TRANSLATION PROJECT

EREN JANBERK GENÇ

BOĞAZİÇİ UNIVERSITY

ISTANBUL

JUNE 2021

Seeking Traces of the Translator’s Invisibility

by

Eren Janberk Genç

Advisor: Şule Demirkol Ertürk

Submitted to the Department of Translation and Interpreting Studies in partial fulfillment of the requirement for the degree of Bachelor of Arts in Translation and Interpreting Studies

Boğaziçi University

Istanbul

June 2021

TABLE OF CONTENTS

Abstract

There have been many overarching theories throughout the history of translation studies that tried to make sense of the role of the translator and the place of translation within the contemporary world. Lawrence Venuti's theory regarding the invisibility of the translator is one such theory that aims at making sense of a large portion of the translational reality. Inspired by both Venuti's theory and his methodology, this paper attempts to find traces of the translator's invisibility in the literary reviews posted on a popular user-driven website known as Goodreads using computational methods. A small dataset of user reviews is scraped from the website, cleaned and analyzed in order to find evidence in support of certain claims derived from Venuti's specification of the phenomenon. A brief exploration of Venuti's ideas is followed by an overview of the methodological leanings of this paper and by a throughout elaboration of the technical specifications. Analysis results are presented and discussed using data visualization techniques, revealing that the phenomenon as presented by Venuti can also be traced in the user-generated translation reviews found on Goodreads. Lastly, a critical discussion of this paper's methodology and the results is offered.

Introduction

There have been many overarching theories throughout the history of translation studies that tried to make sense of the role of the translator and the place of translation. Lawrence Venuti's theory regarding the invisibility of the translator is one such theory that aims at making sense of a large portion of the translational reality. First published in 1986 (Venuti, 1986) as a journal article and then later as a book (Venuti, 2008), Venuti's claims about how the prevalent translational Anglophone norm of fluency, transparency and invisibility hampers the translator and the process of translation have attracted long-lasting attention. His book has been hailed by some as a "modern classic of translation studies" (Delabastita, 2010) and "the closest thing to an academic bestseller the discipline has seen in its recent history" (Delabastita, 2010). It can be claimed that this fame is well-deserved by looking at the wide variety of rigorous discussion, arguing both for and against his theory, that Venuti's work has brought into the field.

A strong side of Venuti's theory is the rigor and clearness with which he collects, analyses and presents evidence. The first chapter of his book, titled Invisibility, lays the groundwork for a deep multi-faceted historical analysis of the phenomenon of translator's invisibility. Venuti's clear framing and articulation of the evidence that he collects serves as a good example of evidence-driven humanities scholarship.

Inspired by both Venuti's theory and his methodology, this paper attempts to find traces of the translator's invisibility in the literary reviews posted on a popular user-driven website known as Goodreads. A small dataset of user reviews is scraped from the website, cleaned and analyzed in order to answer a set of questions derived from Venuti's explanation of the phenomenon. In answering these questions, this paper attempts to provide evidence (or lack of thereof) regarding the translator's invisibility. A brief exploration of Venuti's ideas is followed by an overview of the methodological leanings of this paper and by a throughout elaboration of the technical specifications. Analysis results are presented and discussed using data visualization techniques, revealing that the phenomenon as presented by Venuti can also be traced in the user-generated translation reviews found on Goodreads. Lastly, a critical discussion of this paper's methodology and the results is offered. All of the code and the data that is responsible for the analysis is hosted in the author's GitHub profile. The analysis code is written with the principles of reproducibility in mind, aiming for the easy replication and confirmation of the results.

The Translator's Invisibility

As explained in the second edition of Lawrence Venuti's book with the same name (Venuti, 2008), the translator's invisibility is the name of a translational phenomenon that occurs due to the acceptance of transparency as the determinant of value in translation. The translators who operate in a culture that shares this acceptance translate in a way that strips away all the traces of translational action. They strive to create easily readable texts by using methods such as writing in a standardized English, sticking to orthodox forms of syntax and bringing the source text as close to the target language as possible. In doing so, the translators make themselves "invisible." The result is a translation that does not read as a translation, but as if it were the original.

This illusion of transparency is demanded primarily by readers who value translations to the extent that it gives way to the original meaning (Venuti, 2008, pp. 1 - 2). These readers disregard any value that might be derived from stylistic features of the foreign text. To them, anything that might obfuscate the original meaning is an error. This bottom-up demand in turn influences the value judgements made by literary professionals such as editors, publishers, literary reviewers and even translators themselves. Such a widespread acceptance creates a self-reinforcing feedback loop of values in which the choice of translating with transparency is not even seen as a choice, but as the only way of translating.

According to Venuti, the translator’s invisibility is not just a curious occurrence. It has far reaching consequences that effect the material world of the translator and the broader world in general. Venuti claims that the norm which demands the translators to be left in the dark directly affects the monetary value and the recognition that they get for their work. The obfuscation of the “foreign” and the valorization of transparency also results in some kind of a cultural oppression in which other forms of fluency and value are disregarded and erased.

Venuti’s seminal work has motivated a substantial amount of subsequent research within the translation studies community. His claims have been tested against scopes different from what he has originally presented. Some studies seek the traces of the translator’s invisibility within specialized contexts such as the translation of environmental texts in the Catalan language (Bracho Lapiedra & MacDonald, 2017), the socioeconomical status of Danish in-house translators and its relation to visibility (Dam & Korning Zethsen, 2008) and the visibility of interpreters (Wadensjö, 2008).

Venuti’s work is important not only because of its substance, but also because of the way in which it was conducted. In the first publication that he talks about the translator’s invisibility (Venuti, 1986), Lawrence Venuti chooses to follow a highly empirical approach in elaborating on his broad findings. This tradition continues in the re-iterations of his thesis. The objects of his study and the findings that he presents as evidence are not obfuscated but meticulously defined. A representation of the body of the evidence that Lawrence Venuti suggests can be drawn from the 2008 explanation of his thesis, found in the book The Translator’s Invisibility: A History of Translation. There, Venuti explicitly states that he assembled evidence by sampling reviews from newspapers and periodicals that date to at most sixty years back. His sample includes reviews written by “noted critics, novelists and reviewers.” (pp. 2 - 3). The publications whose reviews he sampled includes literary magazines as well as publications aimed for mass consumption. The gamut of the literary works that the reviews address is also explicitly stated: they are all related to works of fiction who have been written by either European or Latin American writers. The works of fiction represented range from novels to short stories, written in different styles such as “realistic and fantastic, lyrical and philosophical, psychological and political.” (pp. 2 - 3). The 2008 version of the list of the books and authors that the reviews he has covered refers to can be found in Appendix A. The corpus collected for the purpose of this paper includes some of the books and authors related to Venuti’s original list as well.

The strength of the argument for the existence of the translator’s invisibility comes not only from a well-picked and carefully communicated sample, but also from the way in which Venuti dissects his sample. The claim that the Anglophone culture values fluency and transparency over any other literary value is supported primarily by the adjectives and phrases Venuti’s selected reviewers use to describe translations and translators. Appendix B presents a near comprehensive list of the descriptions that the Venuti has found in the 2008 iteration of his study.

Venuti’s evidence-driven approach has inspired a series of translation studies scholarship done on the topic of the translator’s invisibility with a corpora-driven and/or quantitative focus. There are some studies that seek to replicate his findings in other contexts by adopting a similar methodology. The research on the invisibility of the translator in the translation of environmental texts (Bracho Lapiedra & MacDonald, 2017), previously mentioned in this paper, is one such study that employs a corpora-based methodology. Some research even employs more complicated statistical methods to look at the presumed effect of the translator's invisibility on notions such as economic gain, job satisfaction, social capital and happiness (Liu, 2013)

Venuti's mindset can be considered as an early manifestation of a new type of humanities scholarship that has been gaining traction for the last twenty years (Jones, 2016). The methodologies that are championed by this type of humanities scholarship exist under different umbrella names such as "corpus studies", "humanities computing" and "digital humanities." This paper is likewise inspired by the empirical mindset adopted by Venuti in his work and attempts to make use of computational methods in order to present an empirically backed claim. Before talking about the exact methodology and the results of this study, a brief exploration of such computational methods, their history and the reasoning behind their use is needed in order to fully understand the nature, the advantages and the implications of such an outlook to humanities scholarship.

Digital Humanities: Definition and History

Classically, the preferred method of research and interaction for the humanities scholars that concerned themselves with the study of texts was through direct reading. To be more specific, it was a technique called "close reading." Close reading is a form of literary criticism that was developed around the first decades of the 20th century (Hawthorn, 2000). It stands for reading and re-reading in order to uncover different levels of meaning such that a more thorough understanding of the text at hand is reached (Boyles, 2016). The means of reaching such an understanding would be familiar to anyone who has spent any time in literary studies departments: During close reading, the readers "linger over words, verbal images, elements of style, sentences, argument patterns and entire paragraphs and larger discursive units." (James, 2001, p. 93). Close reading may be falsely identified as being "the reading" method.

However, it is but a method of reading (albeit a popular one) that has to be taught beforehand (Jay, 2014, p. 133).

The method of close reading has served humanists well over the years but it is not without its shortcomings. A major problem associated with close reading is the problem of scope: there are simply too many texts and cultural objects to study (Moretti, 2000). This has been true nearly since the end of the eighteenth century, and it is increasingly true now (Wilkens, 2012). The abundance of material to be combed through has created a phenomenon called “the canon” (Wilkens, 2012). When there are too many texts to go through, the scholars often default to focusing on a few works that are deemed important and representative, and then to base all analysis and deductions on the said canon.

From an empirical standpoint, making large-scale claims about the nature of a genre or a period based only on a specific cannon is dubious at best. Even if the scope of research is narrowed down to a single text, perhaps in a study of authorial or translational style, there seems to be an apparent sampling bias. Without establishing a robust procedure for sampling specific textual phenomena or carefully going through the whole of the text in a very close and machine - like manner, any claims made about the nature of the text at hand would be resting on shaky ground. (Rommel, 2004).

The exact nature of this problem is best explained in Moretti’s own words:

> “…the point is that there are thirty thousand nineteenth-century British novels out there, forty, fifty, sixty thousand—no one really knows, no one has read them, no one ever will. And then there are French novels, Chinese, Argentinian, American… Reading ‘more’ is always a good thing, but not the solution. “ (Moretti, 2000, p. 55)

It is apparent that close reading does not serve humanities scholarship well if the aim is to make claims about larger bodies of humanistic phenomena spread over longer periods of time.

It might falter the scholar even within the tight borders of a single text. However, close reading is so internalized within the humanities circles that it is the first thing that comes to mind when one talks of humanities scholarship. It can be speculated that this internalization is not due to choice, but due to the fact that reading and re-reading was the only way through which scholars could meaningfully and rigorously access the contents of a text up until the latter half of the twentieth century when other modes of reading were devised (Jay, 2014, p. 138).

The latter half of the twentieth century is marked with a rapid growth of computers and computing capacity that we are in the midst of even today. The Moore’s Law, an observation that lies underneath this growth, helped replace room-sized computers with much more smaller and much more powerful computing devices (Keyes, 2006, p. 25). The advances in hardware have been closely followed by advances in software. The computers that we have today, even though the software that they use has been written in much less computationally efficient higher-level programming languages (Ousterhout, 1998, p. 30), allow their users to do creative activities that would have been impossible to do with a computer forty years ago.

The combination of powerful hardware and readily available software has created such a catalyst that now, it is claimed that “software is eating the world.” (Andreessen, 2011). Many of the activities that were done manually before the advent of the computers is now done through computers. Similarly, many of the artifacts that existed in the physical world are now readily available in the digital one. Text is no exception to this.

The digitization of text not only brought new methods of access and dissemination, but also new methods related to the study of it. Text as represented within a computer could now be studied with methods other than close reading. Since text is just like any other data to the computer in which it is stored, it can be subject to nearly all kinds of analysis and inquiry that other types of data are subject to. Just as a computer can generate millions of numbers and find patterns in a comparatively short time, so can it parse through text (a whole genre or a single text) and reveal insights that would be tedious or outright impossible to reveal manually.

