UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE COMUNICAÇÃO E EXPRESSÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INGLÊS

Carlos Eduardo da Silva

CORPUS MINING: A NEW PERSPECTIVE ON TRANSLATION STUDIES

Florianópolis
2018

Carlos Eduardo da Silva

CORPUS MINING: A NEW PERSPECTIVE ON TRANSLATION STUDIES

Tese submetida ao Programa de Pós-graduação em Inglês da Universidade Federal de Santa Catarina para obtenção do título de Doutor em Estudos da Linguagem. Orientador: Prof. Lincoln P. Fernandes, Dr.

Florianópolis
2018

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Da Silva, Carlos Eduardo
Corpus mining: A new perspective on translation studies / Carlos Eduardo Da Silva ; orientador, Lincoln Fernandes, 2018.
268 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro de Comunicação e Expressão, Programa de Pós-Graduação em Inglês: Estudos Linguísticos e Literários, Florianópolis, 2018.

Inclui referências.

1. Inglês: Estudos Linguísticos e Literários. 2. Corpus-based traslation studies. 3. Text mining. 4. Corpus linguistics. 5. Parallel corpus. I. Fernandes, Lincoln. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Inglês: Estudos Linguísticos e Literários. III. Título.

Carlos Eduardo da Silva

CORPUS MINING: A NEW PERSPECTIVE ON TRANSLATION STUDIES

Esta tese foi julgada adequada para obtenção do título de "Doutor em Estudos da Linguagem" e aprovada em sua forma final pelo Programa de Pós-Graduação em Inglês: Estudos Linguísticos e Literários.

Florianópolis, 13 de julho, 2018.

Profº Dr. Celso Henrique Soufen Tumolo
Coordenador do Curso

Banca Examinadora:

Profº Dr. Lincoln Paulo Fernandes
Orientador

Profº Dr. Denilson Sell
Universidade do Estado de Santa Catarina (UDESC)

Profº Dr. Marcos Morgado de Oliveira
Universidade Federal de Santa Catarina (UFSC)

Prof.ª Dr.ª Maria Lúcia Vasconcellos
Universidade Federal de Santa Catarina (UFSC)

Prof.ª Dr.ª Rosane Silveira
Universidade Federal de Santa Catarina (UFSC)
(suplente)

ACKNOWLEDGEMENTS

What a long journey I look back on! During my academic life, I have developed a great interest in software engineering and linguistics—especially in areas such as Translation Studies, Corpus Linguistics, and Natural Language Processing. These areas inspired me to ponder a series of questions and ideas related to the combination of technology and linguistics; as a consequence, I have started contemplating whether these subject areas could help provide technological solutions for the investigation of translational phenomena.

The pursuit of a PhD endeavor, especially involving an interdisciplinary approach, is challenging. The use of IT as a key component in this study constitutes an additional degree of complexity, since technology evolves at a rapid pace; following a specific approach can be difficult, therefore, when new areas of research present themselves as worthy of further investigation. I have learned to cultivate patience and to keep my anxiety under control, my objective being to focus on a specific set of techniques and methodological approaches that best serve the development of a cutting-edge thesis in the disciplinary field of Translation Studies.

The achievement of this endeavor would not have been possible without the support of wonderful, confident, assiduous, and trustworthy people who have encouraged me throughout the course of this long project. In addition to numerous people, I also owe much credit to the institutions and scholarship programs that recognized the value of my research project and contributed support. First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Dr. Lincoln P. Fernandes, who believed in my academic capacity and provided tremendous support to me during all phases of this study. His assistance kept me motivated to keep a clear objective in mind and to avoid procrastination and drifting off topic. Professor Dr. Fernandes inspired me to achieve my academic and professional goals in life.

I greatly appreciate the funding support opportunity awarded me by The British Academy via Newton Advanced Fellowships; this support made it possible for me to visit the University of Birmingham (UB) as a visiting PhD candidate. During my time visiting UB, I received scientific support from Professor Dr. Michael Toolan and Professor Dr. Paul Thompson, both faculty members at the Department of English Language and Applied Linguistics. I express my gratitude particularly to Professor Dr. Michael Toolan and his wife Julianne Statham, for hosting me at their welcoming and comfortable home. In addition, I thank Professor Dr.

Gabriela Saldanha, UB Lecturer in Translation Studies, Department of Modern Languages. She graciously spent precious time listening to my hypothesis and giving me invaluable suggestions. I had a wonderful time in Birmingham, Oxford, and then in London. As one who has never previously traveled to an English-speaking country, the chance to visit all of these cities was an unforgettable experience.

This thesis was done at the *Universidade Federal de Santa Catarina* (UFSC)—my alma mater—under the *Programa de Pós-Graduação em Inglês* (PPGI). It was a great privilege for me to study at UFSC. I would like to thank all the professors at PPGI for their assistance, especially Professor Dr. Celso Henrique Soufen Tumolo, Department Head of PPGI. He provided the necessary support when I needed to extend my study period. I would also like to thank the other staff members who kindly and willingly offered assistance every time I needed help navigating bureaucratic procedures.

I also wish to express my appreciation to the readers on my thesis committee: Professor Dr. Denilson Sell from *Universidade do Estado de Santa Catarina* (UDESC); Professor Dr. Marcos Morgado de Oliveira (UFSC); Professor Dr. Maria Lúcia Barbosa de Vasconcellos (UFSC); and Professor Dr. Rosane Silveira (UFSC).

My special thanks for the generous financial support and sponsorship of *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) of the Brazilian Ministry of Education, which provided me a scholarship through the entire period of my research. I also thank all team members of the UFSC research group, *Tradução e Corpora* (TraCor), for sharing their thoughts and ideas. I also thank the UFSC IT team from *Superintendência de Governança Eletrônica e Tecnologia da Informação e Comunicação* (SeTIC) for providing technical support and high-end hosting of infrastructure services for the COPA-TRAD platform. My acknowledgement is not complete without expressing my deepest gratitude and eternal love to my mother Eliane da Silva, who passed away at the beginning of this year after a long struggle against cancer. My mother showed me the meaning of life's joy, even in the face of affliction. I also thank my grandfather, Antônio José da Silva, a great man who passed away last year. Until the last day of my life, I will be grateful to both of them for their unconditional love and devotion, and also for their many admirable virtues—humility, as an example.

Trust the text
John Sinclair, 2004b

ABSTRACT

Drawing on Text Mining (TM) and Corpus-based Translation Studies (CTS), this study addresses the combination of an interdisciplinary methodology for translation research. It focuses on technical and theoretical assumptions necessary to develop the proposed methodological model, henceforth referred to as "Corpus Mining." The hypothesis is that a collaboration between CTS and TM can contribute to the improvement of corpus creation, text processing, and analysis. It is worth noting that the application of Corpus Mining and proposed techniques can lead to the discovery of novel patterns (i.e., linguistic features or, in our case, specific variables typically associated with translated texts). To achieve this goal, the following steps are taken: (a) the utilization of techniques from Text Mining contribute to CTS; (b) an explanation of how Corpus Mining can support the corpus compilation triad (i.e., design, building, and processing) and research based on a parallel corpus; and (c) inclusion of a practical example showing the application of Corpus Mining while developing new tools for COPA-TRAD—*Corpus Paralelo de Tradução*. It is expected that Corpus Mining will contribute to CTS theoretical and practical repertoire by providing a methodological model for the investigation of translated texts through parallel corpus. New horizons to manipulate and analyze translated texts through a set of computerized algorithms that are far from the basic frequency list and pattern matching are available for investigation; it is also expected that Corpus Mining can support or indicate the need for further research in this direction. Results show that Corpus Mining model, set up as a concise, step-by-step guide to support the investigation of translated texts in the light of Text Mining and Corpus-based Translation Studies, can provide means for the analysis of information that is comprehensive and easily accessible.

Key-words: Corpus-based Translation Studies; Text Mining; Corpus Linguistics; Parallel Corpus.

Number of Pages: 268

Number of Words: 55,698

RESUMO

Utilizando-se da Mineração de Textos (MT) e dos Estudos da Tradução com Base em Corpus (ETC), este estudo aborda a combinação de uma metodologia interdisciplinar para pesquisa em tradução. O estudo focaliza-se em pressupostos técnicos e teóricos para desenvolver o modelo metodológico proposto, doravante referido Mineração de Corpus. A hipótese deste estudo é que uma interface entre MT e ETC pode contribuir para a melhoria ao suporte na criação de corpus, processamento automático de textos e análise. Cabe ressaltar, que a aplicabilidade da Mineração de Corpus possibilita a descoberta de novos padrões (isto é, características linguísticas ou no nosso caso, variáveis específicas características do texto traduzido). Para atingir este fim, os seguintes passos são adotados: (a) a utilização de técnicas de Mineração de Textos, Mineração de Dados e Processamento de Linguagem Natural podem contribuir para ETC, (b) como a Mineração de Corpus pode contribuir para a tríade da compilação de corpus (isto é, projeto, construção e processamento) e pesquisas com base em corpus paralelo, (c) a aplicação da Mineração de Corpus em um caso prático, COPA-TRAD—*Corpus Paralelo de Tradução*. Espera-se, que a Mineração de Corpus possa contribuir para o repertório prático e teórico dos Estudos da Tradução com Base em Corpus por prover um modelo metodológico a ser seguido durante a pesquisa em corpus paralelo. Novos horizontes abrem-se para a manipulação de textos traduzidos através de um conjunto de algoritmos computadorizados além de simples listas de frequência e combinações de padrões. Os resultados indicam que a Mineração de Corpus é um guia sucinto para a investigação de tradução à luz de uma metodologia com base em corpus pode prover para o pesquisador informações completas e acessíveis.

Palavras-chave: Tradução com base em corpus; Mineração de Textos; Linguística de Corpus; Corpus Paralelo.

Número de Páginas: 268

Número de Palavras: 55.698

LIST OF ILLUSTRATIONS

LIST OF TABLES

LIST OF ABREVIATIONS AND ACRONYMS

BI – Business Intelligence
CTS – Corpus-based Translation Studies
CM – Corpus Mining
COPA-TRAD – Corpus Paralelo de Tradução
COPA-LIJ – Corpus Paralelo de Literatura Infanto-juvenil
COPA-RAC – Corpus Paralelo de Resumos Acadêmicos
COPA-MET – Corpus Paralelo de Meta Discurso em Tradução
CRISP-DM – CRoss-Industry Standard Process for Data Mining
EAGLES – Expert Advisory Group on Language Engineering Standards
IT – Information Technology
KDD – Knowledge Discovery in Databases
KDT – Knowledge Discovery in Text
MVP – Minimum Viable Product
NLP – Natural Language Processing
ST – Source Text
TM – Text Mining
TS – Translation Studies
TT – Target Text
UI – User Interface
WSD – Word Sense Disambiguation

TABLE OF CONTENTS

1. CHAPTER ONE: Introduction

1.1. Introductory Remarks

This chapter discusses the fundamental elements contextualizing the core idea proposed here, as well as the arguments permeating this study. It is divided into seven sections, as follows: First and foremost, the context of investigation and an overview of COPA-TRAD (Fernandes & Silva, 2016) are presented. Second, the problem and the preliminary context are discussed to situate this research, and then an overview on Corpus Mining is provided in order to depict the proposed concept informing it. The "Scope of the Study" section states the problem, the purpose, and research questions. Next, the research method and its rationale are introduced. Finally, this chapter problematizes the nature of parallel corpus data and lays out the structural organization for the research here envisaged.

1.2. Context of Investigation

Over the years, the amount of digital data produced has exceeded the human capacity to process, analyze, and transform it into knowledge. For instance, from 2005 to 2020, Gantz and Reinsel (2012) estimated that "the digital universe will grow by a factor of 300, from 130 exabytes to 40 exabytes, or 40 trillion gigabytes." During this period, the total amount of the digital universe will "about double every two years" (p. 1). As a matter of fact, our society is evolving toward a knowledge-consumer market. Researchers have to find ways to utilize the amount of data produced (i.e., electronic information) for the purpose of investigating problems from different perspectives, such as those found in Linguistics and, in a narrower sense, Translation Studies. Tymoczko (1998), a well-known scholar in Translation Studies, stated that

> [t]he information age has brought an explosion in the quantity and quality of information we are expected to master. This, along with the development of electronic modes for storing, retrieving, and manipulating that information, means that any discipline wishing to sustain itself in the twenty-first century must adapt its content and methods. Corpus translation studies is central to the way that Translation Studies as a discipline will remain vital and move forward (p. 1).

Furthermore, to be brief, huge amounts of raw data[1] do not create meaning or outcomes by themselves. From a translational perspective, for example, the research outcomes based on electronic data are even more complex[2] as they go beyond the issue of storing vast amounts of data.

The initial aim of this research was limited in scope and inherently inapplicable to providing processing and data visualization tools for COPA-TRAD Version 2, if a diversified theoretical foundation to satisfy the technical requirements was ignored. COPA-TRAD (Fernandes & Silva, 2016) was developed during my MA[3], and the initial concept was suggested by Fernandes (2004). COPA-TRAD parallel text[4] corpus will serve as input data for the application of Corpus Mining methodology and its sophisticated text-processing techniques will be used to harvest useful information[5].

At some point, I realized, the theoretical foundation is inherently interdisciplinary. For this reason, I suggested the name "Corpus Mining" because, as mentioned previously, it amalgamates methods from Text Mining and Corpus-Based Translation Studies. As a result, Corpus Mining is founded on a confluence of such heterogeneous methodology, and it is applicable in parallel corpus research, the central focus of this study. Depending on the requirements of the research, Corpus Mining can be used in its totality or, to satisfy specific parts of the research, just partially.

Although Text Mining is not broadly disseminated in CTS, resources from this technology were applied for the investigation of translated texts, more specifically in phraseology and stylistics. A work worth of mention is the PhD thesis of Ji (2008), who used a parallel corpus in the linguistic pair Spanish - Chinese and "a number of standalone text mining applications," to investigate two versions of Cervantes's Don

---

[1] Data (singular datum) without being processed and analyzed are considered as "raw."

[2] For example, the researcher has to deal with texts in two or more languages.

[3] Held at the *Programa de Pós-Graduação em Inglês at Universidade Federal de Santa Catarina (UFSC)* entitled "Developing online parallel corpus-based processing tools for translation research and pedagogy." Available at: https://tracor.ufsc.br/uploads/ma-carlos-a5.pdf

[4] Parallel text or bitext (this last one, is the term adopted in this study – see Section 2.3.5).

[5] The expression "useful information" refers (here) to the type of information meaningful for the research.

Quixote translation (p. 12). The researcher recognizes "traditional textual analyses" could be renovated with the utilization of "sophisticated corpus techniques" (p. 12). However, the researcher argues that use of these techniques "prioritizes the generation of quantitative textual data," and then he lists some of these quantitative data, such as keyword indexing, collocation patterns, type/token ratio, node word distribution spectrum, mutual information score, etc. (p. 12).

The methodological procedures adopted in the Ji (2008) study are dependent on a set of textual annotations, to find "potential linguistic features" (pp. 46-47). I have the perception, to a certain extent, these techniques and methodological procedures are not purely Text Mining; they could be part of it, but it is not, if compared specifically in the manner I am advocating here and according to authors in the area cited in this study. In Chapter 2, in the same work, Ji (2008) cites the tools he is referring to, Text Mining functions like WordSmith Tools, ParaConc, or Xaira (p. 46).

Other scientific works (e.g., Tsatsoulis, 2013) dealing with Text Mining and CTS share similarities to the one cited above (i.e., the application of text mining techniques to literature). (I provide a more comprehensive discussion about these topics in Chapter 2.) However, what I want to call attention to is the fact pertaining to how Text Mining has been dealt with in CTS. And the gap I want to cover is this: In the applied branch of Corpus-based Translation Studies, based on my investigation, there is no research focused on the use of an interdisciplinary interface of Corpus Mining such as the one advocated here.

In addition, there is no free online parallel corpus with computerized tools available to process and discover new[6] and useful information for the investigation of translational phenomena[7]. Another point I want to call attention is related to the lack of studies on data visualization focusing on Corpus-based Translation Studies investigations. The subject related to data visualization and its procedures

---

[6] Differently from Information Retrieval systems, Text Mining "is not for locating any wanted information in a large collection of texts in response to a query", but the objective "is to infer new knowledge, mostly as convert information about facts, patterns, trends or relationships of text entities, which is hidden in, and hence inaccessible via the comprehension and literary interpretation of, texts" (Chunyu & Jian-Yun, 2015, p. 509).

[7] Singular: phenomena. Plural: phenomenas (Merriam-webster.com, 2018).

for presenting the processed and discovered data is dealt with by the use of Text Mining, too, but will not be covered in this work. Here, the specific focus is on the processing mechanisms and methodological procedures for discovering novel and useful information, facilitating the understanding of translational and linguistic features that are based on the input data (i.e., texts).

I would like to make a parenthetical statement here, so as to briefly explain the comparison of "data" and "information," for me. Based on previous reading, "data" is the raw text[8] content (i.e., without meaning or unprocessed), and "information" is the useful data that emerges from the raw text processing—in other words, meaningful data for the researcher. The above concept surpasses the common approach of how to hypothesize and theorize about a specific subject worthy of investigation. Based on this idea, the researcher does not try to create a preconceived hypothesis or formulate questions; rather, the researcher assumes the position of a data analyst that handles intelligent systems to mine information. Possible clues about a translational phenomena under investigation may emerge[9] from the analyzed corpus. (To illustrate the role of a data analyst, Chapter 4 discusses real examples from COPA-TRAD.)

1.2.1. An Overview of COPA-TRAD

As previously mentioned, during my MA, an online platform based on a parallel corpora[10] was proposed and developed according to specifications from work by Fernandes (2004). The platform is available online for the academic community, and it is named COPA-TRAD (*Corpus Paralelo de Tradução*) (see Appendix A; Appendix B; Appendix C; Appendix D; Appendix E). COPA-TRAD features innovative concepts in an academic parallel corpus—for example, the usage of friendly interface and affordable usability. The adopted concepts do not require

---

[8] *Raw text* stands for keeping the text as it is clean of any other codes (Sinclair, 1991, p. 21).

[9] The use of "may emerge" is due to the fact the text under investigation has to be scrutinized as well as the algorithms should be refined in case no meaningful clues are found at the first attempts.

[10] The term Corpus (plural: corpora) "can be described as a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria" (Bowker & Pearson, 2002, p. 9).

specialized technical skills of the user, because it provides a clean, easy-to-use, and modern visual interface. For instance: reducing and controlling the number of mouse clicks, the amount of information in each page (i.e., page content consistency), and the use of icons, images, and buttons providing the means for the user understand and perform the expected action.

A great number of operations on COPA-TRAD can be performed without reading, for example, user-guides. As a consequence, in the classroom, the online platform can be utilized as a supporting teaching tool for students without previous knowledge on parallel corpus, as well as experienced researchers in the academic area of study. The searching procedures in COPA-TRAD was prepared to be intuitive so the user, with previous knowledge of how to navigate the Internet—such as finding information on search engines like *Google*, *Bing*, *Yahoo*, *Duck Duck Go*, *Baidu* or *Yandex*—is able to use COPA-TRAD. For experienced users, more refined results can be obtained by using wildcards (i.e., special characters with a specific function and meaning for the textual searching tool). In addition to parallel corpora, COPA-TRAD has useful tools available to assist the investigation of texts, as well as assisting researchers who encounter obstacles when trying to add their text in the corpus, due to genre or domain not being included in the COPA-TRAD specification. However, these tools are available to facilitate the creation of lists and corpus on the fly. One tool I would like to mention here is TREETAGGER CLOUD, a web interface for using TreeTagger[11] without knowledge of shell commands (see Appendix F). In addition, COPA-TRAD platform is mobile friendly and usable on smartphones and tablets.

As a commitment to the scientific community, COPA-TRAD was registered under the name of *Universidade Federal de Santa Catarina* (UFSC). The platform consists of a web-based corpus available for free on the Internet. The user interface (UI) as well as the visual elements created to provide information for the user, were planned with a short learning curve in mind. However, Varantola (2002) observes some problems related to this kind of corpus:

> A Web-based corpus is also likely to have the latest terminology, which might be time-consuming to locate elsewhere. The downside is, of course, that

[11] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

> the information is not vetted and there is no guarantee of its quality. The user is thus alone responsible for applying the information further (p. 186).

I can understand Varantola's comments in relation to web-based corpus because during my MA research, I experienced the difficulties related to this kind of technology, especially for the language pair Portuguese-English. Bearing in mind this problem, COPA-TRAD was developed with a set of tools to overcome the abovementioned "downside"—for example, the established rules to maintain the quality of the submitted material. A human moderator checks all textual data submitted to COPA-TRAD.

As mentioned in the introductory paragraphs, two other studies that I conducted provided a solid basis for the research here envisaged. In my specialization course, Text Mining techniques were investigated to find out if they could be applied on a parallel corpus or not. The product of this investigation was a monograph entitled "Using Text Mining Techniques on Parallel Corpora for the Academic Research on Translation Studies: A Case Study." The case study was conducted over COPA-TRAD to test and exemplify how specific techniques from Text Mining could be applied in the platform. This work also shed light on the use of this technological approach to investigate translational phenomena. As a result, it opened the doors for the comprehensive research delineated here, based on a dual interface of Text Mining and Corpus-based Translation Studies. This study reveals how both areas cooperate with each other and what the bond is that connects it.

In relation to platform upgrade and development of new features proposed for COPA-TRAD as well as the dynamic nature of technology, a platform such as the one envisaged here needs constant attention. Perhaps, for scholars in any other linguistic area, this could be viewed as a caveat. However, technology expands at an incredible pace, and upgrades are necessary in order to keep the corpus system relevant and up to date. Following this line of thought, Sinclair (1991) explains that

> [o]nce a corpus is in existence, it needs regular maintenance and upgrading. There are always errors to correct and improvements to be made, adaptations to new hardware and software, and changes in the requirements of users. In addition, there should be constant attention to the retrieval systems, and processing and analytic tools. For

> some time to come, software will improve dramatically and frequently (pp. 22-23).

It is notable that Sinclair wrote these words in 1991 and that now, more than ever, his words are still valid and in accordance with our present technological context. For instance, based on my previous studies, I could find sophisticated methods and techniques based on Text Mining, which can harvest previously unknown and useful information from a parallel corpus without tagging/labeling the textual data. Unfortunately, due to scope and time, these findings are going to be discussed and were implemented in this study. These technological advancements (i.e., "improvements" and new "requirements") were implemented only in COPA-TRAD Version 2 because, after two years of availability to the public, it was possible to track the most common problems and difficulties users faced while using the platform.

As observed in Sinclair's quote above, special attention should be given to the "retrieval system" or searching engine used to find information in COPA-TRAD corpora. The advanced filters and implementation of Text Mining techniques, used to find specific information that go beyond the simple keyword matching, started being developed; some of the initial results are already available to the public. These features and necessary components were previously pointed out in my MA dissertation.

> A suggestion for future research is to create advanced search interface in which the user would be able to select the texts in the corpus in terms of author, year of publication, language variation, translators, etc. Moreover, the alignments of the texts need some refinement, especially as regards the automatization of the whole alignment process. The inclusion of text-mining techniques for finding linguistic patterns that will eventually lead the researcher to new and useful kinds of information about translation phenomena [have to be implemented] (Silva, 2013, p. 124).

These limitations and suggestions cited above were analyzed and improved. The result of this process culminated in this study and the product of it, that is, the Corpus Mining model.

1.3. The Problem and its Current Scenario

Currently, the real challenge of information technology has shifted from how data can be stored to how to make use of this data for creating useful knowledge (Grossman, 2012, p. 164). Grossman's (ibid.) reflection is from a technological point of view but illustrates and affirms the current scenario. In addition, it is not incompatible with other areas, such as Corpus Linguistics (CL), as observed by Varantola (2002). The researcher explains the context under consideration from a linguistic point of view as follows.

> Some ten years ago corpus size was the only thing that seemed to matter, but now when size is no longer an issue and there is an abundance of electronic text available, the "my corpus-is-bigger-than-yours" rhetoric has subsided even in lexicographical contexts. Instead the focus has moved on to a discussion of the adequacy of the corpus for its purpose, onto a "my-corpus-is-smarter-than-yours" rhetoric (p. 174).

The preoccupation permeating the corpus storage mechanisms cannot be denied. As observed by Kennedy (1999), the former linguistic corpora were created entirely manually in pieces of papers (pp. 13-19). Such early corpora consumed a great deal of human resource, time, and money, in order to conduct any linguistic endeavor (ibid.).

With the advent of the first computers, one of the first applications for these "new machines" was in linguistics. This allowed for the compilation of the first electronic corpus (i.e., the Brown Corpus), and from then on, linguistics perception in relation to language has changed (Kennedy, 1999, p. 23). In face of this pioneering advancement, however, I believe the arrival of innovative technological applications in the academic field of linguistics (or in a narrow sense Translation Studies) has suffered from long delays over the last two decades. To my view, specific technology we have today to analyze corpora (from a linguistic or translational perspective) is out-of-date and even limited if compared to recent or ongoing findings in the wide area of Natural Language Processing. Similarly, Gries (2015) asserts that "change in the field of linguistics is slow, and corpus linguistics in particular is limited in two ways" (p. 93).

First of all, the software tools available to analyze corpora are, in most cases, proprietary (i.e., paid), and the researcher does not have

access to the source code in order to adapt and improve it for his/her own needs. Some of these tools are WordSmith Tools, MonoConc PRO, ParaConc and AntConc[12] (Gries, 2015, pp. 93).

Secondly, the use of statistical methods in corpus linguistics relies on known, simple and common formulae, and this constitutes a limitation, since other technical apparatuses are available nowadays and are not being used—for example, the application of Sentiment Analysis[13] in translated literary texts (Gries, 2015, pp. 93-94).

Similar issues, as observed above, are valid for the disciplinary field of Translation Studies. The existing methods to extract information from raw texts are based, in most of the cases, on simple search mechanisms (e.g., full or partial keyword matching).

In the course of my academic life, I have observed that most researchers are still using software such as WordSmith Tools or AntConc, and also adopting pre-existing statistical methods that did not adapt properly for their investigations. In fact, the researcher inputs a keyword into a specific software application, and it automatically searches for exactly the same occurrences, based on that keyword list. The software does not produce any sort of greater in-depth knowledge for a qualitative research. Possible misspelled keywords are not included in the output results; or in the absence of results, the system cannot suggest analogous keywords. Another point worth mentioning is related to annotation, because in the early stages of corpus linguistics, annotations were done manually. For instance, in 1959, the "SEU Corpus was the last major pre-electronic corpus," meaning that tasks like count occurrences had to be done manually (Kennedy, 1999, p. 23). After two years, in 1961, "planning began for the compilation of the first machine-readable corpus for linguistic research" (ibid.).

As previously observed, for a long time, researchers started taking advantage of computer technology and its input devices, such as flatbed or sheet-fed scanners, in order to digitize analog information such as legal documents, texts, or even entire books. In a post-step, these

---

[12] Although AntConc has a "non-commercial or freeware" license there are some restrictions even for the commercial license see laurenceanthony.net/software/antconc/releases/AntConc357/license.pdf

[13] An algorithm derived from Text Mining techniques which classifies a text, commonly known in three categories: positive, negative or neutral. However more different categories are also possible depending on the research objectives and the algorithm applied.

scanned files were converted to text using a technology named OCR (Optical Character Recognition) (Srihari & Hull, 1992); depending on the OCR software, at the same time the text is being scanned, the OCR process is carried out.

Currently, scanners are still in use. However, with the consolidation of digital textual content on the Internet, as well as electronic books in a variety of formats (e.g., PDF, MOBI, DOC, DOCX, ODF, RTF, TXT, EPUB, AZW), the availability of textual content does not involve the manual problem of digitizing physical texts or rekeying (i.e., retyping the text when OCR fails) as it used to be. For instance, buying a book and then scanning and preparing it to be inserted in a corpus may not be the only option available, as it was some years ago. More and more often, it is possible to download books from the Internet or even buy the digital file on websites such as Amazon, the well-known retailer.

Even though advantageous in terms of saving time, this reality raises concern about another problem. As mentioned before, the amount of data available—in this case, text—is enormous. In the face of this reality, to conduct research using parallel corpora, innovative methods or enhanced methods urgently need to be developed and proposed for the scientific community, in order to satisfy an ever-growing demand.

Regardless of the scientific method adopted for a corpus-based study, the application of technology is prone to be accepted "because computers have enabled humans to gather more data than we can digest, it is only natural to turn to computational techniques to help us unearth meaningful patterns and structures from the massive volumes of data" (Fayyad, Piatetsky-Shapiro & Smyth, 1996a, p. 1). For Fayyad, Piatetsky-Shapiro and Smyth (1996b) and in this study, **data** are a fact or a group "of facts," and **pattern** "is an expression in some language describing a subset of the data or a model applicable to the subset" (p. 41). As an example, "pattern" can be a specific rule denominating a recurring linguistic or translational element[14] in a database[15]. From a linguistic

[14] Humans can hardly enough perceive patterns by naked eyes.

[15] "Electronic databases have long been used as a tool in linguistic research […] An electronic database is an obvious and appropriate tool: it can store the attribute values (= grammatical properties) of entities (= languages, constructions, etc.), and execute queries which recover information about the entities meeting a set of criteria. Once an adequate amount of data has been entered into the database for a workable sample of languages, the researcher can quickly and easily find out

point of view, a pattern is a repetition observed when words, sounds, rhythms or structures are reproduced twice or more (Hunston, 2010, p. 152).

Large corpora are humanly difficult to be analyzed manually, since the use of technology enabled the storing of a significant amount of data. Moreover, searching techniques were developed to find specific keywords in the ever-growing textual collection of a corpus. I would like to emphasize the limitations in the most common searching mechanisms available for a textual corpus; or in other cases, a degree of complexity for inexperienced researchers (lacking a technical background) in corpus analysis that may lead to anomalous and imprecise results; or in other cases, technological problems, such as a long period of time to find a keyword and lack of memory to process the request. In other words, during my MA investigation (2013), I found the three relevant parallel corpora for the study of the language pair Portuguese-English have a certain degree of usability or technical limitation, if compared with the technology available today (Silva, 2013, pp. 25-30). In relation to the searching mechanisms, the reviewed corpora systems rely on limited—or in other cases, too complex—searching techniques (e.g., the direct use of regular expressions).

Having in mind such challenges, I now direct attention to Translation Studies—more specifically, to Corpus-based Translation Studies (CTS, henceforth). In my MA study, I conducted investigation of CTS technological tools available for the Translation Studies academic community. During this study, it was possible to find a gap in the method involving the parallel corpus, building a proposed triad: Design, Development, and Processing (Figure 1). As a means to suggest a free and easy-to-use parallel corpus analysis tool, a parallel corpora and platform named COPA-TRAD was proposed[16] (Fernandes & Silva, 2011). The corpus building triad, as mentioned, was proposed by Fernandes (2004) in his PhD research:

> (i) corpus design, where general theoretical issues associated with corpus planning are discussed; (ii) corpus building, where the technical decisions made throughout the corpus compilation are

---

which languages in the sample have any combination of the described properties." (Musgrave, Dimitriadis, & Everaert, 2009, pp. 1-2).

[16] Available at: http://copa-trad.ufsc.br

> described; and (iii) corpus processing, where the hardware, software and set of computational tools used for processing the corpus are specified (p. 104).

The concise structure by Fernandes (2004) for corpus building guided and set out the basis for my previous MA study. However, the objectives of my study were to propose and develop parallel corpus storage and processing tools. Therefore, I have added a new dimension to Fernandes' (ibid.) corpus building triad (i.e., the software-engineering dimension) in order to guide the technical procedures of software development for COPA-TRAD.

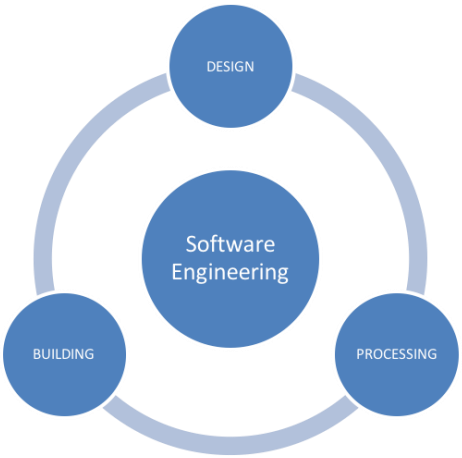


Figure 1. Corpus building triad and Software Engineering coordinating the three elements based on Fernandes (2004).

Regarding the aforementioned gap, I have started my master's study, and subsequently, I developed a deep interest in Software Engineering as an area that could possibly help me find out about theories and methods that could be applied to the development of online parallel corpus-based tools for the investigation of translational phenomena[17].

[17] Study conducted at the *Curso de pós-graduação Lato Sensu em Engenharia e Projetos de Software* at *Universidade do Sul de Santa Catarina* (UNISUL) entitled *Utilização de técnicas de mineração de textos em corpora paralelo para*

Among the theories and methods in Software Engineering, Text Mining seems to provide a new perspective for the study of translation. Although I cited Text Mining technology in my MA study, I did not detail what it consists and how it can be applied in CTS. In this sense, I would like to propose a combination of these two areas (i.e., CTS and Text Mining) and complementary related fields (i.e., Corpus Linguistics and Natural Language Processing) into one perspective and consequently suggest a name for it, which is "Corpus Mining." This set of building blocks is arranged in a standard model for assisting CTS investigation.

Having briefly explained the motivations that have led me to the study at hand and the found gap, this initial chapter introduces Corpus Mining, as a set of procedures and rules embodied in an unifying research model. Corpus Mining provides defined steps for building, processing, and analyzing parallel corpus by means of Text Mining techniques specifically prepared for CTS environment. In the following section, the major definitions comprising Corpus Mining are presented.

1.4. Corpus Mining: An Overview

The detailed discussion of Text Mining and interrelated subjects such as Data Mining and Natural Language Processing (NLP) are going to be covered in the second chapter. Additionally, the following chapter deals with Corpus-based Translation Studies and the role of linguistic-processing tools for direct or indirect application in corpus investigation. However, it is necessary to give at least a concise overview on the concept of Corpus Mining and why and how I came to this claim.

During my previous research, I used an interdisciplinary approach which culminated in the online platform, COPA-TRAD. Since doing so, I noted a dispersion of techniques from this interdisciplinary approach, which may lead to a certain amount of confusion for those involved in the investigation and development of CTS computer-aided tools. While investigating the possible technologies (e.g., Text Mining) to support and to be implemented in translational corpora research, I found, in Software Engineering, a concept embodying Text Mining and Data Mining—a combination also known as Duo Mining (Figure 2). The idea behind a unifying concept named Duo Mining was influential for this study. For the sake of creating a unifying interface between Text Mining

---

auxílio na pesquisa acadêmica em estudos da tradução: Um estudo de caso. Available at: http://tracor.ufsc.br/uploads/monografia-carlos.pdf

and CTS and because of that expand the techniques available for corpus investigation; Duo Mining proves how Text Mining can integrate to other fields. Therefore, the concept and techniques from Duo Mining are used in Business Intelligence (BI) for supporting business decisions, not Translation Studies research.

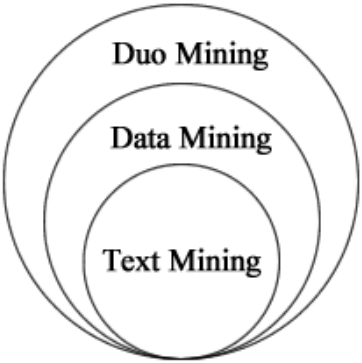


Figure 2. Duo Text Mining (Krishnaiah et al., 2012, p. 423).

Based on this primary observation, I settled upon a Corpus Mining model, aiming to provide an analytical method for creating computational systems for the direct or indirect application of Text Mining techniques in conjunction with CTS. Not to mention the fact that Corpus Mining can be utilized as methodology for the investigation of translational features. As a result, other possible patterns not thought before may emerge from the text itself. Defining what subject is part of the other in computational processing for natural language investigation is a challenging endeavor, from a linguistic and technological point of view. However, in Figure 3, I attempt to map the subjects Corpus Mining interfaces: Corpus-based Translation Studies, Text Mining, Natural Language Processing, and Corpus Linguistics[18].

[18] Corpus-based Translation Studies (CTS) and Corpus Linguistics: both areas may seem interchangeable (in a manner of speaking) but there are historical differences, well discussed in the theoretical literature (see, e.g., Olohan, 2004).

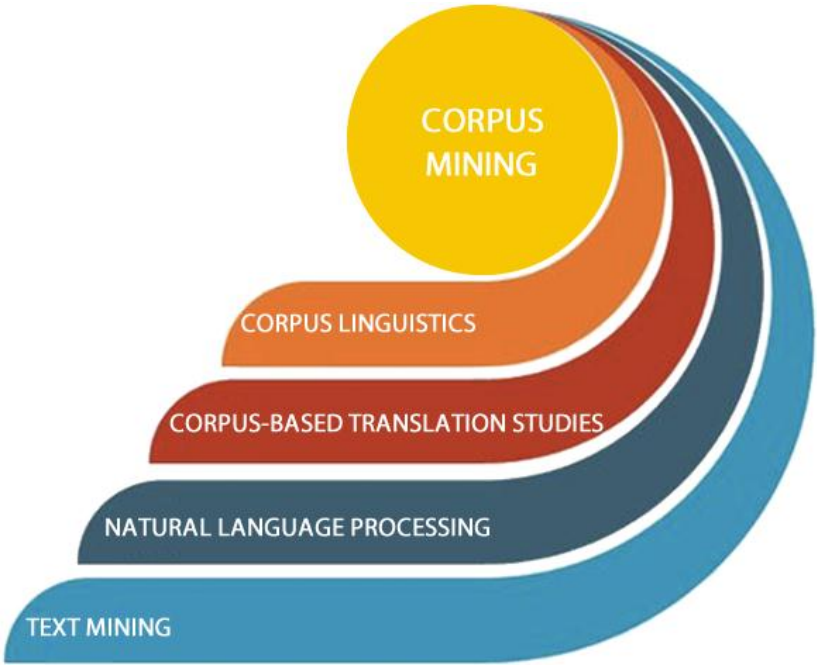


Figure 3. Corpus Mining interplay of subjects.

1.5. Scope of the Study

This study aims to integrate and analyze the theoretical background as well as the techniques from Text Mining and in consequence of it Natural Language Processing along with Data Mining. In addition, CTS and Corpus Linguistics are dealt with to find a common ground to support Corpus Mining. To achieve this purpose I want to demonstrate: (a) the use of techniques from Text Mining, Data Mining and Natural Language Processing contribute to CTS, (b) how can Corpus Mining support the corpus building triad and research on a parallel corpora and (c) Corpus Mining application on COPA-TRAD—*Corpus Paralelo de Tradução* (see Section 1.2.1). I expect that Corpus Mining can contribute to theoretical and practical repertoire of CTS by providing a methodological model to follow throughout parallel corpus research.

