remorse. On the contrary, her eyes shone, more sparkling and joyful than usual under her ebony-black brows.

“Oh, your lovely hair!” said Louise, regretfully.

“Don’t I look a hundred times better like this?” Eugénie asked, smoothing down the few curls left on her now entirely masculine haircut. “Don’t you think I’m more beautiful as I am?”

“Oh, you are beautiful, beautiful still,” Louise cried. “Now, where are we going?”

[...]

“What are you looking at?”

“You. You truly are adorable like that. Anyone would say you were abducting me.”

The crux of their relationship is shown in (6), when Eugénie cuts off her long hair in order to disguise herself as a man. In both versions, Eugénie expresses only pleasure and satisfaction with her newly “masculine” appearance, while Louise mourns the loss of long hair. However, Louise hastens to assure Eugénie several times, using words like
beautiful and adorable (the same translations in both GUTENBERG and BUSS), that she is still physically attractive to the world and to Louise. The subtext in this passage, contrary to the original hypothesis of Victorian censorship affecting GUTENBERG, points toward the classic stereotype of butch-femme contrast in a lesbian relationship.
Chapter 5
Conclusion

Through a combination of corpus-based data analysis and qualitative textual analysis, this study has examined several hypotheses, including fluency as a translational strategy, foreignizing or domesticating choices, and the influence of societal norms on the production of a translated text. The resulting interpretations are grounded in both translation theory and corpus linguistics, as well as the intersection of these two fields in corpus-based translation studies.

5.1 Results

The quantitative analysis presented in Chapter 3, drawing upon techniques from corpus-based translation studies and natural language processing (NLP), finds consistent differences between GUTENBERG and BUSS. Comparing relative frequencies pinpoints words which are more frequent, relative to their own corpus, in one corpus than another; similarly, computing a best-fit constant for the Zipf distribution shows that the differing relative frequencies are systematic. Moreover, comparisons of type-token ratio and paragraph and sentence lengths support the hypothesis that BUSS favors fluency to a greater degree than GUTENBERG does.

These interpretations are further refined through a close reading, presented in Chapter 4, of Chapter 97 across all three corpora. The qualitative analysis confirms greater readability and fluency in the Buss translation, but this pattern does not necessarily correlate with the observed decisions between domestication and foreignization in the two contrasting English translations. With respect to the hypothesis of socially influenced censorship of a lesbian relationship in the novel, the analysis finds some differences in tone but an ultimately similar portrayal through subtext which allows readers to interpret the relationship platonically or romantically as they are inclined.
5.2 Future work

My analysis has been limited here to parallel corpora drawn from multiple editions of *The Count of Monte Cristo*. To facilitate further study of stylistic features, it would be worthwhile to construct a larger corpus of Alexandre Dumas père’s entire published oeuvre, as well as to manually align at least some parts of the corpora. I leave for future work the further application of tools such as part-of-speech taggers and lemmatization (grouping together all inflected forms of a word) from NLP research. For instance, NLP tools would provide more accurate results for sentence boundaries, permit a breakdown of the most common words by morphological or syntactic function, and enable a finer-grained analysis of word frequencies within particular part-of-speech categories.

Moreover, the quantitative analysis given here has left two significant avenues unexplored. My proposed interpretations would be strengthened by rigorous tests of statistical significance in the results. Additionally, the concordance software commonly employed by corpus linguists would provide a local-context-oriented perspective on the data through keyword in context reports and word collocations.

Additionally, my analysis has been grounded almost solely in Western translation theory. There remains a rich tradition of translation criticism in Asia, with particular historical depth in China, which unfortunately fell beyond the scope of this thesis and my literature review. For example, Hung and Wakabayashi (2005) present an excellent overview of Asian translation traditions. Similarly, He (2005) explains Chinese translators’ borrowing (and sometimes misuse) of Venuti’s domestication/foreignization terminology. It would be valuable to extend my qualitative analysis in light of this alternate theoretical tradition, perhaps from a comparative perspective.

In conclusion, this thesis would surely benefit from extensions of both a technical or a critical nature. I have striven to combine hard numbers and theoretical ruminations, giving each side approximately equal weight, and I hope that future work in corpus-based translation studies will do the same.
References


Dryden, John. 1680. From the preface to *Ovid’s Epistles*. In Venuti (2012). 38–42.