A Preliminary Definition

The practices and methodologies that revolve around using computational methods to gather, process, analyze and present textual data represent an alternative to the classical method of close reading (Jänicke et al., 2015). Humanistic research that is done through such methodologies has been called with many names over the years. Literary computing, humanities computing, digital humanities (Kirschenbaum, 2012) and computational humanities (Henrickson, 2019) are just some of the names that have been used.

Currently, the convention adopted by a diverse set of practitioners, institutions, journals and universities is using the name "digital humanities" (Vanhoutte, 2013). This paper will henceforth comply with this convention and use the name "digital humanities." Loosely defined, digital humanities is an academic discipline (Rockwell, 2013) that "… refers to a set of methods and projects that investigate how the pairing of the terms 'digital' and 'humanities' extends one another." (Sheridan, 2016, p. 4).

The definition above does not explicitly talk about close reading or the analysis of text through digital methods. In fact, it seems to include much more than just the computational study of text for humanities research. The reasons behind this apparent lack of emphasis can be found once the name shift from "humanities computing" to "digital humanities" is tracked. There is a reason deeper than the whims and the spirit of the time that lies behind the change of naming that occurred throughout the years. This change is the apparent marker of a deeper philosophical and methodological underpinning which is reflected in the definition of digital humanities we have provided (Svensson, 2013, p. 160). In order to fully appreciate what digital humanities means and to understand the methodological basis of this paper, a careful tracking of the field's history and the name changes that it went through is needed.

History of Digital Humanities

If we take the first use of computational methods in the analysis of textual data for the sake of humanistic study as the starting point of this discipline, then we can say that a clear birth date can be identified. In 1949, an Italian Jesuit priest by the name of Roberto Busa set out to investigate the lexical nature of St. Thomas Aquinas' work. Having heard of computers, he traveled to the United States and sought the help of the IBM for this ambitious project. Over many years, Busa worked with the IBM to complete and re-complete iterations of his work in which he created concordances (a list and a count of words and their co-occurring words) of Thomas Aquinas's texts (Hockey, 2004, p. 4).

Following in Father Busa's footsteps, select literary scholars around the world started to use such methodologies and mustered computing resources to create concordances of their own (Vanhoutte, 2013, p. 126). Although most work done at this time served as being proofs-of-concept, concordances were ultimately used in the automatic authorship attribution of texts. This way, computational methods were used to solve a humanities-related problem that dated back to 1851 (Hockey, 2004, p. 5). No matter how revolutionary it was, the work done in the 50's and 60's was extremely laborious and time consuming. Computing was not yet a common resource and the ways of inputting and processing data were primitive (Rommel, 2004, p. 92).

In the 70's and 80's, using computational methods in the analysis of textual data for humanistic purposes became more widespread. The best evidence regarding the increasing popularity of such methods can be found in the increase of the number of conferences, journals and associations that began to pop up in both sides of the Atlantic, in the USA and in the UK. Over time, these journals and associations started to also accommodate other fields of humanities such as art history and archeology. The history of corpus linguistics is also intertwined with these advances, as linguists frequently contributed to and worked with these associations (Hockey, 2004).

As institutional adoption of such methods became widespread, a scholarly consensus around what to call these practices were formed. Starting from mid-eighties, the name "humanities computing" could be found in the title of many teaching programs, computing centers, journals and academic publications. (Vanhoutte, 2013). Although there was much debate around the name, it ultimately stuck and the field was called such until around 2004 (Vanhoutte, 2013, pp. 119 - 120). During this period, the research emphasis was on using computers and computational methods to model humanities data and our understanding of it (Unsworth, 2013). There was a clear tool – object relationship in the field such that information technology was seen as a tool and the text was seen as an object of study (Svensson, 2013, p. 165). This clear-cut focus on a singular object and the set of methods to study it would later become the main distinction point between humanities computing and digital humanities.

Although computational methods were gaining traction and piquing the interest of institutions and organizations, they were still far away from the mainstream. One reason was the fact that the computers were still a wildly expensive resource. This all changed with the advent of the personal computers in the middle of the 80's. From 85 and onwards, computational methods could now be used by humanities scholars with less resources. This led to a proliferation of new practitioners, and from that proliferation came discipline - wide projects like the Textual Encoding Initiative (TEI) which sought to create a standard around the representation of text data (Hockey, 2004).

The era of personal computers was shortly followed by the advent of the Internet. the Internet has transformed both the digital world and our physical reality thanks to the high-speed dissemination of information that it enables over long distances (Leiner et al., 2009). Although the Internet was always built with interactivity and connectivity in mind, it can be argued that it became even more interactive and connected over time thanks to a set of protocols, ideas and software that have been implemented over it. The Web 2.0, an umbrella term which refers to a specific set of technologies and the philosophy that they embody (Davidson, 2008), has played a special role in this progressive improvement.

The Web 2.0 can be characterized by applications and practices which rely on user participation, such as social media sites, multimedia sharing, blogs, RSS feeds, wikis and podcasts (Andersen, 2007). The combination of these technologies and practices with the ideas of individual production, user generated content, openness and wide-scale participation (Andersen, 2007) has made visible and amplified many old and new forms of media. It also opened up new ways of authorship, collaboration, presentation and research for digital humanities scholars (Sheridan, 2016, p. 4).

With the gradual proliferation of the Web as an everyday tool instead of a “fringe” nerdy activity, many culture – building activities started being transported to the digital world and the web. Novels could now be written on the computer and self-published on the Internet. Music and audio could be hosted online and remixed on various platforms. Visual arts were getting increasingly blended in with the digital. These forms of cultural creation were quickly joined by “digital – born” modes of culture creation such as games, short-form writings (blogs, tweets), data visualizations and audiovisual mashups. Humanities scholars that wanted to study these digital-infused or digital-born methods of cultural creation through classical humanistic methods or through the methods cultivated by humanities computing found themselves a home within the humanities computing circles in the academia (Kathleen, 2012,

p. 12 & p. 14). Thus, humanities computing became associated with a new and theory - focused field of inquiry in which information technology was not seen as a tool, but as the object of study.

In a parallel process, the proliferation of personal computing, the advent of the Internet and the massive boost of computational capacity aroused an increased desire to develop methodologies similar to those developed by the humanities computing core for textual applications. Although objects of inquiry other than the text were always present, they were considered to be fringe (Svensson, 2012). With this boost in desire and the capacity to work with data other than text, other humanities fields found themselves looking for seats alongside the table of humanities computing (Kathleen, 2012, p. 13). This expansion of focus also brought along a set of scholars that concerned themselves with building tools and working on novel methods that aimed to augment the research capacities of other humanities scholars (Laubichler et al., 2016).

The inclusion of these venues of scholarly inquiry and research into humanities computing prompted a desire to rename the field in order to better accommodate the expanded practices that the scholars interested themselves with. Thus, starting from 2004 and onwards, the field came to be known as digital humanities.

A Revised Definition

It is now possible to extract a clearer interpretation of the digital humanities definition that we have provided above. In saying that digital humanities investigates "…how the pairing of the terms 'digital' and 'humanities' extends one another" (Sheridan, 2016, p. 4), we actually mean one or more than one of the following:

- Digital humanities is the study of digitally-represented humanistic phenomena (including but not limited to text) using digital and computational methods.

- Digital humanities is the study of digital – infused or digital – born methods of cultural creation using humanistic methods.
- Digital humanities is the formulation, design, development and the criticism of computational and digital – driven methods for the sake of humanities research.

Digital Humanities and Translation Studies

Translation studies as a field of inquiry stands in a unique spot because its main focus translation enjoys the honor of being one of the first post-WW2 scientific problems tackled by the usage of computational methods (Vanhoutte, 2013, p. 122). In the 50's and 60's, the United States government funded various machine translation projects with a focus on the automatic decoding of Russian and German documents (Martin - Nielsen, 2012, p. 67). Much effort and funding went into this task. However, the results were subpar. In 1966, the infamous ALPAC report (Pierce et al., 1966) concluded that a successful and cost - effective machine translation system for the unaided translation of scientific texts was not possible with the technology of the time (Pierce et al., 1966, p. 19). Although the results obtained by the projects of that era were considered to be unsatisfactory, the pioneering effort that went into the task of automatic machine translation sparked an interest in various academic disciplines like computer science, artificial intelligence, natural language processing and computational linguistics (Poibeau, 2017).

These early attempts of tackling translation through computational means did not regard translation itself as an object of study. The effort that went behind making a machine speak another language was entirely driven by practical reasons, namely wanting to decode Russian messages and documents in the heat of the Cold War (Pierce et al., 1966).

Using computational methods to gain insights about the notion of translation itself came much later than the publishing of the ALPAC report. In fact, mainstream academic interest in corpus – based translation studies can be traced back to 1993 (Laviosa, 2011, p. 13).

In 1993, Mona Baker published an article titled *Corpus linguistics and translation studies: Implications and applications* (Baker, 1993). In this pioneering article, Baker defined and presented the concept of using large-scale corpora to study the phenomenon of translation for the sake of gaining deeper insights. The potential advantages of corpus methodology over other methodologies for descriptive translation studies that Baker lists is very reminiscent of the advantages that humanities computing scholars have stated over the years. To her, basing research on corpora allows the researcher to make empirically stronger claims over a larger portion of the translational reality. It might even allow the theorist to further chase what one might call translation universals (Baker, 1993, p. 242). In her own words:

> “Large corpora will provide theorists of translation with a unique opportunity to observe the object of their study and to explore what it is that makes it different from other objects of study, such as language in general or indeed any other kind of cultural interaction. It will also allow us to explore, on a larger scale than was ever possible before, the principles that govern translational behaviour and the constraints under which it operates. Therein lie the two goals of any theoretical enquiry: to define its object of study and to account for it.” (Baker, 1993, p. 235).

Later on, Baker also laid down the theoretical groundwork around corpus – based translation studies by providing definitions and explaining key metrics (Baker, 1995). In the same work, she even correctly prophesied that the corpus – based approach to translation would prove to be fruitful for cases such as technical translation and the betterment of automated translation systems.

The kind of approach that Baker introduced and many other translation scholars took up can be categorized as being similar to the humanities computing approach in which text is the main study object. Indeed, many projects attempt to reach new insights about the notion of translation through the product of the translation process itself. It seems to be that most projects that lie on the intersection of digital humanities and translation studies follow this vein. However, the umbrella of digital humanities allows for more varied methods of study that contribute to the discussion that occurs where the digital and the translation meets. For example, investigating the phenomenon of "fan subbing" (Lee, 2011) can be considered as being within the scope of digital humanities because the object of study is a digital-born practice. Similarly, designing novel data visualization forms (Alharbi et al., 2020) and/or creating custom-made tools (Cheesman & Roos, 2017) that help translation studies scholars in engaging with research questions in a digital – driven manner can also be considered as digital humanities research.

Methodological Justification

Where Venuti Stands

Whether the name is humanities computing, digital humanities or corpus – based translation studies, one thing is common: all these methodologies possess in their core a focus on conducting verifiable large-scale research based on empirical data and not on intuitions or perceptions (Biber et al., 1994).

In a way, it may seem that digital humanities methodology is the manifestation of a desire that was always present within certain fields of humanistic inquiry. Scholars in translation studies cited compiling corpora and manually searching them as being valid empirical research methodology before the field of computational linguistics took up the term and added the connotation of being digitized to it (Baker, 1995, p. 225). With their addition, what was once “small collections of text which are not held in electronic form and which are therefore searched manually” (Baker, 1995, p. 225) became “any collection of running texts … held in electronic form and analyzable automatically or semi-automatically” (Baker, 1995, p. 226).