1.5.1. Statement of the Problem

The interface between Text Mining and CTS provide insightful perspectives for the investigation of translation and possibly other linguistics field (but not covered here). Conducting research based on

Corpus Mining standard model with eight defined steps and methods for each research phase leads to automatized or semi-automatized processing mechanisms to analyze texts and creation of useful knowledge (i.e., research outcomes). In addition, the technology to analyze and process natural language in text format evolved and I believe it should be incorporated into CTS for the creation of new tools (for the area) and extract fruitful knowledge. Despite this, the technological apparatus available, at present, for CTS community poses a limitation for implementing hypotheses to subjects related to, for example, identifying translational features such as simplification, explicitation, normalization or even translator style (see Baker 1996, 2000). This is due to the fact it requires from the researcher great investment of time to prepare and process textual data (e.g., labeling or language normalization) and to handle the chosen tool for the study. Going further, hidden textual features will not be found with software limited in scope (i.e., processing and analysis possibilities).

In addition, researchers on translation, driven by the nature of the investigation, tend to annotate their corpus manually or automatically. In other words, the text part of the corpus data, are normalized and labeled to provide custom variables for investigation. In case some linguistic or translational evidence is badly labeled or there are linguistic/translational phenomena unknown before the annotation it will be ignored during the processing phase. As a result, this technical issue can be problematic while analyzing the research outcomes. A well-known problem related to labeling is ambiguity, for example, when a linguistic feature has more than one category. The decision of what to do with the ambiguous segment is a problem in manual and automatic labeling. To illustrate, Manning (2011) observes a manual labeling process "depends of many factors" such as researchers "aptitude for the task, how much they are paying attention, how much guidance they are given and how much of the guidance they are able to remember" (p. 173). Another problem is posed by McEnery (2003) where he says "corpus annotation only makes explicit what is implicit, it does not introduce new information" (p. 453). The researcher will not find clues of a certain phenomena or behavior unobserved before that could lead to different conclusions. Again, such issue can undermine an ongoing study if it has not been considered.

The tools available on the market are able to perform searching operation based on the keyword provided by the user. In case, the keyword is similar to other candidate expressions inserted in the text, there is a considerable possibility that it will not match. As a result—to my knowledge—researchers with a software development background

are moving to develop their own software for text analysis and processing. If this is the situation, most of them are creating their own methodology to conduct this kind of scientific endeavor. The adoption of methods without following a comprehensive model may lead to confusion entering in a vicious circle "reinventing the wheel."

Based on the aforementioned assumptions I mapped in Figure 4 the possible causes—to my view—that lead to a limited technology adoption in CTS. The list does not aim to be extensive and the elements pointed out are based on my own experience while working and researching in the field. The Ishikawa (Figure 4) diagram breaks down and organizes the major causes and in each branch it is possible to check the inner causes of each category pointing to the main problem.

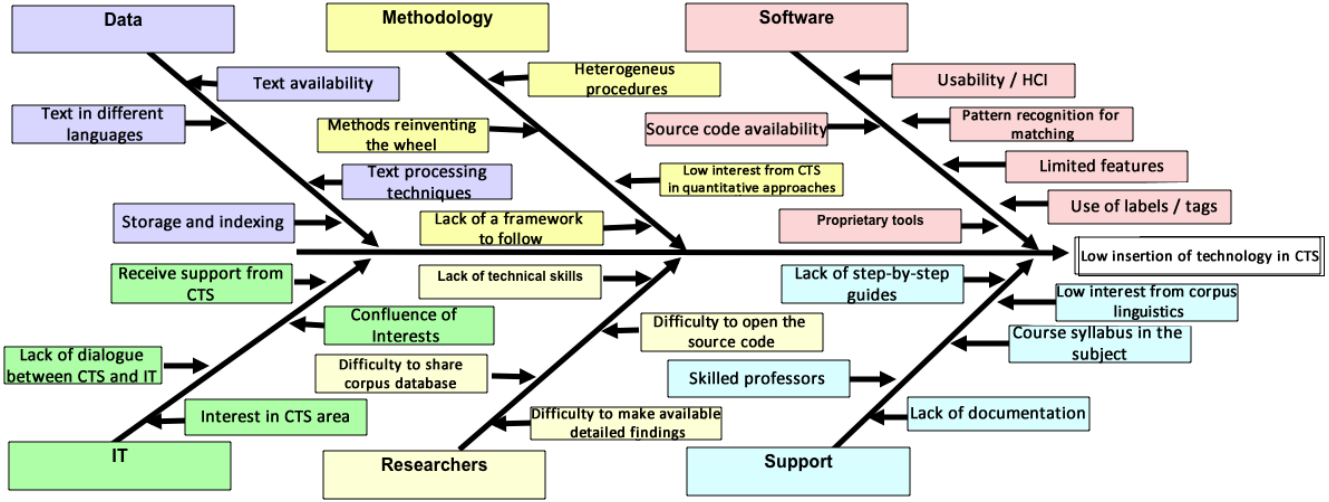


Figure 4. Possible causes for limited application of textual analysis technology in CTS.

Starting from left to the right, on the branches above the continuum, the first category of causes to be mentioned is the one related to data, which in a certain extent it is connected to the other categories and is part of the main problem branch. First of all, text availability is not always possible due to copyright issues and local policies from university or governmental laws. Then, especially in CTS, there are difficulties to find texts in different language which most of these texts have to be aligned in parallel (i.e., originals bound to translations). Fernandes (2004) had to deal with copyright issues while compiling his corpus of Translation Children's Literature, part of COPA-TRAD:

> However, copyright holders are not always willing to cooperate, especially if the request involves granting permission for children's books. Matters become even worse if the children's books in question hold a bestselling status. The reason for this refusal is not always clear, but it certainly has to do with commercial considerations in children's literature (p. 78).

Following the data branch, the second branch holds the causes related to methodology problems, this is a serious issue and it was foreseen by Baker (1993) when she pinpoints that "there is now an urgent need to explore the potential for using large computerized corpora for translation studies" and she continues "what we need is a research methodology and a set of tools that can help us put this program into action" (p. 248) "program" here, is for the investigation of translational features and descriptive translation studies research into action. Baker continues saying the area needs a "suitable methodology and a set of very powerful and adaptable tools," but these necessities are provided and available "from corpus linguistics." To a certain extent, I do agree with Baker, at the time when corpus linguistics was the direct option available to conduct investigation in CTS, remaining few possibilities out of it.

The last branch above the continuum is the one named "software" and lists the problems that might constitute a barrier especially for novice researchers. At this category the causes such as the availability of the software (most of them are paid), the complex display mechanisms (e.g., the user has to guess what action a button does), the data preparation beforehand and the insistence of availability of source code (if the researcher wants to change anything in the algorithm to adapt to his own necessities).

Below the continuum from the left to the right, the first branch named "IT," I am using this expression because it is general and I hope it can include the areas of Computer Science, Natural Language Processing, Statistics, Software Engineering, Knowledge Engineering, Information Systems, Data Science, Big Data, Computational Linguistics amongst others. I perceived the IT branch poses a problem for CTS because there is a lack of dialogue between these two areas as well as there is a low interest from IT in CTS and vice-versa. As observed by Čulo, Hansen-Schirra and Neumann (2011):

> Exchange between the translation studies and the computational linguistics communities has traditionally not been very intense. Among other things, this is reflected by the different views on parallel corpora. While computational linguistics does not always strictly pay attention to the translation direction (e.g. when translation rules are extracted from (sub)corpora which actually only consist of translations), translation studies are amongst other things concerned with exactly comparing source and target texts (e.g. to draw conclusions on interference and standardization effects) (p. 1).

Sometimes the IT tries to develop its own methods on automatic translation and find its own linguistic explanation for some linguistic phenomena and some of the answers they are trying to find it is already investigated by CTS. For this reason, the IT has to receive support from CTS in order to provide better technological solution for scientific community. There should be a confluence of interests between researchers from different backgrounds. On the other hand, the IT researchers have to listen what CTS community thinks about the software developed for linguistic research and include in their analysis this user feedback. Varantola (2002) observed this problem at the time: "likewise, software producers find it hard to explain to the translators what lies behind their designs and user interfaces" (p. 183).

Following the IT branch, the second branch "Researchers" there is a problem with the personnel involved in CTS investigation because some of them find difficult to share their technological experiment or database. Some of them have difficulty to share their corpus database due to copyright issues or due to the time, effort and money they invested in this endeavor. Others find difficulty to open the source code of the

software they developed. It is possible to perceive either, that some scientific papers on CTS explains the translational findings with the tools they created but if any other researcher wants to replicate the experiment he/she will find it difficult to have access to the same tools. Another reason for this problem is related to software literacy, as highlighted in the "Software" branch, translational aid software is complex to handle and because of that researchers or TS students find difficult to use it. Varantola (2002) explains this problem in detail.

> The major problem with advanced software is naturally the users' limited willingness to learn to use them. Software literacy should not be taken for granted. Impatient translators have to be sure that learning to use the software is worth their while, cost effective and does things that they can benefit from. The problem with most language technology applications is that software developers and software users still find it hard to communicate with each other (p. 183).

To indicate a possible solution for the problem caused by software literacy in TS context, Varantola (2002) proposes a solution to include corpus education in the syllabus of TS courses:

> In practical terms, this means that translator trainees need to be taught to understand the various uses of corpora, how to compile them and apply them in an intelligent way. Translator training should thus include courses in the compilation and use of corpus information. Broadly speaking, instruction in corpus linguistics could be divided into courses in corpus use, corpus compilation and corpus analysis software (p. 183).

I share the same point of view and I believe the corpus education should move beyond specific theoretical courses. Corpus and its methodology can be included in TS courses from a more practical perspective, focusing on hands-on activities that could help users to demystify the use of technology and help to eliminate technophobia.

Finally, the third branch below the continuum is the one named "support" and lists the possible problems found in this category. To conduct a linguistic analysis with technological tools the user has to move

between several tools, at this part many researchers move to the Operational System LINUX and MacOS, because it offers powerful tools at command line to process texts, but it requires the knowledge of each command as well as the knowledge of regular expression. For instance, the command line software such as *grep*, *awk* or *sed* require such a technical knowledge from the researcher. Independent of each other, the (a) lack of step-by-step guides, (b) lack of documentation explaining how to use the tools available for translational investigation, (c) the low interest of corpus linguistics to contribute for the expansion of CTS and (d) the limited technical background of some CTS professors to guide their students can constitute a visible problem for a more precise adherence to the technological apparatus we have available today. Varantola (2002) recognizes the learning curve even for well-known software and it can constitute a barrier even for CTS professors when they have to teach the use for their students, or a translator who needs a fast solution and have to learn how to use the technological aid:

> Yet, problems do arise. For instance, the learning threshold for a text analysis tool is much higher than that for dictionaries. Even if the WordSmith Tools functions are a kind of snap-on tools in that particular toolbox, much could be done to make them easier to use (p. 183).

Needless to say, texts are natural language and they are available in heterogeneous formats, most of them are marked up in HTML or XML, others in Microsoft Word® special characters or PDF and there is a need to apply processing techniques in order to obtain clean text for analysis. Furthermore, the problem of storage and indexing, the texts collected have to be stored in a special place and if the amount of texts is considerable large, this can pose a serious problem, especially for indexing the content for the application of data retrieval techniques.

However, we currently have the possibility to take advantage of technological solutions, far from the scope of corpus linguistics. This area has its own challenges and it cannot deal efficiently with the present scenario depicted in the context of investigation. For this reason, I elucidate in the Methodology section, the following causes: the existence of heterogeneous procedures, researchers in CTS are recurring for different procedures in other to fill the technical gap of their investigation; this can be noted on the different techniques they report in their scientific papers. This scenario leads to the following cause, the problem of

"reinventing the wheel" (i.e., creating an apparatus which is already done) with some differences but the same in essence.

Such problems could be avoided by applying a standard model for the project under investigation. As observed by Baker (1993), there is a lack of a model to follow a set of defined steps in a workflow that make possible for the researcher track where he/she is what he/she has to do in order to achieve the goal of the research (p. 248). This model has to be generic as possible in order to fit for the variety of studies under investigation (i.e., it has to be adaptable).

Another problem to address in the methodology section is the low interest of CTS in quantitative approaches. However, quantitative information is interesting and useful to set out a qualitative driven research (Olohan, 2004, p. 86). I recognize the usefulness of quantitative approach and the contribution it offers but most of the techniques available to perform statistical procedures, especially the ones based on likelihood measures poses a certain degree of certainty and uncertainty and because of that exact results cannot be defined, possibly some researchers do not deal with it. Nevertheless, researchers need to accept, deal with statistics and supporting software even for qualitative studies. Olohan (2004) provides a heterogeneous solution, an approach consisting of qualitative and quantitative analysis "is, in most cases, desirable, particularly if fuller descriptions of linguistic and translational phenomena are to be given and reasons suggested for their occurrence" (p. 86).

In order to achieve a plausible solution for the problems and assumptions stated above or at least promote a discussion on the issues related to CTS, this work proposes a standard model named Corpus Mining. The ideas relevant for this model are cleared up in the thesis statement Section 1.5.3.

1.5.2. Purpose of the Study and Research Questions

The purpose of this study is to investigate the interface between CTS and Text Mining with its related technology. Then, apply the findings to a particular online parallel corpus platform (i.e., COPA-TRAD) that aims to investigate translational phenomena. The full implementation of Text Mining for a computer programming point of view is time consuming and this area is evolving fast so what are going to be discussed, in some of the cases, will be nuggets for future work. Therefore, I expect to accomplish the main purpose by improving and developing the available and new tools in COPA-TRAD platform, for example, aligning tool, display

formats, advanced filters, processing mechanisms and graphic reports. To fulfill the purpose, the following research questions are discussed.

- How can CTS and Text Mining be combined to improve a large data analysis of translated texts?
- What kinds of tools can be created from a Corpus Mining perspective?
- What are the advantages of Corpus Mining for Corpus-based study?
- How can Corpus Mining surpass textual annotation in source texts and target texts? Why is that important? How can it be done?

1.5.3. Thesis

The combination of Text Mining techniques with Corpus-based Translation Studies contributes as a method for investigating translated texts. New horizons to manipulate and analyze translated texts through a set of computerized algorithms far from the basic frequency list and pattern matching are available for scientific community. Text Mining technologies such as "information extraction, summarization, categorization, clustering and information visualization" are some of the referred computerized algorithms "to teach computers how to analyze, understand and generate text" (Gaikwad, Chaugule, & Patil, 2014, p. 43).

Different institutional sectors such as academic, industrial, financial, pharmaceutical, political and news media are using Text Mining with encouraging outcomes (Sumathy, & Chidambaram, 2013, p. 31). The utilization of Text Mining in a variety of sectors and knowledge areas apart from Translation Studies just confirms the benefits of the technology, for this reason, I argue for CTS incorporate the proposition of this study. I believe in a two-way cooperation between CTS and Text Mining because both subjects can contribute each other, with its own expertise, for the improvement of automatic text processing and analysis. It is worth noting the application of the proposed method enables the discovery of new patterns (i.e., linguistic features or in our case, specific variables typical from translated texts). Here, I am using the word "new" because by applying the suggested methods, new patterns out of initial hypothesis may emerge from the corpus under scrutiny. The idea of novelty is endorsed by the words of Fayyad, Piatetsky-Shapiro and Smyth (1996b):

> The discovered patterns should be valid on new data with some degree of certainty. We also want

> patterns to be novel (at least to the system and preferably to the user) and potentially useful, that is, lead to some benefit to the user or task. Finally, the patterns should be understandable, if not immediately then after some postprocessing (p. 41).

For this reason, I argue for the aforementioned cooperation and based upon this claim a new horizon is about to come to CTS: Different techniques (e.g., Sentiment Analysis), if compared to CTS usual ones, to manipulate text may lead to new possible answers explaining a given translational phenomena or even lead to new questions. The proposed set of techniques come from Text Mining mainly. I believe these techniques can be applied and adopted for the sake of innovation as well as widening the available technological assets of Corpus-based Translation Studies.

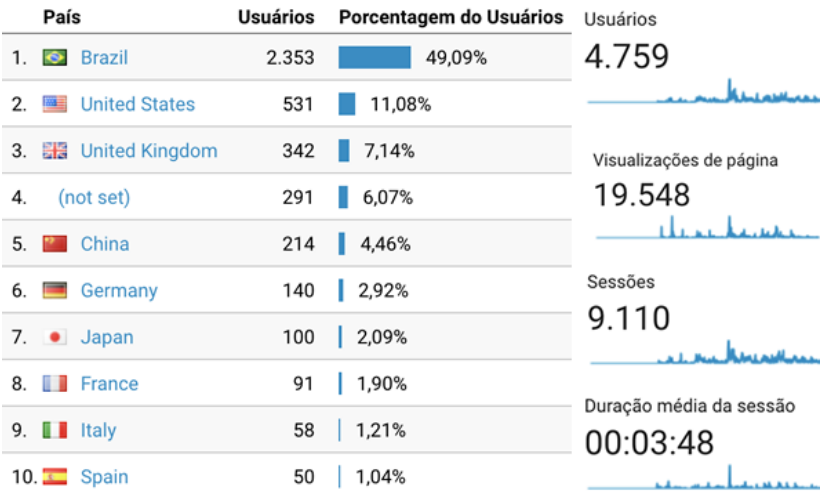
Text Mining techniques enable the analysis of large corpora in an automatic process without losing track or missing significant patterns as could happen if a text were analyzed manually. Naturally, we have to take into account the precision and recall[19] measured from the system being used. Computer algorithms can be trained (e.g., Machine Learning) to process and find variables based on a handmade set of examples with the desired translational features, using statistical measures or defined rules. It all depends on the chosen approach. The application of a twofold approach based on qualitative and quantitative methods as endorsed by Olohan (see Section 1.3.1), for analyzing and processing corpora, to my view, is essential for a comprehensive analysis of translated text which may point to possible new evidences.

To achieve this purpose, in my MA dissertation, I advocated for the employment of an easy to use a web-based corpus platform (i.e., COPA-TRAD) with a display interface enabling the researcher avoid dealing with complex and technical computer programming language and other processing mechanisms, such as the direct use of regular expressions (Silva, 2013, p. 32). Users ranging from undergraduate students to researchers with diverse technological background can manipulate corpus technological tools without prior technical knowledge.

---

[19] Precision when the search does not return too many "wrong" hits (i.e., irrelevant results), called false positives and recall when the search does not miss too many correct items, called false negatives (Lüdeling, Evert & Baroni,2007, p. 11).

As a result, COPA-TRAD has 268 registered users from universities across the globe (Appendix G) and more than 4000 online visits from Brazil and abroad totalizing nearly 20000 page views (Figure 5).



	País	Usuários	Porcentagem do Usuários
1.	Brazil	2.353	49,09%
2.	United States	531	11,08%
3.	United Kingdom	342	7,14%
4.	(not set)	291	6,07%
5.	China	214	4,46%
6.	Germany	140	2,92%
7.	Japan	100	2,09%
8.	France	91	1,90%
9.	Italy	58	1,21%
10.	Spain	50	1,04%

Figure 5. COPA-TRAD web analytics.

COPA-TRAD user access report indicate, among other things, the recognition from translators and researchers and show us how the system is performing. The uncomplicated interface provided the possibility from users with different levels and background use the platform. These report numbers give confidence in the adopted approach applied for COPA-TRAD construction as well as feedback from improvement. The resulting information mined from the corpus can be displayed for analysis through a set of tools available in the platform or even exported in CSV, XML and PDF extending the possibilities of analysis and gain more users. COPA-TRAD first version, also known as MVP (Minimum Viable Product), relied exclusively on pattern matching technology but I found Text Mining offers possibilities for processing texts far beyond the earlier technology. The user experience and acceptance, among other factors, set the groundwork for Corpus Mining and its application in COPA-TRAD. For this reason, I would like to emphasize my belief in a fruitful integration between Text Mining and CTS which provides a new perspective on the model utilized to investigate translational phenomena.

1.5.4. Research Method

The research method adopted in this study is based on the proposed Corpus Mining model. Most of the computational tools developed for COPA-TRAD current version were conceived following each step of the model. Afterwards the demonstration and discussion of the inner steps of the proposed tools, parts of the system are going to be used to show the architecture and how the adopted algorithms work. The proposed and discussed tools in the methodology are research outcomes from the investigation I conducted in my monograph to obtain a Software Engineering post-graduate specialization degree.

For COPA-TRAD context, Text Mining techniques found suitable for Corpus-based Translation Studies were implemented and other ones are still under development or in the backlog awaiting the technical analysis and development. Furthermore, the technological apparatus such as software or algorithms available on the Internet and used specifically for Text Mining or NLP are presented and the applications inside COPA-TRAD are going to be described. The priority for the use of software or any other additional technological solution was and will be based on the availability of the source code (i.e., Open Source[20]) as well as if it is Free Software[21] (i.e., software or any other technological tool with permission to modify, adapt and use the independent of situation).

The additional Open Source software execute, for example, sub-tasks in COPA-TRAD internal processing system. To illustrate, the use of an Open Source software named *Abiword* to convert files from DOC to TXT file format are part of COPA-TRAD automatic alignment tool. In case *Abiword* stops working for any case or is not available in the production environment (i.e., the hosting infrastructure), COPA-TRAD algorithm makes use of a self-hosted solution, developed in PHP, as an alternative tool for file format conversion. Similar fault-tolerant techniques were applied in COPA-TRAD when found necessary.

As already revealed, the methodology of this study is based on a standard model named Corpus Mining and will be deeply explained in the

---

[20] "Open source software is software that can be freely used, changed and shared (in modified or unmodified form) by anyone". From: http://opensource.org

[21] Free software, means, "that the users have the freedom to run, copy, distribute, study, change and improve the software". From: https://www.gnu.org/philosophy/free-sw.html

next section. By doing that, I expect Corpus Mining supports this study as well as researchers interested in conducting scientific translational research on parallel corpus. The design and proposal of Corpus Mining will be based on a set of questions and problems to be considered before entering in any parallel corpus development endeavor. The mentioned questions and problems I advanced in my MA dissertation (Silva, 2011, pp. 58-59) and it will guide the design phase.

In addition, well-known processes especially from Data Mining are used to provide a starting point for the model being proposed. The mentioned processes are Cross Industry Standard Process for Data Mining (CRISP-DM) and Knowledge Discovery in Databases (KDD). These two processes are in a way confusing, but used here according to the definition suggested by Fayyad, Piatetsky-Shapiro & Smyth, 1996a.

1.6. The Nature of Parallel Corpus Data

Before moving forward with this investigation, it is important to clear up the nature of parallel corpus which consists of texts arranged in a special kind of way (as discussed further in Section 2.3.6), and the difference it might have in relation to other kind of data stored in databases (e.g., quantitative data). The availability of corpus data varies according to a set of technical and methodological factors, for example, McEnery and Hardie (2012) clarify in their study that the term text "denotes a file of machine-readable data" the file is in fact "textual in form, so that each file represents, for instance, a newspaper article or an orthographic transcription of some spoken language" (p. 2). Halliday and Matthiessen (2004) provide a wider definition for texts: "when people speak or write, they produce text. The term 'text' refers to any instance of language, in any medium, that makes sense to someone who knows the language" (p. 3).

Here, text share the same concept as the one provided by the cited researchers, however, database is being used instead of electronic files. The use of database maximizes the possibilities to manipulate and attach extra information to these machine-readable data. It important to emphasize possibility of a corpus being compiled with other type of data, such as media—to name video, audio or images (McEnery & Hardie, 2012, p. 2).

The available data for a parallel corpus is entirely texts or snippets of it, indifferent of that, texts are samples of natural language used by humans to communicate and prone to possible errors, language twists or paradoxes and dualities such as ambiguities. As well observed by Gupta and Lehal (2009):

> Humans have the ability to distinguish and apply linguistic patterns to text and humans can easily overcome obstacles that computers cannot easily handle such as slang, spelling variations and contextual meaning. However, although our language capabilities allow us to comprehend unstructured data, we lack the computer's ability to process text in large volumes or at high speeds (p. 60).

As a matter of fact, the problems related to texts in digital form extrapolates the language barrier and other problems of different nature have to be dealt, such as text encoding, text special characters and symbols, specific file formats and the means to process raw data in order to extract clean text for automatic processing and analysis.

Due to aforementioned assumptions, texts in digital form are different from the printed versions. Not just because the medium, but also because the transformations such as the ones already mentioned, as well as the "change in font type, the loss of page boundaries, or changes in the representation of paragraphs" constraints possibly unimportant for many kinds of linguistic analysis, but quite important to understand certain translational features that are lost in the process of digitizing and processing (Barlow, 2004, p. 206). To overcome such caveat, a quantitative and qualitative approach is advocated by this study (see Section 2.2.3).

The techniques to store and organize textual data in a parallel corpus may vary according the scope of the project and the decisions made during the requirement gathering phase. However, the basic unit for any parallel corpus is still texts in natural language, more specifically, texts in two or more languages consisted of originals and translations aligned at word, sentence or paragraph level.

The nature of textual data may vary and there is a possibility the researcher have to deal with such heterogeneous object of study. To keep an open ended corpus, an ever growing database, and in the case of COPA-TRAD, receive contributions from researchers and translators around the world the system have to be elastic in the sense of the data it receives. For instance, Fernandes aligned the first texts provided for COPA-TRAD in his PhD study (2004). The texts contained XML data to label the corresponding paragraphs in the target language. Not to mention a the use of a XML file to retain extra textual information such as data of

publication, author name, translator name, gender, nationality, etc. (Fernandes, 2004, Appendix D). In order to receive such body of textual content and load it in COPA-TRAD database, an especial mechanism able to interpret the XML markup was developed to process such content, so in this case, the system provided some "elasticity" (i.e., the capacity for processing texts in distinct formats) to receive such material. Based on the nature of the heterogeneous data type as part of my investigation in my specialization degree, I investigated (Silva, 2004, p. 35) the three kinds of data:

- Unstructured Data – Texts in natural language created by humans in general or machines in a variety of formats such as TXT, PDF or DOC without any kind of labeling technique to describe any linguistic portion of data just as words or expressions.
- Semi-structured Data – texts marked up with specific labels or tags such as HTML or XML.
- Structured Data – information stored, organized and categorized in database tables, noSQL[22] objects or RDF[23]. The structured data can be quantified and retrieved from database using query languages such as SQL[24].

Having in mind the three categories mentioned above, the nature of texts used for COPA-TRAD relies mainly on unstructured data with some exceptions for semi-structured data as observed by the aforementioned example of Fernandes compiled documents (2004). However, there is a possibility to store and process large quantity of semi-structured data and COPA-TRAD provides such elasticity for this kind of data. The possibility to deal with semi-structured data has to be considered because Bhatia et al. (2011) informed that "much of the web information is semi-structured due to the nested structure of HTML code, much of the web information is linked, and much of the web information is redundant" (p. 443).

As COPA-TRAD being an open-ended corpus some of the texts will be extracted from webpages and it must be a possibility to use texts from webpages encoded in HTML format. At this point "text mining is an important step of knowledge discovery process. It is used to extract hidden information from not-structured or semi-structured data" (Bhatia

[22] Not Only SQL.
[23] RDF stands for Resource Description Framework.
[24] SQL stands for Structured Query Language.

et al., 2011, p. 443). Text mining techniques are used in this study to investigate translated texts as well as well as the NLP techniques that handle the raw text data submitted by users in COPA-TRAD platform.

1.7. Organization of the Study

After introducing the context as well as the objectives, research questions and the significance of the study, Chapter Two discusses the theoretical background and concepts permeating the type of analysis here envisaged, Chapter Three describes the methodology employed in the study. Next, Chapter Four is related to the application of Corpus Mining on COPA-TRAD Version 2 and Chapter Five sets out the final remarks with the recapitulation of the previous chapters, an attempt to answer the research questions and finally the limitations and suggestions for future study.

2. CHAPTER TWO: Theoretical Background

2.1. Initial Remarks

This chapter discusses the theoretical foundations informing the present study. Due to the density of this chapter and the correlation between topics, I decided to organize it into two main sections. The first one is related to linguistic issues discussed from a Corpus Linguistics (CL) and Corpus-based Translation Studies (CTS) perspective. The second section focuses on technological concepts such as Text Mining and Natural Language Processing, as well as specific computerized techniques used in this study. Due to the complexity of the concepts in these two main sections, it is important to highlight the fact that only those concepts directly related to this study will be explored.

2.2. Linguistic Background

In the 1960s, the study of language and translation by means of electronic corpora and computerized processing algorithms was born, and since then, it has expanded at an incredible pace (Kennedy, 1998, p. 13). Specific disciplinary fields comprised of technology and linguistics emerged, sharing a common point of focus (i.e., technology and language) but maintaining individual emphasis (i.e., one addressing software, the others addressing language). In summary, though each field was defined as separate, the common ground shared was to scrutinize human communication. However, the approach used to manipulate linguistic resource may diverge among the different disciplinary fields interested in the aforementioned subject; because of that, some problems, as elicited in Section 1.5.1, took the scene. In fact, divergent or not, being problematic or not, the relevant point to keep in mind is the huge contribution to the study of language that each field produced. Even if in the areas of research where the main objective is not the investigation of linguistic or translational feature, intrinsically speaking, the contribution is perceivable. To name some topics for such areas of research, they might include: development of dialogue agents, machine translation, or text-to-speech synthesis. The key point, here, is that the primary source of data is the linguistic resource, and the cited areas have to deal with it, as observed by Yang and Li (2003):

> The linguistic resource is considered to be an important factor in natural language processing

> applications and information retrieval, which is particularly important for cross-lingual information retrieval in the recent research. […] The linguistic resource includes written and spoken corpora, lexical databases, grammars and terminologies (p. 731).

In addition, Godfrey and Zampolli (1997) explain the application of linguistic resource is important to "building, improving, or evaluating natural language and speech algorithms or systems" (p. 381). (The subject of Natural Language Processing is reviewed in Section 2.3.5.)

As discussed, the techniques developed to process linguistic resource, such as plain text in natural language, contributed to several other studies and have this common ground: scientific understanding of language. The outcomes reached so far were refined and improved over time, and nowadays, they are used in software ranging from automatic text processing to translation-aided tools. For instance, in computer science, the search for the holy grail of machines translating like humans did not succeed—although I have to recognize the advancements in the field, especially in the present moment with the application of a specific Deep Learning technique known as Neural Machine Translation, in the Google Translate system (Wu et al., 2016). Computational Linguistics is one of the disciplinary fields contributing directly to the investigation of the relationship between linguistics and computing, and theorizing and developing systems capable of recognizing and producing information in natural language (Vieira & Lima, 2001, p. 1). The fact that Computational Linguistics is one of the main fields informing and supporting Text Mining (Figure 6) should also be noted.

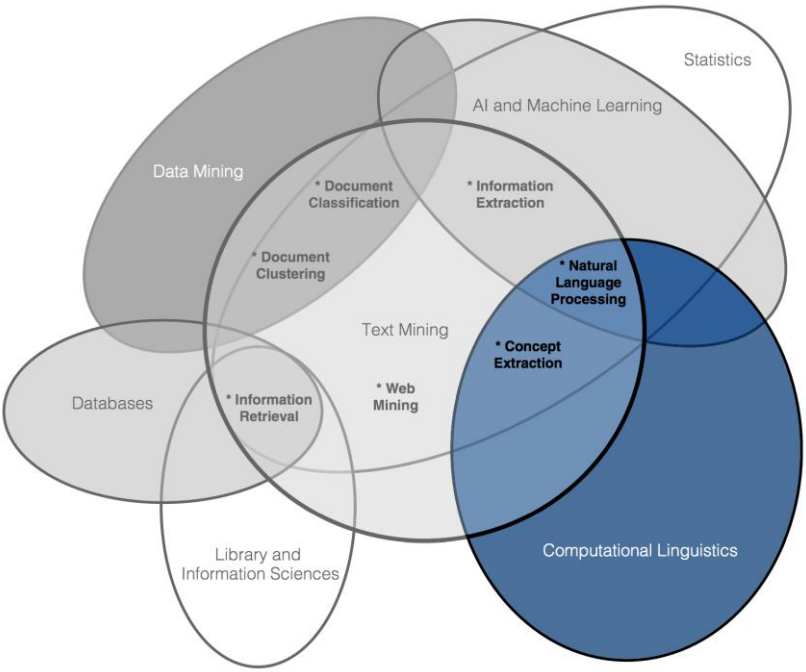


Figure 6. Adapted diagram showing the intersection between Text Mining and Computational Linguistics (Miner et al., 2012, p. 31).

Indeed, the first mention of Computational Linguistics was in 1949, when Warren Weaver suggested the possibility of automatic translation; then the focus on machine translation was abandoned "in favour of more fundamental scientific research in language and language processing" (Kay, 2003, p. xvii). The object of study changed, as well as the name of the field, according to Kay (2003):

> Progression from machine translation to Computational Linguistics occurred in 1974 when Machine Translation and Computational Linguistics was replaced by the American Journal of Computational Linguistics, which appeared initially only in microfiche form. In 1980, this became Computational Linguistics, which is still alive and vigorous today (p. xvii).

I would like to emphasize that the object of study in Computational Linguistics changed for a while, because in the 1980s, "machine translation began to look practical again" (Kay, 2003, p. xvii).

According to Kay (2003), there were two motivations for the development of Computational Linguistics: One was theoretical and "came from the growing perception that the pursuit of computational goals could give rise to important advances in linguistic theory." The second motivation "came from the desire to create a technology, based on sound scientific principles, to support a large and expanding list of practical requirements for translation, information extraction, summarization, grammar checking, and the like" (p. xviii). Although, in both cases, Kay argues that no complete success was achieved "by linguistic methods alone" (p. xviii). Computers are capable of doing far more than just processing math and rules. natural language algorithms have the power to understand human communication, in a certain extent, which is much greater and more complex a task than just solving simple arithmetical counts. Nowadays, computers try to achieve such challenges by extrapolating meaning from huge information databases (see, for example, the role of Big Data).

Therefore, Kay (2003) recognizes Computational Linguistics itself cannot solve the problem of understanding language in context and that the field providing the necessary solution is Artificial Intelligence: "but the task is clearly a great deal more daunting even than building comprehensive linguistic models, and success has been limited" (pp. xviii-xix). For this reason, Kay (2003) recalls the fact Computational Linguistics "gained a reputation for not measuring up to the challenges of technology" and that this produced "frustration and misunderstanding" (p. xix). Due to the lack of an Artificial Intelligence system for linguists seeking to apply their scientific research, they "have been forced to seek a surrogate, however imperfect, and many think they have found it in what is generally known as 'statistical natural language processing'" (p. xix).

The discussion and details encompassing Natural Language Processing (NLP) is going to be dealt in the next section, but to provide a glimpse here, it "associates probabilities with the alternatives encountered in the course of analyzing an utterance or a text and accepts the most probable outcome as the correct one" (Kay, 2003, p. xix). The use of corpora, more specifically the parallel type, has been an object of interest in Computational Linguistics mainly because of the subject of machine translation. McEnery (2003) affirms and confirms that "corpus data are, for many applications, the raw fuel of NLP, and/or the testbed on which an NLP application is evaluated" (p. 448). In addition, Kennedy (1998)

indicates researchers of Computational Linguistics "have been concerned with the use of corpora to develop, among other things, algorithms for natural language processing and the modeling of linguistics theories" (p. 9). Due to this considerable interest in corpus, Othero (2006) makes it clear that Computational Linguistics can be divided into two branches, namely Corpus Linguistics and Natural Language Processing (p. 342).

2.2.1. Corpus Linguistics

A corpus is the primary linguistic resource of natural language processing applications and speech algorithms. The "linguistic resource refers to (usually large) sets of language data and descriptions in machine readable form" (Godfrey & Zampolli, 1997, p. 381). In relation to "corpus" define it as a "set of language data" is a narrow view that do not includes all the intricacies involved with it. For McEnery (2003), the concept of corpus "should properly be applied to a well-organized collection of data, collected within the boundaries of a sampling frame designed to allow exploration of a certain linguistic feature (or set of features) via the data collected" (p. 449). The application of rigorous procedures for the corpus compilation, as suggested by McEnery, reinforces the validity of the study being conducted and the results produced. In some cases, the outcomes from the studies in the area of Corpus Linguistics are quantitative data, due to the aforementioned large sets of language data, but it is not restricted to it, as clarified further on.

Corpus Linguistics investigates language through means of electronic corpora (or corpus, in singular form) and sophisticated computational techniques. This field "is not an end in itself but is one source of evidence for improving descriptions of the structure and use of languages, and for various applications, including the processing of natural language by machine and understanding how to learn or teach a language" (Kennedy, 1998, p. 1). The emergence of Corpus Linguistics did not coincide with the growing availability of computers, but it is undeniable that the field was propelled forward more quickly with the advent of this technology (Kennedy, 1998, p. 2). Computers reduced "much of the drudgery of text-based linguistic description and vastly increasing the size of the databases used for analysis" (Kennedy, 1998, p. 2).

The two fundamental techniques in the field, concordances and frequency data, "exemplify respectively the two forms of analysis, namely qualitative and quantitative, that are equally important to corpus linguistics" (McEnery & Hardie, 2012, p. 2). In like manner, McEnery and Wilson (2001) see the value of "supplementing qualitative analyses

of language with quantitative data" (p. 81). The application of qualitative and quantitative approaches in research is not defended by some scholars, and they tend to criticize quantitative methods. For this reason, the authors argue that a quantitative approach goes beyond the "simple counting," and various sophisticated statistical techniques are used, all of which provide a mathematically rigorous analysis of often complex data (McEnery & Wilson, 2001, p. 81). These different methods are far from complicated, but the field of Corpus Linguistics may seem fuzzy[25] for researchers outside the area. The authors, McEnery and Hardie (2012), provide a suitable definition of Corpus Linguistics, describing its heterogeneous object of analysis:

> It is certainly quite distinct from most other topics you might study in linguistics, as it is not directly about the study of any particular aspect of language. Rather, it is an area which focuses upon a set of procedures, or methods, for studying language (although, as we will see, at least one major school of corpus linguistics does not agree with the characterization of corpus linguistics as a methodology). The procedures themselves are still developing, and remain an unclearly delineated set – though some of them, such as concordancing, are well established and are viewed as central to the approach. Given these procedures, we can take a corpus-based approach to many areas of linguistics (p. 1).

Because of the nature of Corpus Linguistics as a heterogeneous field, the authors reveal that this topic "has the potential to reorient our entire approach to the study of language" and that this field facilitates the exploration of new language theories (p. 1). According to the authors, such new language theories emerge from attested language use. The application of language used to formulate new theories is a shift from prescriptive to descriptive methods of observing language. In addition,

[25] Corpus Linguistics can be difficult to understand to novice researchers specially with subjects related to the field considered complex, for example, the discussion and application of computational algorithms or even statistical calculations.

there should be no difference between "corpus-based" and "corpus-driven," because McEnery and Hardie (2012) sustain:

> For those who accept it, the corpus-based versus corpus-driven dichotomy creates a basic, binary distinction, under which most works of corpus linguistic research can be sorted into one or the other group. However, our own perspective rejects the notion that the corpus itself has a theoretical status, and thus also rejects the binary distinction between corpus-based and corpus-driven linguistics. From this point of view, all corpus linguistics can justly be described as corpus-based (p. 6).