Under this light, it is important to once again reiterate that Lawrence Venuti’s work on the invisibility of the translator can be considered as a precursor to digital humanities scholarship in translation studies. Venuti’s work is based on manual sampling and retrieval from a clearly delimited corpus of reviews. The data that he presents is reminiscent of a concordance search one might conduct using computational methods: it is in the form of excerpts (text data with context). In these excerpts, the emphasis is placed on the modifiers (adverbs and adjectives) describing the translator or the translation. If Venuti was to conduct his research on a larger corpus using computational methods, his research would have been considered as being a textbook example of humanities computing.

Where This Paper Stands

With the history and meaning of digital humanities established and after emphasis is drawn on where Venuti’s work stands in relation to this tradition, the exact methodological background of this paper can be stated. This paper uses digital humanities methodology to tackle a problem that was originally studied in digital humanities-esque manner. It can be seen as a work of digital humanities scholarship because of the following reasons:

- It uses computational methods to collect, clean and analyze textual data.
- It presents its findings with the help of data visualization: a non-classic method of presenting humanities scholarship.
- It reflects on the advantages, disadvantages and the potential shortcomings of the computational methodology that it has adopted.

Now that the exact reasoning and the methodological background supporting this paper is stated, the study itself can be described.

Research Question

Before talking about the object of study, the results and the process through which those results were obtained, a clear articulation of the questions that are directed to the corpus compiled for this paper is needed. As this paper seeks to find traces of the translator's invisibility over a larger and differently scoped corpus, the original claims regarding the existence of the translator's invisibility might serve as useful guidelines.

In the 2008 version of Venuti's book *The Translator's Invisibility,* a passage in which the textual trails of the translator's invisibility are highlighted is present:

> *On those rare occasions when reviewers address the translation at all, their brief comments usually focus on its style* [emphasis added], neglecting such other possible questions as its accuracy, its intended audience, its economic value in the current book market, its relation to literary trends in English, its place in the translator's career. *And over the past sixty years the comments have grown amazingly consistent in praising fluency while damning deviations from it*…[emphasis added] (Venuti, 2008, p. 2)

Three explicit claims can be extracted from this passage:

1. In translation reviews, reviewers do not directly address the translation as often as they address the original.
2. When they do address the translation directly, their comments focus on the style of the translation and rarely on anything else.
3. A "fluent" translation is preferred over everything else.

In order to test these claims, they need to be operationalized into more concrete questions whose answers fully or partially cover the claims they originate from. From these three claims, the questions through whose answers this paper tries to find the trails of the translator's invisibility can be formulated:

1. Is the author, the original book and the act of writing more often addressed by the reviewers than the translator, the translation and the act of translation?
2. What are the most common modifiers (adverbs and adjectives) that are used to talk about the author, the original book and the act of writing when compared to those used to talk about the translator, the translation and the act of translation?
3. What is the sentiment value (positive / neutral / negative) of the most common modifiers that are used to talk about the tokens described above? Are the modifiers that modify the author, the original book and the act of writing more positive overall than those that modify the translator, the translation and the act of translation?

The first question maps directly to the first claim made by Venuti. Knowing the exact number of times the translator, the translation or the act of translation is mentioned and comparing those numbers to the number of times the author, the book and the act of writing is mentioned can conclusively prove or disprove the first claim. The second question wholly covers the second claim and adds more. By looking at the modifiers that are used to describe each token of interest, one can learn not only if the focus is on the style or not but also what the focus is on if it is not on the style. The third question does not directly map to the third claim but seeks to expand on the answers of the second question by trying to discover the value attributed to the results of the second question.

Dataset Collection and Cleaning

Dataset Collection

Although Venuti based his corpus on reviews coming from literary journals, magazines and newspapers, this paper utilizes an alternative source of translation reviews. The reviews analyzed in this study were drawn from an online social platform called Goodreads.

Goodreads is a social book cataloging website centered around books and reading. In Goodreads, users can create their own personal account and use it to leave reviews and comment on books. Being a book cataloging website, it also supports features such as reading list creation and curation. In fact, there are many book lists created both by the users themselves and the editorial team employed by Goodreads. Launched in 2007, Goodreads claims to be “the world’s largest site for readers and book recommendations.” (*About Goodreads*, 2021). The social platform is currently owned by the internet giant Amazon who acquired the website in 2013 (Olanoff, 2013). It is said that Amazon, the biggest book vendor in the world, bought Goodreads to complement the services it offers around books and e-books and to strengthen its book sales by providing Amazon users with a book-related social media platform (Vinjamuri, 2013).

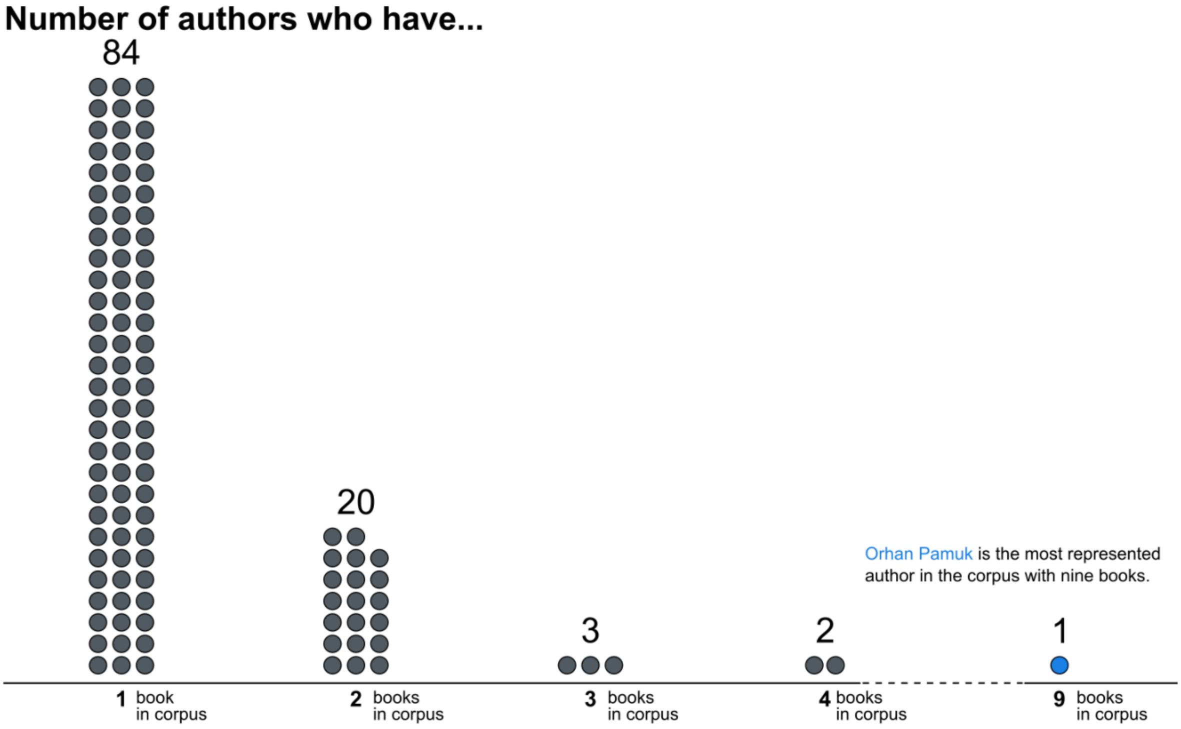
The main reason why this study decided to compile a corpus from Goodreads rather than computationally scanning print-based media is ease of access. Goodreads stands as a centralized place where many book reviews are located. This centrality allowed for the implementation of a computational strategy through which a large number of reviews could be gathered with minimal manual interference needed. If the reviews were to be drawn from disparate online sources, the computational strategy would have had to be modified for each different source.

A total of 150 books to gather reviews from were selected. As the total compute time of the analysis increased in proportion to the number of books investigated, the number of books were capped at 150. The full list of the selected books and their authors is presented in

Appendix C. While the majority of the authors were represented with one book in the corpus, some authors had more than one book featured. The dot plot below displays a categorization of authors based on the number of books they are represented with in the corpus.

Figure 1

Categorization of the authors in the study based on the number of books they are represented with in the corpus.



Note. A dot plot showing the categorization of authors based on the number of books they are represented with in the corpus. The x axis (bottom axis) omits the values between four and nine because there are no authors present with that many books. 84 authors are represented with one book, 20 authors are represented with two books, 3 authors are represented with three books and 2 authors are represented with four books. Orhan Pamuk, a Nobel prize – winning Turkish novelist, is the most represented author in the corpus with nine books.

In Venuti’s study, the criteria for the selection of books are explicitly stated. The books that he selected for his study were non-English fiction books written by European and Latin American writers, justified on the grounds that these nationalities were “the most translated into English.” (Venuti, 2008, p. 2). Similar criteria were adopted in the selection of books for this study. All the books whose reviews were analyzed are non-English fiction books of different narrative styles. There exists a difference in the range of nationalities represented in the dataset: authors from the Middle East, Africa and Asia were included alongside authors from Europe and Latin America. The authors whose books Venuti included in his corpus and the very books that he investigated are also represented in the corpus presented by this paper to a degree. These authors and books are highlighted in the table presented in Appendix C. A further elaboration on the sampling method adopted by this study can be made. The exact books selected for this study were taken from various reading lists found on Goodreads. The reading lists bore titles such as “Top Fiction Translated into English.” These lists were taken as a proxy of the metric popularity.

A web scraping script written in Python was used to target the 150 books and to gather the reviews from them. Python is the programming language in which the whole project is written. It is a high-level, general-purpose programming language (*General Python FAQ — Python 3.9.5 Documentation*, 2021, "What is Python good for?" section) that is in wide use today. In fact, Python ranks among the top three most popular programming languages according to different indexes (*PYPL PopularitY of Programming Language*, 2021; *TIOBE - The Software Quality Company*, 2021). This popularity is well deserved because Python boasts a vibrant ecosystem of programming libraries through which users can handle many different tasks, including those relevant to digital humanities methodology such as textual analysis, data visualization and geospatial mapping. It is especially accessible to digital humanities scholars and students because it is considered to have a natural language-like, straightforward syntax.

Everything from the initial web scraping to the cleaning, processing, testing, analysis and the preliminary visualization of data is handled purely or partially in Python for this project. The full technical specification and the reproduction guide can be found on the GitHub page in which the source code is hosted.

Dataset Cleaning

The data whose analysis results are presented in this paper was scraped on the date of 18/04/2021. After the scraping process, a dataset of 49,274 reviews drawn from 150 books was obtained. Metadata about the reviews such as scraping date, review date, reviewer name, reviewer ID and rating was also collected. Additional metadata that enabled the tracking and testing of the reviews as they went through various states of processing and analysis were also added to the dataset. However, not all the reviews in the first iteration of the corpus were subject to analysis. After the initial scraping, the corpus went through a multi – stepped process of data cleaning, data quality testing and data processing. During the iterative process of cleaning and testing for data quality, some reviews were deemed unfit for the data quality standards adopted for this project. The reviews that were dropped from the initial corpus are documented in a separate folder which is also present in the GitHub page on which the source code of this project is hosted. The following is a brief summary of the reasons why certain reviews were dropped:

1.) Some reviews were found to contain data that was corrupted, wrongly encoded or outright missing. Reviews where the review data itself or the accompanying metadata was found to be faulty were dropped from the dataset.

2.) Even though the scraping script intentionally targeted English reviews only, some reviews were found to be written partially or completely in languages other than English. In such cases, the parts of the review that were written in languages other than English or the entirety of the review were dropped.

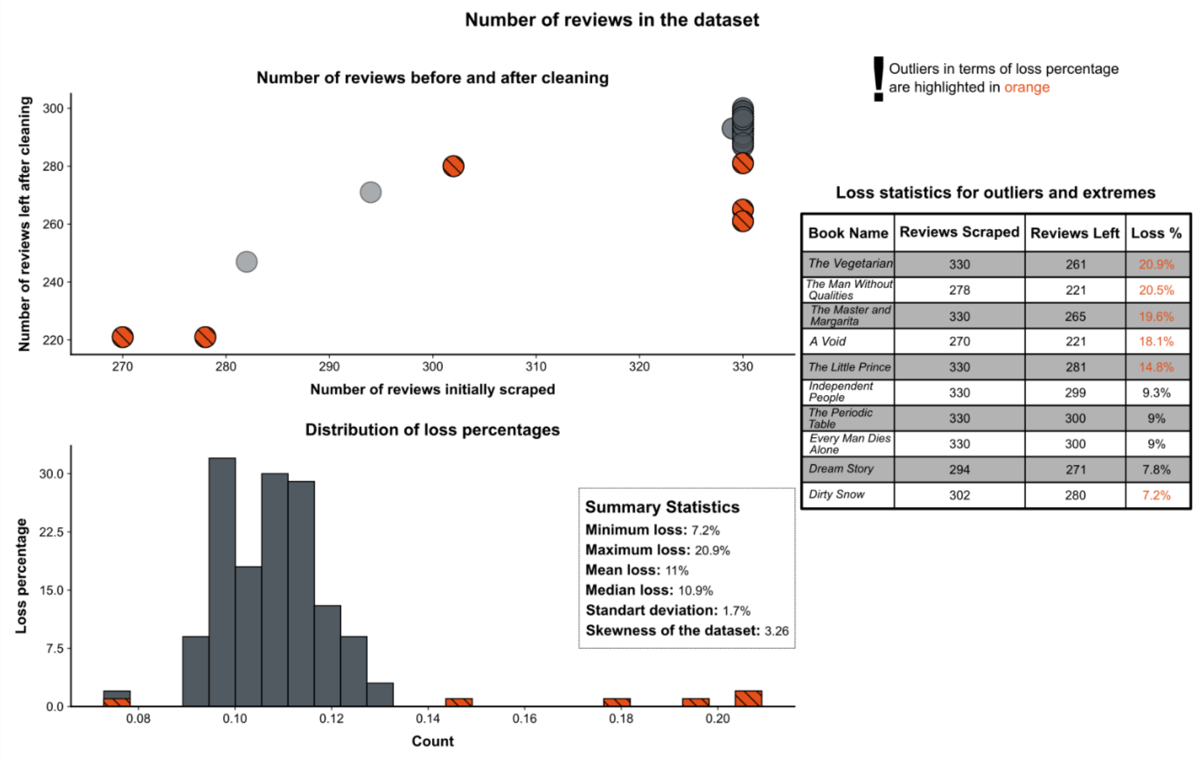
3.) Some reviews were detected to be duplicates of the reviews that already existed in the corpus. The duplicate reviews were dropped and only one copy of the reviews whose duplicates were found was left.

4.) Some reviews and/or portions of reviews were found to be too short to offer any meaningful insight when subject to computational analysis. The cutoff point was three words. The reviews or the part of the reviews that fell below the cutoff point were dropped from the original dataset. The cutoff point was justified based on English sentence structure. It was determined that a sentence needs to be at least three words to encode any meaningful modifier (adverb or adjective) information, such as “a good book” or “a great translation.”

After the iterative cleaning process, a dataset of 43,849 English reviews consisting of 433,828 sentences (or 9,423,988 words) drawn from 150 books was obtained. This means that 5,425 reviews were dropped. Although the degree of loss varied among the books whose reviews were scraped, all selected books were still represented in the corpus after cleaning. The maximum number of reviews left for a single book after cleaning was found to be 300, whereas the minimum was found to be 221. The small multiples visualization below offers details on the variance of loss due to data cleaning.

Figure 2

A small multiples plot offering details about the reviews before and after the cleaning process.



Loss statistics for outliers and extremes

Book Name	Reviews Scraped	Reviews Left	Loss %
The Vegetarian	330	261	20.9%
The Man Without Qualities	278	221	20.5%
The Master and Margarita	330	265	19.6%
A Void	270	221	18.1%
The Little Prince	330	281	14.8%
Independent People	330	299	9.3%
The Periodic Table	330	300	9%
Every Man Dies Alone	330	300	9%
Dream Story	294	271	7.8%
Dirty Snow	302	280	7.2%

Summary Statistics
Minimum loss: 7.2%
Maximum loss: 20.9%
Mean loss: 11%
Median loss: 10.9%
Standart deviation: 1.7%
Skewness of the dataset: 3.26

Note. A small multiples plot consisting of a scatterplot, a histogram and a table. The scatterplot (upper left, with dots) plots the number of reviews before and after cleaning. It can be clearly seen that 330 reviews were the maximum number of reviews obtained for many books. The histogram below the scatterplot shows the distribution of loss percentages. Some outliers in terms of the percentage of books lost during the cleaning process were observed. These outliers are highlighted with the color orange across all visualizations in this figure. The small table to the right of the figure offers a closer look at the outliers and the extremes.

It is apparent from the visualization presented above that the maximum number of reviews initially drawn was 330 for many books whereas the maximum number of reviews remaining after cleaning was 300. This occurrence requires a special explanation: the front-facing interface offered by Goodreads displayed only 330 reviews at maximum for each of the books represented in its database, no matter what the actual number of reviews was. This placed a hard limit on the number of reviews per book that could not be surpassed. The potential implications of this hard limit is discussed in the Discussion section.

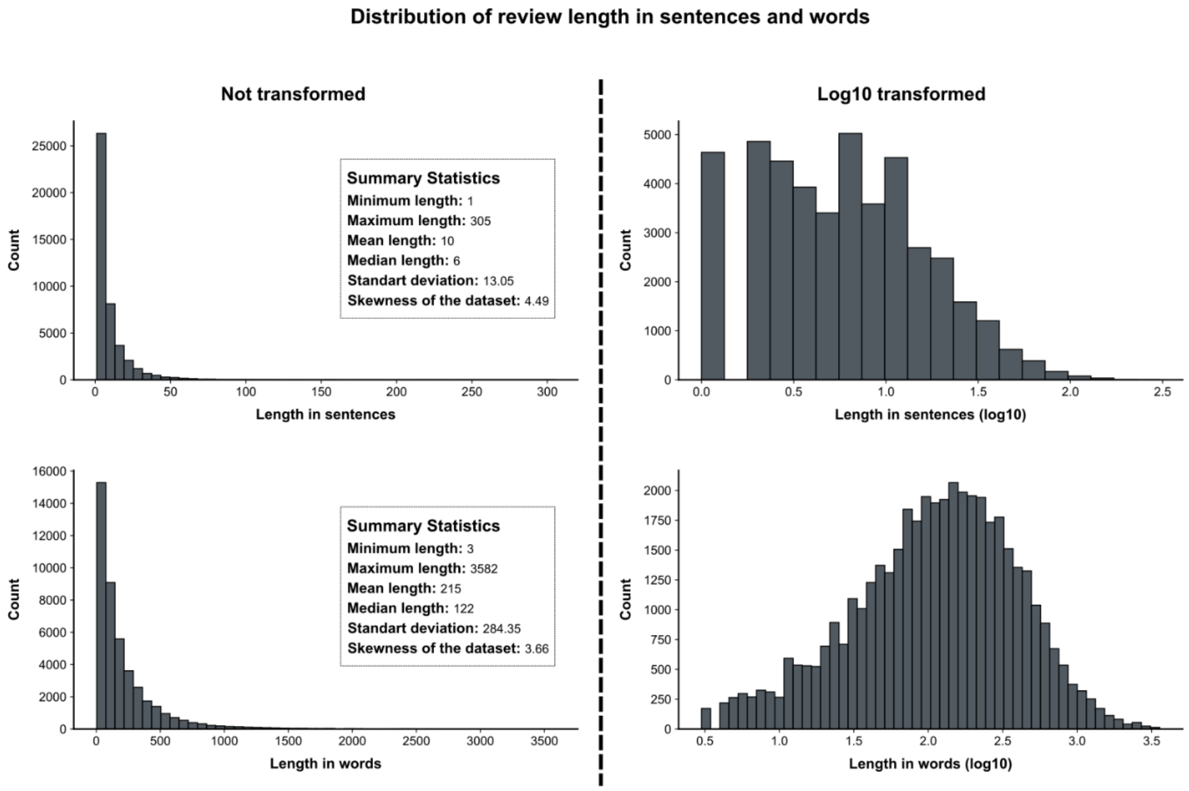
A mean loss of 11% was observed due to the process of cleaning the dataset. However, outliers[1] that lost far more or far less reviews exist. There are six books that can be considered as outliers in terms of loss percentage. Five of them can be considered to have lost more books than the average with 14.8%, 18.1%, 19.6%, 20.5% and 20.9% loss. One book is observed to have lost far less reviews than the average with 7.2% loss. The lowest and the highest of these outliers also represent the minimum and the maximum percentage of books lost. The distribution of loss percentages is heavily skewed to the right with a skewness of 3.26. This signals that although there are some extremes that deviate from the mean (more losses than average), the majority of the books suffered lower than around 12% loss as a result of the cleaning process. Considering the number of reviews left after the cleaning process, the remaining data was deemed enough for the scope of this analysis.

Another measure which was used to describe the dataset was the length of each individual review. Review length was calculated on both the sentence level and the word level.

Figure 3

Distribution of review lengths, calculated in sentences and words.

[1] As per common statistical practice, any values that are 1.5 IQR (Interquartile Range) more than the end of the third quartile of the dataset and any values that are 1.5 IQR less than the start of the first quartile of the dataset are considered to be outliers.



Note. A small multiples plot consisting of four histograms. The histograms on the left column show the non-log transformed distributions of review length. The histograms on the upper row show the non-log transformed and log transformed distribution of review length in sentences whereas the histograms on the lower row do the same for review length in words.

The distribution of review lengths was found to be heavily skewed to the right in both length in sentences and length in words. In fact, the skewness of the distributions was calculated to be 4.49 and 3.66 respectively. The histograms on the left column of Figure 3 are not able to demonstrate much detail because of this skewness. The histograms on the right column show more detail about the shape of the distribution thanks to a Log10 transformation. Nonetheless, on average the reviews were found to be 10 sentences and 215 words long whereas the extremes could go up to 305 sentences and a stunning 3582 words. Although the distribution of review lengths were found to be extremely skewed, this was not deemed as a potential problem. The reviews themselves were not taken as the actual unit of analysis. Since they were split up into individual sentences and then further into abstract syntax trees and token –

modifier pairings, review length was not determined to be an important metric for the outcome of the analysis.

Data Processing and Analysis

Data Processing

The cleaning processes described above were performed before or after several steps of data processing. The reviews as they were scraped from the Goodreads website were not suitable for analysis that would reveal the answer to the questions directed towards the dataset. Thus, they were subject to a series of computational manipulations called as “data processing.”