I concur with the opinion of the authors, and because of that, the term "corpus-based" permeates this study, as well as the use of "corpus" in Translation Studies—namely, Corpus-based Translation Studies.

Regarding Corpus Linguistics, the amount of data being processed is relevant for the reason that it produces more generalized results based on statistical measures. However, we cannot guarantee total accountability of a language feature, as we will see further on in this paragraph. Consequently, alternative methods such as the application of statistics and specific algorithms, in order to quantify, for example, linguistic patterns and a likelihood that statistical measures are applied in order to deduce a linguistic behavior in a broad sense and find "what is likely to occur in language use" (Kennedy, 1998, p. 8). For this reason, "the advantage of processing a text corpus is that it is possible to obtain context-specific information about syntactic structures and the usage of words in a given language" (Li & Yang, 2006, p. 632).

Kennedy (1998) explains that the Corpus Linguistics object of study, "like all linguistics, is concerned primarily with the description and explanation of the nature, structure and use of language and languages and with particular matters such as language acquisition, variation and change" (p. 8). It is important to call attention to the fact that the resulting description of language should not be induced or manipulated to fit the researcher's hypothesis. As observed by McEnery and Hardie (2012), who advise avoiding induced results from the linguistic hypothesis being tested, they offer this: the "selection of examples to favour those examples that fit the hypothesis, and no screening out of inconvenient examples" (p. 15). Filtering out or ignoring examples "from the corpus that do not fit the hypothesis under investigation" can result in contradictory

hypothesis, and the "use of corpus data would be fatally undermined" (McEnery and Hardie, 2012, p. 15). On the other hand, McEnery and Hardie recognize the criticism involving the utilization of corpus, because in essence, "the corpus itself is necessarily a finite subset of a much larger (and in principle non-finite) entity, language"; but, they argue that "total accountability in corpus linguistics must be moderated." The total accountability is restricted to the corpus under investigation and not language entirely (p. 15). The criticism related to issues such as total accountability is not exclusive to Corpus Linguistics; for example, in the area of astronomy, they "theorise on the basis of the subset of the Universe that is visible to them," and over time, this subset is expanded with the support of new technology (p. 15).

Apart from the linguistic outcomes, Corpus Linguistics contributed to the emergence of Corpus-based Translation Studies (CTS) (see Section 2.2.2). Nevertheless, there are issues between Corpus Linguistics and CTS beyond the scope of quantitative data criticized by some scholars. As Kenny (2001) recognizes, "corpus linguistics thus take an empirical approach to the description of language," and this, to a certain extent, is similar to CTS, although the problem is that corpus linguists "insist upon the primacy of authentic, attested instance of use" (p. 50). In other words, translated texts are not included in the repertoire of objects for being investigated in Corpus Linguistics because it may not seem "authentic" but, rather, artificial. In like manner, Fernandes (2006) denies the relevance of linguistic corpora in Corpus-based Translation Studies; since these types of corpora "do not provide realistic models for trainee translators," the absence of translated text cannot provide for trainee translators "what procedures are being used by professional translators" (p. 92). Malmkjær (1998) proposes a different view on the issue. In her viewpoint, "linguists and translators ought to be the best of friends; their areas of interest, however one wants to look at them, and however they may differ, have language and linguistic activity at the centre" (p. 2). For this reason, the next section discusses CTS in depth, as well as its implications for Translation Studies and for those who are interested in conducting research in the field.

2.2.2. Corpus-based Translation Studies

Corpus-based Translation Studies (CTS) is the main theory informing this study. CTS emerged from descriptive studies which in turn started to be adopted in Translation Studies. Indeed, CTS grew out of corpus linguistics and assumed a role of protagonist in Descriptive Translation Studies to the point Tymoczko (1998) alleges "CTS at the

same time marks a turn away from prescriptive approaches to translation toward descriptive approaches" (p. 652). The descriptive approaches provide the means to investigate the process and the product of translation taking into consideration the smallest details (such as choosing the text by the individual translator) as well as the largest cultural patterns intrinsic and extrinsic to the text (Tymoczko, 1998, p. 653).

For Olohan (2004) Translation Studies is a discipline focused on the study of translational phenomena "from a diverse of angles and perspectives" and it is interested in: (a) translator role mediation; (b) the process of translation; (c) the translation product; (d) the causes and effect related to translator activity and so on (p. 5). Finally, the author emphasizes that "what translation studies focuses on" is a hotly debated issue (Olohan, 2004, p. 5). Rather than comparing translations with its corresponding originals, Translation Studies and clearly CTS are focused on the understanding of the inner mechanisms in a translated text, the context of production and under what circumstances the translated text was produced. In the same fashion, Laviosa (1998) emphasizes the aim of CTS is "to understand the specific constraints, pressures, and motivations that influence the act of translating and underlie its unique language" and not simply investigate the "third code" (p. 474). In a subsequent paper Laviosa (2011) details CTS object of study:

> First, the object of study consists of authentic samples of language use rather than idealized entities; linguistic regularities are regarded as probabilistic norms of behaviour rather than prescriptive rules; language patterns reflect and reproduce culture. Moreover, both corpus linguistics and DTS adopt a comparative research model in which descriptive hypotheses that make claims about the probabilistic generality of a given phenomenon are put forward, and texts are examined across corpora representing different language varieties, for example, translated versus non-translated language, original texts and their translations, different text types or different modalities within the same language, and so on (p. 14).

The change of paradigm from prescriptivism to descriptivism set up the scenario for the emergence of CTS and the creation of a methodological, theoretical and systematic or scientific model to

investigate translational phenomena and its features (Olohan, 2004, pp. 5-11). Olohan's (Ibid.) arguments are appropriate for researchers interested in corpus methods applied to translation and the shift to descriptivism made it possible. The scholar keeps clear this point and ensures "corpus-based studies in translation are clearly aligned with the descriptive perspective" (Olohan, 2004, p. 10).

Mona Baker, who was an advisee of John Sinclair and was inserted in Corpus Linguistics context, was one of the first to employ corpus tools for investigating translation. The seminal paper entitled "Corpus Linguistics and Translation Studies: Implications and Applications" was published by Baker "in 1993 as part of a collection of research articles in honour of John Sinclair" (Laviosa, 2011, p. 13). According to Laviosa (2011), two years later in Corpora in Translation Studies: An Overview and Some Suggestions for Future Research, published in Target journal, the first idea claimed by Baker (1993) in her seminal paper "was further developed by suggesting specific research projects involving the design and analysis of parallel, bi/multilingual and, above all, monolingual comparable corpora" (p. 13). Due to this fact Laviosa (2002) recognizes the endeavor and contribution of this scholar for the emergence of CTS: Mona Baker "well deserves the affectionate title of mother of Corpus-based descriptive Translation Studies" (p. 18). Since the seminal papers published by Mona Baker (1993) starting in the early 1990's CTS gained its momentum and nowadays is a solid area of Translation Studies.

Corpus-based Translation Studies is not interested in generating numbers to testify a possible hypothesis. According to Fernandes (2006), "the issue of corpus size in CTS becomes a relative one in the sense that qualitative aspects sometimes may be more relevant than quantitative ones" (p. 88). However CTS recognizes the contribution of quantitative approach if combined with a qualitative research not only based on the text itself whether is original or not but the social and extra linguistic conditions of the translator why he or she translated in this way and not that way (Olohan, 2004, pp. 8-9, 22, 86). The combination of different methods point "corpus work as a research methodology" and can be applied "with a range of theoretical frameworks, assumptions, tools and concepts" and here I include Text Mining and its related techniques (Olohan, 2004, p. 9). In like manner, Tymoczko (1998) believes in CTS openness "rather than a convergent approach to the theory and practice of translation" (p. 656). Instead creating a division between different areas and theories Olohan (2004) calls for a unification:

> The relevance of these perspectives for a discussion of corpora in translation studies relates to the fact that, on the one hand, research using corpora is generally seen as linguistic, empirical, quantitative, data-based or data-driven. However, it has also been billed as an approach with potential to unify opposing perspectives (pp. 8-9).

This is possible by means of corpus work and adopting the product (i.e., the translated text) as the primary source of investigation. With this in mind, the application of corpus methods and techniques contribute to "describe translation and the behaviour of translators" extra textual information that contributes to understand the translated text (Olohan & Baker, 2000, p. 142). This change of paradigm from original or source text to translated or target text focus on the object of study "not in terms of its equivalence to source texts but as valid object of study in its own right" (ibid.). The focus on the translated text is an important pace for Translation Studies in general because scholars specially from Linguistics tend to consider translated text the same as the original with variation between then and possibly because of that, translated text has been excluded from scientific endeavors such as the British National Corpus (Olohan, 2004, p. 13).

The methods employed to investigate translation in CTS context might vary according to the nature of the research being envisaged. Depending on the hypothesis under consideration, more than one method can be employed occasionally. This situational context opened the gates to conduct research using CTS as a method even if the research itself is concerned to study a translational phenomena from another TS theoretical field. In view of this extensive applicability in the area of Translation Studies Olohan adds that "corpus methodology clearly has some applicability within the broad theoretical framework of DTS" (i.e., Descriptive Translation Studies) because "it provides a method for the description of language use in translation" (Olohan, 2004, p. 17). In particular to this premise, Olohan (2007), observes the primary use application of CTS:

> Corpus-based methods of analysis in translation studies have been used primarily in relation to literary or journalistic texts to study translation strategies, linguistic patterning in translations, creative or conventional use of language by

> translators, simplification or explicitation in translation, translator's style, etc. (p. 139).

In addition, Olohan points out areas not included in the applied branch of Translation Studies that started to use CTS methods and analytical framework to investigate "aspects of scientific translations" (cited above) such as "corpus linguistics, corpus-based translation studies, descriptive translation studies, studies of rhetoric and style, discourse analysis and so on" (p. 140). Likewise the research method already mentioned, my objective is to include Corpus Mining in the methodological apparatus being proposed here in order to contribute to the investigation of translation, especially at a technical level. This is extremely important since "as with any scientific or humanistic area of research, the questions asked in CTS will inevitably determine the results obtained and the structure of the databases will determine what conclusions can be drawn" (Tymoczko, 1998, p. 654). As observed by Tymoczko (1998) all the aspects permeating a CTS research are important and relevant because they have direct impact in the results and conclusions even the model designed for the database and how the data are being stored.

Unfortunately scholars in CTS such as Malmkjær (1998) exposes "some literature on corpus linguistics suggests a worrying tendency to try to divide the labour in such a way that the mind's share is minimal or expended during the corpus construction phase" (p. 7). Malmkjær (1998) has the impression that the only "for texts' inclusion in the corpus" is the digital format able for machine processing (p. 7). As observed, this is not the case in CTS since extra textual information are highly valuable to understand translation and due to this reason this kind of data have to be part of the texts compounding the parallel corpus. Differently from the practices exposed by Malmkjær textual content do have impact in the design of the database.

The application of a CTS method for the type of research being conducted "can reveal common features of translated text and thus provide insight into the translation process" (Olohan & Baker, 2000, p. 141). In fact, trying to understand what happens inside the head of the translator (commonly considered as a "black box") is not possible. As a matter of fact, the objective of CTS is not that one, but try to understand the translation process based on (a) the corpus analysis and (b) comparing with other translated text snippets to find similar patterns and then come up with an outcome. For that reason, a corpus-based analysis "can contribute to the description or characterization of individual translators' linguistic behavior" (p. 141). Of course the application of CTS has some

limitations as well, for example, Tymoczko (1998) observed these "studies are products of human minds, of actual human beings, and, thus, inevitably reflect the views, presuppositions, and limitations of those human beings" (p. 654). For this reason, even when working with large corpora the researcher has to avoid the use of generalizations but rather create hypothesis and conclusions always based in the domain of his corpus and the size of his corpus, among other attributes, to keep clear the comprehensiveness of the study. Based on the aforementioned assumptions the pioneering work of Corpus-based Translation Studies require the adaptation of new technology and innovative methods to support, expand and improve the discipline (Tymoczko, 1998, p. 658).

2.2.3. Quantitative vs. Qualitative Research

The use of the term "versus" in the title begs a question of whether quantitative and qualitative research are really "opposed" at all. Some researchers promote qualitative over quantitative research, while others are the opposite. Still others, as I do, promote a combination of the two, as I will explore over the following report.

Generally speaking, a quantitative approach can help one interpret events or the objects of study in terms of numbers and mathematical formulae. However, quantitative methods to investigate corpus for translational purposes are "regarded as limited in their usefulness" in application; but this "does not mean, of course, that they are of no use" (Olohan, 2004, p. 22).

Olohan calls for applying qualitative and quantitative methods of research and analysis to corpus numerical and statistical data as well. She indicates each method of analysis can complement each other rather than oppose each other (Olohan, 2004, p. 22). A similar point of view is shared by Baker (2004) who advocates that "we need to take a closer look at the data and get a feel for the texts and what is happening in them, as well as the people who produce these texts, in order to move beyond low-level description to situated explanation" (p. 183). Thus, beyond investigating the product but also, we have to understand the processes behind its context and the social interactions involved in the translation's production.

Baker (2004) stresses that "corpus-based research in principle takes textual material as a starting point, but this does not mean that it necessarily ignores or sets out to downplay the human element" (p. 184). A comprehensive investigation would provide the necessary information for understanding the circumstances in which the text was produced. Hence, Baker (2004) also adds that this approach is not or should not "be

seen as a free-standing methodology that does not need to be complemented by other methods of research" (p. 184).

Gries and Wulff (2012) consider the quantitative method as an "evolution of the field towards more empirically rigorous" approach (p. 35). Gries and Wulff (2012) further support their point of view by saying that "after a long reign of generative approaches to grammar and their largely intuitive grammatically judgment" the interest in "probabilistic theories" raised substantially (p. 35). The authors also emphasize the idea that "many areas of applied linguistics" are using "quantitative tools, and translation studies are no exception" (p. 35). In this sense, they give clues qualitative methods did not evolve or that they are not rigorous, which would be is completely false.

For Ji (2012) the quantitative methods have problems which qualitative ones do not. Ji (2012) recognizes "the lack of systematic descriptions of quantitative methods" and this situation "posits a serious problem for the theoretical development" of Corpus-based Translation Studies (p. 53).

The point Ji (2012) wants to call our attention to is that research is the attempt to bring in quantitative research variables broadly used in qualitative research such as social, cultural, stylistic and textual representation of a translation (p. 55). We sometimes need both to form a fuller picture of what we study.

For the researcher qualitative variables "remain essentially under-explored" from a quantitative approach (Ji, 2012, p. 55). Ji also adds other variables such as social and cultural context along with history can be analyzed and quantified through being monitored for "statistically significant correlations" among the data (pp. 55-56).

Based on these ideas Ji (2012) proposes the investigation of three types of relationship: (a) "source text and the target text relationship"; (b) "the translation and the target social and cultural context"; (c) "translated language vis-à-vis the target language in general" (p. 70).

In the third type the author indicates the existence of a difference between translated and original text in the same language, subject to being investigated as translationese (i.e., the third code). For Laviosa (1998) the objective of Corpus-based approach is not the investigation of a "third code" by itself but "to understand the specific constraints, pressures, and motivations that influence the act of translating and underlie its unique language" (p. 1). It is possible to perceive the view shared by this author in relation to other kinds of methodology by selecting excerpts from the text at random without any systematic procedure to conduct the selection.

On the other hand, McEnery and Wilson (2001) point out that a simplistic approach to working with quantitative data consisting of classifying items according to a specific model and then performing "an arithmetical count of the number of items," words or any other linguistic feature relevant for the project (p. 82). The comparison of frequency lists and linguistic features in both languages (source and target) it is a common quantitative method used in corpora (McEnery & Xiao, 2007, p. 4).

Rather than relying on one research methodology to the detriment of potentials we can gain from the other, it would be fruitful if we adopted a mixed-methods approach (i.e., both qualitative and quantitative). A simple analogy to understand this approach, from my point of view, is to think the quantitative analysis as a "compass" because it will give the researcher the direction or point what direction to follow; meanwhile the qualitative can be viewed as a "magnifying glass" by maximizing the possible clues pointed out by the first approach.

2.3. Technical Background

The technical theoretical background dealt in this section are key concepts employed along this study. It is a discussion of key concepts related to this study and aims to support the technical aspects of Corpus Mining as well as the methodological procedures discussed in Chapter 3. This section, discusses in a narrow sense the concepts and technicalities related directly to the proposed subject and it is far beyond the scope of this study provide a comprehensive analysis or a review of the discussed disciplinary fields. However a technical discussion is necessary "for the first-stage examination of a text" which is the processing mechanisms such as lemmatizers, taggers, and parsers and other advanced techniques applied to create useful knowledge from text (Sinclair, 1987, p. 84).

2.3.1. Knowledge Discovery in Databases and Data Mining

Knowledge Discovery in Databases (KDD) is a multistep process utilized for transforming vast amount of raw data into useful knowledge, to put it simply, "make sense of data" (Fayyad, Piatetsky-Shapiro & Smyth, 1996b, p. 37). To achieve this purpose, one critical step in KDD process is Data Mining (DM). Data Mining techniques play an important role in knowledge discovery, and as observed, it is a critical phase in KDD. Whereas an important part of researchers identify DM as a phase of a large process (KDD), there are others who adds DM in an especial place, a disciplinary field (Fayyad, Piatetsky-Shapiro, & Smyth, 1996b, pp. 39; North, 2012, p. 3). Both sides are proper, since the first one apply

DM techniques in KDD inner process and they work in an intersection between fields, though the other, are focused specifically in the intricacies of the disciplinary field. For Fayyad, Piatetsky-Shapiro and Smyth (1996b) in their seminal paper KDD and DM are defined as follows:

> In our view, KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data (p. 39).

KDD as discussed, is a multistep process or end-to-end it involves data gathering, storage, and how it can be accessed as well as how data are going to be processed through DM algorithms and visualized mainly by ordinary users, without prior technical knowledge (Fayyad, Piatetsky-Shapiro and Smyth, 1996b, p. 40). In other words, KDD process is applied to turn "low-level data into high-level knowledge" (Goebel & Gruenwald, 1999, p. 21).

The core element in the KDD process undoubtedly is Text Mining techniques which can be viewed as an interactive step since it can change depending on the planned goals and outcomes (Goebel & Gruenwald, 1999, p. 22). The nature of DM is quite diverse as well since it is a confluence of areas such as statistics, logic, artificial intelligence, machine learning and data management systems (North, 2012, p. 3). For this reason, Krochmal and Husi (2018) recognize the interdisciplinary nature of Data Mining (p. 233). Due to this diverse field the "success of a data mining project depends on the proper mix of good tools and skilled analysts" (Wirth & Hipp, 2000, p. 30).

DM techniques are used for data manipulation and production of useful information from databases as well as what I am going to discuss—Text Mining (see Section 2.3.2). In fact, Text Mining emerged from DM, but the kind of data each one process is quite different – unstructured Data (i.e., text in natural language) opposed to structured data (i.e., numerical and statistical data as well as relationships among elements from databases).

Data Mining as a whole, have been used for years in companies of several activities especially retailers, marketing agencies, call centers, and accounting. The techniques to process, analyze, and understand data can be traced back to 1980s but it was in late 1990s started to flourish due mainly the adoption of it in large companies (North, 2012, p. 5). The origins of DM refer to Artificial Intelligence a complex area dealt

specifically by disciplines of IT and Statistics, but it gained popularity to the point user friendly applications running on desktop PCs can perform DM tasks (Moscarola & Bolden, 1998, p. 406; Youzhi, 2010, p. 459).

As a matter of fact, Data Mining reveals "a large number of implicit, previously unknown and potentially non-trivial value of information" (Youzhi, 2010, p. 459). The primary objective of DM is derive suitable patterns (previously unknown) from structured information, for example, relational databases or data warehouse[26] systems. To derive useful knowledge from structured data, computer algorithms are key point in any Data Mining application. In addition, these algorithms are used to extract patterns from large datasets[27] (Fayyad, Piatetsky-Shapiro, & Smyth, 1996b, pp. 39-40). For North (2012) DM enables people to locate and interpret the patterns in order to support the decision making (p. 3). DM do not look backwards in the past, but also and mainly forwards by predicting future patterns and practices and for this reason DM techniques are extremely valuable for companies (Sucharita, Rao, Satya & Rajarajeswari, 2017, p. 38). The final results or the information created, possibly novel, from the application of Data Mining techniques can be both, quantitative or qualitative, this last one less frequent but also possible (Krochmal & Husi, 2018, p. 233).

In order to extract, manipulate and analyze data, DM apply two different models, one is supervised and the other unsupervised, it can be divided in three stages, namely: (a) Cleaning Data (i.e., remove noise and duplicates, solve conflicting and truncated data—when possible); (b) Extracting (i.e., selecting convenient data suitable for the desired analysis); (c) Transforming and Loading (i.e., preparing and organizing data to suit for a particular DM algorithm) (Sucharita et al., 2017, p. 38; Chary, Reddy & Bhuahan, 2018, p. 1341; Simoudis, 1996, p. 26). For Sucharita et al. (2017) the unsupervised model "deal with finding the

---

[26] Data warehouse (DW) refers to specialized software used for data integration, storing, reporting and data analysis. This technology is largely used in enterprise environment to support business decisions. "A data warehouse ensures that disparate data is integrated consistently under a single data model and is cleaned in the process" (Simoudis, 1996, p. 30). For Fayyad, Piatetsky-Shapiro and Smyth (1996b) DW supports KDD in two main phases: data cleaning and data access (i.e., provide uniform methods for accessing data) (p. 40).

[27] "A collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer" (Sucharita, Rao, Satya & Rajarajeswari, 2017, p. 42).

clusters from the given data where there is no teacher" (i.e., training examples without labels/tags) and the supervised "deal with classification of the data for making predictions" (i.e., training examples with labels/tags linking to classes) (p. 38).

As the same manner KDD has an interactive process, Data Mining has a well-established and solid process known as CRISP-DM (CRoss-Industry Standard Process for Data Mining). CRISP-DM is a cyclic and interactive process where the completion of one step leads to the next one or in case any problem is found is also possible to return backwards do the necessary correction and continue once again. An important aspect as observed by North (2012) is the conceptual nature of CRISP-DM which enables the application in any kind of data and tool (p. 5). This flexibility leads DM projects to a "less costly, more reliable, more repeatable, more manageable, and faster" deployment (Wirth & Hipp, 2000, p. 30). CRISP-DM subject is covered in-depth in next section.

2.3.2. The CRISP-DM Methodology

As observed in previous section, CRISP-DM[28] is a well-established and solid process applied to Data Mining projects. CRISP-DM is an acronym for *CRoss-Industry Standard Process for Data Mining*, and by observing the name, it is possible to perceive this process was developed inside the enterprise environment. Wirth and Hipp (2000), however, suggest that "the process model is independent of both the industry sector and the technology used." The first version of this well-known process was conceived "in late 1996 by three 'veterans' of the young and immature data mining market" (p. 29). These three veterans, at the time, belonged respectively to DaimlerChrysler (then Daimler-Benz and now Daimler AG)[29], SPSS (then IBM)[30], and NCR[31] (Chapman et al., 2000, p. 3). The creation of CRISP-DM was important as Wirth and Hipp (2000) explain:

---

[28] "The CRISP-DM process model is being developed by a consortium of leading data mining users and suppliers: DaimlerChrysler AG, SPSS, NCR, and OHRA. The project was partly sponsored by the European Commission under the ESPRIT program" (Wirth & Hipp, 2000, p. 29).

[29] Germany - https://www.daimler.com/en/

[30] USA - https://www.ibm.com/analytics/data-science/predictive-analytics/spss-statistical-software

[31] USA and Denmark - http://www.ncr.com

> In the market, there is still to some extent the expectation that data mining is a push-button technology. However, this is not true, as most practitioners of data mining know. Data Mining is a complex process requiring various tools and different people. The success of a data mining project depends on the proper mix of good tools and skilled analysts. Furthermore, it requires a sound methodology and effective project management. A process model can help to understand and manage the interactions along this complex process (p. 30).

The last section also cleared up that CRISP-DM is a cyclic and interactive process, constituted of well-defined steps. These steps are important for a reliable, feasible, and valid Data Mining project. In addition, CRISP-DM process provided a solid starting point for the creation of Corpus Mining model which combines methods and techniques from Corpus-based Translation Studies and Text Mining.

According to Chapman et al. (2000), CRISP-DM has a hierarchical process (Figure 7) constituted at four levels of abstraction named: Phases; Generic Tasks; Specialized Tasks; Process Instances (p. 9). The first level is the most important for the whole project, since it describes each step in any Data Mining project. These steps are going to be discussed in the next paragraphs. Chapman et al. (2000) make clear each step from the first level have "several second-level generic tasks," and the second level is named "generic" because the aim of it is to be general enough with the objective to provide a possibility of application in all Data Mining projects (p. 9). Following a top-bottom approach, the third level describes how the actions from the previous level have to be carried out, for instance, "at the second level there might be a generic task called clean data. The third level describes how this task differed in different situations, such as cleaning numeric values versus cleaning categorical values" (Chapman et al., 2000, p. 9). Finally, Chapman et al. describe the fourth step which is "organized according to the tasks defined at the higher levels, but represents what actually happened in a particular engagement" not what happens in general (i.e., a detailed description and decisions of how a specific task is applied) (p. 9).

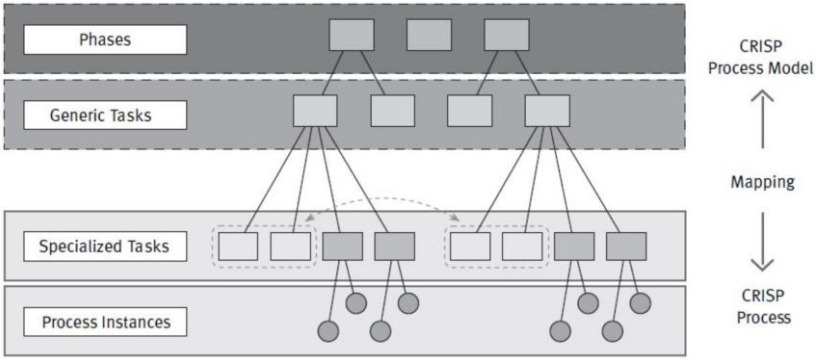


Figure 7. CRISP-DM four level hierarchy (Chapman et al., 2000).

The first level defines the six steps in a cyclic process for the development of any Data Mining project (Figure 8). In addition, the arrows indicate the most frequent path to follow and the results of one step leads to the other. According to Wirth and Hipp (2000), the outer circle "symbolizes the cyclic nature of data mining itself" and the process is not "finished once a solution is deployed" (p. 32).

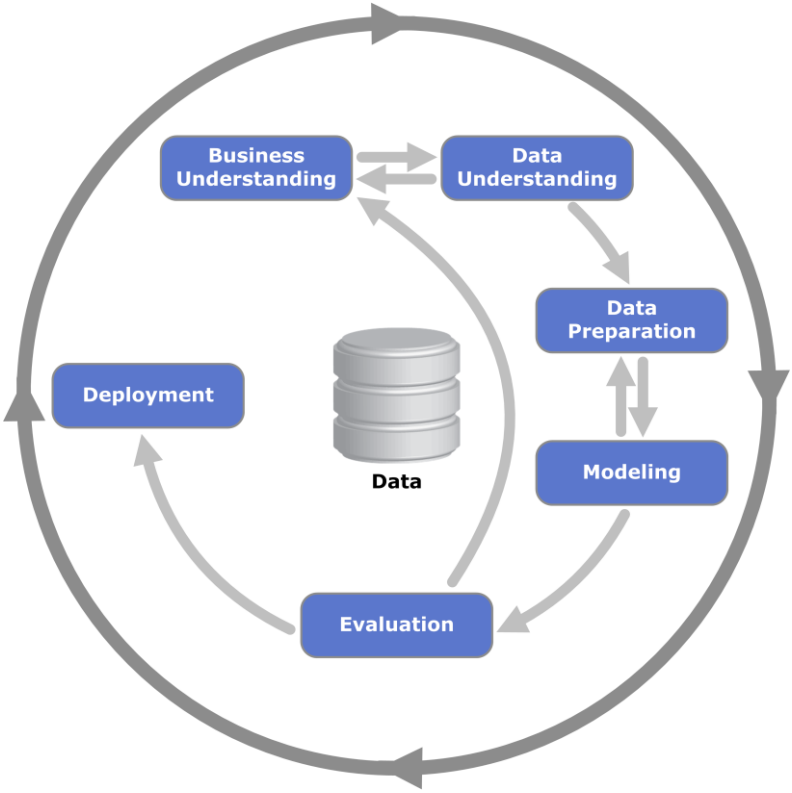


Figure 8. Phases of the CRISP-DM process model (Chapman et al., 2000).

The first phase is the **Business Understanding**, as the name suggests, this step deal with the comprehensive understanding of the business the objectives and requirements and "then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives" (Chapman et al., 2000, p. 13). From a Corpus Mining perspective, a similar approach is adopted, since it is necessary to understand all the intricacies, necessary for the correct creation of a parallel corpus.

Then, following the process in clockwise direction, the next one is **Data Understanding,** which deals with a understanding of the collected data in the initial step as well as "to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information" (Chapman et al., 2000, p. 14). From a Corpus Mining view,

there is a phase aimed to understand the collected information and organized variables such as the chosen domain, text genre, directionality, etc.

The third step **Data Preparation** is one of the longest step since it deals with all the data preparation for the construction of the final dataset for Chapman et al. (2000), the "data preparation tasks are likely to be performed multiple times and not in any prescribed order" (p. 14). This third step indicate the importance of preparing the data before inserting in a database. Corpus Mining has a similar phase, but even more complex, since is dealing with text in more than one natural language as discussed further.

The **Modeling** step defines, selects, and applies the modeling techniques based on the form of the data (Chapman et al., 2000, p. 14). Wirth and Hipp (2000), give a clear example by comparing the Data Preparation and Modeling, for them, "one realizes data problems while modeling or one gets ideas for constructing new data" or models (p. 34). For Wirth and Hipp (ibid.), the Modeling phase consists of the following tasks: (a) Selecting modeling techniques; (b) Generate test design; (c) Build model; (d) Asses model.

Soon after, the next step is the **Evaluation** as the name suggests, at this point the constructed model and obtained results are evaluated before the final deployment. Another task at this phase is to check and review all the actions applied to construct the models to make sure "it properly achieves the business objectives" (Chapman, 2000, p. 14).

Finally, the last step, **Deployment** is the real utilization of the entire generated knowledge. At this phase, the obtained information have to be organized for utilization by all the interested people in charge to take business decisions. Here, the visualization mechanisms to read and understand the acquired knowledge are developed, this process "can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise" (Chapman et al., 2000, p. 14). The tasks involved in the Deployment step are listed by Wirth and Hipp (2000): (a) Plan deployment; (b) Plan monitoring and maintenance; (c) Produce final report (final report and final presentation); (d) Review project (experience and documentation) (p. 34).

This detailed discussion about CRISP-DM was important to understand how data are transformed in order to create knowledge. One point to keep in mind is that the whole process deals with structured data such as quantitative information. However, this process leads to a well successful example of the application of mining techniques and

contributes as a solid starting point for the proposed Corpus Mining model.

2.3.3. Theoretical Foundations of Text Mining

The interplay between data and knowledge, have set the ground for a quite number of scientific work, especially the ones related to technology directly or indirectly. The term "data" is a generic concept because it can refer to a set of digital content (i.e., music, video, text, etc.) in computers but here, refers specifically to texts in electronic format. Even being in electronic format computers cannot, for example, understand text meanings, its content, linguistic intricacies or build a cohesive and coherent relationship. Similarly simple or "low level formatting units" such as unrecognizable characters, uppercase, lowercase and the decision of what constitutes a word or sentence boundary are "difficult to process automatically" (Manning & Schütze, 1999, p. 123). For a hypothetical text if we manage to ask a computer, for example, the subject, plot, characters and their relationships, dates, place, mood, how many times something or someone was mentioned in the text, we know computers will not do that without specific complex algorithms.

Bearing this idea in mind, it is possible to conclude at a basic level that texts stored in computer files are as useful as texts in printed format. For this reason, researchers started to develop techniques capable of extracting knowledge (i.e., useful information) from text. For "useful information" or "potentially useful" is related to the outcomes or "samples to be found for an application generate a benefit for the user" (Hotho, Nürnberger & Paaß, 2005, p. 2). In addition, Frawley, Piatetsky-Shapiro and Matheus (1992) in their seminal paper expose a growing problem at the time and still a challenge today, "a growing gap between data generation and data understanding" (p. 57), The authors urge for the development of "advanced techniques for intelligent data analysis," not relying just in simple statistical algorithms (Frawley, Piatetsky-Shapiro & Matheus, 1992, p. 57). At the time the mentioned research gave indication of the "mining" techniques we have today.

It is a common fact among those who know computer basics the advantage computers and software offers to handle digital text. One of the basic operations in any text processor, such as Microsoft Word® is the ability to search for any piece of text the user wants to find also known as pattern matching which "means matching predefined sequences of text with user text" (Patel, Soni & Vallabhbhai, 2012, p. 244). Text with a few pages or hundreds and thousands of pages are completely uncovered with

a mouse click. This simple search and find mechanism illustrates us the possibilities and the power computers offer us.

From now on, this section is going to discuss a selected group of techniques available to handle and manipulate text in digital format. These techniques are used to support computers "understand" text since complex or simple linguistic structures require knowledge that goes beyond the text itself. The use of understand between quotes is intentional due to the fact "understand" here is different if compared to human cognition. This perception is also recognized by Patel, Soni and Vallabhbhai (2012): "however, although our language capabilities allow us to understand unstructured data, we lack the computer's ability to process text in large volumes or at high speeds" (p. 243). Computers understand human language is one of the most complex and challenging tasks computers and processing algorithms can do and different approaches to achieve such endeavor have been made since modern computing era.

From now on, this section presents Text Mining as one of the mentioned approaches used to harvest knowledge from electronic texts. Knowledge is a polymorphic term but Frawley, Piatetsky-Shapiro and Matheus (1992) give a suitable explanation by clarifying that "a pattern that is interesting (according to a user-imposed interest measure) and certain enough (again according to the user's criteria) is called knowledge" (p. 58). Tan (1999) see Text Mining as the "next wave of knowledge discovery" and highlights the point Text Mining "has a very high commercial values" and "a commercial potential higher than that of data mining" (p. 65). With regard to Text Mining, the idea behind it is to "mine nuggets" valuable to user, this analogy emerges from the name "Text Mining" itself. This analogy is not just peculiar but serves as an illustration of what is being done with texts as well and the possible outcomes. Continuing the analogy with the real world, the desired and positive outcomes like in mining is not always guaranteed.

Here, for "mining" I can highlight the processing and analysis computer algorithms and for "nuggets" I would like to highlight the several possible patterns or running standard linguistic occurrences found in text (Hearst, 1999, p. 3). In this sense, Text Mining is a set of techniques applied to process and extract information from text and find a relevant pattern "which is novel and unknown earlier" for those who are interested in the text being investigated (Patel, Soni & Vallabhbhai, 2012, p. 243). A similar definition is shared by Miner et al. (2012) who emphasized that Text Mining is a "broad umbrella term" covering a "range of technologies" to process text (p. 30). Patel, Soni and

Vallabhbhai also mention Text Mining as Knowledge Discovery in Text (KDT) which deals with text analysis supported by computers (p. 243). Text Mining is also named as Text Data Mining and Knowledge Discovery from Textual Databases (Tan, 1999, p. 65).

The techniques used to extract information from texts is not always clear and for this reason Kao and Poteet (2005) calls for a "scientific assessment of what linguistic concepts and NLP techniques are beneficial for what text mining applications" (p. 2). This interface among concepts and techniques is intrinsic in Text Mining and due to this heterogeneous nature various areas are using this technology.

The wide application of Text Mining in many disciplinary fields reveals its interdisciplinary facet: Text Mining is a rather new "interdisciplinary field in the intersection of the related areas information retrieval, machine learning, statistics, computational linguistics and especially data mining" (Hotho, Nürnberger & Paaß, 2005, p. 1). Tan (1999) adds to the list of fields, among the cited above, text analysis, information extraction, clustering, categorization, visualization, database technology (p. 65). For Kao and Poteet (2005), "Text Mining is more recent, and uses techniques primarily developed in the fields of information retrieval, statistics, and machine learning" (p. 1). These authors advocate the purpose of Text Mining "is not to understand all or even a large part of what a given speaker/writer has said" but the main objective is to extract "patterns across a large number of documents" (Kao & Poteet, 2005, p. 1).

In relation to patterns identification, Text Mining does not come down to extract these patterns based solely on search and find mechanisms. Witten (2004) observes that "text mining appears to embrace the whole of automatic natural language processing and, arguably, far more besides" (p. 2). For Tan (1999), Text Mining does not extract ordinary patterns but rather "interesting and non-trivial patterns or knowledge" meaningful for the investigation (p. 65). As observed by Tan (ibid.) and also confirmed by Witten (2004), Text Mining "attempts to glean meaningful information from natural language text" by using techniques from NLP and far more such as Linguistics. The use of the verb "attempts" is an important point to highlight because when dealing with text in natural language, researchers have to deal with probabilities and approximations in a given situation (p. 1).

In case the guarantee of expected results is not possible, researchers have to refine their methods in order to adapt for the text under investigation. For this reason, Text Mining involves more than finding patterns. For Witten (2004), Text Mining is being "loosely characterized

as the process of analyzing text to extract information that is useful for particular purposes" (p. 1).

Additionally, texts are complex to handle digitally because they are unstructured (as opposed to structured information in databases), amorphous and complex to deal with algorithms (Witten, 2004, p. 1). Tan (1999) adds "text data that are inherently unstructured and fuzzy" and he emphasizes the role of language "text mining involves a significant language component" (pp. 65 & 69). For instance, one specific algorithm planned and developed to process and analyze text from financial domain in American English will not work properly for texts from the same domain in Brazilian Portuguese or any other language. Manning and Schütze (1999) give an interesting example comparing English and Finish, they say in "English, a regular verb has only 4 distinct forms, and irregular verbs have at most 8 forms" to prepare a list for training an algorithm the researcher can list all word forms while in Finish "verb has more than 10,000 forms" which in a way makes it difficult for the application of the same approach like in English (pp. 82-83).

Dictionaries or word lists and training models have to be compiled for each language under analysis, for example. The extraction of information or maybe the "mining of information" is not an isolated process. The analogy of "mining" can be brought again here: in order to find a valuable nugget a process of finding, separating, evaluating and refining takes place. The analogy is not so different from Text Mining, because as well observed by Witten (2004) "just as data mining can be loosely described as looking for patterns in data, text mining is about looking for patterns in text" (p. 2).

Finding occurrences of patterns in text is a basic issue when discussing Text Mining, of course, it is not possible to deny the relevance of this particular technique. For Miner et al. (2012) the point in common or the main object of Text Mining is "turn text into numbers" from distinct nature, counting, statistical, likelihood and other, but the point for these authors is to transform text (unstructured data) into structured format able to perform and apply analytical algorithms to a large collection of text (p. 30). When extracting or finding information from text the outcomes are not hidden far from human eyes or hidden in a secrete place. With regard to this subject, it is important to emphasize due to the fact some authors might give the impression the information is "hidden" for example in KDD, KDT (or text mining) Hotho, Nürnberger and Paaß (2005) say the objective is to "finding hidden patterns and connections" in texts as an example (p. 2). Otherwise, Witten (2004) observes this situation precisely:

> With text mining, however, the information to be extracted is clearly and explicitly stated in the text. It's not hidden at all-most authors go to great pains to make sure that they express themselves clearly and unambiguously-and, from a human point of view, the only sense in which it is "previously unknown" is that human resource restrictions make it infeasible for people to read the text themselves (p. 2).

As revealed by Witten the information is not hidden, it is already stated in the text, but sometimes not so explicitly in the sense that complex algorithms have the power to create correlations and relationships among several texts and offers answers not previously thought about by researchers. Theoretically, humans can do that but it would take time, workforce, intuition and money. To illustrate this scenario, a classic example is given by Kennedy (1998) who cited the SEU Corpus (Survey of English Usage Corpus) the time lapse to collect transcriptions and perform processing "took over 25 years and was not completed until 1989" (p. 19).

Text Mining is a collection of techniques applied to generate knowledge from textual data. It is unclear what techniques should be part of Text Mining or not and Kao and Poteet (2005) call the attention to this fact informing "what the field needs now is a sober scientific assessment of what linguistic concepts and NLP techniques are beneficial for what text mining applications" (p. 2). At this study the focus is on translation and parallel corpus so that the techniques not just from Text Mining but Corpus-based Translation Studies and Corpus Linguistics are organized in a standard model of Corpus Mining. This model enables the researcher to avoid the use of corpus "merely to prove the obvious or give confirming quantification where none is really needed, in short, to engage in the type of exercise that after much expense of time and money ascertains what common sense know anyway" (Tymoczko, 1998, p. 658). Among the possible techniques available in Text Mining, Patel, Soni and Vallabhbhai (2012) list the most used ones involved in this knowledge generation such as:

- Text processing;
- Tokenization;
- Stop word removal;
- Stemming;

- Text mining methods;
- Interpretation and evaluation (p. 243).

The listed techniques above are dealt in the next chapters of this study. Of course, it is necessary to bear in mind that the list above may vary depending on the investigation being carried out or the author views, but it is a satisfactory, concise starting point. For instance, Tan (1999) divides Text Mining procedures in two components namely Text Refining applied to transform unstructured text "into an intermediate form" (i.e., semi-structured) and the second one Tan refers to Knowledge Distillation that perform analysis to "deduces patterns or knowledge from the intermediate form" (p. 65). (Most of these items are going to be discussed in methodology section as well as expanded and inserted in the proposed model of Corpus Mining.)

The cited techniques are important for Text Mining since it "needs special preprocessing methods to convert textual data into a format which is suitable for data mining algorithms" (Hotho, Nürnberger, Paaß, 2005, p. 2). Here the suitable format is the structured one. In addition the techniques are able to perform many tasks, for example, find and extract structured information already available in the text. Witten (2004) provides a reasonable list of structured information present in text and able to be extracted such as:

> An important form of text mining takes the form of a search for structured data inside documents. Ordinary documents are full of structured information: phone numbers, fax numbers, street addresses, email addresses, email signatures, abstracts, tables of contents, lists of references, tables, figures, captions, meeting announcements, Web addresses, and more. In addition, there are countless domain-specific structures, such as ISBN numbers, stock symbols, chemical structures, and mathematical equations (p. 10).

As also revealed by the cited structured data, secondary information can be extracted by creating relationship among various attributes identified in the text or by count how many times these attributes are mentioned in the text. Thus, producing useful information for the parts interested in the analysis. The use of these techniques from other areas working specially with structured data poses some barriers when working with text for this reason "an adaptation of the known data mining algorithms to text data is usually necessary" (Hotho, Nürnberger, Paaß, 2005, p. 5). The authors

continue discussing that the application and adaptation of text mining techniques is to structure text (Hotho, Nürnberger, Paaß, 2005, p. 10). When the information is structured as much as possible facilitates the storing in database, indexing and perform statistical analysis.

Text mining techniques have shed light on natural language understanding by processing large quantity of texts in a short time and it made possible for researchers investigate language phenomena not thought before and predict possible generalizations. The interdisciplinary field of Text Mining proved to be practical not just to implement its own techniques but to investigate text from various domains and genres. The upcoming challenges and shortcomings from Text Mining is undeniable since it is a wide field defined with differences among the authors, but still fruitful. The opportunities available to translation go beyond this study but it is expected this work contributes to the discussion. The use of Text Mining in CTS undoubtedly corroborates the invitation made by Tymoczko (1998) "to begin to envision the widest possible range of corpora" or in other words to find different ways corpora can be used for both conceptual and methodological aspects (p. 658).

2.3.4. Duo Mining

While the literature on Data and Text Mining is quiet abundant, only a few studies are dedicated to the investigation of Duo Mining. Nevertheless the topic is worth mentioning to show different views and the field versatility for providing possibilities to apply the techniques with different approaches and/or domains. As I have discussed in the last section, Text Mining already encompasses techniques from Data Mining, but its object of analysis is text and not structured data as in Data Mining. Moreover, the idea behind Text Mining as some authors argue, is to convert text to structured information and store it, for example, in databases, then apply Data Mining algorithms. For this reason, however, Text Mining is a multidisciplinary field "it can be difficult even for text mining experts to concisely characterize it" (Miner et al., 2012, p. 30).

For Miner et al. (2012) the field is considered a "'Wild West' of analytics" due to the various technologies competing with each other, not to mention the fact that the stages of maturity among the techniques also contribute to the impossibility of a precise definition (p. 30). The point in question is the name definition and the coverage of Text Mining can be different or adapted and modified to fit for a researcher needs or a disciplinary field. To show these intricacies, this section discusses Duo Mining which is characterized as the "combination of data and text mining" (Veni, Praveena & GanaPriya, 2013, p. 124). The problem found

is the wide availability of data stored in both formats structured and unstructured as well as the need to cross-check the information from different sources based on this idea just a combination of the two methods can do that (Krishnaiah, Sekhar, Rao & Prasad, 2012, p. 423). For this reason, researchers have noted the combination of Text Mining and Data Mining "is worth more than the sum of its parts" because this combination can include more than structured data and as consequence provide better insights and significant patterns (Chary, Reddy & Bhuahan, 2018, p. 1337).

While discussing the new developments in the field Veni, Praveena, and GanaPriya (2013) suggested Text Mining as being part of a large process "that go beyond simple searching methods" (p. 127). This large process was defined by the authors as Duo Mining (ibid.). A similar point of view is shared by Freeda (2015) who explains that Text Mining is concerned just for the management of unstructured data and once prepared or structured in a database most of the knowledge extracted can be done through Data Mining (p. 50). Krishnaiah et al. (2012) recognize the similarity between Text and Data Mining due to the fact both "mine" data in a broad sense, but what both technology analyze is different (p. 424).

As a result, Freeda (2015) claims that researchers "have discovered that [Duo Mining] is a combination that is worth more than the sum of its parts" (p. 52). In a nutshell, this idea seems plausible because this study indicates another possible combination, Text Mining and Corpus-based Translation Studies named Corpus Mining.

2.3.5. Natural Language Processing

Natural Language Processing (NLP) has been in the academic and enterprise community for decades and is present in two key disciplinary fields Linguistics and Computer Science (Kao and Poteet, 2005, p. 1; Miner, 2012, p. 37). Precisely speaking, Witten (2004) reveals NLP came into existence "in the late 1940s and early 1950s," as discussed in Computational Linguistics section (p. 4). It is possible to perceive the field started almost at the same time of modern computing. As observed, researchers at the time where trying to find the holy grail of automatic translation and the roots of NLP go deeper in this subject (Witten, 2004, p. 2). In an attempt to create automatic translation without taking into account the complexity of a language, researchers at the time relied in simple techniques such the direct translation of word for word. The fiasco of such approach may have lead NLP for some discredit, so to speak, for a period of time. However, even the research done at the time contributed

for the field in many ways and after researchers left the automatic translation aside to focus in other kind of language processing.

Most part of the techniques, theory and methods are roughly the same and prove to be useful with excellent outcomes to process language with distinct objectives out of automatic translation. This historical synopsis may lead Witten (2004) to recognize NLP "dominated in its infancy by unrealistic ambitions and swinging in childhood to the other extreme of unrealistically artificial worlds and trivial amounts of text, has matured" (p. 3). Furthermore, Witten (2004) also made clear that even for algorithms and technology with the capacity to process natural language at an "illiterate child's level" the technical development, computing power and programming language proved to be "astonishingly sophisticated medium that does not succumb to simplistic techniques" (p. 2).

To date, NLP is a well-established and solidified ground for language research if compared to Text Mining or even Data Mining. Due to the fact the aforementioned subjects tries to create knowledge and in a certain extent, understand language or create meaning based on real world data, these subjects are overlapped in key features. For this reason, some scholars tend to say NLP and Computational Linguistics are different terms for the same research field just because both "process human language in terms of its meaning" (Sattikar & Kulkarni, 2012, p. 6). However this assumption may raise questions since the most accepted idea is that Computational Linguistics is a compounding part of NLP or vice-versa, since NLP techniques can be used in Computational Linguistics as well.

To be precise, NLP is an area from Computer Science involved in the "language study focused on software development, applications and specific computational systems" (Othero, 2006, p. 343). In addition, NLP techniques are largely used in Text Mining especially in the first phases of it, to manipulate and process text at a linguistic level. This sort of confusion is natural since NLP is so rich that depending on the study being carried out the researcher have to decide which parts from NLP to use.

To understand language, one of the main objectives of NLP is to find and develop computational models to process, analyze and understand distinct linguistic phenomena. The linguistic categories to develop such computational models are listed by Bender (2013) in her book, where the scholar provided a non-exhaustive list of subjects interesting to investigate hand-in-hand with NLP (Table 1).

Table 1. *A non-exhaustive sample of linguistics branches (Bender, 2013, p. 1).*

Subfield	Description
Phonetics	The study of the sounds of human language
Phonology	The study of sound systems in human languages
Morphology	The study of the formation and internal structure of words
Syntax	The study of the formation and internal structure of sentences
Semantics	The study of the meaning of sentences
Pragmatics	The study of the way sentences with their semantic meanings are used for particular communicative goals

The subfields are notably extensive and the author emphasizes her focus in the book is morphology and syntax (also known as morphosyntax) which have more relevancy in NLP, but she adds similar books could or should be written addressing the other categories (Bender, 2013, p. 2). These categories are important for NLP in many ways, for example, supporting the test structure this is due so the position of words in a sentence adds a non-linear structure by encoding "information about the relationships between words," strictly speaking, create meaning (Bender, 2013, p. 5). This is a key point to have in mind since Text Mining techniques avoid the use of such approach, word order and sentence structure, focusing more in bag-of-words[32], stemming or n-grams. For this reason, NLP "developed various techniques that are typically linguistically inspired, i.e., text is typically syntactically parsed using information from a formal grammar and a lexicon, the resulting information is the interpreted semantically" and used to extract patterns and useful information (Kao & Poteet, 2005, p. 1).

To process language in various degrees of analyses NLP is divided in two groups namely **shallow** and **deep**. For Kao and Poteet (2005) **deep** is related to "parsing every part of every sentence and attempting to account semantically for every part" whereas **shallow**

[32] In simple terms, a special kind of frequency list. The bag-of-words is a model where "all the structure and linear ordering of words within the context is ignored […] a bag is like a set, but allows repeated elements" (Manning & Schütze, 1999, p. 237).

involves “parsing only certain passages or phrases within sentences or producing only limited semantic analysis” and it can use statistical approach for word sense disambiguation (WSD) (p. 1).

The division is necessary because each one consumes different computer power processing, time and a degree of complexity. Moreover, the separation is more suitable because each part can be applied specifically for project needs. It is important to note Text Mining apply widely shallow techniques more related to the ones similar in information retrieval and may avoid a deeper and cognitive techniques (Witten, 2004, p. 2). One of the reasons behind the use of shallow techniques in Text Mining is uncovered by Kinyon (2001), she argues deep techniques are “costly to develop, costly to run and often yield errors, because of lack of robustness of wide—coverage grammars and problems of attachment” (p. 330). Generally speaking, shallow techniques do not consider text structure such as word order but rather it uses a more simplistic approaches but efficient such as the use of n-grams, conversion of all words to its basic root in stemming process and the concept of bag-of-words. The application of such model is not error free at all, for example, while processing sentences like “Dear, dear, this is not a happy cup...” or “I suppose Bane'll be happy”[33] are more likely to indicate something positive than negative. The words “dear” appear two times while happy just once in the first sentence. For this reason, shallow techniques use training sets and probabilistic techniques to determine with more precision the results.

NLP provides the basis for Text Mining and according to Miner (2012) it provides “useful input useful input variables for text mining such as part of speech tags and phrase boundaries” (p. 37).

2.3.6. Bitext Alignment – The Dataset for Corpus Mining

Bitext and textual alignment plays an important role in Corpus-based Translation Studies (CTS) and Natural Language Processing (NLP). A common core shared by these two areas is undoubtedly the parallel corpus[34], and its main compounding part: the aligned texts, also

[33] Examples extracted from COPA-TRAD.

[34] Defined by Tiedemann (2011) as a “collections of bitexts” (p. 2). For Baker (1995) “a parallel corpus consists of original, source language-texts in language A and their translated versions in language B” (p. 230).

known as bitext or even parallel text[35]. In a broad view, bitext refer to texts in one language along with its corresponding translations in one or more other languages—used firstly in Translation Studies, and then "in a larger community with many other applications in mind" (Tiedemann, 2011, p. 1). To be more precise, Tiedemann (2011) defines a bitext according to the following formalism: $B = (B_{src}, B_{trg})$ where B_{src} is the source language or original text, B_{trg} is the target language or translated text, and B is the correspondence or the connection of each element, in one direction or the other (p. 7). An important point to emphasize, according to Tiedemann (2011), is that bitext is not strictly involved with originals and translations, but it can refer as well, to a "wider range of parallel resources," for example, several translations for only one original text or even translations in the same language but from different historic periods, can be considered bitext as well (p. 1). To illustrate, Figure 9 shows a possible scenario of this "wider range of parallel resources."

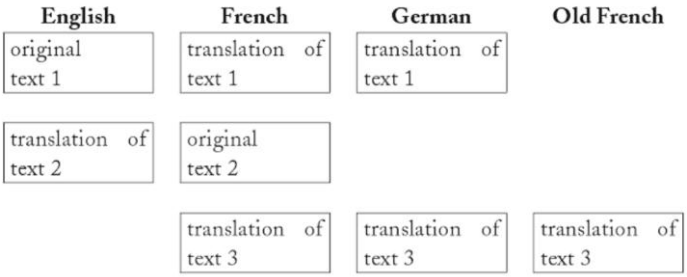


Figure 9. Possible setup of several bitexts from the same parallel corpus (Tiedemann, 2011, p. 1).

The approach for bitext analysis varies according to the methodology adopted and the field the research is inserted. In the area of CTS, especially in qualitative research, there is a tendency to analyze bitext extracted from a parallel corpus manually. On the other hand, NLP uses bitext mostly to train translational algorithms to perform processing operations such as automatic translation, word sense disambiguation

[35] According to Tiedemann (2011) "parallel text is often used synonymously with the term bitext. Unfortunately, this creates a confusion with the research on translation theory and terminology" (p. 1). For this reason this study adopts the term "bitext."

(WSD), terminology extraction, machine learning, etc. Independent of being manual or automatic, it would be a problematic challenge define the boundary between CTS and NLP though, an interdisciplinary view grouping these areas would be fruitful (Olohan, 2004, pp. 8-9).

The alignment tools and techniques provided by NLP is an important resource to the advancement of research in CTS. The aforementioned resource, facilitates and accelerates the acquisition of bitext, necessary to build parallel corpora. There are different techniques in textual alignment technology. However, the definition of **text alignment** (Figure 10) is quite uniform. In a nutshell "the sentence alignment task is to identify correspondences between sentences in one language and sentences in the other language" (Gale & Church, 1993, p. 3). For Tiedemann (2011), the property of correspondence between the two texts coupled together is an important characteristic and for him **alignment** "is the task of making this correspondence explicit, which makes it a central task in processing bitexts" (p. 1). This automatic process is challenging because the relation 1:1 is not always maintained. For example, Gale & Church (1993) found the relation 1:1 between sentences in the linguistic pair English-French, on the other hand they made clear the possibility of one sentence in English correspond to two sentences 1:2 or more in French language 1:n (p. 177). Other alignment type is also possible, such as the 0:1, for example, a sentence added in translated version with no correspondence in the original text (Koehn, 2010, p. 57). In addition, further challenging cases such as substitution, deletion, insertion, contraction, expansion and merger "impose a set of slope constraints" to deal algorithmically (Church & Gale, 1991, p. 43).

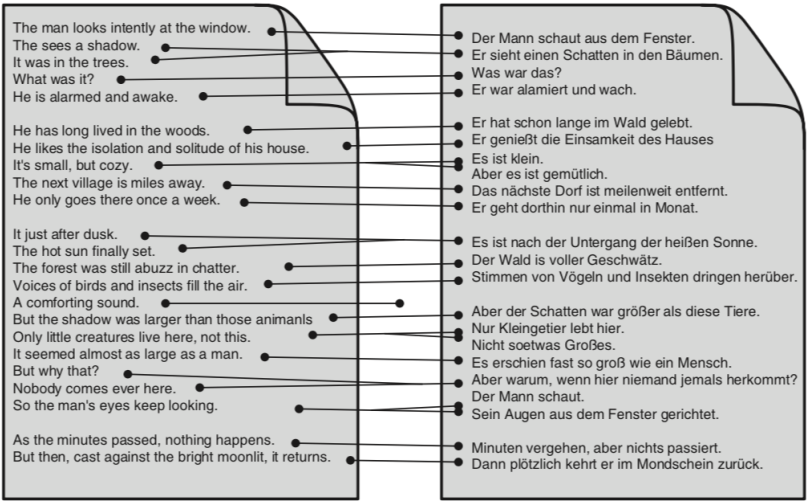


Figure 10. Sentence alignment, the process of finding occurrences of sentences that are translations of each other across texts in different languages (Koehn, 2010, p. 56).

In relation to alignment types, Santos (2011) divides the textual alignment in two different levels: **sentence** and **word** (p. 117). Although in Santos definition for text alignment there is no direct mention to "sentence" but "blocks or tokens" (ibid.). Token is a string of characters and commonly referred to as words. Block is an open term and refers to sentence, paragraph, chapters or entire sections (p. 127). The level of granularity for textual alignment assigned for each research defines what is considered a "block." In relation to segmentation or defining the boundaries of each block Tiedemann (2011) states that

> [t]he type of segmentation of a text into basic text elements determines the type of alignment that can be established. In the examples given so far, we have already seen different kinds of segmentations, for example, a division into sentences (sentence boundary detection) and a division into word tokens (tokenization). Many other types are possible, for instance, paragraphs, N-grams, syntactic constituents, morphemes or characters. Depending on this segmentation we will talk about document alignment (base elements=documents), sentence alignment (base elements=sentences),

> word alignment (base elements=words) and so forth (p. 9).

Tiedemann (2004) the conceiver of OPUS[36] parallel corpus, adopts a more restrictive point of view on the alignment methodology by inserting it specifically in the area of NLP (p. 212). Nevertheless, as observed in the first paragraph, an interdisciplinary approach considering the contribution from CTS would be fruitful.

Caseli and Nunes (2004), propose two different alignment types or techniques applied in the creation of bitext. The first one, is the **sentence level alignment**, in which sentences from the original and translated text are paired side by side, for example, "What are the Black Riders?" → "O que são os Cavaleiros Negros?". The alignment tool based on its internal algorithm decides the exact match between the sentence in source language and the sentence in target language. There are different algorithms to determine this correspondence, for example, measuring sentence lengths. Translational phenomena such as omissions or additions in the translated text are the kind of problem that has to be handled automatically by the computational tool while aligning a text (Santos, 2011, p. 127; Gale & Church, 1993, pp. 180-181; Chen, 1993, p. 10). To illustrate, Tiedemann (2011), states that "it is sometimes convenient to add empty elements in order to allow *empty alignments* corresponding to deletions/insertions" (p. 7).

The second type proposed by Caseli and Nunes (2004) is the **lexical alignment**, which is done through words, for example, in "magic" → "mágica" (pp. 1-2). The problem of word inversion, omission or addition in the translated text constitutes a challenge for researchers. The algorithm has to decide which word in the original corresponds for the correct or most approximate one in the translation (Santos, 2011, p. 127; Kay & Röscheisen, 1993, p. 121). In order to achieve an alignment at word level (i.e., low level alignment), Santos (2011) suggests "higher level alignment is performed first, which allows to obtain improved results" (p. 118). However, the lexical alignment deserves a special attention, because it is employed in the construction of bilingual dictionaries (word for word) to provide linguistic information for automatic sentence aligners (Tiedemann, 2012, p. 2215).

Last paragraph presented details related to lexical and sentential alignment. It is also important to mention here the method used by

[36] http://opus.nlpl.eu/

different alignment algorithms because they may vary in their technical approach. As a result, this variance reflects on the final product (i.e., the bitext and its alignment quality). The alignment method can be divided in statistic or heuristic: "statistic approaches estimate alignment probabilities whereas heuristic approaches use associative measures derived either from corpora, or external sources such as dictionaries" (Foo, 2007, p. 1).

The statistic method relies on the basic idea of sentence size which "longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences" (Gale & Church, 1993, p. 179). Gale and Church (ibid.) explain that a probabilistic score is assigned "to each of proposed pairs" then a special algorithm uses this score to find the maximum likelihood alignment of sentences. In other study, Church and Gale (1991) explain that they found "the length of a text is highly correlated with the length of its translation," in other words, "a longer sentence will be translated into a longer sentence (or perhaps two sentences whose total length is not very different from the original sentence), whereas a shorter sentence will be translated into a shorter sentence" (p. 41). This method is based "on the idea that when a sentence corresponds to another, the words in them must also correspond" (Santos, 2011, p. 119). The application of this method presented a success rate of 80% and Gale and Church recognize that this technique works better in large corpora, to improve the rate the authors proposed a lexical approach not covered in the study (pp. 182-184). Gale and Church algorithm counted the number of characters to calculate the score.

Another study proposed by Brown, Lai and Mercer (1991) used a similar technique, but their algorithm counted the number of words (tokens), here the tool used to identify and count tokens has to be robust with a low error rate or the sentence alignment would be a disaster (p. 169). The authors (ibid.) advocate the use of anchor points or linguistic clues (such as cognate words) available in the source and target text to improve the accuracy of the alignment. The use of anchor points here relies on a statistical method since these variables are used on a probabilistic algorithm.

The second alignment method, the heuristic, takes advantage of extra sources such as bilingual dictionaries "to help establish the correspondences" (Santos, 2011, p. 124). For example, Chen (1993) proposes an algorithm that generates on the fly a word-to-word dictionary to be used as lexical clues to improve the overall quality alignment ratio (p. 10). This algorithm is capable of handling large omissions in the

translated text since it can identify where the omission starts and where it ends based on the lexical information provided by the dictionary (ibid.). The use of idiomatic expressions, collocations or fixed expressions can be used in the alignment process since in most of the cases "these kind of expressions cannot be translated word by word" so they can work as anchor points (Santos, 2011, p. 125).

Chen (1993) proposed his alignment algorithm in 1993 and computing power at the time was scarce and expensive thus he made clear that "substantially greater computing power is required before these approaches become practical" (p. 15). However, from the past ten years computer power is not a problem as it was in the past. Consequently new complex and improved methods have been proposed such as the hybrid one. Varga et al. (2005) proposed a hybrid sentence-level aligner named Hunalign that uses heuristic and statistic methods. They designed a hybrid algorithm that uses the dictionary and length-based technique which "successfully amalgamates the two" methods (p. 592). The authors created a 50 million-word Hungarian-English parallel by using Hunalign tool (ibid. p. 590). What makes Hunalign suitable for different text types is the possibility of using custom dictionaries but if no dictionary is provided Hunalign creates one on the fly (ibid. p. 596). Performance it is a striking feature in Hunalign since it was developed in C++ language instead of scripting languages that do not have a compiled version of the code such as Python or slower compiled codes written in Perl (Santos, 2011, p. 126; Varga et al., 2005, p. 593).

The creation of bitext through automatic aligners algorithms are key components for the creation of parallel corpora. As a result, researchers have to deal with the task of text alignment, because they are convinced that bitext "produce a number of lexical resource that could be of great value to their research" (Church & Gale, 1991, p. 40). However, the issue of technicality is one point elicited by Gerdes (2010), for researchers that decide to align their parallel corpora manually, the level of technicality—required for using textual alignment tools constitutes a barrier. Nevertheless, new studies on the area are being conducted and friendly alignment tools are being developed, especially the ones on the Internet such as *InterText*[37], and the one discussed here AUTO ALIGNER, a translational tools available in COPA-TRAD platform, which could facilitate and approximate the low level or too complex

[37] http://wanthalf.saga.cz/intertext

aligners algorithms to researchers, who tend to align their parallel corpora manually. Section 3.4 discusses AUTO ALIGNER from a technical point of view.

2.4. Final Remarks

This Chapter discussed the theoretical background supporting this research. Discuss an interdisciplinary theory is challenging, but the attention was directly on the topics addressed in this study. Next chapter, the procedures of Corpus Mining model are discussed with real examples showing how the technical layer was conducted while designing and developing COPA-TRAD Version 2.

3. CHAPTER THREE: Procedures

3.1 Initial Remarks

Bitext is a special kind of data and demands appropriate translational methods to process and analyze it. In fact, translation research involves a set of knowledge from different areas. As observed by Mossop (1994), people who are not involved or study translation may think "translation is basically a linguistic exercise and that communication through translation is a straightforward process," which is not the case (p. 405). For instance, extratextual information and key variables (such as context of production and reception, and translator behavior) specific to the area of Translation Studies have direct impact upon the final analysis of a translated text. When investigating translated texts by means of parallel corpora, the researcher needs technological skills in order to carry out his investigation. As well-observed by Danielsson (2004), the researcher "do not need to be a Jack-of-all-trades to become a corpus linguist," but after years in the field the researcher may notice he/she had to learn a bit of everything (p. 225). For the purpose of investigating bitext in this particular context, Corpus Mining seems an adequate and useful model.

Based on the aforementioned assumptions, this Chapter 3 discusses the methodological concepts and practicalities comprising each phase of Corpus Mining model, and the systematization of it. The design, implementation, and an in-depth discussion of this concise model as well as its potential application are delineated to create COPA-TRAD platform. To do so, practical techniques and technological tools are used to demonstrate the application of Corpus Mining in the design and development of COPA-TRAD Version 2 (Section 1.2.1). However, the initial sections, are going to discuss the key terms and concepts utilized alongside the study, and the technical elements applied to test the validity and reliability of Corpus Mining model. Then, the application of this model in COPA-TRAD tools is discussed. Although the utilization of a Corpus Mining model in other project may be possible, this subject is not going to be discussed here. Finally, the interface of Text Mining techniques and Corpus-based Translation Studies (CTS) shows how it can help analysts to become more consistent in terms of their results since this interface deals with a special kind of data that is adopted in this particular study (i.e., bitext).

3.2 Experimental Setting

Practical experiments were conducted in two different environments, before and after deploying COPA-TRAD Version 2: one in the development server and the other the production server[38], which in reality is a cluster of virtual machines powered by Vmware 6.5[39]. The development server is a PC with an Intel Core 2 Duo 3.0Ghz CPU, 500 Gb internal hard drive and 4 Gb of RAM. The Operating System is OpenSUSE Leap[40] 42.1 running Apache 2, PHP 5.4 and then 7.2 and MariaDB as an Open Source substitute for MySQL. The production environment is much more technologically advanced and it presents the following setup:

- 1 proxy (NGINX 1.4.6) load balancing to forward requests (3Gb RAM);
- 4 web nodes running Apache 2.4.7 (4Gb RAM each);
- 1 NFS node for file and configuration sharing between Apache nodes (8Gb RAM);
- 1 node for database MySQL 5.6.30 (10gb RAM);

The virtualization host (i.e., the physical machine running the virtual ones) has the following specification:

- PowerEdge R710 with Intel(R) Xeon(R) CPU X5690 @ 3.47GHz - 24 cores and 140 Gb RAM;
- The NFS, where Sphinx search engine is running, has 8 cores and 8 GB RAM;

The operating system for the environment is Linux Ubuntu 14.04-x86_64 and all the connection to COPA-TRAD platform, is secured with a valid HTTPS encrypted and authenticated using TLS 1.2 (a strong protocol), ECDHE_RSA with P-256 (a strong key exchange), and AES_256_GCM (a strong cipher) – Figure 11.

---

[38] maintained by SeTIC (*Superintendência de Governança Eletrônica e Tecnologia da Informação e Comunicação*) at UFSC – http://setic.ufsc.br/.
[39] https://www.vmware.com/
[40] https://en.opensuse.org/Portal:Leap

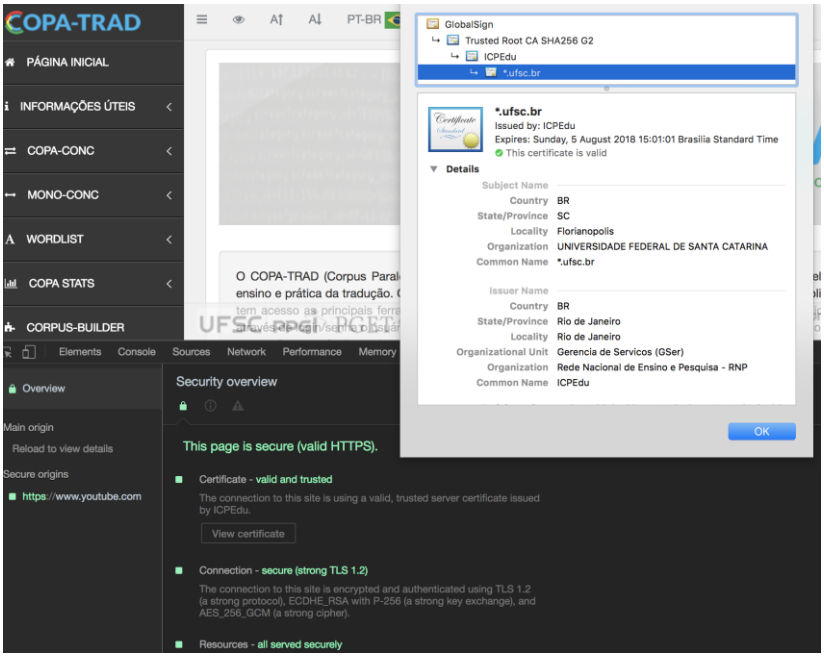


Figure 11. COPA-TRAD HTTPS certificate information.

Most part of the processing mechanisms and Text Mining algorithms were implemented solely in PHP programming language, which constituted an extra degree of complexity, since researchers in the area tend to utilize Python and R programming language. As a result, most of the algorithms available and implemented, for the academic community, are in these programming languages. However, the code developed for COPA-TRAD Version 2 is in PHP which showed stability and, in most part of the time, fast throughput in production environment. The development server, could run the code, but depending on the size of the text being processed, and analyzed, the average time took between 20 and 40 minutes, sometimes even more. In some cases, the process was repeated over and over, to adjust the algorithms and adapt to PHP programming language which constituted an exhausting process.

The dataset for testing the programing code implementation was the same from COPA-TRAD database, as well as other Children's Fantasy literature such as The Lord of the Rings (not part of COPA-TRAD at the moment). This last one, was utilized mainly for testing Text Mining techniques and automatic alignment algorithm. Due to hardware

limitations in development server, as listed above, this research phase consumed more time than previous planned.

3.3 The Systematization of Corpus Mining Model

In this section, the Corpus Mining model is introduced in order to systematize the interface between Text Mining and Corpus-based Translation Studies (CTS). Additionally, Corpus Mining is organized in a concise step-by-step guide to the investigation of translation through the use of a corpus-based methodology. In a similar way, CRISP-DM also involves a step-by-step guideline to conducting data mining. This is an interactive process, since in each key phase, it is possible to evaluate the results and then review for corrections and adjustments of the algorithms, then repeating this process once again.

There are similarities between Corpus Mining and CRISP-DM, but differences are perceivable, as well; the phases reflect the process of a corpus-based translational investigation. Furthermore, the use of this model is applicable for both qualitative and quantitative research. In fact, Corpus Mining was created to fulfill the combination of these two approaches (see Section 2.2.3), but the use of one or the other individually may be possible (despite not being covered in this study). The nature of Corpus Mining is empirical because it was created based on the experience acquired while developing COPA-TRAD and researching[41] potential technology that could be applied directly to translated text investigation. This empirical nature is a key component, since all the proposed phases were tested in a practical situation: COPA-TRAD Version 2.

During the design and development phases of COPA-TRAD Version 2, both the tasks and the techniques utilized in the construction were organized in phases ranging from planning to final utilization. In light of this organization and the results, a set of recommendations were added to each phase, along with practical examples. Implications and applications of Corpus Mining are addressed specifically for COPA-TRAD Version 2 development, but the generic nature of this model may

---

[41] Study conducted at the *Curso de pós-graduação Lato Sensu em Engenharia e Projetos de Software at Universidade do Sul de Santa Catarina (UNISUL),* entitled *Utilização de técnicas de mineração de textos em corpora paralelo para auxílio na pesquisa acadêmica em estudos da tradução: Um estudo de caso.* Available at: https://tracor.ufsc.br/uploads/monografia-carlos.pdf

also indicate the possibility of its application being used in other Translation Studies projects.

Building on the techniques and theory of Text Mining and Corpus-based Translation Studies, the process of Corpus Mining follows an approach similar to the previously mentioned areas, when utilizing texts in natural language (e.g., selecting texts, normalizing, extracting information, indexing). The similarities exist at a superficial level, but there are differences, as well; for this reason, each phase of Corpus Mining will be discussed in detail. According to the research questions presented in the introductory chapter, the proposed model offers a set of phases that can be utilized in its totality or separately (e.g., a researcher might choose to conduct only the linguistic-processing phase).

The practical experiments are explained as each phase of Corpus Mining is discussed; then these experiments are employed in the design and development of COPA-TRAD Version 2. The ongoing evolution of Corpus Mining and the developmental cycle for COPA-TRAD are both important to keep in mind, as well as the fact that this model has been elaborated on and planned for specific use in Corpus-based Translation Studies. After this, all the steps are presented and described to show how Text Mining techniques can be utilized in combination with Corpus-based Translation Studies. Figure 12 displays the phases of Corpus Mining workflow in a sequential order.

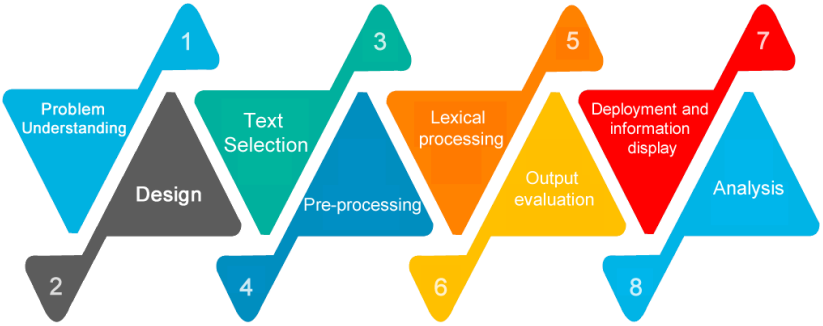


Figure 12. Corpus Mining phases.

Each phase of the workflow is discussed in separate sections comprised of its core concept, internal tasks, and practical example from COPA-TRAD Version 2.

3.3.1 Problem Understanding

The first phase of Corpus Mining is the most important one, since it deals specifically with requirements, information-gathering, and an understanding of the first idea (which is a problem usually related to a particular translational phenomena). The explicit definition of the translational phenomena is important, since it will "facilitate the task of determining the database structure," as well as define the linguistic processing techniques (Brown, Tiberius, Chumakina, Corbett, & Krasovitsky, 2009, p. 148). Additional tasks are then required in order to gather the necessary information to prepare a concise project. The tasks consist of defining: (a) the problem; (b) necessities; (c) requirements; (d) domain/genre; (e) directionality; (f) objectives; (g) hypothesis; (h) sampling.

The meaning or the explanation of the tasks can be implied by the name of each, but two deserve special attention. The first task is "directionality," which defines the direction of the linguistic pair being adopted—for example, Brazilian-Portuguese source-texts and English translations, English source-texts and Spanish translations, etc. The second task has to do with "sampling," which provides the first glimpse of the texts and may indicate possible information that these texts can provide. In this particular task, sampling can be prepared manually, and size is not an issue at this point. In addition, there is no need to prepare the complete list of the texts at this point, because the refinements and specific decisions are dealt with in the next phase. To prepare the sampling, the researcher can either extract information from an existing corpus or select some possible texts according to what has been defined in the tasks already mentioned in this phase.

To create this initial preview, the utilization of any translational software is suitable—such as WordSmith Tools[42], AntConc[43], CasualConc[44], etc. For the purpose of the first samples in this study, I selected CasualConc (Figure 13); even in the monolingual version, interesting information can be extracted. I also tried the parallel version of CasualConc, named CasualMultiPConc[45] but it did not run on MacOS High Sierra 10.13.4.

---

[42] http://www.lexically.net/wordsmith/
[43] http://www.laurenceanthony.net/software.html
[44] https://sites.google.com/site/casualconc/
[45] https://sites.google.com/site/casualconc/utility-programs/casualmultipconc

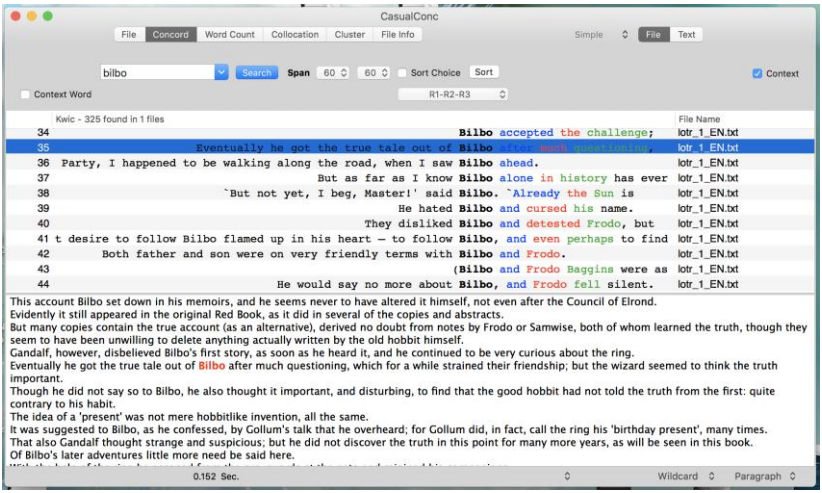


Figure 13. Utilization of CasualConc to prepare the sampling.