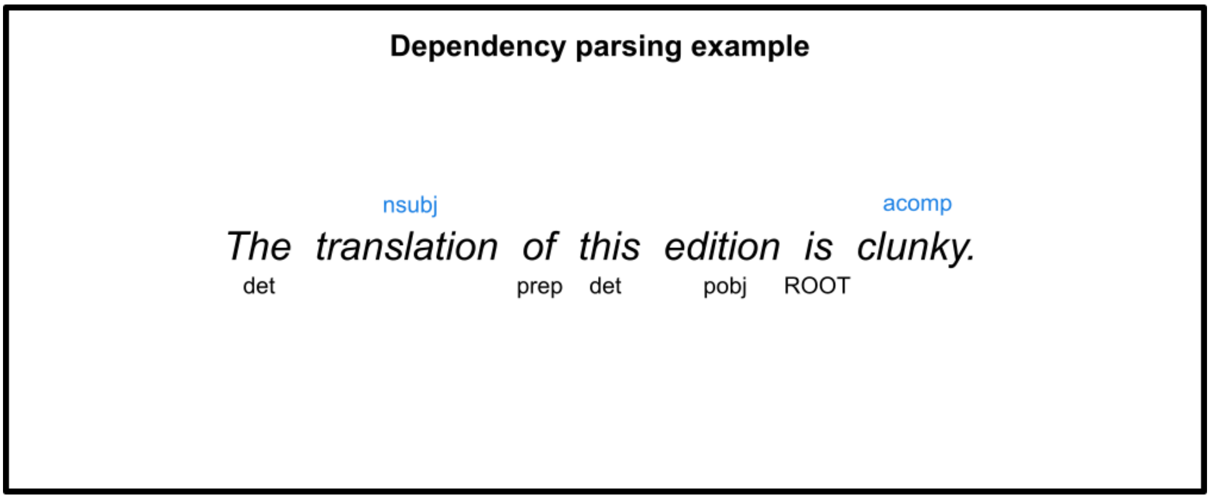
During processing, reviews were first broken down into sentences. The reason why the reviews were broken down into individual sentences is because the modifier detection algorithm employed for this study worked well at the sentence level. After this, each sentence was tagged according to whether or not it contained any direct reference to the author, the book, the act of writing, the translator, the translation or the act of translation. The occurrence of any lemmas of the words specified above were considered as a direct reference. The tagging was done through utilizing a pattern expression language known as Regular Expression (or RegEx for short) within the Python programming language. A comprehensive list of the regular expressions utilized and the example, non-exhaustive list of words that they have matched can be found in Appendix D. It is important to note here that further processing (dependency parsing) was done after tagging to disambiguate certain matched tags.

After tagging, it was revealed that 1.02% of all sentences contained a direct reference to the translator, the translation or the act of translating whereas a bit over 22.71% of all sentences contained a direct reference to the author, the original book or the act of writing. 0.7% of all sentences contained a direct reference to both, and 75.56% of all sentences did not contain any direct reference to either. This statistic stands as the first piece of evidence towards the theory that the predictions made by this paper (derived from Venuti's theory) hold true for this specific corpus.

After the reviews were tokenized into individual sentences and the sentences were tagged, further processing and analysis was conducted. Each sentence was first subjected to a process known as "dependency parsing." Dependency parsing is the process of forming an abstract grammar tree in which the relationship in between the words within a sentence are made explicit. Through dependency parsing, the list of sentences was transformed into a large array of tokens, "head" tokens (main tokens) and the relationship in between the two. Figure 5 below offers a sample sentence from the dataset that was collected for this analysis, tagged in the dependency parsing format.

Figure 4

Dependency parsing example.


Dependency parsing example
nsubj
acomp
The translation of this edition is clunky.
det
prep
det
pobj
ROOT

Note. After sentences were tagged in the format presented above, the variations of the following algorithm were employed to the modifiers of the tokens of interest: "Find me all the 'acomp' (adjective) tagged tokens whose 'nsubj' (noun subject of the sentence) is translation".

Data Analysis

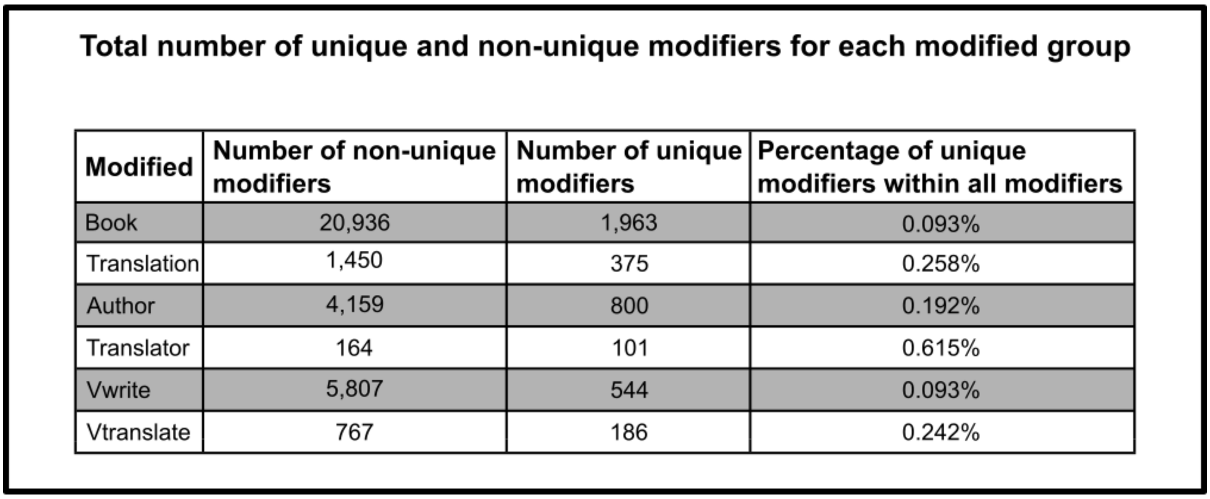
With the dependency tree, the relationship in between the tokens of interest (book – author – act of writing, translation – translator – act of translation) and their modifiers (adverbs or adjectives) could be tracked. This allowed for the answering of the main questions posed by this study.

Unique and non-unique modifiers for each token of interest

The first analysis done after dependency tagging was counting the number of times each token of interest was modified by a modifier. This, along with the data about the number of times each token of interest was mentioned, provided answers towards the first question formulated by this study.

Figure 5

Total number of unique and non-unique modifiers for each modified group.



Total number of unique and non-unique modifiers for each modified group

Modified	Number of non-unique modifiers	Number of unique modifiers	Percentage of unique modifiers within all modifiers
Book	20,936	1,963	0.093%
Translation	1,450	375	0.258%
Author	4,159	800	0.192%
Translator	164	101	0.615%
Vwrite	5,807	544	0.093%
Vtranslate	767	186	0.242%

Note. A table showing the total number of unique and non-unique modifiers for each modified group. The column "modified" represents each token of interest. The modified categories "Vwrite" and "Vtranslate" stand for the verb "to write" and for the verb "to translate."

What is made apparent by this table is the fact that the tokens representing the original book, the author and the act of writing were modified far more times than the tokens representing the translation, the translator and the act of translating. In fact, the token "book" is modified around fourteen times more than the token “translation”, the token “author” is modified around a staggering twenty-five times more than the token “translator” and the process of writing is modified around seven times more than the process of translating. This data is in line with the number of explicit references that was previously reported. The general trend shows that the works themselves (the tokens “book” and “translation”) were modified more than the verbs describing either writing or translation. In turn, the tokens representing the authors and the translators were modified less than the other two token types. The percentage of unique modifiers within all modifiers seems to be higher for the tokens related to translation: the modifiers of the tokens representing the translation, the translator or the act of translating were described with a higher percentage of unique words than other tokens. However, it can be speculated that this percentage is just a function of the total number of tokens and thus is not at all relevant.

After the total number of modifiers (unique and non-unique) was determined, the top twenty unique modifiers of each token pairing of interest were analyzed in an attempt to provide answers for the second and third questions formulated by this study.

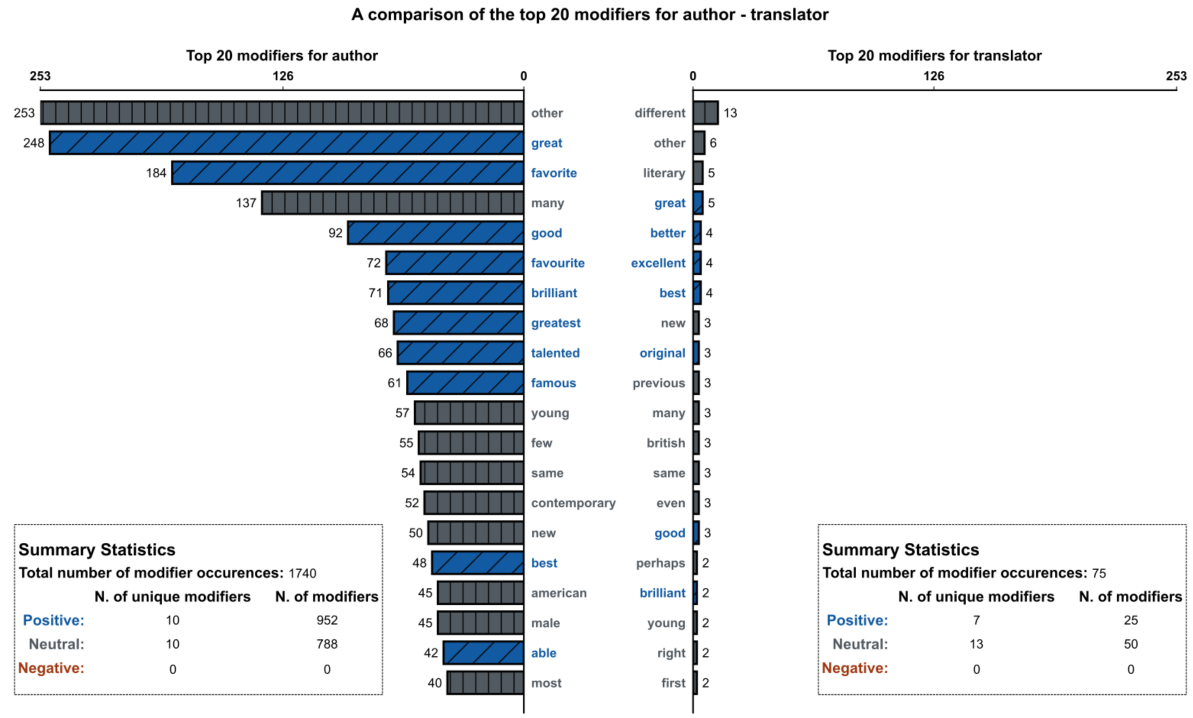
Top twenty modifiers for each token of interest pairing

For this analysis, the top twenty unique modifiers (sorted in a descending fashion with the most used modifier on top) for each token of interest were listed. These twenty modifiers were then categorized into three bins (positive, neutral, negative) by the author according to their perceived sentiment value.

The first token pairing under investigation was the author – translation pairing. As predicted with the statistics presented above, the occurrence count of the top twenty modifiers of the token “author” was more than the occurrence count of the top twenty modifiers of the token “translator.” The token “translator” was modified a total of 75 times (25 positive, 50 neutral, 0 negative) with the top twenty modifiers when compared to the token “author” which was modified a total of 1740 times (952 positive, 788 neutral, 0 negative). It can be said that the sentiment of the top twenty modifiers used to modify the token “author” was overall more positive than the sentiment of the top twenty modifiers used to modify the token “translator.” The dual axis horizontal bar chart below shows which were the top twenty modifiers for each token.

Figure 6

A comparison of the top 20 modifiers for author – translator.

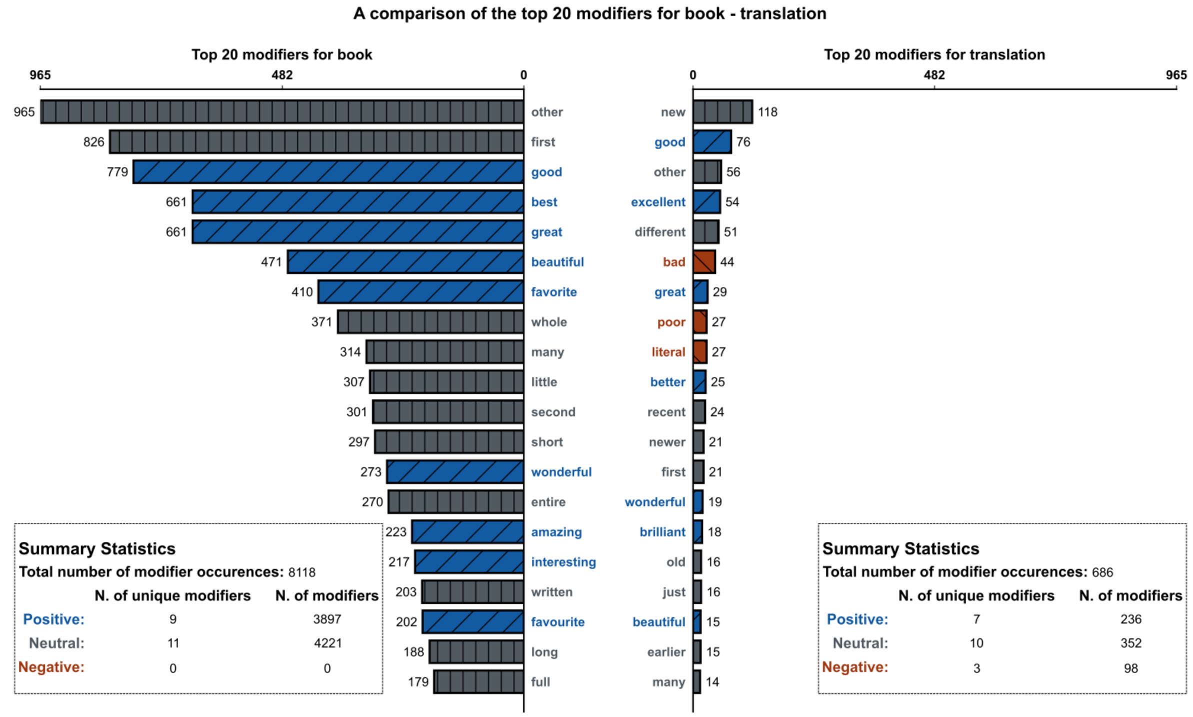


Note. Two horizontal bar charts displaying the top twenty modifiers for the tokens “author” and “translator”. The y-axis of each horizontal bar chart lists the top twenty unique modifiers for each token whereas the x-axis reveals the number of times they were modified. Modifiers with positive sentiment are encoded in blue, modifiers with neutral sentiment are encoded in gray and those with negative sentiment are encoded in orange.

The second token pairing to be studied was the book – translation pairing. In a pattern similar to the author – translator pairing, the occurrence count of the top twenty modifiers for the token "book" was more than the occurrence count of the top twenty modifiers for the token "translation." The token "book" was modified 8118 times (3897 positive, 4221 neutral, 0 negative) with the top twenty modifiers whereas the token "translation" was modified 686 times (236 positive, 352 neutral, 98 negative). Interestingly, the analysis revealed that the top twenty modifiers for the token "translation" included a total of 98 negative modifiers made up of 3 negative unique modifiers whereas the token "book" was not modified in a negative way.

Figure 7

A comparison of the top 20 modifiers for book – translation.



Summary Statistics
Total number of modifier occurences: 8118

	N. of unique modifiers	N. of modifiers
Positive:	9	3897
Neutral:	11	4221
Negative:	0	0

Summary Statistics
Total number of modifier occurences: 686

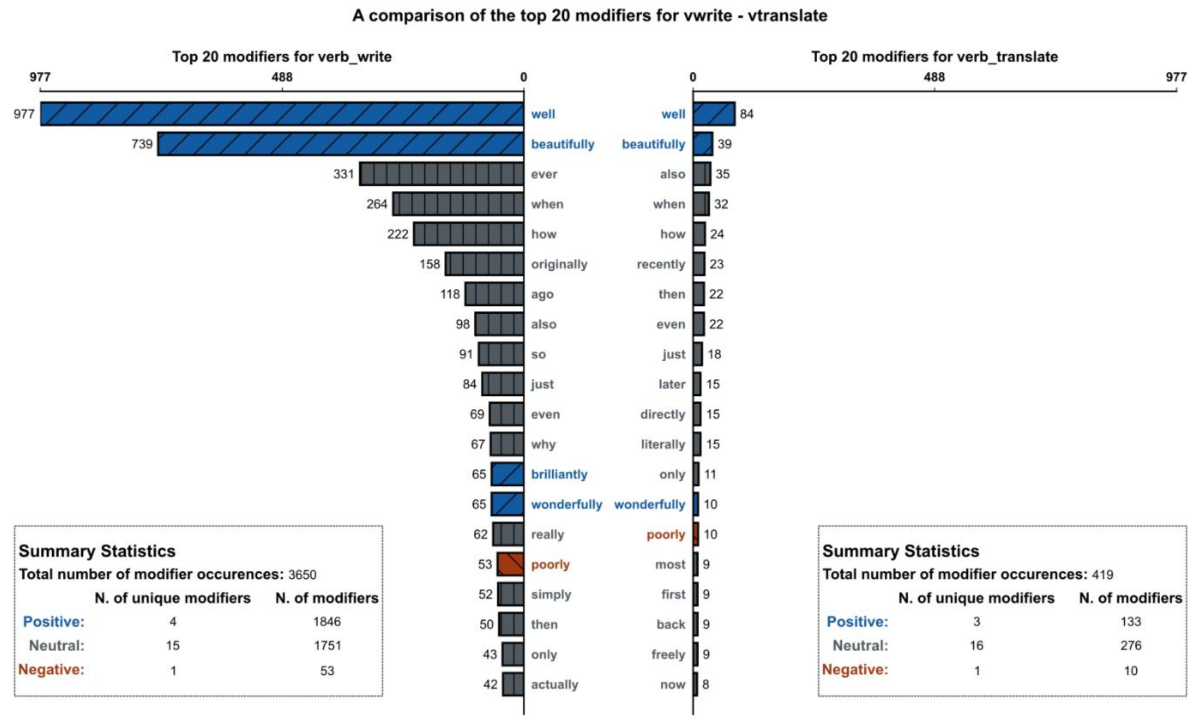
	N. of unique modifiers	N. of modifiers
Positive:	7	236
Neutral:	10	352
Negative:	3	98

Note. Two horizontal bar charts displaying the top twenty modifiers for the tokens "book" and "translation". The y-axis of each horizontal bar chart lists the top twenty unique modifiers for each token whereas the x-axis reveals the number of times they were modified. Modifiers with positive sentiment are encoded in blue, modifiers with neutral sentiment are encoded in gray and those with negative sentiment are encoded in orange.

The last token pairing under scrutiny was the verb “to write” and the verb “to translate”. The occurrence count of the top twenty modifiers for the tokens related to the word “to write” was significantly higher than the occurrence count of the top twenty modifiers for the tokens related to the word “to translate.” The former was modified 3650 times (1845 positive, 1751 neutral, 53 negative) whereas the latter was modified only 419 times (133 positive, 276 neutral, 10 negative). While both of the tokens were modified by at least one negative unique modifier, the modifiers related to the verb “to translate” were more negative in general. Special attention needs to be paid to the modifiers “directly”, “literally” and “freely” found under the top twenty modifiers for the verb “to translate.” These modifiers hold a special place within the translation criticism vocabulary because they describe very specific translational styles. The existence of these modifiers might serve as evidence supporting the claim that comments regarding a translation most often explicitly mention the translation’s style over anything else.

Figure 8

A comparison of the top 20 modifiers for vwrite – vtranslate.



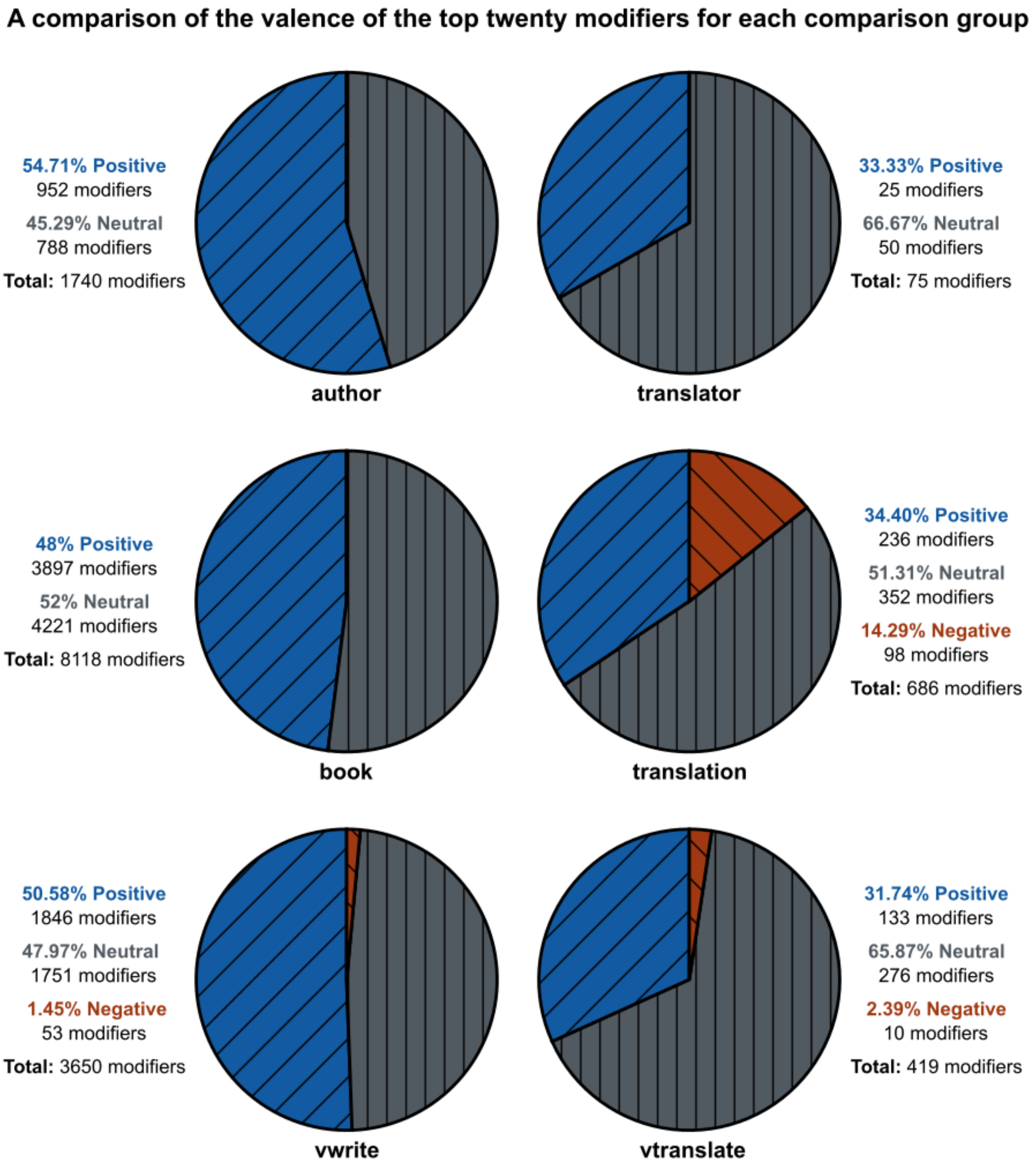
Note. Two horizontal bar charts displaying the top twenty modifiers for the tokens related to the verb “to write” and “to translate”. The y-axis of each horizontal bar chart lists the top twenty unique modifiers for each token whereas the x-axis reveals the number of times they were modified. Modifiers with positive sentiment are encoded in blue, modifiers with neutral sentiment are encoded in gray and those with negative sentiment are encoded in orange.

Comparative sentiment distribution for each comparison group

While the analysis and the visualizations above conclusively answered the second question regarding the most common modifiers for each token of interest, it only partially answered the third question regarding the sentiment value of the modifiers. In order to answer the third question in a more precise way, a different analysis was conducted.

Figure 9

A comparison of the valence of the top twenty modifiers for each comparison group.



Note. Six pie charts presenting the sentiment value of the top twenty modifiers used to modify their respective tokens. Modifiers with positive sentiment are encoded in blue, modifiers with neutral sentiment are encoded in gray and those with negative sentiment are encoded in orange.

The six pie charts above, presented in groupings of two, gives a concise answer to the third question. By looking at the percentage distribution of the sentiment value attributed to each modifier, one can say that the tokens representing the author, the original book and the act of writing were modified overall more positively than those representing the translator, the translation or the act of translation. In most cases, the latter set of tokens directly had more negative modifiers when compared to the former set. In the case of author – translation comparison pair, the modifiers modifying the token author were more directly positive than those modifying the token translator even though both did not have any negative modifiers.