Just as an example, while preparing the sampling, the researcher can identify in Figure 13 some special characters that he/she must be aware of while processing or cleaning the text. The utilization of an acute accent, rather than quotation marks, for instance, can interfere with token extraction (see Section 3.3.5). These issues may have a direct negative impact while processing a text, since the algorithm responsible for cleaning this text might need adjustments, adding new characters that do not initially appear on this algorithm cleaning list.

For instance, in the COPA-TRAD AUTO ALIGNER tool options for special characters, *"‘","’","‘","’","…","—","–"* are replaced with *"'","'","'","'","...","-","-"*; after this process, the regular expression */[^(\x20-\x7F)\x0A]*/* is executed to remove all characters not in the hexadecimal range 0x20 to 0x7E[46].

3.3.2 Design

The design is the second phase of the Corpus Mining model. The decisions taken at this level, have direct impact upon all subsequent phases. A similar point of view while creating a corpus is well-observed by Sinclair (1991):

[46] For detailed list: http://www.asciitable.com/

> The beginning of any corpus study is the creation of the corpus itself. The decisions that are taken about what is to be in the corpus, and how the selection is to be organized, control almost everything that happens subsequently. The results are only as good as the corpus (p. 13).

The design phase has to do with the planning as well as the decisions related to corpus building. The tasks in this phase are the following: (a) corpus size; (b) selection criteria; (c) copyright permissions; (d) texts; (e) type of corpus. Most of the tasks from this phase need necessary comments to clarify them. The first one is the delimitation of corpus size which can be small, medium, large and open-ended. The size definition comes first on the list just because of the impact upon the next tasks. Then the selection criteria is utilized to define the scope of text selection. The selection is a long list and the most important items are listed in Table 2. Fernandes' (2009) "A Portal into the Unknown: Designing, Building and Processing a Parallel Corpus" presents a comprehensive list of extralinguistic information that can be used in corpus compilation.

Table 2. *A non-exhaustive list of selection criteria based on Fernandes (2009, pp. 25-26).*

List of extralinguistic information			
Dates	Location	Languages	Languages Varieties
Texts	Status	Translators	Authors
Gender	Nationality	Representativeness	Size
Genre	Special Feature	Translation Process	Mother Tongue
Prizes	Word-count	Copyright	Mode

When dealing with literary texts, the copyright issue is a sensitive point to deal since the country laws have to be checked. The issue of copyright is a factor that may impact in the next task which is the one named simply as "text" which can be full texts or extracts. The copyright have impact in this task since due to some policies just part of the text can be used (i.e., not integrally.) Table 3 lists the most common types of corpus for the last task in this second phase. This study main focus in on parallel corpus since it is the most common applied for CTS research, but other combination are possible as well. For example, the utilization of a monitor corpus in combination with parallel corpus to investigate if the language

employed in translated texts are the same or have the same patterns in originals texts from the same language.

Table 3. *List of most common type of corpus.*

Type of corpus			
Parallel	Comparable	Monolingual	Bilingual
Multilingual	Static	Dynamic	Monitor
Balanced	Imbalanced	Learner	Ad-hoc
General Purpose	Web-based		

In the case of COPA-TRAD, some texts were scanned and others downloaded from the Internet. In relation to copyright issues, a decision taken for COPA-TRAD is to never provide the whole text as a unique file, but chunks of it are presented in a ranked form established by the search engine (i.e., complete texts, but only in chunks.) However all the processing and analysis are conduct with full-texts as well as the frequency lists, etc. The size of COPA-TRAD corpus is defined as open-ended since the database is constantly growing with its users contributions.

3.3.3 Text Selection

The third phase is related to necessary actions to obtain the texts in electronic format. Some of the selected texts maybe available online for downloading, the other just in printed version, so scanning is necessary. To do so, the first task is capturing, constituted of the following actions: Scanning, Downloading, Typing, Rekeying and Speech-to-text. The most frequent actions here, are Scanning and Downloading. Rekeying is a process most associated when OCR[47] fails to recognize text characters properly.

The second task, named technicalities has to do with the following actions: OCR, text extraction, filename extension and charset. The OCR actions is related to scanning but is not exclusive to it since PDF files with scanned texts like images need OCR recognition as well. Text extraction has to do mainly convert the text in TXT file format, since scanned files are commonly in RTF format and other ones in PDF, etc. A command utilized for conversion in Linux environment (here OpenSUSE

[47] Optical Character Recognition.

Leap) is the *pdftotext*[48] which converted texts with considerable quality. Filename extension for texts, which is an unstructured kind of data, the most suitable format is TXT in Section 1.3 several file formats were cited. In case a text is obtained in any format (such as the ones listed in Section 1.3) a conversion to TXT is required. The final action at this phase is the role of charset. The UTF-8 is the standard version. In case a text is in other kind of charset such as ISO 8859-1, Windows-1252 a conversion to UTF-8 is necessary as well.

Some texts collected for COPA-TRAD in PDF format were simply converted to TXT with the following commands in shell:

```
$ cd folder_with_all_pdf_files
$ pdftotext ./*
```

The conversion to UTF-8 is also possible to do it through software like Notepad++ or similar ones. In the next section, I discuss how the selected texts were pre-processed in order to be further lexically processed.

3.3.4 Pre-processing

The pre-processing phase deals with a set of technical procedures and the aim is to sanitize the selected texts. The objective in the pre-possessing phase is to obtain the text free from any other extra content. Here, it is worth noting the observations of Sinclair (1991) in relation to the text format suitable for corpus research:

> Once a text has been selected for study, the first decision is in what way it is to be re-created inside the computer. It may seem a simple enough process, to reproduce a text inside the machine, but in practice not all the features of a text are coded. Different, features are picked out according to the needs of the work. For most general processing the text is kept to a very simple format—usually a single long string of letters, spaces, and punctuations marks. The letters of the alphabet, punctuations marks, and the word spaces are called *characters*; the distinction between upper and lower case is preserved. Page and line numbers are kept only for reference purposes, and other layout

[48] https://linux.die.net/man/1/pdftotext

> setting, and type-face information is [sic] discarded (p. 28).

The text format described by Sinclair is the same as advocated here, with a minor difference: the pages and line numbers are not preserved since there is no need to preserve references in such way (in case of COPA-TRAD specifically for the reason it uses a different approach to keep references). This phase is necessary since texts even in TXT and UTF-8 without BOM[49] can present the following problems:

- Different kinds of break-lines such as CR (Carriage Return) or LF (Line Feed);
- Texts badly digitalized;
- Microsoft Word characters;
- Extra white-spaces;
- HTML or XML tags;
- Figures and other extra-textual information;

For this reason, this phase is constituted with three actions: sanitization, tools and structural normalization. Sanitization deals with the available techniques applied to refine text and remove the most common noise, in fact, reduce at the maximum. The noise can vary according to selected text, but it can be, for example, unrecognized characters due to scanning and OCR failure or even Microsoft Word special characters as well as other symbols.

Tools are extra algorithms to conduct this pre-processing phase and structural normalization is the action necessary for text alignment: While conducting the experiments with automatic alignment, this study found that texts (originals and translation) structurally similar are aligned with higher quality if compared to the ones which are not. The action for keeping texts structurally similar consists of removing all extra blank spaces in the source and target text (this process is executed automatically in COPA-TRAD.) The removal of extratextual content, such as images, table of content, prefaces, translator notes and other paratextual information are removed at all, for COPA-TRAD these procedures are done manually. Except for the last actions, the other ones listed here are dealt by the available tools in COPA-TRAD platform.

[49] BOM refers to byte-order mark.

3.3.5 Lexical Processing

Lexical processing is the phase where most of Text Mining and Natural Language Processing techniques are conducted to find "nuggets" (i.e., useful information). In this sense, producing content to store in a database for later analysis and the real creation of knowledge. The actions in this phase are: tokenization; noise removal; tagging/annotation; lemmatization; stemming; aligning; statistics; indexing. These actions have fundamental importance, since, "data entered into a database with little thought or attention to its categorisation are at best useless, and in the worst case harmful if used to make spurious generalizations" (Brown et al., 2009, p. 117). The objective of tokenization is related to the creation of tokens, more specifically in segmenting a text into *word-forms* not merely *words* as advocated by Sinclair (1991):

> Note that a word-form is close to, but not identical to, the usual idea of a word. In particular, several different word-forms may all be regarded as instances of the same word. So *drive*, *drives*, *driving*, *drove*, *driven*, and perhaps *driver*, *drivers*, *drivers', driver's, drive's,* make up ten different word-forms, all related to the word drive. It is usual in defining a word-form to ignore the distinction between upper and lower case, so *SHAPE*, *Shape* and *shape* will all be taken as instances of the same word-form. This convention no doubt blurs a few hundred useful distinctions, like *polish* and *Polish*, but obliterates many thousands of word-forms which have an initial capital letter merely because they begin a sentences (p. 28).

The process of tokenization seems a trivial aspect, but defining boundaries of a word-form is a complex procedure (see discussion in Tóth, Farkas, & Kocsor, 2008, pp. 465-466) and demands decisions that may include or exclude possible tokens[50] (Figure 14).

---

[50] See the case of hyphen and apostrophe which need special attention (Sinclair, 1991, p. 28). However, not discussed in detail by Sinclair (ibid.), hyphen and apostrophe are parts of design decisions: whether to keep them or not. For example, in COPA-TRAD, these elements are kept and are quickly selected from the corpus with simple query such as "*SELECT * FROM copa_word WHERE*

Figure 14. Words boundary, example from Hansards of the 36th Parliament of Canada extracted from COPA-TEJ (COPA-TRAD corpus of law texts).

After the process of tokenization a set of undesired characters may appear on the final list and the removal of this content is necessary too. Noise removal can be compared with the action of sanitizing as mentioned in the previous phase, but here with a more specific connotation. Defining what to remove, at this moment, may vary according to the selected texts, for this reason the most suitable approach is to open the list and find for undesired patterns. For example, roman numerals may appear in the word list and the user may not want them so he/she can prepare a specific code to remove it. Then the process of tagging/annotation to identify part-of-speech (POS)—such as nouns, verbs, adverbs, prepositions, etc.

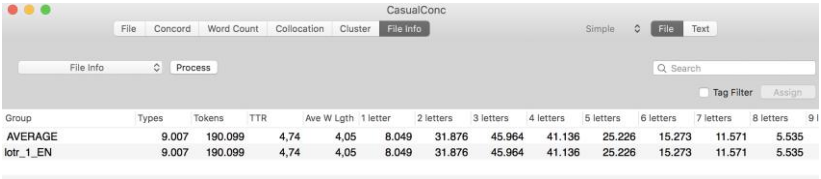
The common aspect in tagging as the name suggests is adding tags across the text, but for me a different approach seems more suitable: rather than tagging a text I prefer to create lists of POS and store them in a database. With this approach I can keep the texts clean and raw. This decision, however, was taken in the Design phase and may vary in different studies. Lemmatization is a special kind of word normalization since it checks beforehand part-of-speech and them the normalization process (when necessary) and then categorization. Stemming, is a less complex process with the objective to reduce a word to its root also known as stem.

The aligning task (Section 2.3.6) is the process of creating bitext the key component for any parallel corpus, this process can be conducted manually, automatically, semi-automatically and automatically with human supervision. Statistics is the application of statistical algorithms in order to create probabilities, estimations or simply counting like word frequency, as an example, Figure 15 presents simple statistics generated automatically by CasualConc software. The application of customized statistical processing mechanisms are dealt at this part as well. The

---

wrd_word LIKE "%\'s';" which produces results such as *artemis's, family's, anyone's, man's there's, he's, tutankhamen's*, etc.

statistical task was added in this part because different types of counting can be done while tokenizing a text, stemming, etc. As a result a huge amount of time and computer processing power can be saved, desired features specially for medium and large parallel corpus.



Group	Types	Tokens	TTR	Ave W Lgth	1 letter	2 letters	3 letters	4 letters	5 letters	6 letters	7 letters	8 letters	9 l
AVERAGE	9.007	190.099	4,74	4,05	8.049	31.876	45.964	41.136	25.226	15.273	11.571	5.535	
lotr_1_EN	9.007	190.099	4,74	4,05	8.049	31.876	45.964	41.136	25.226	15.273	11.571	5.535	

Figure 15. CasualConc stats for Lord of the Rings – The Fellowship of the Ring.

The process of indexing, as the name suggests, is related to creating indexes of the generated content as well as the whole texts in the corpus. Most of common software databases have special mechanisms to manage it, but other options are still available such as the utilization of Sphinx Search[51] (as discussed in Silva, 2013) or the development of a custom indexer software.

In COPA-TRAD the tokenization process is performed by an internal tool named COPA-TOKENIZER, "this multilingual (i.e., it supports not just English but any other Latin-based language) tool processes the word extraction, statistics as well as the storing in the database" (Silva, 2013, p. 82). Noise removal is also conducted to remove special typos. In addition, for COPA-TRAD version 2, the tagging/annotation and lemmatization are conducted automatically by TreeTagger an especial software that "assigns to each word in a sentence the part of speech that it assumes in the sentence" (Charniak, 1997, p. 34). Alignment is performed by Hunalign and the indexing by Sphinx search engine tool named simply as "indexer"[52]. Figure 16 displays the final result after this phase in COPA-TRAD, the table showed is the "copa_word" with an unique id, the extracted word, the MD5 hash of the same word (this is a specific feature taken for COPA-TRAD in order to keep a standardized version of the word to perform comparisons, etc.), the part-of-speech identification (this one vary according to the dictionary utilized in TreeTagger), lemma (provided by TreeTagger as well),

[51] http://sphinxsearch.com/
[52] http://sphinxsearch.com/docs/current/ref-indexer.html

language identification (in this case English), language variation (in this case English – Ireland) and finally word frequency. It is necessary to emphasize that the columns/attributes for this table were organized to satisfy COPA-TRAD necessities established according to Corpus Mining previous phases.

wrd_id	wrd_word	wrd_hash	wrd_tag	wrd_lemma	lg_id	lgv_id	wrd_frequency
1	artemis	01dac4a9afa6561565d10b5b16e9dcdb	NCMP	<unknown>	1	50	2614
2	fowl	863e5f582e99730838f47c53e92767eb	NCMP	<unknown>	1	50	679
3	prologue	b6be887ccda6478673baa12a1edfb385	NP	<unknown>	1	50	3
4	how	db88a0257c220dbfdd2e40f6152d6a8d	WRB	how	1	50	1227
5	does	5440e70c43cc02aba90d879c888e6e09	VVZ	do	1	50	697
6	one	f97c5d29941bfb1b2fdab0874906ab82	CD	one	1	50	2765
7	describe	0d16430e9f0aa8879ba9dac2540543fc	VVP	describe	1	50	20
8	various	16875aa2b5eed3e388dcceaa36f56214	JJ	various	1	50	107
9	psychiatrists	45b6a81e510da4f31f5b0fecd7b77261	NNS	psychiatrist	1	50	1
10	have	b42dad5453b2128a32f6612b13ea5d9f	VH	have	1	50	6192
11	tried	c6b526b2fe3d4d1318bf9d7c41744854	VVN	try	1	50	233
12	and	be5d5d37542d75f93a87094459f76678	CC	and	1	50	24577
13	failed	26934eb377001f66e37289a5c93fe284	VVD	fail	1	50	63
14	the	8fc42c6ddf9966db3b09e84365034357	DT	the	1	50	68966
15	main	fad58de7366495db4650cfefac2fcd61	JJ	main	1	50	88
16	problem	0861a099e9593791de261ebb86e75eac	NN	problem	1	50	285
17	is	a2a551a6458a8de22446cc76d639a9e9	VBZ	be	1	50	10181
18	artemis's	69a0cba86914fcf014a54c28680ba64e	NULL	NULL	1	50	116
19	own	b515e18aa3fbe7d264d7ca5a95ef73e1	JJ	own	1	50	641
20	intelligence	90b884933d23c4d57ca01dbe2898d405	NN	intelligence	1	50	20
21	he	6f96cfdfe5ccc627cadf24b41725caa4	PP	he	1	50	8967
22	bamboozles	4606dde83adcc16179c498ebda4b7794	VVZ	bamboozle	1	50	1
23	every	83ab982dd08483187289a75163dc50fe	DT	every	1	50	684
24	test	098f6bcd4621d373cade4e832627b4f6	NN	test	1	50	59
25	thrown	357498e03a88a17f471aba561d9fe677	VVN	throw	1	50	38
26	at	7d0db380a5b95a8ba1da0bca241abda1	IN	at	1	50	4774
27	him	664d242a7528bf4230386c9ac1a437f8	PP	him	1	50	2643
28	has	3309a7a7941818e131b4dfb9a6349914	VHZ	have	1	50	2563
29	puzzled	6b8cf6b2baedcde8275905077d4d96bd	VVD	puzzle	1	50	16
30	greatest	1130a355aa0b0e797b53081162097ad1	JJS	great	1	50	69
31	medical	7cbdd4e997c3b8e759f8d579bb30f6f1	JJ	medical	1	50	66
32	minds	b13628525c873fb808e56fb489742ea0	NNS	mind	1	50	35
33	sent	789183b7e98646c11d5f0544c8f3c4c9	VVD	send	1	50	165

Figure 16. COPA-TRAD database table "copa_word" extract.

3.3.6 Output Evaluation

In order to identify possible errors and maximize accuracy, reliability and validity is necessary an evaluation and perform a detailed analysis of the obtained results from the previous phase. The proposed actions for this phase are: search nodes; sorting; quantitative procedures; qualitative procedures. The search nodes are specific terms to locate in the database through a query language or any other searching mechanisms (translational features inquired by the researcher in first phase). For instance, the nodes can be, lexical items, syntactic patterning, frequent recurring features, infrequent unusual features, etc. While searching for nodes, several sorting options can be applied, such as alphabetically, sort

by frequency, most common, least common, or any other selection criteria as defined in design phase. These sorted lists are more useful than simple reference as suggested by Sinclair (1991):

> The main use of alphabetical lists is for reference, but they are occasionally useful as objects of study. They are often helpful in formulating hypotheses to be tested, and checking assumptions that have been made (p. 31).

Next, the quantitative procures has to do with evaluations related to numerical results such as investigate frequency lists and the distribution of feature across texts. Finally, qualitative procedures, this one, aligned with descriptive analysis and focused to study features such as co-text, context, genre, translator profile, author profile, translational features such as omissions, additions, etc.

In COPA-TRAD the search for nodes as well as sorting, is implemented on top of Sphinx search engine, which provides advanced searching mechanisms as well as sorting. Sphinx returns only the references/ids of the found items but the textual information have to be retrieved from the database yet—which is done through SQL. The quantitative procedures is a complex task performed by COPA-TRAD platform, where several counting and statistical algorithms are applied to generate suitable results. The visualization of these procedures are discussed in next chapter, but here the inner processing mechanisms are going to be dealt. The quantitative information in COPA-TRAD, up to the moment, are the following ones: (a) frequency lists—general and specific; (b) type, token and ratio (TTR)—general and specific; (c) Zipfian distribution—only for books in COPA-TRAD Version 2.0; (d) Standard Deviation; (e) standardized type/token ratio (STTR); type token ration percentage; Additional quantitative techniques are planned for COPA-TRAD Version 3.0 but is not going to be dealt here.

Frequency lists, as the name suggests, are simple calculations, for every extracted element from text, the system increments plus one. This process is carried out through MySQL query language, in case an existing element is already stored (i.e., *on duplicate entry*) an increment to is done. The ratio (based on type and token) in COPA-TRAD, is available for each language and individual texts. The process is performed based on the total of each element stored in database, since no duplicate elements are stored, the quantity of types is easily calculated. Then with the number of types

and tokens the ratio is calculated according to the formula in Figure 17. In addition, the TTR percentage is also available.

$$\text{Type-Token Ratio} = \left(\frac{\text{number of types}}{\text{number of tokens}} \right) * 100$$

Figure 17. Type Token Ration formula.

Zipf [53] distribution is a special kind of data, and its results are utilized to plot a graph for each COPA-TRAD book. The graphs are generated once, and then the final graph is stored as an image to minimize power and processing consumption. The general context of Zipf's law are considered by Manning and Schütze (1999):

> In his book Human Behavior and the Principle Effort, Zipf argues that he has found a unifying principle, the Principle of Least Effort, which underlies essentially the entire human condition (the book even includes some questionable remarks on human sexuality!). The Principle of Least Effort argues that people will act so as to minimize their probable average rate of work (i.e., not only to minimize the work that they would have to do immediately, but taking due consideration of future work that might result from doing work poorly in the short term). The evidence for this theory is certain empirical laws that Zipf uncovered, and his presentation of these laws begins where his own research began, in uncovering certain statistical distributions in language (p. 23).

Manning and Schütze (1999) explain that the total number of types occurs according to its position in a ranked frequency list (in descending order)—Zipf's law demonstrate that "both the speaker and the hearer are trying to minimize their effort" (pp. 23-25). In addition to what have been mentioned, the Zipf's law graphs generated in COPA-TRAD can be a resource for comparing vocabulary density between texts: is possible to compare between translators and authors, translators and translators, authors and authors, how they are using language—this kind of graph can

[53] Linguistics professor and eponym of Zipf's law (Zipf, 1932).

serve as an evidence supporting much larger investigations. Although, the behavior of language use is an intrinsic aspect of human mind (according to Zipf,) in literally texts, for stylistic reasons, the writer may decide to use words not so common as the ones in day-by-day language. Table 4 is a real example from COPA-TRAD database listing words according to its frequency and rank as indicated by Zipf's law.

Table 4. *Zipf's law on COPA-TRAD corpus.*

Word	Frequency	Rank	Word	Frequency	Rank
the	68966	1	I	11754	8
a	34402	2	It	11429	9
to	33149	3	Is	10181	10
of	31279	4	Harry	9704	11
and	24577	5	For	9186	12
in	18976	6	Was	9184	13
that	14271	7	He	8967	14

Standard deviation (i.e., SD or even the Greek letter σ "sigma") is applied to investigate dispersion of frequencies (i.e., common distance between each frequency type) from COPA-TRAD texts (Appendix H). The SD is an accurate indices of variability and it is "used to show how spread or concentrated distributions are" (Riazi, 2016). According to Riazi (2016) "the standard deviation is the extent of dispersion around the mean and is the main measure of variability that is usually reported with the mean in research reports." To put it simply, SD "tells you how tightly all the various examples are clustered around the mean in a set of data" (Niles, n.d.). On COPA-TRAD the SD is calculated by text, due to this characteristic, is possible to compared how many variation (i.e., the amount of variation) between texts and for COPA-TRAD the deviation of each text shows the difference between the tokens of each type. This complex procedure follows the steps:

- Get the list of available texts;
- Get each frequency list of types by text;
- For each book, the individual frequencies (number of word occurrences) are added in a list (e.g., [500,150,20,30,60…]);
- SD is calculated over the generated previous list, and the final result is rounded (PHP function "round") with four digits after the decimal point;

- The final result is stored in database and the algorithm keeps looping until the last book in the list.

To estimate the SD the following procedures are carried out:

- Count the total number of elements in the book frequency list;
- If the total is zero (i.e., no values) an error is generated;
- The general mean of the book is calculated: the sum of each frequency divided by the total.
- For each individual frequency, the deviation is calculated by subtracting the frequency value out of the general mean;
- Then each value is squared and summed in the final list;
- The SD is calculated by the square root of final list divided by the total number of elements;

For Scott (2001) the classic TTR is meaningless because it "varies very widely in accordance with the length of the text" (p. 126). For this reason, the researcher proposes a more accurate method: the standardized type/token ratio (STTR) where "the ratio is calculated for the first 1,000 running words, then calculated afresh for the next 1,000, and so on to the end of the text or corpus" (Scott, 2001, p. 126). The STTR is calculated on COPA-TRAD by books.

First, the algorithm reads the text, and breaks it into lines/sentences then, tokens are extracted from each line. Next, the algorithm checks if the number of sentences if exact 1000, in case there are less tokens another line/sentence is added until the completion of 1000 tokens. If more than 1000 tokens are extracted the system cut the exceeding part and saves into memory for the next interaction. The TTR is calculated for every thousand tokens and them summed with other TTRs in a total ratio variable. This process continues until the end of file. When finished, the final step is to divide the total ratio per number of interactions/number of chunks (i.e., how many 1000 cycles) and then the final score is saved in database. Texts with less than thousand tokens are calculated but the result is always zero.

3.3.7 Deployment – Information Display

After the output evaluation a final interpretation of the outcomes can indicate the most suitable tool to analyze the data. Once the processing and display tools are developed the researcher may decide to make them available. This process is also known as deployment and comes after the development and depending on the nature of the system difference exists, for example, local software or online the process to

deploy each of them are different. In addition, in cases the deploy is for an update of the system just the modified code may be changed.

To achieve this purpose the following tasks in this phase are: interpretation/assessing; processing tools; display tools; availability. Interpretation/assessing of the obtained results to check if the accuracy, if the results are expected according to the design phase. The processing tools, possibly were used to process the bitext as described in the last phases, but here is the adaptation of the processing system for the final use, for example, prepare the code to work in easy steps for unexperienced users.

The last task leads to display tools which is the user interface (UI) such as buttons, color pallet, images, text, layout, in other words, display format for a better user experience as well as provide reliable methods to analyze information. Results and other related information emerged from the interpretation/assessing can be utilized to indicate the best visual approaches to adopt and develop. For instance, in COPA-TRAD, the whole platform can be dived in two major layers: backend and frontend. The backend correspond to the processing mechanisms not directly available to the user. The frontend is the part of the platform available for the user which is possible to operate. The frontend is the layer that "talks," so to speak, to the backend. The deployment of COPA-TRAD was done in an online environment as described in Section 3.2, and consisted of the preparation of its configuration file, database setup and server setup. Then, sequential updates were/are made through a process of replacing edited files.

3.3.8 Text Analysis

Text analysis has to do with the academic research and the assets exploited in this phase can be done through the tools proposed in the last phase. At this final phase, the tasks involved are: concordances; wordlists; frequency lists; collocational information; validity; final report; feedback. Concordances[54] are generated from the results in the corpus as well as wordlists and frequency lists. Then the collocational information as the name suggests is the investigation related to collocations which is a linguistic feature largely studied in linguistics as well as Translation Studies, this last one in both languages—originals and translations.

---

[54] "A concordance is a collection of the occurrences of a word-form, each in its own textual environment", in plain English, word surrounded by nearby context. (Sinclair, 1991, p. 32).

Validity is a special task and its aim is to confront the results and analysis with the hypothesis, claim, problem and research questions. Then the final report is created presenting all the knowledge harvested from the corpus. To illustrate, COPA-TRAD provides a set of different types in wordlists, such as: hapax legomena; general list; acronym list; proper noun list; 2-grams; 3-grams; 4-grams; 5-grams. For each type of wordlist a distinct processing procedure takes place. General list is a straightforward process of extracting tokens from text. The processing mechanism for extracting hapax legomena is basically the same of previous one, but here just words with frequency of only one are mined. Then the acronym and proper noun list are the first prototypes to mine information from text without the use of tags/annotation. The first results presented some noise, but the results obtained so far are promising. For the two cited tasks, the text mining techniques utilized are the combination of specific regular expressions to identify possible nuggets. The acronym extraction do the following steps:

- Parse each alignment/sentence of a text;
- The whole alignment/sentence in uppercase is ignored;
- Long sequences of uppercase are ignored, for example, “CHAPTER 1.”
- Repeating words in sequence and uppercase are ignored, for example, “HA HA HA” or “EH EH EH”;
- Words with the cedilla “Ç” are ignored;
- If passed, Microsoft Word symbols and other typos are removed;
- Then only uppercase, uppercase with dots or uppercase with numbers are permitted, for example, “S.P.U.G.”;
- Next, all acronyms candidates are added in a provisory list;
- The provisory list is parsed and each item extracted;
- If empty the element is ignored;
- Spaces at beginning and end are removed;
- Elements with diacritic are ignored;
- Less than six characters are ignored;
- Only numbers are ignored;
- Only alphanumeric greater or equal to two (ignoring dots);
- Stop words from a predefined list are ignored;
- Saves the remaining acronyms.

The process of extracting proper nouns is much more complex than extracting acronyms. Proper nouns appear in a variety of formats and different approaches have to be employed in order to extract possible candidates. In addition, COPA-TRAD algorithm supports six languages which demanded a degree of complexity while prototyping and developing. The tool interacts with the database in order to check duplicates and update quantitative information for later analysis. Another, feature worth mentioning is the possibility of enabling stop words lists for six languages, this procedures improve the accuracy of proper noun identification. Possible candidates are identified with specific conjunctions according to the training set[55] in Table 5. The list is far from complete, in fact, completeness is not the objective. The list was generated based on the texts information, and the initial investigation outlined according to Corpus Mining model. An important point to emphasize is that, the training set can be expanded, as new texts are added to COPA-TRAD, and possible new patterns appear. Besides the utilization of a training set, stop word list other procedures are applied to find the desired "nuggets." Next chapter information extracted through this method are analyzed with some major examples.

Table 5. *Training set applied to COPA-TRAD learning algorithm for proper noun identification.*

	English	**Portuguese**	**Spanish**	**French**	**Italian**	**German**
Start	the, mr, mrs, ms, dr, mstr, miss, sir	senhor, sr, sra, senhora, dr, vex, srta, vsa, prof	señor, sr, señora, senhora, sra, señorita, srta,	madame,sr, mesdames, mademoiselle, monsieur, messieurs, veuve, docteur, docteurs,	signora, signorina, signore, signorine, dottore, signor, dottor, dottores	herr, dame, frau, herrn, fräulein, doktor, geehrter,

[55] A set of most common/standard examples applied as a kind of cross-reference in order to identify possible and desired patterns in text. In addition a training set can be a set of "manually tagged text to learn the regularities of tag sequences" (Manning & Schütze, 1999, p. 345).

			don, doña, doctor, doctora	professeur, professeurs, maître, maîtres, monseigneur	sa, egregio, onorevole, chiarissimo	sehr geehrter herr, sehr geehrte frau, lieber herr
Middle	of, the, and	de, da, do, di, dos	de la, y, de, de las, dos	du, de, le, la, les, des	di, della, dell, de, dello, dei, dall, dagli, da	der, von
Dot	mr, mrs, ms, dr	sr, sra, dr, vex, srta, prof	sr, sra, dr, vex, srta, prof, d, profa, dra	mme, mmes, mlle, mlles, m,mm, vve, dr, drs, pr, prs, me, mes, mgr	sig, dott, avv, chia, dr, egr, vve, dr, prof, sigg, proff, chiar	dr, frl, fr

A key point while extracting proper nouns, is to remove all the unnecessary information such as non-alphabetical characters. The following characters are removed:

PHP programming language.

```
1 array('/','\\','\"',"'",'“','”',"'",',',' ','<','>','?',';',':','[',']','{','}','|','=','+','-','_',')','(','*','&','^','%','$','#','@','!','~','`','.','«','»','0','1','2','3','4','5','6','7','8','9');
```

Line breaks (i.e., newline) an special character, usually hidden, are ignored as well. Since there is a possibility of these line breaks staying between sentences which could indicate two or more words rather than one. These characters are the following ones:

PHP programming language.

```
1    array(" ", "\n", "\t", "\r", "\r\n");
```

In addition punctuation are utilized to identify the beginning of a new sentence usually in uppercase which indicates a possibility of the word not being a proper noun. Finally words if less or equal to two characters are ignored, this procedure is necessary to reduce the noise. Each proper noun identified are stored in COPA-TRAD database.

The identification of n-grams is a well-known process and on COPA-TRAD the list of n-grams are generated by text/book. Details of n-grams process are not discussed here. Finally, Figure 18 presents Corpus Mining breakdown structure.

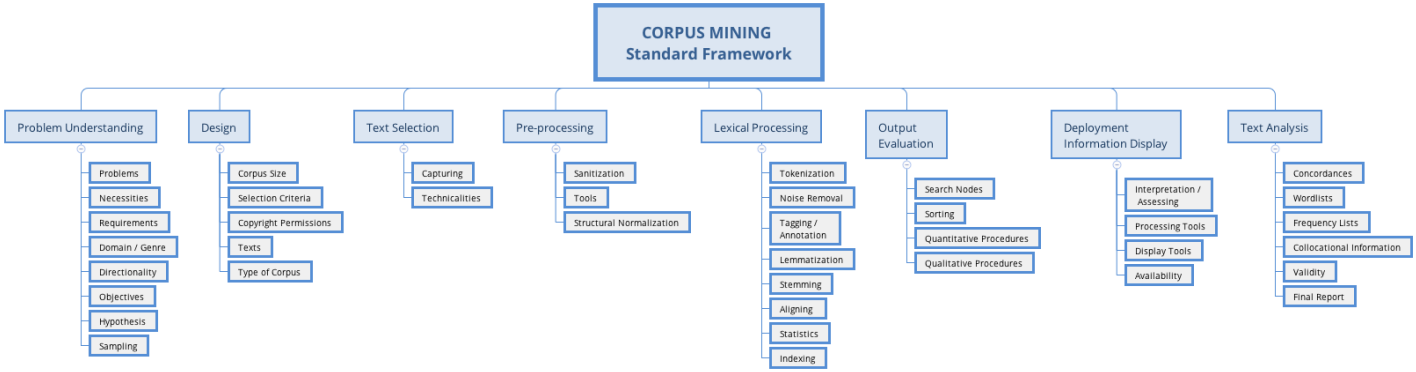


Figure 18. Proposal of Corpus Mining model (Work Breakdown Structure – WBS).

3.4 COPA-TRAD Linguistic Resources: AUTO ALIGNER

To support Corpus Mining phases in COPA-TRAD, a special tool to assist in the automatic alignment was developed: AUTO ALIGNER. As reviewed in the next Chapter, AUTO ALIGNER is a tool to perform automatic alignment between originals and translations. This technology, was incorporated in COPA-TRAD version 2, and its main objective is to facilitate the creation of bitext for COPA-TRAD parallel corpus. Not just the creation of bitext, but AUTOALIGNER performs various processing tasks to prepare the text before the alignment. For this reason, the tool can be utilized for aligning texts as well as to pre-process texts, as delineated in Corpus Mining model. In general lines, two texts (original and translation) are uploaded to AUTO ALIGNER, then the tool performs a set of processing techniques and make the final version available for download. The results is available in Microsoft Excel format (XLS), just because in first column, contains the original text and the second column, the translated text.

According to the discussion in the introductory chapter, COPA-TRAD is an online bidirectional parallel corpus developed as a result of the author M.A. research based on the concept proposed by Fernandes PhD thesis (2004). Gerdes (2010) observed that "many researchers on translation end up aligning large parts of corpora manually, lacking tools with basic heuristics to simplify this task" (p. 257). These observations were valid for COPA-TRAD first version, where manual textual alignment approach was being used for creating bitexts. For example, the first texts provided to COPA-TRAD were aligned manually by Fernandes (2004) during his PhD. The results of this manual process produced a shortcoming because it was slowing down the availability of fresh bitexts in the corpus as well as limiting the overall size of it. Three experiments were conducted to find out the aligner tool that would fit better for COPA-TRAD, and based on the found results, it was chosen one automatic aligner that uses a hybrid algorithm which allows the use of custom dictionaries especially in English-Portuguese pair (or vice-versa).

The hybrid sentence alignment tool chosen to create bitext for COPA-TRAD is Hunalign (Varga et al., 2005). This tool was chosen because it process literary texts with great accuracy. The score that measures the alignment quality in Hunalign varies in a range between 0 to 1. The 0 score means a text poorly aligned whereas the score near one means a bitext well aligned with minor errors to be corrected manually (if necessary). The chosen text to test Hunalign was the source English text "*The Lord the Rings: The Fellowship the Ring*" and its Brazilian Portuguese translation "*O Senhor dos Anéis - A Sociedade do Anel*." As

discussed in Corpus Mining section, the texts were organized and pre-processed with the aim to create similar TXT files in both languages. Since the files were in PDF format, with figures, prefaces, translator initial words, summaries, and appendixes, the application of Corpus Mining initial steps were necessary. Finally the TXT files achieved a structurally similar appearance (Figure 19).

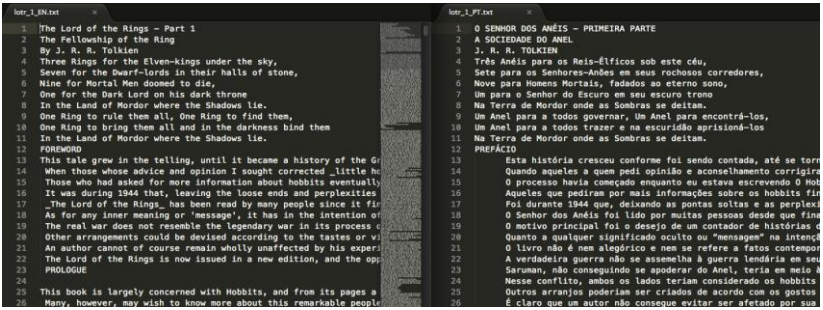
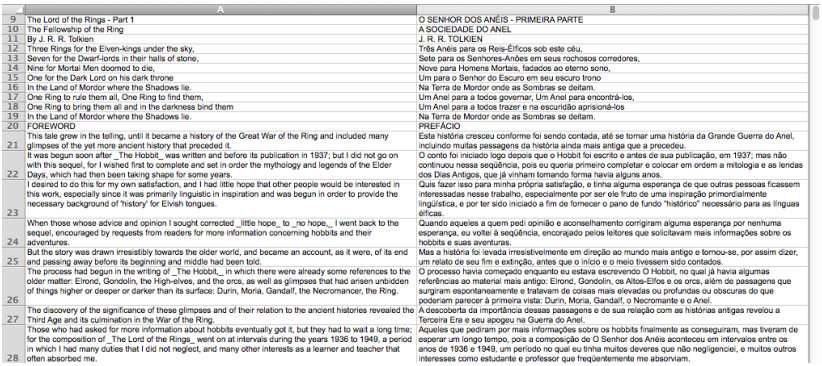


Figure 19. Source and target texts side-by-side before the alignment process.

In the **first** prototype scenario, Hunalign was tested without a dictionary and the final score was 0.2 as a result the bitext was messy with blank spaces two or more sentences together with no relation to their translated version, etc. The **second** prototype scenario, Hunalign was tested with a dictionary of English-Portuguese of most common words, and the score was 0.7—this quality is already acceptable with minor mistakes. Finally, on the **third** prototype scenario, Hunalign was tested with a custom dictionary created from a word list extracted from the very same text, with specific vocabulary from Tolkienian style and the score reached 0.845. In a second step the final bitext was transferred to a spreadsheet on Apache OpenOffice Calc[56] for manual inspection and correction. The procedure was doable to conduct, and provided 11,520 sentences well aligned and ready for insertion in COPA-TRAD parallel corpus (Figure 20). The discussed procedure are still under investigation for further refinements in the process, but the results reached so far, are encouraging because similar outcomes were achieved with texts from different authors and translators in COPA-TRAD text domain.

[56] http://www.openoffice.org/pt/product/calc.html



	A	B
9	The Lord of the Rings - Part 1	O SENHOR DOS ANÉIS - PRIMEIRA PARTE
10	The Fellowship of the Ring	A SOCIEDADE DO ANEL
11	By J. R. R. Tolkien	J. R. R. TOLKIEN
12	Three Rings for the Elven-kings under the sky,	Três Anéis para os Reis-Élficos sob este céu,
13	Seven for the Dwarf-lords in their halls of stone,	Sete para os Senhores-Anões em seus rochosos corredores,
14	Nine for Mortal Men doomed to die,	Nove para Homens Mortais, fadados ao eterno sono,
15	One for the Dark Lord on his dark throne	Um para o Senhor do Escuro em seu escuro trono
16	In the Land of Mordor where the Shadows lie.	Na Terra de Mordor onde as Sombras se deitam.
17	One Ring to rule them all, One Ring to find them,	Um Anel para a todos governar, Um Anel para encontrá-los,
18	One Ring to bring them all and in the darkness bind them	Um Anel para a todos trazer e na escuridão aprisioná-los
19	In the Land of Mordor where the Shadows lie.	Na Terra de Mordor onde as Sombras se deitam.
20	FOREWORD	PREFÁCIO
21	This tale grew in the telling, until it became a history of the Great War of the Ring and included many glimpses of the yet more ancient history that preceded it.	Esta história cresceu conforme foi sendo contada, até se tornar uma história da Grande Guerra do Anel, incluindo muitas passagens da história ainda mais antiga que a precedeu.
22	It was begun soon after _The Hobbit_ was written and before its publication in 1937; but I did not go on with this sequel, for I wished first to complete and set in order the mythology and legends of the Elder Days, which had then been taking shape for some years.	O conto foi iniciado logo depois que o Hobbit foi escrito e antes de sua publicação, em 1937; mas não continuou nessa seqüência, pois eu queria primeiro completar e colocar em ordem a mitologia e as lendas dos Dias Antigos, que já vinham tomando forma havia alguns anos.
23	I desired to do this for my own satisfaction, and I had little hope that other people would be interested in this work, especially since it was primarily linguistic in inspiration and was begun in order to provide the necessary background of 'history' for Elvish tongues.	Quis fazer isso para minha própria satisfação, e tinha alguma esperança de que outras pessoas ficassem interessadas nesse trabalho, especialmente por ser ele fruto de uma inspiração primordialmente lingüística, e por ter sido iniciado a fim de fornecer o pano de fundo "histórico" necessário para as línguas élficas.
24	When those whose advice and opinion I sought corrected _little hope_ to _no hope,_ I went back to the sequel, encouraged by requests from readers for more information concerning hobbits and their adventures.	Quando aqueles a quem pedi opinião e aconselhamento corrigiram alguma esperança por nenhuma esperança, eu voltei à seqüência, encorajado pelos leitores que solicitavam mais informações sobre os hobbits e suas aventuras.
25	But the story was drawn irresistibly towards the older world, and became an account, as it were, of its end and passing away before its beginning and middle had been told.	Mas a história foi levada irresistivelmente em direção ao mundo mais antigo e tornou-se, por assim dizer, um relato de seu fim e extinção, antes que o início e o meio tivessem sido contados.
26	The process had begun in the writing of _The Hobbit_, in which there were already some references to the older matter: Elrond, Gondolin, the High-elves, and the orcs, as well as glimpses that had arisen unbidden of things higher or deeper or darker than its surface: Durin, Moria, Gandalf, the Necromancer, the Ring.	O processo havia começado enquanto eu estava escrevendo O Hobbit, no qual já havia algumas referências ao material mais antigo: Elrond, Gondolin, os Altos-Elfos e os orcs, além de passagens que surgiram espontaneamente e tratavam de coisas mais elevadas ou profundas ou obscuras do que poderiam parecer à primeira vista: Durin, Moria, Gandalf, o Necromante e o Anel.
27	The discovery of the significance of these glimpses and of their relation to the ancient histories revealed the Third Age and its culmination in the War of the Ring.	A descoberta da importância dessas passagens e de sua relação com as histórias antigas revelou a Terceira Era e seu apogeu na Guerra do Anel.
28	Those who had asked for more information about hobbits eventually got it, but they had to wait a long time; for the composition of _The Lord of the Rings_ went on at intervals during the years 1936 to 1949, a period in which I had many duties that I did not neglect, and many other interests as a learner and teacher that often absorbed me.	Aqueles que pediram por mais informações sobre os hobbits finalmente as conseguiram, mas tiveram de esperar um longo tempo, pois a composição de O Senhor dos Anéis aconteceu em intervalos entre os anos de 1936 e 1949, um período no qual eu tinha muitos deveres que não negligenciei, e muitos outros interesses como estudante e professor que freqüentemente me absorviam.

Figure 20. Resulting bitext organized in a spreadsheet.

I believe that using a hybrid alignment tool in association with a custom dictionary (prepared from the text to be aligned) is possible to achieve a significant high quality alignment in the final bitext. Furthermore, a manual inspection to check and correct bad alignments can be carried out easily by putting the bitext on a spreadsheet.

After this brief description this section approaches to the technical procedures and methods applied for AUTO ALIGNER construction. The tool is divided in four steps: (a) texts upload; (b) linguistic processing; (c) text alignment; (d) spreadsheet generation and download. The focus here is on items b and c since the other are common procedures in any online platform.

3.4.1 AUTO ALIGNER – Linguistic Processing

The linguistic processing in AUTO ALIGNER is applied mainly for text normalization and a special Text Mining technique is part of this task. The subtasks at this part, are the following verification and correction algorithms: (a) tags/annotation; (b) file format; (c) charset; (d) BOM; (e) text structure; (f) language detection. Additionally, there was ideas to identify which text is the original and which one is the translation automatically. However, due to philosophical and theoretical issues, this task could not be done. I leave the translation automatic identification as a suggestion for future research in last chapter. The suggestion accompanies a possible approach elaborated by me to accomplish this action or at least to create an initial prototype. At the first step of AUTO ALIGNER process, a series of verifications is carried out. The first one is file checking and it supports the PDF, DOC and DOCX, if any of this kind is identified the system tries to convert automatically. For this

purpose, the system checks if the Linux option *pdftotext* is available and if so the following code is triggered:

PHP programming language.

```
1    $return = shell_exec("pdftotext -layout " . $file);
```

In case the option above is not available, the AUTO ALIGNER system loads a PHP library named *pdf2text*[57] to prepare the conversion, but with less quality (i.e., not all kinds of PDF can be converted). In case the file is DOC another kind of conversion takes place like the same manner of PDF firstly the system tries a Linux command *Abiword*:

PHP programming language.

```
1    shell_exec("abiword --to=txt " . $file . " -o " . $outputText);
```

In case *Abiword* is not available another PHP function is applied "read_doc_file"[58] this function I prepared a custom code in order to properly convert Microsoft Windows ASCII characters according to UTF-8 table (Figure 21).

[57] Available at: https://pastebin.com/dvwySU1a
[58] Available at: https://github.com/joeblurton/doccounter/blob/master/class.doccounter.php

Unicode code point	character	UTF-8 (in literal)
U+2600	☀	\xe2\x98\x80
U+2601	☁	\xe2\x98\x81
U+2602	☂	\xe2\x98\x82
U+2603	☃	\xe2\x98\x83
U+2604	☄	\xe2\x98\x84
U+2605	★	\xe2\x98\x85
U+2606	☆	\xe2\x98\x86
U+2607	☇	\xe2\x98\x87
U+2608	☈	\xe2\x98\x88
U+2609	☉	\xe2\x98\x89
U+260A	☊	\xe2\x98\x8a
U+260B	☋	\xe2\x98\x8b
U+260C	☌	\xe2\x98\x8c
U+260D	☍	\xe2\x98\x8d
U+260E	☎	\xe2\x98\x8e
U+260F	☏	\xe2\x98\x8f

(...)

Figure 21. UTF-8 encoding table and Unicode characters[59].

The final file checking procedure is DOCX which tries firstly run a third-party Perl script[60]:

PHP programming language.

```
1   $return = shell_exec("perl " . APPPATH .
    "third_party/docx2txt.pl " . $text);
```

With the TXT version of the texts the system continues to the next step which is tags/labels removal. The process removes all tags enclosed by "<" and ">", customized tags different than the above ones are not

[59] https://www.utf8-chartable.de/unicode-utf8-table.pl?start=9728&number=128&names=-&utf8=string-literal

[60] Available at https://sourceforge.net/projects/docx2txt/

removed. However for most of the text found on Internet or any other text database this method is reliable because it can remove HTML as well as XML. Other kinds of markups can be removed manually. After this process the system proceed with text encoding conversion, in case is not in UTF-8, the process is carried out through a PHP native function named "mb_convert_encoding." The next step is checking if the text has BOM and then proceed with its removal. When text encoding and BOM removal is finished, the system executes the text structural organization by extracting each line (e.g., a sentence) from the text and then removing the following special characters from the line beginning and ending:

Table 6. Special characters removed by PHP trim function (trim, n.d.).

Special Characters		
" "	(ASCII 32 (0x20))	An ordinary space.
"\t"	(ASCII 9 (0x09))	A tab.
"\n"	(ASCII 10 (0x0A))	A new line (line feed).
"\r"	(ASCII 13 (0x0D))	a carriage return.
"\0"	(ASCII 0 (0x00))	The NUL-byte.
"\x0B"	(ASCII 11 (0x0B))	A vertical tab.

After the cleansing a new standard line ending is merged with the line and this line is written in a new file. This process continues over and over until finishing parsing text lines. Finally, the old file is removed from the server and the new file receive the necessary reading and writing permission. After this process, AUTO ALIGNER system tries to guess both texts languages, since the alignment tool, works for texts in the linguistic pair[61] (Portuguese and English or vice-versa). This process, is automatic through a Text Mining algorithm for language identification which uses language models (i.e., a special kind of dictionary) to assess the text. Parts from the text are compared with parts from the training set, in case of equality, the score receives an increment (plus one). For the tests conducted this algorithm demonstrated consistent accuracy. After this process original and translation as identified manually due to the issues already mentioned.

[61] Other languages are planned for AUTO ALIGNER in future version.

3.4.2 AUTO ALIGNER – Bitext Alignment

Section 2.3.5 discussed the theoretical aspects of bitext alignment and the software embedded in AUTO ALIGNER, Hunalign. The current version of AUTO ALIGER, only supports the alignment with a default dictionary, but new versions this limitation will be solved. The use of a dictionary is important because "not all languages have the same grammatical features" such as morphological and syntactic features that is, word order and redundancy (Yang & Li, 2003, p. 734). As a result, the first step of the alignment process is to identify the directionality in order to load in memory the correct dictionary: *en-pt.dic* (original in English and translation in Portuguese) or *pt-en.dic* (original in Portuguese and translation in English). These dictionaries are simple TXT file with 1:1 word alignment. Figure 22, is a snapshot of the English—Portuguese training dictionary. Due to Hunalign internal algorithm, the target language (Portuguese) comes first and then the source language (English). With the availability of the standard dictionary, the alignment procedure follows according to the code:

PHP programming language.

```
1  shell_exec($aligner_bin . ' -text ' . $dict . ' ' . $source_text . ' ' .
   $target_text . ' > ' . $aligned_text);
```

On server side the text is saved in simple TXT file with a delimiter character between each part of alignment (i.e., source & target combined).

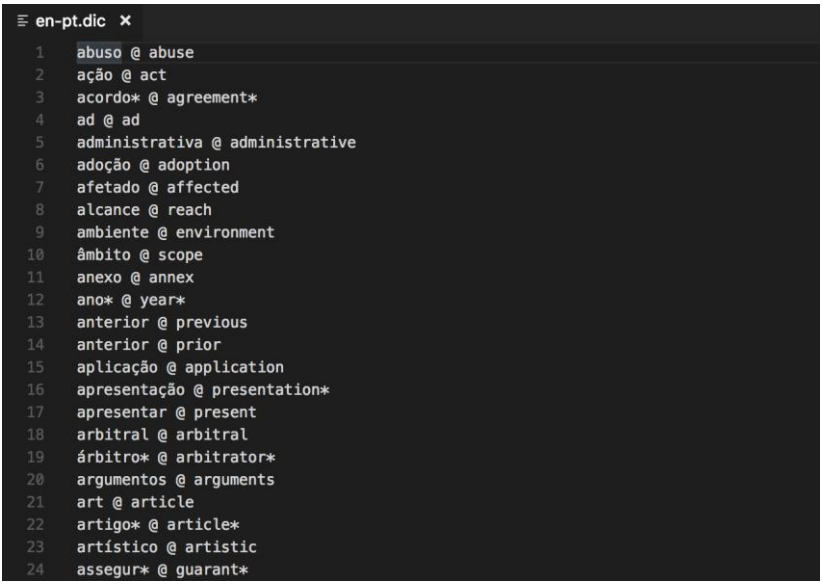


Figure 22. Hunalign English – Portuguese default training dictionary.

The final step is the availability of the bitext in Excel format. At this part, AUTO ALIGNER reads the generated TXT file, converts the sentences from UTF-8 to UTF-16LE (just because Excel difficulty in recognizing certain characters) and appends to final XLS file for downloading. The Excel XLS file is divided in two columns: the first one contains the source text and the second one the translated one. In case the user wants to add each text in TXT file, he/she can simply copy the entire column and past in Notepad++.

3.5 Closing Remarks

This chapter discussed Corpus Mining model and its procedures. In addition, the application of these procedures and concepts are going to be demonstrated in a real case scenario in next chapter. COPA-TRAD Version 2 tools and its internal processing mechanisms, namely AUTO ALIGNER. I would like to emphasize that other tools were developed during the period, for example, Smart Search which uses sentiment analysis to categorize all COPA-TRAD corpus entries in Positive, Neutral and Negative, but due to time and scope of this study the an in-depth discussion of this tool can be further explored in new studies. As already mentioned, the next chapter discusses the application of Corpus Mining model on COPA-TRAD Version 2 tools, from an user perspective.

4 CHAPTER FOUR: Applying Corpus Mining on COPA-TRAD Version 2

4.1 Initial Remarks

A preliminary analysis of COPA-TRAD tools was commenced during my MA dissertation (Silva, 2013, pp. 85-116). The focus of this chapter is on the tools developed for COPA-TRAD Version 2. The intention here is to discuss the tools in the light of Phase Eight from the proposed Corpus Mining model (Figure 23). The discussion in the following sections presents the tools, providing a practical context and showing some of the Text Mining techniques implemented by the COPA-TRAD platform in action.

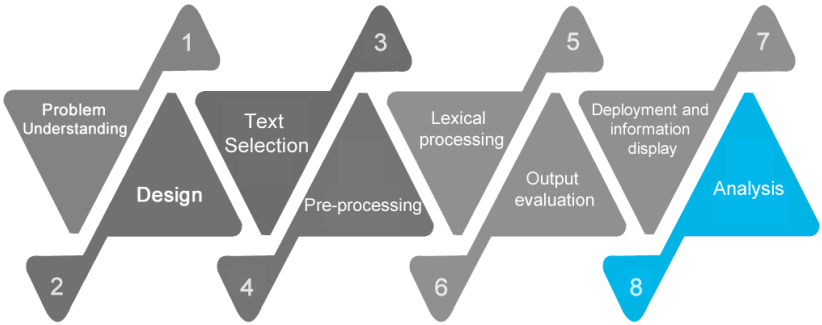


Figure 23. Corpus Mining model phase eight.

First of all, from the point of view of an end user, I present a tool named AUTO ALIGNER, and I explain how it works in its current version. Then, the Advanced Search feature of the COPACONC parallel concordancer is presented, its use also explained. Finally, the other tools to be discussed include: COPACONC Expert Search, WORDLIST (new version), TREETAGGER CLOUD, and COPA STATS.

4.2 AUTO ALIGNER: COPA-TRAD Automatic Alignment Tool

The development of COPA-TRAD's AUTO ALIGNER tool (Figure 24), its algorithm, and internal processing mechanisms have been extensively discussed in Chapter 3. A point worth emphasizing here is that AUTO ALIGNER runs entirely in the cloud; that is to say, there is no necessity to download and install any sort of software to use this particular tool on a PC. This feature is attractive, especially for users of tablets and smartphones, since these gadgets are usually limited in terms of storage capacity and processing. AUTO ALIGNER could be developed

to work seamlessly in just one interaction with the user. For example, the user uploads the texts and clicks on the "align" button.

However, because AUTO ALIGNER (as with other tools) is intended for the classroom environment, it has a didactic appeal, especially pertaining to Translation Studies courses. The objective is to support the process of teaching about corpus investigation, textual alignment, and how texts are processed. The tool was conceived to be user-friendly, and each processing step is visually displayed for the user. I would like to make the point here that this particular tool is not just intended for advanced researchers with technical knowledge of parallel corpora. In fact, I believed that this tool can empower general users (students, professors, translators) interested in parallel corpora, those who have always found it difficult to handle the processing of textual alignment. I expect that this feature can contribute to the popularization of parallel corpus investigation in academia.

For this reason, AUTO ALIGNER works, considering the four basic steps discussed, as follows: (a) texts upload; (b) analysis and lexical processing; (c) textual alignment; (d) final result/download. Needless to say, AUTO ALIGNER works on the back end with a set of complex techniques (see Section 3.4) at programming level. In addition, the application of Text Mining techniques and training, set by means of a specific algorithm used to automatically identify the language of each submitted text, is an important feature that works without the user's intervention. As previously discussed, this feature guarantees that the texts submitted are written in the languages (i.e., English and Brazilian Portuguese) supported by the system.

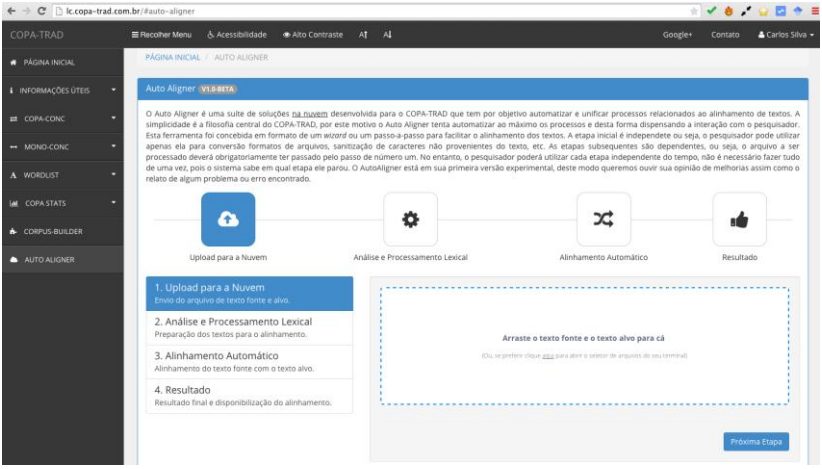


Figure 24. AUTO ALIGNER main screen.

AUTO ALIGNER is capable of analyzing and checking the submitted texts for possible errors (i.e., problematic textual information not necessary for the final alignment, such as excess of blank spaces, tags/labels, and so on). In case the uploaded files are in different file formats, such as JPG, SWF, MP4, an error message pops up, informing the user to upload files according to the specified format. The upload process as well as the validation mechanisms are always triggered when the second file is sent. The only error that appears when the first file is uploaded is the one related to the file format. In case the user tries to upload a third file, the system does not allow this, triggering a specific error message to pop up. These procedures work without the web page being refreshed, and they are run in the background so that they can quickly provide feedback to the user.

AUTO ALIGNER processes are comparatively simple, rather than complex, especially for users who find similar tools difficult to use and inaccessible, either due to the complicated nature of the installation of these tools or the unavailability of the training set (linguistic models/dictionaries) in the desired languages. AUTO ALIGNER, from a user's point of view, is divided into the four steps illustrated in Figure 25.

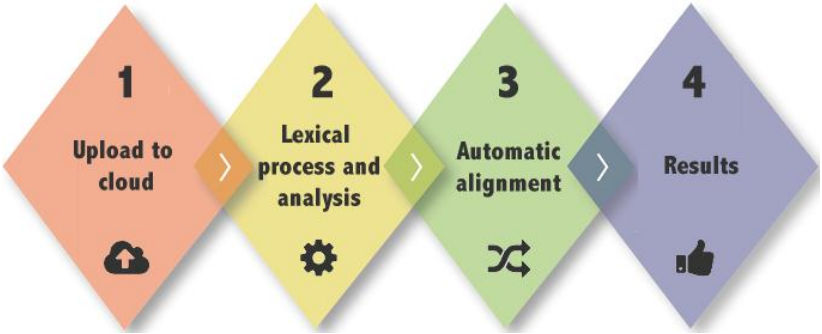


Figure 25. AUTO ALIGNER process workflow.

Figure 25 illustrates the four steps of AUTO ALIGNER that users deal with while using the tool. The following subsections will discuss in detail each step that is presented to the user while using the tool. I would like to stress that AUTO ALIGNER can be utilized to process texts according to the available features in Step 2, which is far from the first tool's objective, which is alignment. To illustrate, if a user wants just the text conversion and cleaning, he/she can use AUTO ALIGNER to do this job. However, if the user wants the bitext, he has to go to the final step. At this point, the processed and cleaned files are available for download. Now the focus is on how to obtain better results with AUTO ALIGNER, and to use the tool, the first thing the user has to do is to log in, since this is the only means available to use the tool. The log-in is necessary to keep the access by users under control, as well as being part of planning for the next version: store the aligned files, and associate them with the user account in order to keep the files in a private cloud that is always available whenever the user needs access.

4.2.1 AUTO ALIGNER First Step

According to Figure 26, the first step is the availability of the texts to be processed and aligned. At this part, the user has to upload the texts he/she has wants to align. The allowed file formats to be uploaded are TXT, DOC, DOCX and PDF. At this first version of AUTO ALIGNER just the four file formats are allowed and the PDF file should be already converted with the OCR (*Optical Character Recognition*) technology. For the next version of the Auto Aligner, the OCR technology will be available and embedded in the tool and it will work transparently for the user (i.e., the system will try to convert to text even PDF scanned as images).

Figure 26. Auto Aligner first step screen.

After uploading the files to COPA-TRAD cloud environment, the first step screen shows a success message. AUTO ALIGNER shows the files in the dashed rectangle area with the file sizes in Kb as well. The button to proceed to the next step is enabled and stretched indicating to the user what he/she has to do (Figure 27).

Figure 27. AUTO ALIGNER first step with the files already uploaded.

By observing Figure 26 as well as Figure 27 it is possible to perceive the design is simple and intuitive. To precede the upload to the cloud user can simply drag the files from his personal folder and drop in the indicated area (i.e., dashed rectangle). Alternatively, users can click on the dashed rectangle and a dialogue box opens showing the files from the local computer to be uploaded. This screen shows a short explanation centered

in the blank space so it calls the user attention. Bellow, the upload box there is a button, but in case the user tries to click on it, nothing happens because it is disabled by default. The button only works after step one is completed. After the upload and the success of it the following window (or panel) pops in, as observed in Figure 28. The screen is simple and was developed to be intuitive and informative (i.e., the task list is accompanied with individual explanations) as possible.

4.2.2 AUTO ALIGNER Second Step

Step two as exemplified in Figure 28, is the one related to the application of lexical processing techniques and textual normalization to prepare the text for bitext alignment. Before going deep the explanation of this step, I would like to discuss, in a nutshell, the AUTO ALIGNER main screen and how it works when the user triggers the event when he/she clicks on "next button."

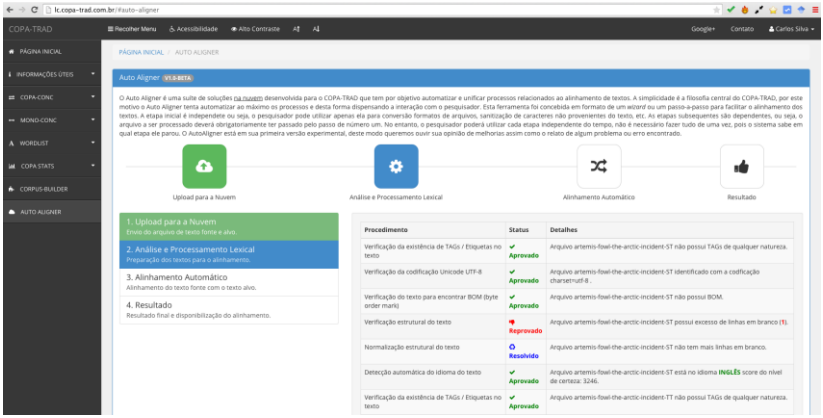
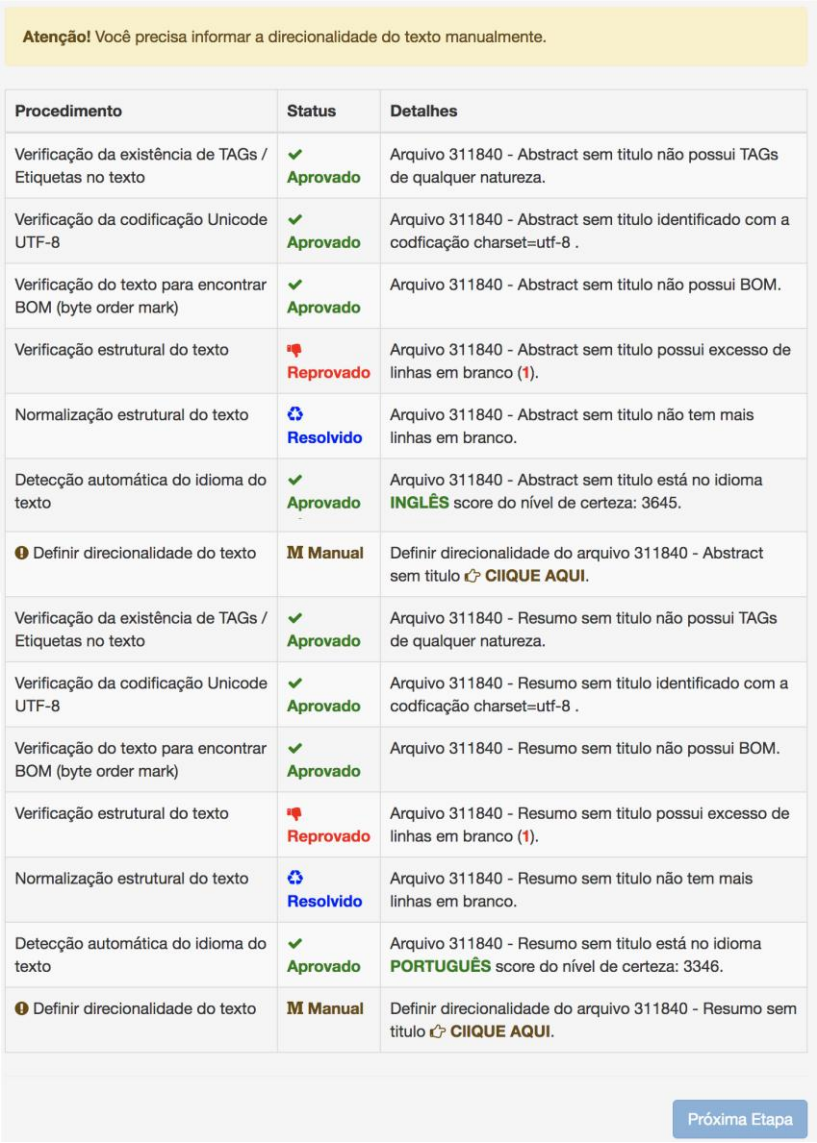


Figure 28. AUTO ALIGNER main screen showing its state after first step.

The focus on Figure 28 is on the green and blue squares in the continuum line and the blue and green rectangle at the left side panel. In order to make the tool user friendly for novice users each step of AUTO ALIGNER process is signaled visually to the user. As observed in Figure 28, the first square and rectangle are blue. While the user moves to the next step the previous step becomes green signaling everything worked as expected and succeeded. The current step, is signaled to the user with blue color and this behavior continues until the last step.

After this brief explanation of AUTO ALIGNER and how its events are triggered when the user moves forward to next step, I want to

discuss now, the second step in detail. For that reason the table in Figure 29 demonstrate the most common tasks of this step, the information were prepared to be didactic as possible, due to the reasons already discussed.



Atenção! Você precisa informar a direcionalidade do texto manualmente.

Procedimento	Status	Detalhes
Verificação da existência de TAGs / Etiquetas no texto	✔ Aprovado	Arquivo 311840 - Abstract sem titulo não possui TAGs de qualquer natureza.
Verificação da codificação Unicode UTF-8	✔ Aprovado	Arquivo 311840 - Abstract sem titulo identificado com a codficação charset=utf-8 .
Verificação do texto para encontrar BOM (byte order mark)	✔ Aprovado	Arquivo 311840 - Abstract sem titulo não possui BOM.
Verificação estrutural do texto	👎 Reprovado	Arquivo 311840 - Abstract sem titulo possui excesso de linhas em branco (1).
Normalização estrutural do texto	♻ Resolvido	Arquivo 311840 - Abstract sem titulo não tem mais linhas em branco.
Detecção automática do idioma do texto	✔ Aprovado	Arquivo 311840 - Abstract sem titulo está no idioma **INGLÊS** score do nível de certeza: 3645.
❗ Definir direcionalidade do texto	M Manual	Definir direcionalidade do arquivo 311840 - Abstract sem titulo 👉 **CIIQUE AQUI**.
Verificação da existência de TAGs / Etiquetas no texto	✔ Aprovado	Arquivo 311840 - Resumo sem titulo não possui TAGs de qualquer natureza.
Verificação da codificação Unicode UTF-8	✔ Aprovado	Arquivo 311840 - Resumo sem titulo identificado com a codficação charset=utf-8 .
Verificação do texto para encontrar BOM (byte order mark)	✔ Aprovado	Arquivo 311840 - Resumo sem titulo não possui BOM.
Verificação estrutural do texto	👎 Reprovado	Arquivo 311840 - Resumo sem titulo possui excesso de linhas em branco (1).
Normalização estrutural do texto	♻ Resolvido	Arquivo 311840 - Resumo sem titulo não tem mais linhas em branco.
Detecção automática do idioma do texto	✔ Aprovado	Arquivo 311840 - Resumo sem titulo está no idioma **PORTUGUÊS** score do nível de certeza: 3346.
❗ Definir direcionalidade do texto	M Manual	Definir direcionalidade do arquivo 311840 - Resumo sem titulo 👉 **CIIQUE AQUI**.

Próxima Etapa

Figure 29. AUTO ALIGNER second step showing a table with possible results.

The alignment process could be done without the text preparation, but the final result (i.e., the bitext) could have a wide range of missing sentences or sentences badly aligned. To avoid such problem, during my investigation and development of the prototype, I could perceive the final quality of the aligned text increases positively if the text is pre-processed or pre-organized, as discussed in the methodological chapter. This pre-processing step is not aimed to be exhaustive, but the execution of simple tasks can increase the quality of final product (i.e., bitext). As result, the second step, was developed to provide the initial tasks before the bitext alignment. Better results can be obtained with the application of this step in the alignment process.

The table in Figure 29 was prepared to be didactic as possible and it is divided in three columns namely "procedure," "status" and "details." The first column lists the procedures/tasks taken while processing the texts. These tasks may vary according to the submitted text structured and its content. The internal algorithm of AUTO ALIGNER decides what to do: when a possible problem is identified AUTO ALIGNER tries to solve the problem automatically. In case a problem cannot be solved automatically a button near in "status" appear for the user intervention. The second column, shows the possible three states of a specific task: "approved," "failed" and "solved." The last column explains in detail the results of each procedure. This last column presents detailed information explaining each step to the user. Table 7, presents all possible tasks available at this version of AUTO ALIGNER and an explanation of them.

Table 7. *AUTO ALIGNER procedures executed in step two.*

Procedure	**Explanation**
File format checking	Auto aligner tries to identify automatically the file format of the two sent files.
Automatic conversion from PDF to TXT	In case the file is identified as PDF the systems runs the automatic conversion from PDF to TXT.
Automatic conversion from DOCX to TXT	In case the file is identified as DOCX the system runs the automatic conversion from DOCX to TXT.

Automatic conversion from DOC to TXT	In case the file is identified as DOC the system runs the automatic conversion from DOC to TXT.
Text analysis to check if texts have tags/labels such as HTML or XML markup	If tags or labels are found all of them are purged to keep just the text itself.
Analysis of text charset encoding	Carries out an analysis to check if the text files are in Unicode UTF-8 charset encoding. If not, the system converts the files automatically to UTF-8.
Textual analysis to find if the text was encoded in UTF-8 with BOM (Byte order marker)	If the BOM is found in the text file it is removed automatically.
Text structure analysis	Verifies if the system has blank lines. In case blank lines are identified they are removed. The new line characters are kept.
Automatic language identification	The system extracts three text samples from the text being analyzed one from beginning, one from the middle and one from the end. After that Auto Aligner identifies automatically the language of the text. This procedure is based on likelihood formulae.

The procedures listed in Table 7 are executed automatically without user intervention. The only manual intervention at this step is to inform which text is the source and which one is the target (Figure 30). As mentioned before, the system cannot identify originals and translations. The user has to click on the informed link, then a modal[62] opens where the file name, text snippet and option selection (with two options original and translation) are displayed. The user selects the desired option and finally clicks on the "save" button. The same procedure is carried out for the second text, the only difference is that the option selection has only one

[62] A child of screen that opens in front of the main screen.

option the already selected item for the first text disappear. Then the button to proceed to third step is enabled like the first version.

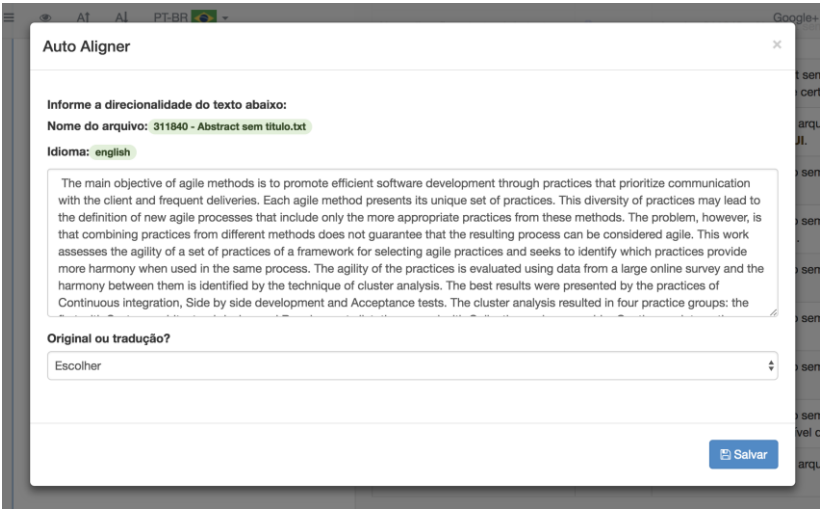


Figure 30. AUTO ALIGNER screen asking the user if the text is original or translation.

Experienced researchers in parallel corpus may admit some of the tasks listed in Table 7 are complex to handle manually (e.g., executing individual software to solve possible found problems). The steps may be complex to handle manually, because, users have to find and install in his local PC the most suitable software to convert a text, deal with another software or elaborate a regular expression to match and remove the tags and labels, deal with software to convert the files to UTF-8 and finally remove the blank lines which can be problematic in long streams of texts. These possible problems are cited just to illustrate what AUTO ALINGER is capable of fulfilling automatically.

4.2.3 AUTO ALIGNER Third Step

The chosen tool for aligning texts, Hunalign, is capable of generating dictionaries on the fly. In addition, in case a customized dictionary based on the text is provided, there is a chance to obtain better results: As discussed in last chapter, I conducted a series of tests in order to obtain better results and I found customized dictionaries created from each text, are the best choice to achieve high quality bitext, but this finding is not available in the current version of AUTO ALIGNER. The

discussion in last chapter, explained Hunalign overall quality ranges from 0 to 1, for example, values such as 0.2 (badly aligned), 0.8 or even 0.9 (best aligned). The third step (Figure 31) comprises tasks related specifically to bitext alignment process. Before the bitext alignment, the training set/dictionaries are loaded for the use according to the language (in this version just Portuguese and English). These dictionaries support the bitext alignment process. The dictionaries were prepared according to COPA-TRAD most common words (words from the list with high frequency).

Figure 31. AUTO ALIGNER third step.

AUTO ALIGNER third step screen is mainly informational and does not require action from the user. There only option in the screen is the button which moves forward to the final step. In future versions of AUTO ALIGNER, there is a plan to provide the user the chance to download, check and correct the bitext in order to fix missing alignments and return back to the system. As a result, this procedure will make possible for the user submit the bitext for COPA-TRAD corpus directly from AUTO ALIGNER. This is possible, because the text would be already checked and fixed manually, to guarantee high quality bitext.

4.2.4 AUTO ALIGNER Fourth and Final Step

The fourth step, is the final one in AUTO ALIGNER (Figure 32). At this final part, the bitext has already been processed and aligned. Here, the bitext can be downloaded in Excel. The indicated file available for downloading is organized in three distinct columns: the paired sentences (i.e., source and target language) are stored in two columns side by side, facilitating the manual analysis if the user decides to do it, the third column is just informational and presents the quality score generated by Hunalign for the specific sentence alignment. The name of the columns

are: (a) Source Text; (b) Target Text; (c) Score. The final panel, at present has just a message asking the user to download the final alignment. However in future versions, a message inviting the user to analyze his text, solve some possible alignment problems and submit it to COPA-TRAD will be displayed. Some of the planning for future version are discussed in the next and last chapter.

Figure 32. AUTO ALIGNER final step.

4.3 COPACONC Advanced Search

Apparently, the first ever computer concordancer was developed by Luhn in 1960 who also coined the term "KWIC"[63] (Manning & Schütze, 1999, p. 35). Nowadays, important breakthrough was done in the area and COPACONC Advanced Search (Figure 33) is no exception. This COPA-TRAD tool was designed to provide a comprehensive set of detailed searching filters to retrieve information from COPA-TRAD parallel corpus. The available filters were planned and developed to cover a wide variety of translational variables. Due the number of filters, COPACONC advanced search options are divided in three groups/subpanels and the user can navigate through the options by scrolling down and up. This section is organized as it follows: the first group deals with the logical operators or Boolean query syntax (at this part the main text box for writing keywords is covered) as well as language and subcorpora options. Secondly, paratext filters are covered. The third group is named "directionality" and restricts results according to the following options: (a) any direction; (b) original → target; (c) target

[63] KWIC is an acronym for Key Word in Context (see https://copa-trad.ufsc.br/#monolingual-concordancer).