Results

Now that the analysis is explained and the data is presented, the questions posed by this study can be conclusively answered:

1.) It was found that the reviewers indeed addressed the author, the original book and the act of writing more than they addressed the translator, the translation and the act of translating. This claim is supported by the number of times each token of interest was mentioned and by the number of modifiers (unique and non-unique) each token of interest was modified with.

2.) While the most common modifiers used to describe the tokens of interest were correctly determined, the extent to which they denoted a remark about the style of the translation or the original could not be determined. A significant amount of the top twenty modifiers of each token was found to be neutral in value and not directly related to style. However, the list of top twenty modifiers does include some elements such as “good”, “great”, “brilliant”, “original”, “poor”, “poorly”, “literal” and “bad” that can be interpreted as being related to the style of authorship or translation.

3.) The modifiers describing the translator, the translation or the act of translation were found to be more negative overall than the modifiers used to describe the author, the original book and the act of writing. This claim is supported by looking at both the non-normalized number of negative / neutral / positive modifiers within the top twenty modifiers of each token of interest and at the respective normalized ratio of these modifiers.

Discussion

A few remarks about the methodological choices of this study and their possible implications can be made.

Close Reading versus Computational Analysis

It is obvious that the use of computational methods allowed for a larger scope of research than the scope first presented by Venuti. However, this does not necessarily mean that the efficiency of this study or the net information extracted by this study is greater than that of Venuti's. While computational methods cast a wider net, the understanding that they offer is ultimately of a different scope than what one would get through a careful close reading practice. It is also important to note that close reading and distant reading are not mutually exclusive. These two methods of reading are synergistic: distant reading can be used to get a glimpse of the big picture and interesting leads can then be followed by close reading.

False Positive and False Negatives

In addition to the criticism outlined above, it can also be said that the computational methods used in this study had the risk of giving a false impression of the underlying data. The method of information extraction used by this study, namely using RegEx tagging and grammatical dependency parsing, is not foolproof. There might be cases where an excerpt was tagged as relevant when it was not (noise, false positive) and cases where an important excerpt was overlooked (silence, false negative). It can be argued that these errors of information extraction would not have happened if more classical methods of textual study were selected, such as close reading and manual extraction. This study offers no additional data about the extent of possible false positives and false negatives. Such an analysis would require a "truth oracle": a foolproof method to tag all sentences (manually or with another algorithm) with their appropriate tags. The actual results of this study would then need to be compared to the results obtained through this oracle.

The problem of the Canon

A big advantage of computational methods over more classical ones is the ability to go beyond the canon. While this study gathers and analyses more data than a classical humanistic study, it is not exhaustive. One technical bottleneck previously described was the fact that only 300 reviews per book could be scraped from the Goodreads database, no matter what the actual number of reviews was. This leaves us with a dataset that can actually be thought of as another form of canon: the dataset is a subset of reviews belonging to the selected books, the selected books are a subset of all the translated books on the Goodreads website, and the Goodreads website is only a place out of many where translation reviews can be found. This subset-of-a-subset selection approach resembles a canon because it is still not universal enough and subject to another set of selection biases, albeit smaller in scope. Tackling the problem of universality through gathering all the reviews that exist out in the world is not realistic. What could have been employed instead is a better sampling methodology, perhaps inspired by the established research methodology of social sciences, which allows for a more educated sampling strategy that draws from disparate-but-complementary data sources.

Ready-Made Tools versus Customized Tools

A major part of the total effort put into this study was dedicated to the codebase that made the scraping, analysis and visualization possible. However, there are many pieces of end-user friendly software that could have been utilized to tackle each computational task independently. One major advantage of such tools is the significantly less time it takes for a scholar to use them when compared to creating one’s own computational tools. What they lack is customizability and an inability to fully understand what is going on behind the user interface provided. Customizability here is taken as the ability to expand and constrict the analysis as necessary. For example, AntConc is a widely adopted freeware corpus analysis tool (Anthony, 2020) that offers a wide range of corpus analysis options. However, the options are ultimately limited: if a scholar wants to do an analysis that is not in the tool, she has to go on an implement it in some other way. Another point of concern is the opacity of the software: although the tool is freeware, it is not open source. Therefore, it is impossible to know what algorithms are actually utilized under the hood. Although laying trust on the maintainer of AntConc might be justified in many cases, this project chose not to and instead made the underlying source code of the analysis available. Another good example to illustrate this tradeoff is the business-centric data visualization platform Tableau. Tableau is an industry standard business intelligence tool that is in wide use today. It provides a robust set of data visualizations to the user and the ability to fine tune most parameters. However, it too is ultimately limited in the number of data visualizations it allows because it is not a data-driven free-hand drawing tool. In contrast, the Python package Matplotlib (Hunter, 2007) that made the visualizations of this project possible offers far more freedom when it comes to the minute details. While the author of this study sees the time spent writing code as a fruitful learning opportunity, ready-made corpus analysis software might also have been partially utilized in such a project. The aforementioned software tool AntConc might have been used to detect the modifiers that relate to the tokens of interest. The process of web scraping might have also been delegated to an external tool, such as Octoparse (*Octoparse*, 2021). The effectiveness of these tools for the use case of this paper in unknown because they have not been tested. In the software toolkit decision process, customizability and the clarity of the underlying processes were prioritized and thus a custom script was written.

Anonymous Reviewers

The reviewer metadata as presented by Goodreads did not provide any demographic metadata as to who exactly the reviewers were. As a consequence, several questions that might actually give more context about the dataset at hand are left unanswered. Below is an inexhaustive list of demographic metadata that is not found in the dataset:

- It is known that all the reviews left after the cleaning process are in English. However, what is the actual country of origin of the reviewers? Knowing the country that they come from might have allowed the analysis to make judgments about things like intercultural power relations and preferences.
- What is the professional status of the reviewers? Investigating if Venuti's theory holds true for both professional critics and for book aficionados might have added another layer of value to the analysis presented by this paper.

The metadata categories described above could have been inferred from the data already scraped. For example, reviewer names could have been used to predict nationality. Pattern extraction could have been used on the actual reviews themselves to discover signs of profession and/or nationality. However, these venues of analysis were not pursued by this paper because they were deemed to be unreliable and out of scope.

The remarks about the anonymity of the reviewers are not exactly a criticism. It can be argued that by focusing on reviews and reviews alone, a conclusion that transcended any underlying demographic patterns was reached. However, making claims about the universality of the results is futile unless the underlying demographics are unearthed.

Conclusion

Venuti's theory about the invisibility of the translator left a profound mark on the translation studies landscape that still triggers further scholarship many years after its first inception.

Originally formulated based on the data obtained by studying literary reviews from newspapers and magazines, the translator's invisibility was found to be a real phenomenon that had real life effects on how the translators carry out their professions and on how the readers interact with translations and the concept of translation. Inspired in part by the strictly empirical methodology adopted by Venuti, this paper attempted to find traces of the translator's invisibility in a newer and differently scoped corpus using computational methods. To this end, a corpus of 49,274 reviews were collected from a book-centered social media website known as Goodreads. The cleaning, processing, analysis and the visualization of this dataset confirmed the predictions made by Venuti's theory. In the corpus, references to the books of interest, the authors of those books and the processes through which they were written were vastly more prominent than the references to the translations, the translators and the process of translating of those books. Not only the references to the former were more prominent than the references to the latter, but they were also demonstrably more positive in sentiment overall. As predicted by Venuti, a significant proportion of the reviews left on the translations fixated on the "style" of translation.

References

About Goodreads. (2021). https://www.goodreads.com/about/us

Alharbi, M., Cheesman, T., & Laramee, R. S. (2020). AlignVis: Semi-automatic alignment and visualization of parallel translations. *2020 24th International Conference Information Visualisation (IV)*, 98–108. https://doi.org/10.1109/IV51561.2020.00026

Andersen, P. (2007). *What is Web 2.0?: Ideas, technologies and implications for education* (Issue 1). JISC Bristol.

Andreessen, M. (2011, August 20). Why software is eating the world. *Wall Street Journal*. https://www.wsj.com/articles/SB10001424053111903480904576512250915629460

Anthony, L. (2020). *AntConc* (3.5.9) [Computer Software]. Waseda University. https://www.laurenceanthony.net/software

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini - Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). John Benjamins Publishing Company.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for further search. *Target*, *7*(2), 223–243. https://doi.org/10.1075/target.7.2.03bak

Biber, D., Conrad, S., & Randi, P. (1994). Corpus-based approaches to issues in applied linguistics. *Applied Linguistics*, *15*(2), 169–189.

Boyles, N. (2016). Closing in on close reading. In M. Scherer (Ed.), *On developing readers: Readings from educational leadership* (pp. 89–99). ASCD.

Cheesman, T., & Roos, A. A. I. R. (2017). Version Variation Visualization (VVV): Case studies on the Hebrew Haggadah in English. *Journal of Data Mining and Digital Humanities, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages*. https://halshs.archives-ouvertes.fr/halshs-01532877

Davidson, C. N. (2008). Humanities 2.0: Promise, perils, predictions. *PMLA*, *123*(3), 707–717.

General Python FAQ — Python 3.9.5 documentation. (2021, June 13). https://docs.python.org/3/faq/general.html

Hawthorn, J. (2000). *A glossary of contemporary literary theory* (4th ed.). Bloomsbury Academic.

Henrickson, L. (2019). Humanities computing, digital humanities, and computational humanities: What’s in a name. *3:AM Magazine*. https://www.3ammagazine.com/3am/humanities-computing-digital-humanities-and-computational-humanities-whats-in-a-name/

Hockey, S. (2004). The history of humanities computing. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities* (pp. 3–19). Blackwell Publishing.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

James, J. (2001). *Sourcebook on rhetoric*. Sage Publishing.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). On close and distant reading in digital humanities: A survey and future challenges. *EuroVis (STARs)*, 83–103. http://dx.doi.org/10.2312/eurovisstar.20151113

Jay, P. (2014). Aesthetics, close reading, theory, and the future of literary studies. In *The humanities "crisis" and the future of literary studies* (pp. 115–142). Palgrave Macmillan. https://doi.org/10.1057/9781137398031_6

Kathleen, F. (2012). The humanities, done digitally. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 12–15). University of Minnesota Press.

Keyes, R. W. (2006). The impact of Moore's Law. *IEEE Solid-State Circuits Society Newsletter*, *11*(3), 25–27. https://doi.org/10.1109/N-SSC.2006.4785857

Kirschenbaum, M. (2012). What is digital humanities and what's it doing in English departments? In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 3–11). University of Minnesota Press.

Laubichler, M., Peirson, E., & Damerow, J. (2016). Software development & transdisciplinary training at the interface of digital humanities and computer science. *Digital Studies/Le Champ Numérique.* https://www.digitalstudies.org/articles/10.16995/dscn.17/print/

Laviosa, S. (2011). Corpus—Based Translation Studies: Where does it come from? Where is it going? In A. Kruger & K. Wallmach (Eds.), *Corpus—Based translation studies: Research and applications* (pp. 13–32). Continuum International Publishing Group.

Lee, H.-K. (2011). Participatory media fandom: A case study of anime fansubbing. *Media, Culture & Society*, *33*(8), 1131–1147. https://doi.org/10.1177/0163443711418271

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., Postel, J., Roberts, L. G., & Wolff, S. (2009). A brief history of the Internet. *SIGCOMM Comput. Commun. Rev.*, *39*(5), 22–31. https://doi.org/10.1145/1629607.1629613

Martin - Nielsen, J. (2012). “It was all connected”: Computers and linguistics in early Cold War America. In M. Solovey & H. Cravens (Eds.), *Cold War social science* (pp. 63–78). Palgrave Macmillan. https://doi.org/10.1057/9781137013224_4

Moretti, F. (2000). Conjectures on world literature. *New Left Review*, *1*, 54–68.