→ original. Before moving forward, COPACONC Advanced Search filters, as shown in Figure 33, can be hidden by clicking on the green button located at the top.

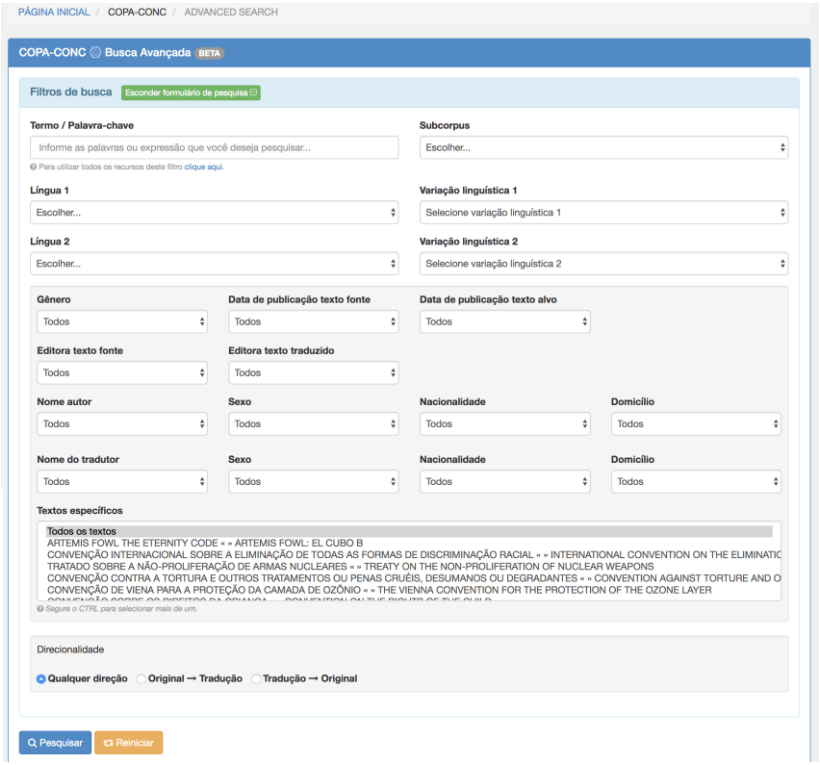


Figure 33. COPACONC Advanced Search screen.

According Figure 33, the "Specific Text" select box is dynamic, that is, according to the filters selected above, the texts titles in this select box are included or excluded from the list in real time. In addition, there is the possibility to select more than one text as well, by holding the CTRL (control) button on keyboard and clicking each title individually. The next sections are going to deal with each subpanel separately.

4.3.1 First Group of Filters

The first group has six filtering operations (Figure 34). The filters are the following ones: (a) Search/Keyword (b) Language 1; (c) Language 2; (d) Linguistic Variation 1; (e) Linguistic Variation 2; (f) Subcorpus.

Figure 34. COPACONC Advanced, first group of filters.

The selection of subcorpus triggers an event and updates the "specific text" field list. In addition, according to the selection of Language 1 and 2 the Linguistic Variation (Figure 35) updates itself with the available options referred to the languages and the quantity in the corpus.

Figure 35. Linguistic Variation for English.

The fields in this first group are very concise but just the first one which deserves more attention.

The first filter is the "term/keyword." The available filtering operations are based on Sphinx Search queries. Several searching possibilities are available to be used in this field. However, it can be complex for novice researchers in the field to deal with such searching mechanisms. Hence, each searching field works similar as a Google search (i.e., the field accepts simples queries as well as complex ones). For instance, users can type what they want like: (a) "said", (b) "-harry" and (c) "hermine AND ron." These examples, are considered valid search query by Sphinx and according to the examples, the results would be all

entries with “said,” all entries except the ones with “harry” and entries with “Hermine and Ron.” Combinations are possible as well for example “hermine AND ron -harry” all entries with “Hermine and Ron” except the ones wich contains “harry.”

The search field accepts keywords separately as well as full phrases. In case, the user decides to inform full phrases it should be enclosed in double quotes, for exact search. This feature works in the same way like commercial search engines like Google, Bing or Yahoo. When the phrase is informed without the double quotes, COPA-TRAD search engine will match each word separately whatever the position is in the text (i.e., ignoring the order text words where informed). To illustrate, a user searching for adverbs he/she just types each one separated by a space: ***dangerously carefully nicely easily***. In case the user wants to investigate the construction VERB + ADVERB specifically he/she may type like “***said carefully***.” As a result, the occurrences for the last searching expression (i.e., “said carefully”) would be like in Table 8.

Table 8. *Entries from COPACONC based on the search query “said carefully.”*

English	Portuguese
He strode to the cabin door, opened it and bowed Dumbledore out. Fudge, fiddling with his bowler, waited for Hagrid to go ahead of him, but Hagrid stood his ground, took a deep breath and said carefully, 'If anyone wanted ter find out some stuff, all they'd have ter do would be ter follow the spiders. That'd lead 'em right! That's all I'm sayin'.'	E dirigiu-se à porta da cabana, abriu-a, fez um gesto largo indicando a porta para Dumbledore. Fudge, manuseando seu chapéu-coco, esperou Hagrid passar à sua frente, mas Hagrid continuou firme, inspirou profundamente e disse com clareza: – Se alguém quiser descobrir alguma coisa, é só seguir as aranhas. Elas indicariam o caminho certo! É só o que digo.
Janet drew a long, quiet breath and managed not to tell him to stay and be turned into a frog then. She made a very ingenious face at the mirror and counted up to ten. “Cat,” she said carefully, “we really are in such	Janet respirou lenta e profundamente, e conseguiu não dizer a ele para ficar e virar sapo. Ela fez uma careta muito engraçada e contou até dez. – Gato, estamos realmente numa encrenca tão grande que: não

a mess here that I can't see any other way out. Can you?"	consigo enxergar outra saída. Você consegue? – ela disse cuidadosamente.
"Paolo said, carefully patient, "I can't hold on much longer. Do you think you can have another try?" The answer from Renata was a sobbing shudder."	Paolo, com cautelosa paciência. disse: – Não vou conseguir me segurar muito tempo mais. Acha que pode fazer outra tentativa? A resposta de Renata foi um gemido e um tremor.

It is worth noting the last row in Table 8 the highlighted match has a comma character. This happened because Sphinx ignores such characters it only consider alphanumeric words. The researcher has to be aware of it because it can be useful or a problem. Other searching possibilities are still available and the comprehensive analysis of it was dealt in my MA dissertation, since this field works at the same way of COPACONC Simple Search (Silva, 2013, p. 89).

4.3.2 Second Group of Filters

The second group (Figure 36) of filters are related specifically to paratext for either, originals and translated texts. The common filter for both kind of text is the one entitled "genre" and it contains the text genres available in the corpus. Then for original and translated texts the filters are the following ones: (a) Publication Date; (b) Publishing house; (c) Author Name; (d) Gender; (e) Nationality; (f) Home address. The only difference is that for translated texts there is an additional field to select specific translators. Like the "corpus" filter, these filters triggers an internal event which updates the "specific text" field list.

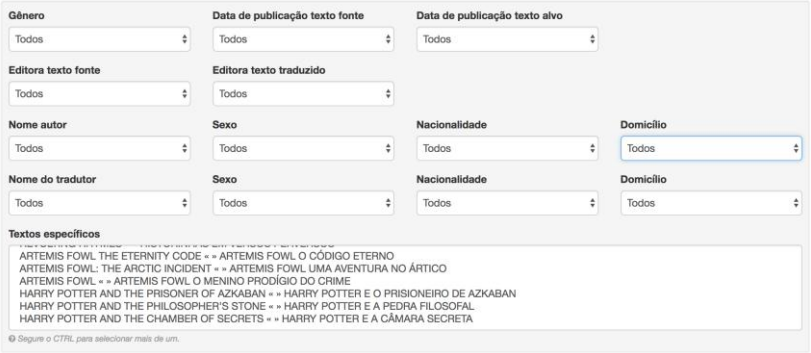


Figure 36. COPACONC Advanced Search, second group filters.

4.3.3 Third Group of Filters

The third group of filters (Figure 37) is the one related to directionality. The options available, as discussed earlier, are: (a) any direction; (b) original → target; (c) target → original. Fernandes (2009) illustrates this feature "for instance, in a corpus containing text originally written in L1 [language 1] and their respective translation in L2 [language 2] the translation is in just one direction, so the corpus is unidirectional" (p. 20). Other possibilities are still available as observed by Fernandes (2009) such as a corpus made up "of texts originally written in L1 and their translations in L2 plus originals in L2 and their translations in L1" the corpus is bidirectional (p. 20). Finally Fernandes (2009) defines the multidirectional corpus "especially when more than two languages are involved and their translation direction is not centered on L1, but on interaction among all the languages constituting the corpus" (pp. 20-21). For this last definition, the "any direction" filter contemplates this feature as well.

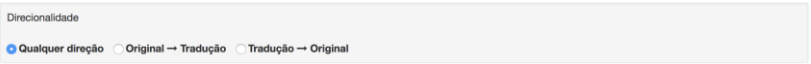


Figure 37. COPACONC Advanced Search third group of filters.

4.4 COPACONC Expert Search

The experimental COPACONC Expert Search[64] (Figure 38) is the first attempt to make this tool available for public use. It offers the

[64] https://copa-trad.ufsc.br/#copaconc-smart-search

direct utilization of Text Mining techniques, applied to a parallel corpus for translation investigation. The Expert Search works similarly to Simple Search; however, the filters are different, and most notably, this tool do not use Sphinx Search, since the searching features are different from the search algorithm provided by a traditional search engine. Due to the processing speed expended to solve complex statistical procedures, the only option available in COPACONC Expert Search is to choose books individually. There is no option to select all the books available in the corpus at the same time. If an attempt is made to select all the books at the same time, the loading time is consequently longer. Specific improvements to address this current limitation are expected in forthcoming versions.

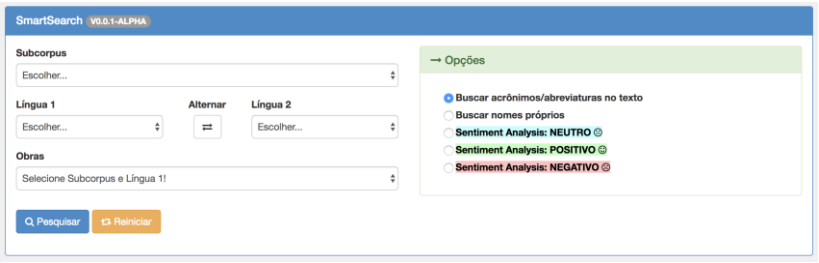


Figure 38. COPACONC Expert Search filters.

The filters available in this tool are the following ones: (a) acronyms/abbreviations; (b) proper names; (c) neutral sentences; (d) positive sentences; (e) negative sentences. The first two filters (a and b) were already described in the methodology chapter, and the last three are based on a Text Mining technique known as "Sentiment Analysis." For Sentiment Analysis filters, a visual resource is applied to indicate the estimated connotation[65] (i.e., the perception), using these colors: cyan for neutral, green for positive, and red for negative. The colors are still present in the final results. A point to keep in mind is that the sentiment analysis is carried out in relation to Language 1; because of that, it is possible to analyze whether or not the same connotation exists in Language 2. In addition, it is necessary to emphasize that the Sentiment Analysis is based on a training set of positive, negative, and neutral

[65] "An idea or meaning suggested by or associated with a word or thing; The set of associations implied by a word in addition to its literal meaning" (TheFreeDictionary.com, 2018).

examples and statistical approximations. Thus, the score reached so far provides for reasonable accuracy, but it is not always correct in its totality—i.e., some entries in the list may be defined with false positives.

Figure 39. COPACONC Expert Search neutral sentences in Language 1 (left column).

According to this message, Figure 39, it should be noted that the left column lists all the entries for Language 1 (original), and the right column lists the translated version. Also of note is that in the left column, all the sentences have the cyan color in the background, indicating the neutral option was selected. However, it should be noted as well, in the right column, that there is one sentence with a green background which was defined by the algorithm. The reason for this situation is explained by the presence of the word "success," which added extra weight to the positive score. Negative sentences in the translated version are identifiable, as well, in Figure 40. The presence of "infernal" may have a negative connotation in Portuguese.

Figure 40. COPACONC Expert Search: neutral sentences on the left and the presence of a negative sentence on the right.

Other possibilities are available too; for example, here is the same text, but now appearing with the negative filter selected. A snippet of the results is shown in Figure 41.

Figure 41. COPACONC Expert Search: negative sentences on the left and the presence of a positive sentences on the right.

In the message Figure 41 the presence of a positive sentence in the translated text is also evident. Through analysis, it is possible to conclude that the words do not have a negative connotation, whereas in the original text, the sentence was scored as negative, possibly because of the word "inconspicuous." The word may retain a neutral connotation, but through time, it has been perceived in a negative sense, as revealed by an in-depth investigation: Doing a more qualitative analysis related to the word "inconspicuous," some conclusions can be drawn. While checking the word in Google Books Ngram Viewer[66], it is noted that this word was utilized more often during the periods of war, especially World War I, World War II, and also in the Cold War (Figure 42).

[66] https://books.google.com/ngrams

Figure 42. Google Books Ngram Viewer research for the word "inconspicuous."

Additionally, checking the word in Google Trends[67] verifies that "inconspicuous" is mainly related to Word War II (Figure 43).

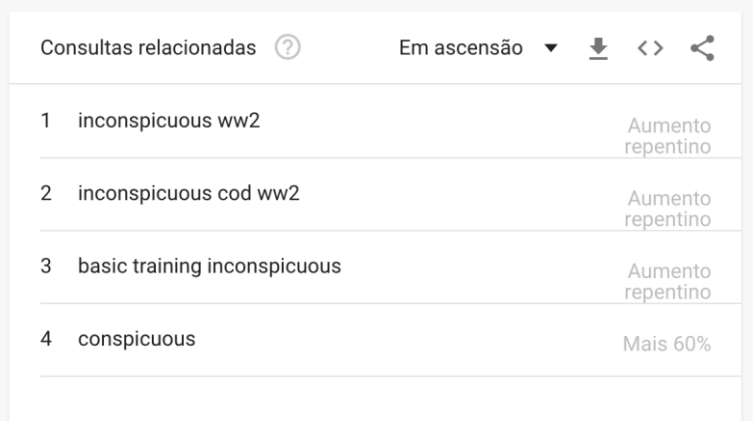


Figure 43. Google Trends worldwide research for the word "inconspicuous."

[67] https://trends.google.com/

By checking the word in a proper linguistic corpus, such as the British National Corpus[68] (BNC), it is noted that the word is more likely to occur in a negative context, as indicated in the initial KWIC results in Figure 44.

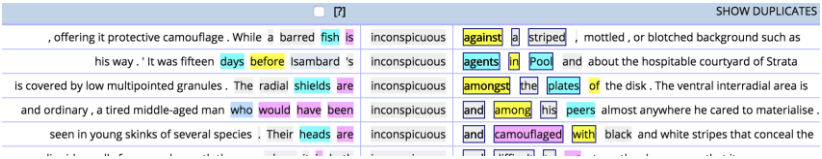


Figure 44. Results for "inconspicuous" in BNC.

Moreover, the discussed word and sentence are linguistic resources from the book entitled "Artemis Fowl: The Eternity Code," by Eoin Colfer; this is a book with a science-fiction plot in which the vocabulary was well selected by the author. However, for some reason, in the translated version, the appeal of the science-fiction scenario may have gotten lost in this passage, since the chosen word was "invisible," a word with a neutral profile. Of course, more comprehensive research is necessary to verify the negative connotation of the word in the original text, as well as the translational decisions (see discussion in Zethsen, 2006); but this brief investigation demonstrates how Text Mining techniques can expand the research possibilities in Corpus-based Translation Research.

4.5 WORDLIST

WORDLIST is a COPA-TRAD tool featuring a wide range creation of wordlists[69] based on some restrictions. These restrictions are related to the possibility of applying stopwords[70] lists for removal (see Appendix I), as well as defining the minimum word frequency. The types available, at the moment, are hapax legomena[71], general list of words,

[68] https://corpus.byu.edu/bnc/

[69] For Barlow (2004) "Probably the most radical transformation of a text used in linguistic analyses is to, in effect, rip it apart to produce a wordlist" (p. 207).

[70] The stopwords lists were created based in three different sources, mainly, high frequency words in COPA-TRAD, MySQL default stopwords list and finally the stopwords lists available for free on the website https://www.ranks.nl/stopwords. Depending on the language, some list may be longer than others but from time to time the stopwords list are updated in COPA-TRAD. Appendix I lists all the stopwords being used at the moment of this study.

[71] Single word in the whole text, that is any word with frequency equals to one.

acronyms list, proper noun list, 2-gram list, 3-gram list, 4-gram list, 5-gram list. All the lists contain frequency, which is a measure of number of word occurrences. It is also possible to order the list in ascending, descending, alphabetical and reverse alphabetical order.

There are two extra features for WORDLIST which are the possibility to show the POS (part-of-speech) as well as the lemma. These two extra features were created with the support of TreeTagger which was integrated to work seamlessly with COPA-TRAD platform. The available features are shown in Figure 45 alongside with COPA-TRAD default filters such as subcorpus, language and text selection. After selecting the desired options the user has to click on the "search" button and the table list is generated on screen. In case, the user desires/needs to check the stopwords in use to restrict the final list of words he/she can choose the language and then click on the button "Check Stopwords" a modal window appear and presents all the stopwords in XML format (in case the user wants to import the list or simply check what is included in the list).

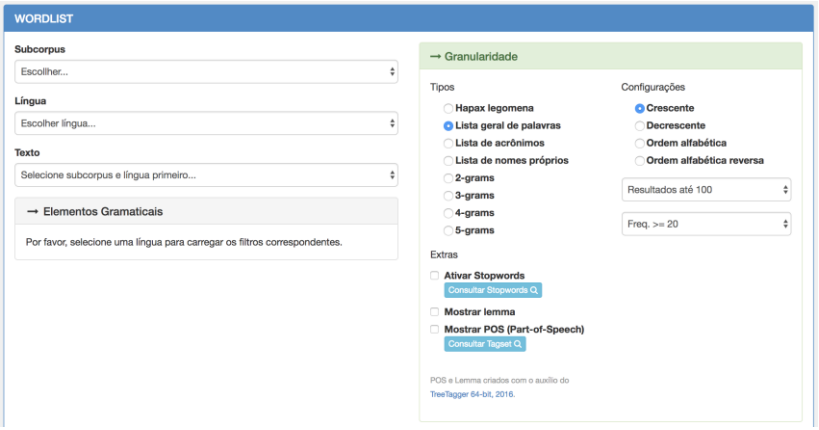


Figure 45. COPA-TRAD WORDLIST filter options.

Another interesting point to highlight is the part-of-speech feature, when a language is selected the "grammatical elements" box (below the text selection box–see Figure 45) loads with all the available labels. For example, when the English language is selected the filter options updates itself according to Figure 46, and indicated in red square number 1. In case none of the POS are selected WORDLIST generates a list with all the available elements and when one or more elements is/are selected specific lists based on these elements are generated. To add in

the list the selected POS it is necessary to check the option "Show POS," as highlighted in the red square number 2 in Figure 46. A key point in this POS feature is that depending on the selected language the POS is different, whether due to the language specific features and distinct intricacies or due to the dictionary/language model utilized by TreeTagger. By observing the elements in Figure 46 is possible to note the POS are defined in specific "codes" such as CC, CD, DT, JJ, RB, NNP, etc. These codes are the same defined in the language models and to know more about each one is possible to check the complete list by clicking on the button "Check Tagset."

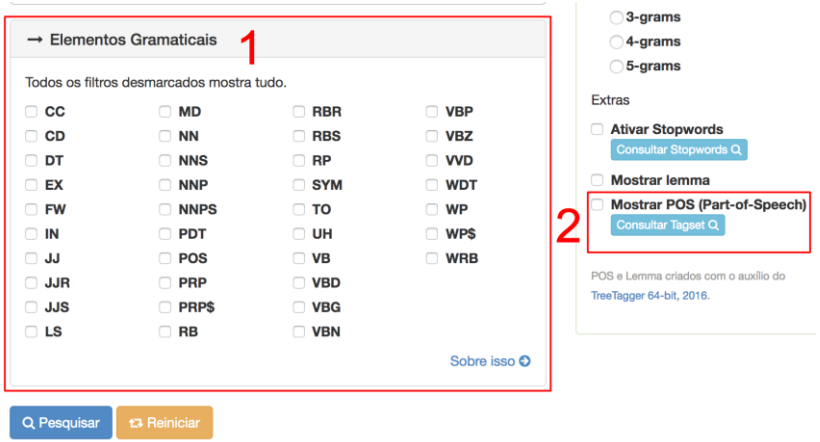


Figure 46. COPA-TRAD WORDLIST filter options when the English language is selected.

For the English language the tagset available were proposed by Penn Treebank[72] (Santorini, 1990) and the ones in use on COPA-TRAD are listed in Table 9. However, for the other COPA-TRAD languages the tagset are different, and based on other proposals like the models for Portuguese that were based on EAGLES[73] (Expert Advisory Group on Language Engineering Standards). Similar approach were used for French, German, Italian and Spanish.

[72] The complete list is available at: https://www.clips.uantwerpen.be/pages/mbsp-tags

[73] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Portuguese-Tagset.html

Table 9. Tagset for English Language.

Tagset for English Language	
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun (prolog version PRP-S)
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present

VVD	Verb, past tense
VBD	Verb, past tense
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun (prolog version WP-S)
WRB	Wh-adverb

To illustrate how results are generated in the final list, I selected the following options: (a) COPA-LIJ; (b) English Language; (c) Harry Potter and the Chamber of Secrets; (d) general list of words; (e) Ascending order; (f) Frequency more than 30; (g) Show POS; (h) JJ – Adjective; (i) JJR – Adjective, comparative; (j) JJS – Adjective, superlative. Part of the first results are the ones listed in Figure 47.

	Palavra	POS	Frequência
1	several	JJ	31
2	best	JJS	32
3	common	JJ	34
4	full	JJ	34
5	able	JJ	35
6	past	JJ	35
7	second	JJ	36
8	loud	JJ	36
9	many	JJ	37
10	headless	JJ	38
11	dear	JJ	38
12	enough	JJ	40
13	new	JJ	41
14	few	JJ	44
15	whole	JJ	45
16	hard	JJ	47
17	sure	JJ	48
18	inside	JJ	48
19	own	JJ	49
20	small	JJ	51
21	old	JJ	53
22	black	JJ	54
23	magic	JJ	54
24	little	JJ	56
25	first	JJ	65
26	open	JJ	69
27	other	JJ	72
28	good	JJ	76

Figure 47. COPA-TRAD WORDLIST results for three different types of adjectives from the book "Harry Potter and the Chamber of Secrets."

Similar lists are possible to generate for all the texts available in COPA-TRAD corpus. In case the user wants to import the final list, he/she can print, Export in CSV, Export in XML and Export in PDF.

4.6 TREETAGGER CLOUD

One of the objectives when COPA-TRAD was launched was to provide tools for use in classroom. With this objective in mind, one of the tools available since Version 1 is the CORPUS-BUILDER which is a do it yourself tool for creating disposable corpus on the fly. In other words, the user can paste his text and analyze it through the procedures available. This tool has a didactic appeal because there is no need to create an account, send the text, wait the approval to and the processing and availability to investigate it. There is no need to wait 24 to 48 hours, everything happens in real time.

By having in mind the concept of do-it-yourself and real time for COPA-TRAD Version 2, a new tool was made available in this important category, which is TREETAGGER CLOUD (Figure 48). This new tool, enables the researchers/students to process any text through TreeTagger technology, without the obligation and the whole process of submitting a text to COPA-TRAD. In addition, this tool is suitable for texts not in COPA-TRAD domain and genre which may not be part of COPA-TRAD corpus. The utilization of CORPUS-BUILDER and TREETAGGER CLOUD open new possibilities of investigation.

TREETAGGER CLOUD works similarly with the POS option in WORDLIST, but it is available for any kind of text. According to Figure 48, the number one area (highlighted in red) is the place where the user can paste his text under analysis. Then the number two is the area where the final processed text is shown. In number three the user has to select the correct language of text to load the correct linguistic models, and in number four the user can choose the output format, that can be: (a) List; (b) CSV; (c) XML. The last two ones are more suitable to import in other translation software such as CasualConc, AntConc, etc. The utilization of the tool is straightforward, to illustrate how it works, I selected the text already used to text all the tools for COPA-TRAD Version 2, which is "*The Lord the Rings: The Fellowship the Ring.*"

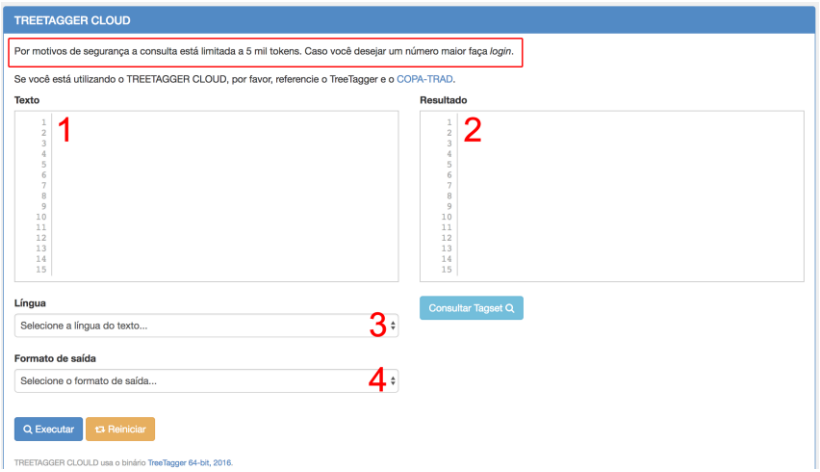


Figure 48. COPA-TRAD TREETAGGER CLOUD tool.

The first thing to do, is to copy the entire text or the part under analysis and paste it in the field "text." Then select the English language, the output format, in this case, the "list" one and finally click on the blue button "run." The output will appear in the right box "result" as presented in Figure 49. A point to have in mind is that the user have to be logged-in to send texts with more than 5000 characters. This procedure was adopted for security reasons since TREETAGGER CLOUD consumes a lot of processing power and some bad intentioned users may want to send huge text to overload COPA-TRAD infrastructure. However, for many translational tasks 5000 tokens is an acceptable number and the user can process in parts in case he/she does not want to log-in. After the processing, the user can copy the final result and use the data in any other tools he decides is applicable to his/her research project.

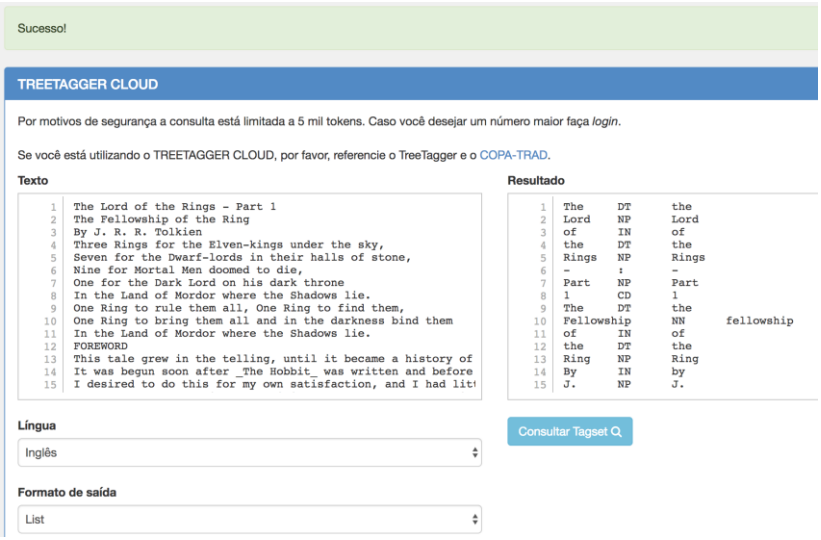


Figure 49. COPA-TRAD TREETAGGER CLOUD in action POS tagging the text “The Lord the Rings: The Fellowship the Ring” in real time.

Just to illustrate the possible use of the resulting POS list generated in TREETAGGER CLOUD. After copying the list to a TXT file and opening it on CasualConc the user can click in the “concord” option and ask the software to find all “JJ” occurrences (Figure 50). All the adjectives are listed for analysis. By doing similar investigations, it is possible to analyze the linguistic features as listed in Table 9.

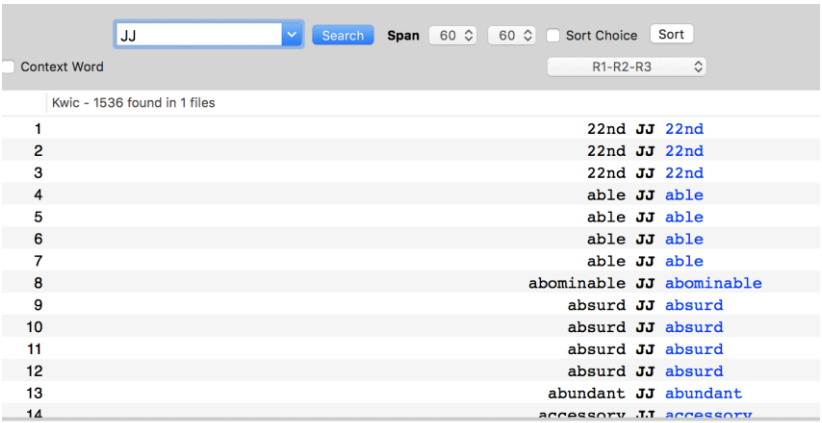


Figure 50. Results from TREETAGGER CLOUD for the text "The Lord the Rings: The Fellowship the Ring" under analysis in CasualConc.

4.7 COPA STATS

COPA STATS is a tool available since the first version of COPA-TRAD (Silva, 2013, p. 110). Due to this reason, just the new features are going to be dealt in this section. There are two new features in COPA STATS, the first one is the Zipf's Law Graph (in logarithmic scale), discussed with a micro-analysis in this section, and the second one is the COPA-TRAD texts detailed information (Appendix H) not discussed here, but Section 3.3.6 provided information about the implementation of it.

The theory and the purpose, methodological aspects behind Zipf's Law and how it was implemented was already discussed in Section 3.3.6. For this reason, discussion here is related to what is available in COPA-TRAD. However, an explanation of how Zipf's law can be applied to text and why is it useful is expressly necessary, since the Zipfian distribution is omnipresent in language and describes word behavior in texts (Quasthoff, Richter & Biemann, 2006, p. 1801; Manning & Schütze, 1999, p. 544). In addition the Zipfian distribution is useful as a measure for text use of words richness (Manning & Schütze, 1999, p. 25). As discussed, the position of a word in a list is defined by its frequency. Manning and Schütze (1999) demonstrate that "if we count up how often each word (type) of a language occurs in a large corpus, and then list the words in order of their frequency of occurrence, we can explore the relationship between the frequency of a word *f* and its position in the list,

known as its rank *r*" (p. 23). This linguistic phenomena is much more perceivable when plotted in a graph, as the one available in COPA STATS.

For the graph, Manning and Schütze (1999) say "Zipf's law predicts that this graph should be a straight line with slope -1," this ideal straight line is plotted in COPA STATS graphs as a reference and it is indicated with the blue color. However, in real life examples, the straight line is not always possible and seems in bad fit for low and high ranks: "there a few very common words, a middling number of medium frequency words, and many low frequency words" (Manning and Schütze, 1999, pp. 24-25). This linguistic phenomena was also perceived by Hovy (2003) who states that "a few words occur very often, fewer words occur somewhat often, and many words occur infrequently" (p. 586)—See Figure 51. The occurrence of the aforementioned linguistic phenomena, is perceivable in COPA-TRAD texts as the ones discussed in this section.

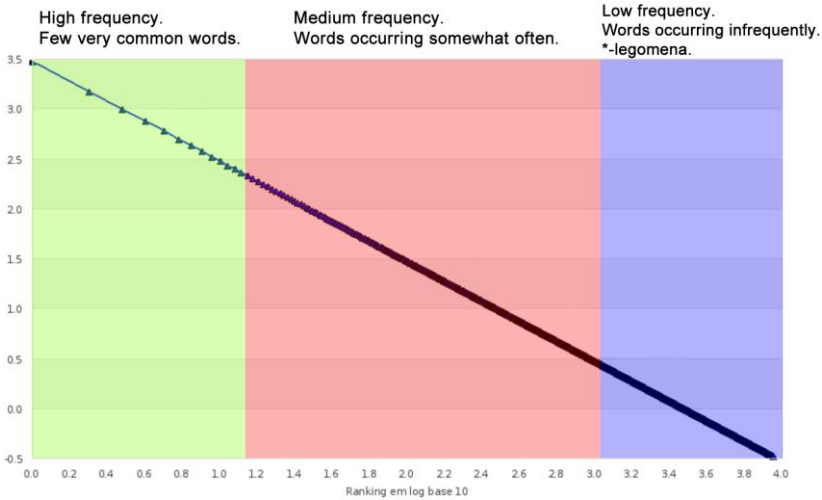


Figure 51. Representation of words incidence in a generic Zipf's law graph generated by COPA-TRAD.

The practical examples, are from COPA-TRAD as well as the graphs. Two different examples are going to be discussed:

- Harry Potter and the Chamber of Secrets (English – UK) « » *Harry Potter e a Câmara Secreta* (Portuguese – Brazil);

- Artemis Fowl The Eternity Code (English – Ireland) « » *Artemis Fowl O Código Eterno* (Portuguese – Brazil).
- Artemis Fowl The Eternity Code (English – Ireland) « » *Artemis Fowl: El Cubo B* (Spanish – Spain).

As observed in the aforementioned section, the graphs generated by COPA-TRAD may support the investigation of vocabulary density between translators and authors, translators and translators, authors and authors, performance, as well as stylistic decisions, language use, and so on. This analysis may provide possible clues indicating such behavior which can lead for a qualitative research investigating what causes lead for actual textual content, if it was intentional or induced by publishing house editors, etc. Figure 52 presents the first graph for the book "Harry Potter and the Chamber of Secrets," the source text/original. According to these graphs, as observed in Figure 52, the extremities are identified according to the phenomena stated by Manning and Schütze (1999) as well as Hovy (2003)—the impossibility of a perfect straight line for low and high ranks. The high incidence of common words is perhaps more probably to occur in all provided examples due to the target audience of these texts who are basically children and adolescents. However it is undeniable the incidence of hapax legomena (such as dis legomenon, tris legomenon and tetrakis legomenon).

The blue line is the ideal curve according to Zipf's law and its presence is used for comparison reasons. The most interesting curve is the red line since it is the actual one generated from "Harry Potter and the Chamber of Secrets" text, the medium frequency words occur in a more patterned way according to the theory as well the head and tail of the red line. This phenomena visually perceivable in the graph was noticed by Scott and Tribble (2006), while working with wordlists because "they have a small number of high frequency items at the head and an enormous tail of hapax legomena (words which occur once only in a corpus)" (p. 11).

By checking the COPA-TRAD tool WORDLIST, it is possible to check these words, for example, the head the most common words are "harry, a, to, and, the" and for the tail, mostly constituted of hapax legomena a total of 3,152 were found such as "hooting, sweetums, smeltings, purple-faced, spellwork, peskipiksi" (words randomly selected). Now moving to the same text, but the official translation in Brazilian Portuguese language by Lia Wyler.

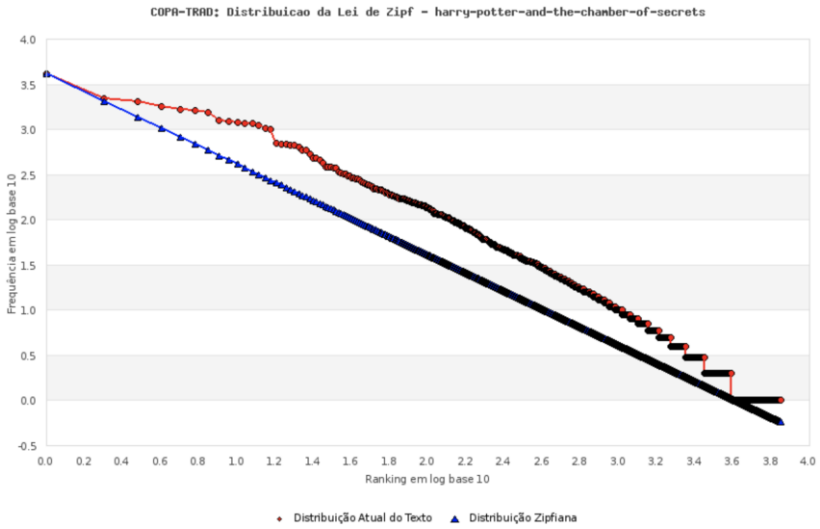


Figure 52. Graph of Zipf's distribution in logarithmic scale for "Harry Potter and the Chamber of Secrets."

Then, Figure 53, presents the graph for the book "*Harry Potter e a Câmara Secreta*" the target text/translation. It is possible to perceive some similarity with the English version, however the head bulge in the middle, then the medium frequency words are almost a straight line and the tail has an incidence of hapax legomena as well. While verifying the text in WORDLIST tool, it is possible to analyze the words compounding the head such as "*de, a, que, o, e, harry*" and for the tail "*inteligência, censura, passasse, instintivamente, cáqui*" (words randomly selected). Now, moving to another book already available in COPA-TRAD another approach is applied to data of same nature.

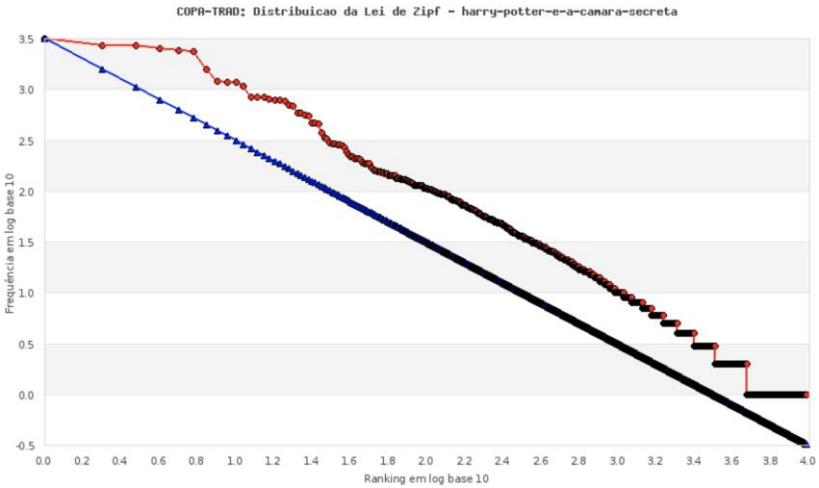


Figure 53. Graph of Zipf's distribution for "Harry Potter e a Câmara Secreta."