Octoparse. (2021). [Computer Software]. Octopus Data.

Olanoff, D. (2013, March 28). Amazon Acquires Social Reading Site Goodreads, Which Gives The Company A Social Advantage Over Apple [Magazine]. *TechCrunch*. https://techcrunch.com/2013/03/28/amazon-acquires-social-reading-site-goodreads/

Ousterhout, J. K. (1998). Scripting: Higher level programming for the 21st Century. *Computer*, *31*(3), 23–30. https://doi.org/10.1109/2.660187

Pierce, J. R., Carroll, J. B., Hamp, E. P., Hays, D. G., Hockett, C. F., Oettinger, A. G., & Perlis, A. (1966). *Language and machines: Computers in translation and linguistics*. National Academy of Sciences. https://www.nap.edu/read/9547/chapter/1#ix

Poibeau, T. (2017). The 1966 ALPAC report and its consequences. In *Machine translation* (pp. 75–89). MIT Press.

PYPL PopularitY of Programming Language. (2021, January 6). https://pypl.github.io/PYPL.html

Rockwell, G. (2013). Is humanities computing an academic discipline? In M. Terras, J. Nyhan, & E. Vanhoutte (Eds.), *Defining digital humanities: A reader* (pp. 13–33). Ashgate Publishing Company.

Rommel, T. (2004). Literary studies. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities* (pp. 88–96). Blackwell Publishing.

Sheridan, M. P. (2016). Recent trends in digital humanities scholarship. In D. Dejica, G. Hansen, P. Sandrini, & I. Para (Eds.), *Language in the digital era: Challenges and perspectives* (pp. 2–13). De Gruyter Open.

Svensson, P. (2012). Beyond the big tent. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 36–49). University of Minnesota Press.

Svensson, P. (2013). Humanities computing as digital humanities. In M. Terras, J. Nyhan, & E. Vanhoutte (Eds.), *Defining digital humanities: A reader* (pp. 159–186). Ashgate Publishing Company.

TIOBE - The Software Quality Company. (2021, January 6). https://www.tiobe.com/tiobe-index//

Unsworth, J. (2013). What is humanities computing and what is not? In M. Terras, J. Nyhan, & E. Vanhoutte (Eds.), *Defining digital humanities: A reader* (pp. 35–48). Ashgate Publishing Company.

Vanhoutte, E. (2013). The gates of hell: History and definition of digital | humanities | computing. In M. Terras, J. Nyhan, & E. Vanhoutte (Eds.), *Defining digital humanities: A reader* (pp. 119–156). Ashgate Publishing Company.

Venuti, L. (2008). *The translator's invisibility: A history of translation* (2nd ed.). Routledge.

Vinjamuri, D. (2013, March 29). Three Hidden Benefits of The Amazon Acquisition of Goodreads. *Forbes*. https://www.forbes.com/sites/davidvinjamuri/2013/03/29/three-hidden-benefits-of-the-amazon-acquisition-of-goodreads/?sh=64df2837731f

Wilkens, M. (2012). Canons, close reading, and the evolution of method. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 249–258). University of Minnesota Press.

Appendices

Appendix A

List of authors and books investigated by Lawrence Venuti.

Author Name	Book
Albert Camus	The Stranger (1946)
Adolfo Bioy Casares	A Russian Doll (1992)
Ana Maria Moix	Dangerous Virtues (1997)
Françoise Sagan	Bonjour Tristesse (1955)
Gabriel Garcia Marquez	One Hundred Years of Solitude (1970)
Gianni Celati	Appearances (1992)
Heinrich Böll	Absent Without Leave (1965)
Ismail Kadare	The Successor (2005)
Italo Calvino	Cosmicomics (1968)
Jose Saramago	The Double (2004)
Mario Vargas Llosa	In Praise of the Stepmother (1990)
Michel Houellebecq	The Elementary Particles (2000)
Milan Kundera	The Book of Laughter and Forgetting (1980)
Orhan Pamuk	My Name is Red (2001)

Appendix B

Modifiers used to describe translation and translators in Venuti’s research.

Modifier	Modified
Natural	Prose
Brilliant	Prose

Crisp	Prose
Elegant	Style
Lovely	Prose
Excellent	Translation
Flawlessly	Translated
Fluent	Translation
Faithful	Translation
Not idiomatic	Translation
Flows crisply	Translation
Disconcerting British	Accent
Wooden	Translation
Careless	Translation

Appendix C

List of authors and books represented in the corpus.

Authors and books that are found in Venuti's original study are highlighted in Yellow.

Author Name	Book
Albert Camus	The Stranger
Albert Camus	The Plague
Aleksandr Solzhenitsyn	One Day in the Life of Ivan Denisovich
Alessandro Baricco	Silk
Alexander Dumas	The Count of Monte Cristo
Alexander Pushkin	Eugene Onegin
Andrey Kurkov	Death and the Penguin
Anna Gavalda	Hunting and Gathering

Antal Szerb	Journey by Moonlight
Antoine de Saint-Exupéry	The Little Prince
Antoine Laurain	The Red Notebook
Anton Chekhov	The Cherry Orchard
Arthur Schnitzler	Dream Story
Arto Paasilinna	The Year of the Hare
Arundhati Roy	The God of Small Things
Ayşe Kulin	Last Train to Istanbul
Boris Pasternak	Doctor Zhivago
Bruno Schulz	The Street of Crocodiles
Carsten Jensen	We, The Drowned
Daniel Kehlmann	Measuring the World
Dino Buzzati	The Tartar Steppe
Elena Ferrante	The Story of the Lost Child
Elena Ferrante	The Days of Abandonment
Elfriede Jelinek	The Piano Teacher
Emile Zola	Germinal
Erich Maria Remarque	All Quiet on the Western Front
Erlend Loe	Doppler
Eugene Vodolazkin	Laurus
Fernando Pessoa	The Book of Disquiet
Franz Kafka	The Trial
Franz Kafka	The Metamorphosis
Franz Kafka	The Trial

Fredrik Backman	A Man Called Ove
Fredrik Backman	My Grandmother Asked Me to Tell You She's Sorry
Fyodor Dostoyevsky	Crime and Punishment
Fyodor Dostoyevsky	The Brothers Karamazov
Gabriel Garcia Marquez	One Hundred Years of Solitude
Gabriel Garcia Marquez	Love in the Times of Cholera
Georges Perec	A Void
Georges Simenon	Dirty Snow
Georges Simenon	Pietr the Latvian
Giuseppe Tomasi di Lampedusa	The Leopard
Günter Grass	The Tin Drum
Gustave Flaubert	Madame Bovary
Halldor Laxness	Independent People
Han Kang	The Vegetarian
Han Kang	Human Acts
Hans Fallada	Every Man Dies Alone
Harry Mulisch	The Assault
Harry Mulisch	The Discovery of Heaven
Haruki Murakami	The Wind-Up Bird Chronicle
Haruki Murakami	Norwegian Wood
Haruki Murakami	Kafka on the Shore
Haruki Murakami	Sputnik Sweetheart
Heinrich Böll	The Clown
Henryk Sienkiewicz	Quo Vadis

Herman Koch	The Dinner
Hermann Hesse	Siddharta
Hermann Hesse	Steppenwolf
Hiromi Kawakami	The Nakano Thrift Shop
Imre Kertesz	Fatelessness
Isabel Allende	The House of Spirits
Italo Calvino	If on a Winter's Night a Traveler
Italo Calvino	Invisible Cities
Italo Calvino	The Baron in the Trees
Italo Svevo	Zeno's Conscience
Ivan Goncharov	Oblomov
Ivan Turgenev	Fathers and Sons
Ivo Andric	The Bridge on the Drina
Janne Teller	Nothing
Jean Giono	The Man Who Planted Trees
Jenny Erpenbeck	Go Went Gone
Johann Wolfgang von Goethe	The Sorrows of Young Werther
John Ajvide Lindqvist	Let the Right One In
Joris - Karl Huysmans	Against Nature
Jose Eduardo Agualusa	A General Theory of Oblivion
Jose Saramago	Blindness
Jose Saramago	Baltasar and Blimunda
Jostein Gaarder	Sophie's World
Kamel Daoud	The Meursault Investigation

Karel Capek	War with the Newts
Katarina Bivald	The Readers of Broken Wheel Recommend
Knut Hamsun	Hunger
Knut Hamsun	Growth of the Soil
Kobo Abe	The Woman in the Dunes
Laszlo Krasznahorkai	Satantango
Laura Esquivel	Like Water for Chocolate
Leo Tolstoy	War and Peace
Leo Tolstoy	Anna Karenina
Magda Szabo	The Door
Magda Szabo	Abigail
Maj Sjöwall	The Man Who Went Up in Smoke
Michael Ende	The Neverending Story
Michel Houellebecq	Submission
Miguel de Cervantes Saavedra	Don Quixote
Mika Waltari	The Egyptian
Mikhail Bulgakov	The Master and Margarita
Mikhail Bulgakov	Heart of a Dog
Milan Kundera	The Unbearable Lightness of Being
Milan Kundera	The Book of Laughter and Forgetting
Mo Yan	Red Sorghum
Muriel Barbery	The Elegance of the Hedgehog
Nikolai Gogol	Dead Souls
Nikos Kazantzakis	Zorba the Greek

Nikos Kazantzakis	The Last Temptation of Christ
Nino Haratischwili	The Eighth Life
Olga Tokarczuk	Drive Your Plow Over the Bones of the Dead
Olga Tokarczuk	Flights
Orhan Pamuk	My Name is Red
Orhan Pamuk	Snow
Orhan Pamuk	The Museum of Innocence
Orhan Pamuk	Istanbul: Memories and The City
Orhan Pamuk	The Black Book
Orhan Pamuk	The White Castle
Orhan Pamuk	The Red-Haired Woman
Orhan Pamuk	A Strangeness in My Mind
Orhan Pamuk	The New Life
Paolo Giordano	The Solitude of Prime Numbers
Patrick Süskind	Perfume: The Story of a Murderer
Paulo Coelho	The Alchemist
Paulo Coelho	Veronika Decides to Die
Peter Hoeg	Smilla's Sense of Snow
Primo Levi	The Periodic Table
Robert Musil	The Man Without Qualities
Sandor Marai	Embers
Sayaka Murata	Convenience Store Woman
Sigrid Undset	Kristin Lavransdatter
Sofi Oksanen	Purge

Stanislaw Lem	Solaris
Stieg Larsson	The Girl with the Dragon Tattoo
Stieg Larsson	The Girl Who Played with Fire
Stieg Larsson	The Girl Who Kicked the Hornet's Nest
Thomas Mann	The Magic Mountain
Thomas Mann	Death in Venice
Umberto Eco	The Name of the Rose
Umberto Eco	Foucault's Pendulum
Umberto Eco	The Prague Cemetery
Umberto Eco	Baudolino
Un - su Kim	The Plotters
Valeria Luiselli	The Story of My Teeth
Victor Hugo	The Hunchback of Notre-Dame
Voltaire	Candide
W.G. Sebald	The Rings of Saturn
W.G. Sebald	The Emigrants
Yasunari Kawabata	Snow Country
Yevgeny Zamyatin	We
Yoko Ogawa	The Housekeeper + The Professor
Yu Hua	To Live
Yukio Mishima	The Sailor Who Fell from Grace with the Sea
Yuri Herrera	Signs Preceding the End of the World

Appendix D

Regular Expressions used for sentence tagging and the example patterns that they match.

RegEx	Matched Pattern
`\b[Bb]ook[\w+]?\b`	Book, book, Books, book
`\b[Aa]uthor[\w+]?\b`	Author, author, Authors, author
`\b[Ww]r[io]t\w+\b`	Write, write, Wrote, wrote, Written, written, Writing, writing, Writer, writer
`\b[Tt]ranslat\w+\b`	Translation, translation, Translations, translations, Translator, translator, Translators, translators, Translate, translate, Translates, translates, Translated, translated, Translating, translating