Just for the sake of comparison between authors and translators now follow the graphs for the books "Artemis Fowl The Eternity Code" (original – English) and "*Artemis Fowl O Código Eterno*" (translation – Portuguese). In addition, I added a second translation in Spanish to compare different translators of the same book. Figure 54, is the graph for the original book in English. The head almost follow the Zipf's law with a slight bulge, the medium frequency words are more straight and the tail with a high incidence of hapaxes. Words constituting the head are, for example, "the, a, to, of, and, was, in, you, I" and the hapax in the tail are "prologue, puzzled, millennium, weaponry, khaki" among others.

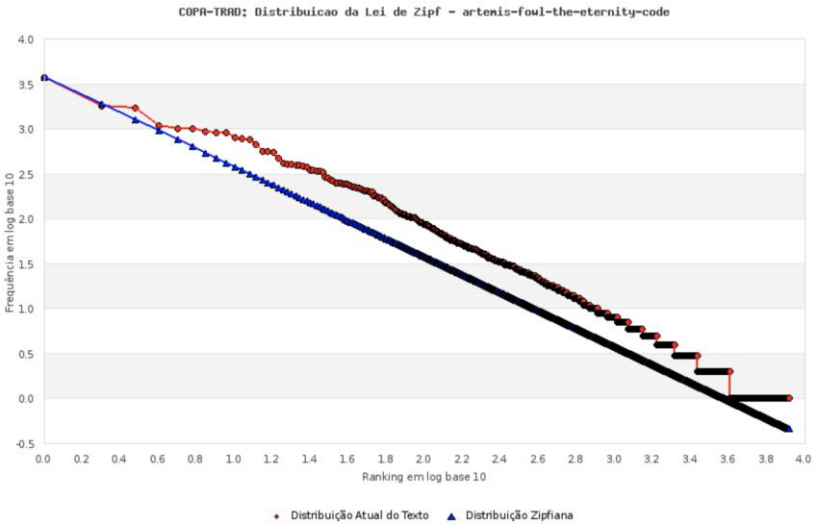


Figure 54. Graph of Zipf's distribution for "Artemis Fowl The Eternity Code" (original – English).

For Brazilian Portuguese (Figure 55), the graph below represents the Zipf's distribution for the book "*Artemis Fowl O Código Eterno*," translated by Alves Calado. Words are more distributed in the head with higher frequency, if compared to English, the medium frequency seems standardized across languages and the tail is possible to note the hapax legomena incidence. The most common words in head are "*o, que, não, um, para, se, com, eu*" and for the tail "*prólogo, descrever, tentaram, fracassaram, omoplatas, estreitava*" among others. Now moving to Spanish another translation for the same book.

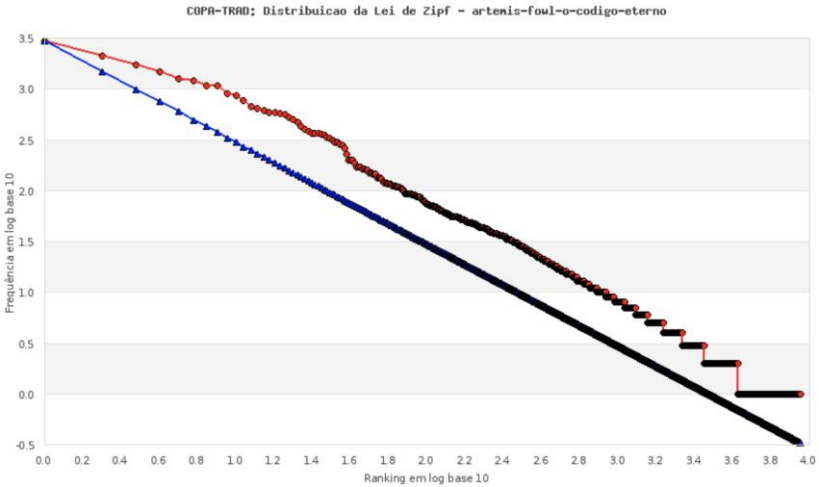


Figure 55. Graph of Zipf's distribution for "Artemis Fowl O Código Eterno" (translation – Brazilian Portuguese.)

Spanish version of "Artemis Fowl: El Cubo B" translated by Ana Alcaina the medium frequency words are less used and the head, the most common words follow a similar pattern like then English version, possibly this translation is much more closer to the original. The most common words in the head are "*de, la, que, el, a, en, y, un, no, se*" and for the tail, mostly constituted of hapax, there are words such as "*perversos, tonta, cursi, lío, dan, seda, destrozado, estirándose, blancanieves*" and others.

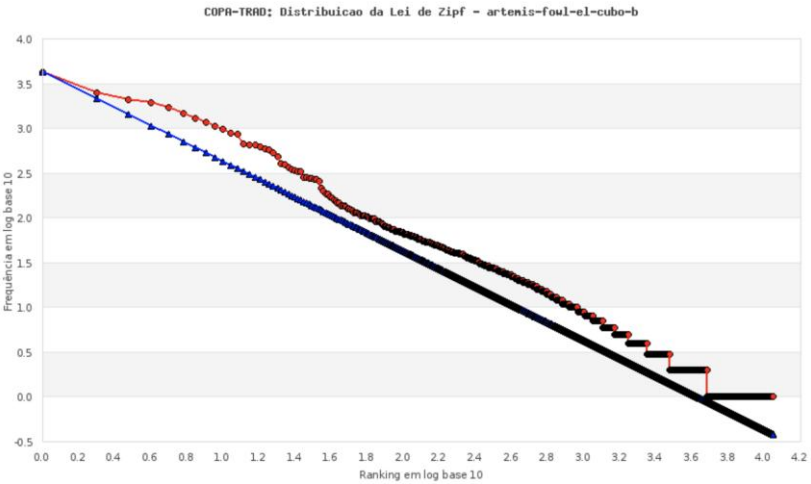


Figure 56. Graph of Zipf's distribution for "Artemis Fowl: El Cubo B" (translation – Spanish).

The three last graphs from Artemis Fowl book were juxtaposed for comparison in Figure 57. The red line represents the original book in English, the green one for Brazilian Portuguese version and the yellow line for the Spanish version. A more comprehensive analysis is necessary but here this micro-analysis showed the potential of this tool as observed by Scott and Tribble (2006) "regularities which can be useful to language researchers searching for patterns of importance in their own text corpora" and "word-lists offer an ideal starting point for the understanding of a text in terms of its lexis" (p. 31). According to Zipf's law graphs it is possible to note wordlists obey mathematical rules and re-arranging them in different types is possible to discover distinct patterns (Scott & Tribble, 2006, p. 31).

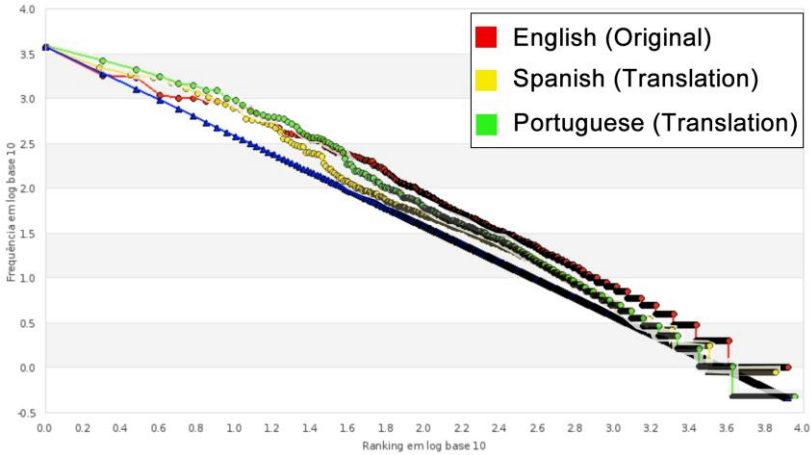


Figure 57. Juxtaposition of the discussed three graphs for Artemis Fowl The Eternity Code.

Preliminary analysis of this relatively small sample of investigation, indicates worthwhile outcomes for further research. The data to investigate the topic are easily accessible at COPA-TRAD platform by means of the COPA STATS tool.

4.8 Closing Remarks

This chapter has offered a real case scenario by demonstrating the application of Corpus Mining model on COPA-TRAD Version 2. First of all, a brief section discussed and introduced the chapter, then AUTO ALIGNER and its fourth steps were reviewed. The next tool developed for COPA-TRAD Version 2 is the COPACONC Advanced Search a discussion of how the user can take advantage of many filters available was done. After this discussion another auxiliary tool from COPACONC was presented, COPACONC Expert Search, this tool utilizes Text Mining techniques directly to discover useful information from parallel corpus. Another tool discussed was the WORDLIST which is a powerful resource in COPA-TRAD since it is possible to generated wordlists from many different forms. The TREETAGGER CLOUD, one of the same category of CORPUS-BUILDER, that is, do-it-yourself tools is an important resource for those who need POS (part-of-speech) processing, but do not know how to utilize, especially important for use in classroom since they are 24 hours available for use. Finally, the last tool review was COPA STATS which utilize several different

quantitative approaches to provide means to investigate translated text. At this part a micro analysis based on Zipf's law was presented and possibly can be expanded in future research. Now the next chapter, presents the concluding remarks of this study.

5 CHAPTER FIVE: Concluding Remarks

5.1 Initial Remarks

This chapter summarizes the final results and outcomes related to this study. A brief summary of each chapter is presented, and then the motivations for this study are briefly discussed. Next, the research questions are revisited and answered based on the results obtained from both the analysis and the Corpus Mining model delineated in Chapter 3—Methodology. In the following section, limitations and suggestions for further research are enumerated as possible future subjects of investigation. When researching an interdisciplinary field, there is no possibility of doing a comprehensive scientific research study of all desired subjects. However, this study sheds light on the major findings that may open new avenues of investigation into Corpus-based Translation Studies. Due to the complexity and the techniques utilized, each COPA-TRAD tool will need future research (see Section 5.5). Finally, in order to exemplify the application of advanced Text Mining and Natural Language techniques, I share the already completed plans and prototypes that will be part of COPA-TRAD Version 2[74].

5.2 Recap of the Study

This section recaps the present study, which is divided into four parts. A brief summary of each chapter (except this chapter) is presented in the following four subsections.

5.2.1 Summary of Chapter One: Introduction

Chapter One, the introduction, sets out the context of investigation, dealing with a discussion of the current technological scenario and the wealth of data available for investigation—even for Translation Studies. In the context of the investigation section, another key area of focus is an overview on COPA-TRAD, the tool utilized to present the practical applicability of Corpus Mining model. The problem and its current scenario is covered in a long section in the introductory chapter that discusses most of the problems I found while conducting research on CTS. Then these problems were organized in an Ishikawa diagram, as shown in the Figure 4.

---

[74] Advancements in COPA-TRAD are documented on https://copa-trad.ufsc.br/#copa-trad-changelog

Based on the problems found and the current scenario, a solution for key problems is proposed and introduced—Corpus Mining. An overview of the initial ideas is presented, along with an explanation showing why the name of "Corpus Mining" was chosen. In an interdisciplinary research study that combines text mining and linguistic processing technology with corpus, the name amalgamates the whole concept. After the presentation of Corpus Mining, the scope of the study is outlined by its major elements: (a) statement of the problem; (b) purpose of the study and related research questions; (c) thesis; (d) research method. A discussion of the data present in parallel corpus introduces the reader to the kind of data this study involves. Finally, the organization of the study is delineated through the presentation of the topics addressed in this research.

5.2.2 Summary of Chapter Two: Theoretical Background

Chapter Two, entitled "Theoretical Background," deals with the discussion of the theory and definition of the topics adopted in this study, as the title suggests. The initial remarks present the organization of the chapter, separated into two main categories: (a) The Linguistic Background; (b) The Technical Background.

The Linguistic Background addresses the main area of this study, which is Corpus Linguistics, then narrowing down the next section, which deals with Corpus-based Translation Studies. Next, two distinct approaches adopted in scientific research are presented, quantitative and qualitative approach. Rather than presenting these as opposing factors, the focus of this study combines both approaches.

The Technical Background discusses the theory as it relates to the technical subjects of this study, and special attention is devoted to ensuring that the discussion is easy to understand, even for people without a technological background. The first topic considered in this second section is the interrelationship between Knowledge Discovery in Databases, Data Mining, and CRISP-DM, since these subjects work together in this study, from a Corpus Mining perspective. Then a special section is dedicated to CRISP-DM model since it is the most well-known process for the Data Mining application. Next, the major technological subject of this study is set out, which is the theoretical foundations of text mining, listing citations from seminal papers by the first researchers involved in this area; they coined the terms "Text Mining" and "Knowledge Discovery in Text"—also known as KDT. Next, Duo Mining, a technique related to Data Mining and Text Mining, is presented. It lays out possibilities for how mining techniques can be combined and

how each part interacts. Another important subject discussed is Natural Language Processing, considered the father of all automatic linguistic processing.

Finally, as an expansion of the initial ideas discussed in the 1.6 section: "The Nature of Parallel Corpus Data," the last theoretical section, presents the Bitext Alignment—The Dataset for Corpus Mining along with the theory behind the topic of bitext, as well as the major seminal papers that were referenced. Following the standard structure of this study, the chapter closing remarks are presented.

5.2.3 Summary of Chapter Three: Procedures

The third chapter, "Procedures," is the major chapter of discussion and detailed presentation of the assumptions related to Corpus Mining model. Technical discussions are offered as a way to show how the implementation of Natural Language Processing and Text Mining techniques function in a real-world situation. These discussions are inserted as part of a major step-by-step guide to a Corpus Mining model. Before delving into the compounding phases of Corpus Mining, the basis for the proposed methodology is established. First of all, the Experimental Setting section presents the technological instruments utilized to conduct the technical procedures. The Systematization of Corpus Mining Model section discusses the methodology suggested here, and then each part of it is discussed. At this point, it is worth presenting, once again, the phases of Corpus Mining model as depicted in Figure 58.

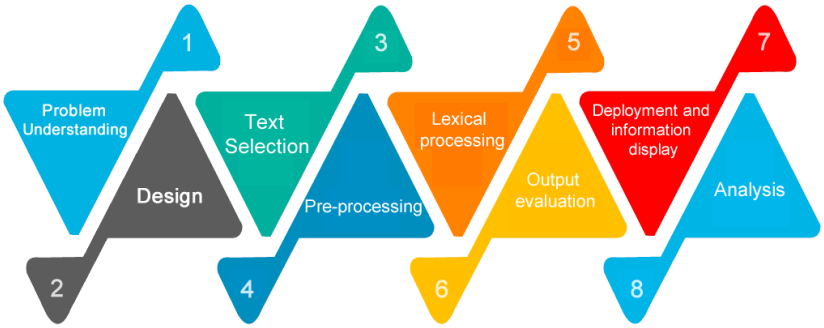


Figure 58. Phases of Corpus Mining model.

After the description and explanation of the phases, another section on the methodological chapter deals with a practical example from the COPA-TRAD Platform: the technical issues involved in the development of the AUTO ALIGNER tool.

5.2.4 Summary of Chapter Four: Applying Corpus Mining on COPA-TRAD Version 2

Chapter Four discuss the major tools of COPA-TRAD Version 2 according to Corpus Mining Phase Eight. After the initial remarks, the AUTO ALIGNER, COPA-TRAD Automatic Alignment Tool section discusses the tool not from a technical level, but from a final user point of view. To do so, the aforementioned section is divided in four subsections, based on the number of steps needed to process and align a text in AUTO ALIGNER. Next, two recent new tools, part of COPACONC suite, are discussed. The first one discussed is COPACONC Advanced Search, where the sections are organized according to the filters type. The second tool discussed is COPACONC Expert Search, a tool with a special display mechanism (i.e., the use of colors in the results) which applies Text Mining techniques in a directly. In Expert Search, visual resources are utilized in order to provide new concepts regarding presenting bitext for analysis.

COPA-TRAD Version 2 included important updates and new tools, and three more tools are analyzed. The tool known as WORDLIST has been available since COPA-TRAD Version 1. However, it received important updates and new features, such as the adoption of TreeTagger to identify POS (part-of-speech), and the creation of lists according to selected linguistic features in six different languages, namely: Portuguese, English, Spanish, German, Italian, and French. The other tool discussed is TREETAGGER CLOUD, a tool in the same category as CORPUS-BUILDER, indicated as useful in the classroom. Like WORDLIST, the utilization of TreeTagger to process and label texts on the fly is applied. Tools to process and build corpus and wordlists on the fly are categorized in this study as do-it-yourself.

5.3 Research Questions Revisited

In section 1.5, the purpose of the study and the related research questions are presented. Now these questions are revisited, the answers to them provided in previous chapters and then summarized here, as follows:

How can CTS and Text Mining be combined to improve a large data analysis of translated texts? The first question pertains to the combination of Corpus-based Translation Studies (CTS) and Text Mining, examining how this combination can support the analysis of large quantities of data. As discussed in the methodology (Chapter 3), as well as the analysis (Chapter 4) Corpus Mining is a model that lays out the

steps traveled from the initial idea to the final results. The use of CTS techniques and the Text Mining automatic patterns identification and extraction is a fruitful way to harvest large quantities of bitext in due time.

What kinds of tools can be created from a Corpus Mining perspective? The second question is related to the possible tools that can be created through Corpus Mining perspective, and it was mostly answered in the methodology chapter, as well (Chapter 3). The phases of Corpus Mining can lead to final decisions and clues that indicate the most suitable procedures required to analyze the data. As an example, the practical experiments were conducted in COPA-TRAD, which led to Version 2 of the platform. As presented, tools such as AUTO ALIGNER, COPACONC (Advanced Search and Expert Search), the WORDLIST, and COPA STATS are some possibilities. However, other tools can be proposed, as well, and some discussion related to this subject is provided in the section on suggestions for future research.

What are the advantages of Corpus Mining for Corpus-based study? The third question looks into the advantages of using Corpus Mining for CTS. Based on the problems outlined in section 1.3, Corpus Mining proposes a solution for the preparation, processing, and analysis of CTS. The advantages pertain not just to the creation of parallel corpus, but also to a new tool to support the investigation. A new way to look at data may indicate possible new hypotheses and variables not thought of before. A plethora of information can be gleaned for further research to be conducted in a quantitative way, as well as a qualitative way.

How can Corpus Mining surpass textual annotation in source texts and target texts? Why is that important? How can it be done? The fourth question is more technical, but I will try to be specific and clear about my answer. Corpus mining provides a set of guidelines as well as a practical example, based on COPA-TRAD Version 2. As delineated in Chapters Three and Four, it is possible to process and analyze without textual annotation, mainly because of the use of advanced algorithms with the support of predefined models and dictionaries/training sets. These predefined materials are utilized as input in order to teach a processing system how to identify possible patterns—that is, how to find the most likely match and provide interesting patterns to the user. This is important because the user will have access to certain types of data not previously possible, as well as saving time in the process. The ways that such procedures can be performed are extensively discussed in the methodology chapter—and as I have already said in this paragraph,

through the use of models and training sets, as well as through the use of probabilistic algorithms specially designed to process bitext.

5.4 Scientific Contributions and Findings

The findings are preliminary, but the present study raises significance and relevance that is twofold. The existence of this twofold nature is related to the reason this study presented a solid theoretical background for conducting research in CTS, as well as contributing to a practical perspective supported by Corpus Mining model. This perspective enables the application of advancements in the area of automatic text processing and understanding through a concise process that is organized into well-defined phases. In addition, the proposed model provides support for researchers interested in conducting investigation in parallel corpora, since a model is more repeatable and manageable. The use of texts in electronic format, *per se*, provides possibilities for discovering new evidence; but processing and analyzing an electronic text through the proposed techniques, especially the ones used by Text Mining, undoubtedly presents the possibility of harvesting information hidden from human eyes. As observed by Sinclair (2004a):

> Substantial collections of language texts in electronic form have been available to scholars for almost forty years, and they offer a view of language structure that has not been available before. While much of it confirms and deepens our knowledge of the way language works, there is also a fascinating area of novelty and unexpectedness – ways of making meaning that have not previously been taken seriously. Further, in studying corpora we observe a stream of creative energy that is awesome in its wide applicability, its subtlety and its flexibility (p. 1).

As already stated, the use of Text Mining techniques can contribute to what Sinclair (2004a) observed in the citation above by offering a view of language not thought of before, while providing "novelty and unexpectedness." In addition, Corpus Mining is expected to have a wide applicability that is not discussed in this study but is possibly a subject to be investigated in further studies and expanded. (For example, here, the focus is on parallel corpus, but the application of Corpus Mining can be beneficial in other types of corpus such as the monolingual ones.)

Additionally, this study contributes to a methodological perspective, since it offers a set of easy-to-use guidelines from Corpus Mining. Moreover, the tools developed for COPA-TRAD Version 2 supported the empirical investigation of translation practices associated with specific linguistic elements, patterns, and other translational phenomenas. As a result, Corpus Mining can be viewed as a valuable set of methodological procedures for use by novice researchers who are learning how to carry out descriptive research in Translation Studies, as well as being of value to advanced scholars in the field.

All in all, Version 2 of COPA-TRAD, developed with the support of Corpus Mining and discussed in this study, can be an excellent resource for professional translators, translation students, and Translation Studies researchers who are focusing on translational phenomenas such as those found in children's literature (since this genre is the main focus in COPA-TRAD). In terms of implications, the applicability of this online parallel corpus consists, for example, of translation research and pedagogy. The present research is also expected to benefit translators, as well as under- and postgraduate Translation Studies students. It can improve and develop their awareness about what possibilities are available for conducting an investigation into CTS through utilization of the Corpus Mining model.

5.5 Limitations of the Study and Suggestion for Future Research

It is an undeniable fact that all scientific endeavor comes with a package of limitations—and even problems. The reason behind this fact is complex and thus hard to explain, but some of the issues that arose from this study are discussed here. Due to time, technological limitation, and personnel involved, some tools will need to wait to be explored in future work. The Corpus Mining model was tested only in COPA-TRAD Version 2. In addition, all the processing techniques, the results obtained, and the generic nature of the model indicate that it can be adopted in other projects, as well as different types of corpus, but these possibilities have not been tested, so far. The model has been designed to be generic, but refinements are needed. The interdisciplinary nature of this study involves the investigation and theoretical discussion of not only linguistic/translational theory, but also, technological theory.

Topics have been organized and separated to preserve the clarity and coherence of the text; but some topics, as well as natural language processing techniques, could not be explored in a systematic way. Deciding what to include and exclude in this study has been a complex

endeavor, but this procedure had to be followed due to constraints by various sources such as technical considerations, time, etc.

Further investigation can be done on the model provided by Corpus Mining, as well as the tools developed for COPA-TRAD Version 2. For the model, the limitations raised in the previous paragraph are a good starting point. Besides Text Mining other technology to investigate bitext, the investigation of the model can be added to the list; CTS is also deserving of serious investigation—examples being the use of Machine Learning, as well as the free API provided by IBM Watson.

Regarding COPA-TRAD, suggestions for future work are very extensive. However, suggestions here are delimited to the topics addressed by this study. The tool discussed is AUTO ALIGNER. Adding the possibility of creating a custom dictionaries/training set based on the input texts is an important feature, since it can increase the quality of the final bitext. To do so, the steps could include the following: (a) text tokenization (original and translation); (b) extraction of less frequent keywords; (c) automatic translation of the extracted keywords in Google Translate environment (or any other machine translation software or API—Yandex, Watson, etc.); (d) creation of the dictionary from source language to the target language; (e) availability of the dictionary to Hunalign software.

In addition, AUTO ALIGNER is not identifying originals and translations automatically; this feature could be very interesting to investigate. The language checker already present is an important feature to identify text language, and it possibly could be used in a specific part of a large process of identifying originals and translations. The possibility of using monitor corpus (of original texts in COPA-TRAD-supported languages) could be an asset used to compare with the text and count, then determining, in the final estimation, whether the text is a translation or not.

I made some effort to prepare a monitor corpus for Brazilian Portuguese (see Appendix J) in order to conduct some prototypes, but it is necessary to do the same for other COPA-TRAD languages (English, Italian, Spanish, French, German.) In addition, another suggestion is to investigate Baker (2007) distinctive features of translated text, such as explicitation, simplification, normalization and standardization, as well as more palpable characteristics, such as lower type-token ratio and the overall text size (p. 12). Researching translational features such as the one mentioned, in order to find out which one is possible to convert algorithmically for the development of a system capable of identifying, with approximations, which one is the original or the translation—this

would be an interesting topic. Another suggestion for AUTO ALIGNER is to integrate it in the submission text panel so that the user can align his text and send it to COPA-TRAD; besides doing this, the possibility of keeping it saved on the cloud for further investigation is an interesting feature to add, as well. Such features, will provide much easier ways to the user to contribute for COPA-TRAD's ever-growing database.

BIBLIOGRAPHY

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). Amsterdam: John Benjamins.

Baker, M. (1995). Corpora in translation studies: An overview and suggestions for future research. *Target*, 7(2), 223-243.

Baker, M. (2000a). Towards a methodology for investigating the style of a literary translator. *Target*, 12(2), 241-266.

Baker, M. (2000b). Linguistic perspectives on translation. In France, Peter (Ed.). *The Oxford Guide to Literature in English Translation* (pp. 20-26). Oxford: Oxford University Press.

Baker, M. (2004). A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics* 9(2), 167-193.

Baker, M. (2006). Translation and conflict: a narrative account. London: Routledge.

Baker, M. (2007). Patterns of idiomaticity in translated vs. non-translated text. *Belgian Journal of Linguistics*, 21(1), 11-21. DOI: 10.1075/bjl.21.02bak

Baker, M., & Saldanha, G. (2009). Routledge encyclopedia of translation studies (2nd ed.). London/New York: Routledge.

Barlow, M. (2004). Software for corpus access and analysis. In Sinclair, J. McH. (ed.), *How to Use Corpora in Language Teaching* (pp. 205-221). Amsterdam: John Benjamins.

Bender, E. M. (2013). Linguistic fundamentals for natural language processing: 100 Essentials from morphology and syntax. In Graeme Hirst (Ed.), *Synthesis Lectures on Human Language Technologies* (Vol. 20, pp. xvii-166). San Rafael: Morgan and Claypool Publishers.

Bhatia, A., Kumar, A., Kumar, V., & Khan, I. (2011). Analysis of pattern recognition (text mining) with web crawler. In *International Transactions in Applied Sciences*, 3(3), 435-450. ISSN: 0975-3761.

Brown, P., Lai, J., & Mercer, R. L. (1991). Aligning sentences in parallel corpora. *29th Annual Meeting of the Association for Computational Linguistics* (169-176), Berkeley, California.

Brown, D., Tiberius, C., Chumakina, M., Corbett, G., & Krasovitsky, A. (2009). Databases designed for investigating specific phenomena. In Everaert, M., Musgrave, S., Dimitriadis, A. (Eds.), *The Use of Databases in Cross-Linguistic Studies*, (pp. 117-156). Berlin: Mouton de Gruyter.

Caseli, H. M., & Nunes, M. G. V. (2004). Corpus paralelo e corpus paralelo alinhado: propriedades e aplicações. In *Estudos Linguísticos*, 33, 1-6.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 step-by-step data mining guide (SPSS). The CRISP-DM consortium.

Charniak, E. (1997). Statistical techniques for natural language parsing. *AI Magazine*, 18(4), pp. 33-44.

Chary, N. C., Reddy, K. B., & Bhuahan, P. V. (2018). Duo-Mining techniques in knowledge discovery process in data base. In *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)*, 3(1), 1337-1343. Retrieved from http://ijsrcseit.com/CSEIT1831123

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of Annual Meeting Of The Association For Computational Linguistics* (pp. 9-16).

Chunyu, K., & Jian-Yun, N. (2015). Information retrieval and text mining. In C. Sin-wai (Ed.) *The Routledge Encyclopedia of Translation Technology* (pp. 494-535). London/New York: Routledge.

Church, K., & Gale, W. (1991). Concordances for parallel text. *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research* (pp. 40-62).

Čulo, O., Hansen-Schirra, S., & Neumann, S. (2011). Parallel corpora: annotation, exploitation, evaluation. In O. Čulo, S. Hansen-Schirra & S. Neumann (Eds.), *Special Issue of Translation: Computation, Corpora, Cognition*, 1(1), 1-5. Retrieved from http://www.t-c3.org/index.php/t-c3/issue/view/1

Danielsson, P. (2004). Simple perl programming for corpus work. In Sinclair, J. McH. (ed.), *How to Use Corpora in Language Teaching* (pp. 225–246). Amsterdam: John Benjamins.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery: an overview, advances in knowledge discovery and data mining. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (Eds.), *American Association for Artificial Intelligence* (pp. 1-36). Menlo Park: MIT Press.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37-54.

Fernandes, L. P., & Silva, C. E. (2016). COPA-TRAD versão 2.0 (corpus paralelo de tradução). Retrieved from http://copa-trad.ufsc.br.

Fernandes, L. P. (2004). Brazilian practices of translating names in children's fantasy literature: a corpus-based study. Unpublished Doctoral Dissertation, Universidade Federal de Santa Catarina, Florianópolis.

Fernandes, L. P. (2006). corpora in translation studies: revisiting Baker's typology. *Fragmentos: Revista de Língua e Literatura*, 30, 87-95. Florianópolis.

Fernandes, L. P. (2009). A Portal into the unknown: designing, building and processing a parallel corpus. *CTIS Occasional Papers* (Vol. 4, pp. 16-36). Manchester.

Foo, J. (2007). An overview of bitext alignment algorithms. Retrieved from http://www.ida.liu.se/~jodfo/gslt/bitext-alignment-jody.pdf

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI Magazine*,13 (3), 57-70. 10.1609/aimag.v13i3.1011.

Freeda, D. S. (2015). A Review paper on data mining techniques with duo mining. In *International Journal of New Innovations in Engineering and Technology (IJNIET)*, 3 (1), 50-55.

Gale, W., & Church, K. (1993). A Program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1), 75-102.

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications (0975 – 8887)*, 85(17), 42-45.

Gerdes, K. (2010). Poverty driven bilingual alignment. In R. Xiao (Ed.), *Using Corpora in Contrastive and Translation Studies* (pp. 257-280). Cambridge: Cambridge Scholars Publishing.

Goebel, M., & Gruenwald, L. (1999). A survey of data mining and knowledge discovery
software tools. *ACM SIGKDD Explorations Newsletter*, 1(1), 20-33. 10.1145/846170.846172.

Godfrey, J. J., & Zampolli, A. (1997). Language resources. In R. Cole (Ed.), *Survey of the state of the art in human language technology* (pp. 381–408). Cambridge/New York: Cambridge University Press.

Gries, S. T. (2015). Some current quantitative problems in corpus linguistics and a sketch of some solutions. In *Language and Linguistics*, 16(1), 93-117. Retrieved from http://lin.sagepub.com/content/16/1/93.full.pdf?ijkey=yNHyJGSvYhY4n1T&keytype=finite

Gries, S.T., & Wulff, S. (2012). Regression analysis in translation studies. In Oakes, M.P., Ji, M. (eds.) *Quantitative Methods in Corpus-Based Translation Studies. Studies in Corpus Linguistics* (Vol. 51, pp. 35–52). Philadelphia: John Benjamins.

Gupta, V., & Lehal, G. S. (2009). A Survey of text mining techniques and applications. *Journal Of Emerging Technologies In Web Intelligence*, 1(1), 60-76.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An Introduction to functional grammar* (3rd ed.). London: Hodder Arnold.

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational linguistics on Computational Linguistics* (pp. 3-10), Morristown: Association for Computational Linguistics. ISBN: 1-55860-609-3

Hovy, E. (2003). Text summarization. In R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics* (pp. 583-598). Oxford: Oxford University Press.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *LDV Forum – GLDV Journal for Computational Linguistics and Language Technology*, 20 (1), 19-62.

Hunston, S. (2010). How can a corpus be used to explore patterns?. In A. O'Keeffe and M. McCarthy (eds.) *Routledge Handbook of Corpus Linguistics* (pp. 152-166). London: Routledge.

Ji, M. (2008). Phraseology in corpus-based translation studies. A stylistic study of two contemporary Chinese translations of Cervantes's Don Quijote (Unpublished doctoral thesis). Imperial College London, London, United Kingdom.

Ji, M. (2012). Hypothesis testing in corpus-based literary translation studies. In Oakes, M. P., Ji, M. (eds.) *Quantitative Methods in Corpus-Based Translation Studies*. Studies in Corpus Linguistics (Vol. 51, pp. 53-72). Philadelphia: John Benjamins.

Kinyon, A. (2001). A language-independent shallow-parser compiler. In *Proceedings of 39th ACL Conference* (pp. 322-329). Toulouse, France.

Krishnaiah, V. V. J. R., Sekhar, D. V. C., Rao, K. R. H., & Prasad, R. S. (2012). Predicting the diabetes using duo mining approach. In *International Journal of Advanced Research in Computer and Communication Engineering*. 1(6), 423-431.

Kao, A., & Poteet, S. (2005). Text mining and natural language processing: introduction for the special issue. *ACM SIGKDD Explorations*, 7, 1-2.

Kay, M., & Röscheisen, M. (1993). Text-translation alignment. *Computational Linguistics*, 19(3), 121-142.

Kay, M. (2003). Introduction. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics*. New York: Oxford University Press Inc.

Kennedy, G. (1998). *An Introduction to corpus linguistics*. London/New York: Longman.

Krochmal, M., & Husi, H. (2018). Knowledge discovery and data mining. In: Vlahou, A., Mischak, H., Zoidakis, J. and Magni, F. (eds.) *Integration of Omics Approaches and Systems Biology for Clinical Applications* (pp. 233-247). Hoboken: John Wiley & Sons, Inc. ISBN 9781119181149 (doi:10.1002/9781119183952.ch14)

Koehn, P. (2010). *Statistical machine translation*. Cambridge: Cambridge University Press.

Laviosa, S. (1998). The corpus-based approach: A new paradigm in translation studies. *Meta*, 43(4), 474–479.

Laviosa, S. (2002). *Corpus-based translation studies: theory, findings, applications*. Amsterdam: Rodopi.

Laviosa, S. (2011). Corpus-based translation studies: Where does it come from? Where is it going? In Kruger, A., Wallmach and Munday, J. (Eds.) *Corpus-based translation studies: Research and applications* (pp. 13-32). London/New York: Continuum.

Li, K. W., & Yang, C. C. (2006). Conceptual analysis of parallel corpus collected from the Web. *Journal of the American Society for Information Science and Technology*, 57(5), 632-644.

Lüdeling, A., Evert, S., & Baroni, M. (2007). Using web data for linguistic purposes. In M. Hundt, N. Nesselhauf & C. Biewer (ed.), *Corpus Linguistics and the Web* (pp. 7-24). Amsterdam: Rodopi.

Malmkjær, K. (1998). Love thy neighbour: Will parallel corpora endear linguists to translators?. *Meta*, 43 (4), 534-541.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics?. In A. Gelbukh (Ed.), *Computational Linguistics and intelligent text processing – 12th International Conference CICLing* (pp. 171-189).

McEnery, T. (2003). Chapter 24 – Corpus Linguistics. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics*. New York: Oxford University Press Inc.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, A. M., & Wilson, A. (2001). *Corpus linguistics: an introduction*. Edinburgh: Edinburgh University Press.

McEnery, A. M., & Xiao, R. Z. (2007). Parallel and comparable corpora: What are they up to? In G. James, & G. Anderman (Eds.), *Incorporating Corpora: Translation and the Linguist (Translating Europe)*. Clevedon: Multilingual Matters.

Mehler, A., vor der Brück T., Gleim, R., & Geelhaar, T. (2014). Towards a network model of the coreness of texts: An experiment in classifying Latin texts using the TTLab Latin Tagger. In Biemann C., & Mehler A. (Eds.), *Text Mining From Ontology Learning to Automated Text Processing Applications* (pp. 87-112). Berlin/New York: Springer.

Merriam-webster.com,. (2018). Phenomena. Retrieved 10 June 2018, from https://www.merriam-webster.com/dictionary/phenomena

Miner, M., Delen, D., Elder, J., Fast, A., Hill, T., & Nisbet, B. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Waltham: Academic Press.

Moscarola, J., & Bolden, R. (1998). From the data mine to the knowledge mill: Applying the principles of lexical analysis to the data mining and knowledge discovery process. In Żytkow J. M., Quafafou M. (eds) *Principles of Data Mining and Knowledge Discovery*. *PKDD 1998*. (Vol. 1510). Berlin: Springer.

Mossop, B. (1994). Goals and methods for a course in translation theory. In M. Snell-Hornby, F. Pöchhacker, & K. Kaindl (Eds.), *Translation Studies: An Interdiscipline* (pp. 401-409). Amsterdam/Philadelphia: John Benjamins.

Musgrave, S., Dimitriadis, A., & Everaert, M. (2009). Introduction. In Everaert, M., Musgrave, S., Dimitriadis, A. (Eds.), *The Use of Databases in Cross-Linguistic Studies* (pp. 1-12). Berlin: Mouton de Gruyter.

North, M. (2012). *Data mining for the masses*. ISBN: 9780615684376 0615684378

Niles, R. (n.d.). Statistics Every Writer Should Know. Retrieved from http://www.robertniles.com/stats/stdev.shtml

Olohan, M., & Baker, M. (2000). Reporting that in translated english: Evidence for subconscious processes of explicitation?. In *Across Languages and Cultures*, 1(2), 141-158.

Olohan, M. (2004). *Introducing corpora in translation studies*. London/New York: Routledge.

Olohan, M. (2007). The status of scientific translation in translation studies. In *Journal of Translation Studies*, 10(1), 131-144.

Othero, G. A. (2006). Linguística computacional: uma breve introdução. In *Letras de Hoje*, 41(2), 341-351. Porto Alegre: EDIPUCRS.

Patel, F.N., Soni, N., & Vallabhbhai, S. (2012). Text mining: A brief survey. In *International Journal of Advanced Computer Research*, 2 (4), 243-248.

Quasthoff, U., Richter, M., & Biemann, C. (2006). Corpus portal for search in monolingual corpora. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk & D. Tapias (eds.), *LREC*, (pp. 1799-1802). European Language Resources Association (ELRA).

Riazi, A. M. (2016). *The Routledge encyclopedia of research methods in applied linguistics: quantitative, qualitative, and mixed-methods research*. Milton Park, Abingdon: Routledge.

Santorini, B. (1990). *Part-of-speech tagging guidelines for the penn treebank project*. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Sattikar, A. A., & Kulkarni, R. V. (2012). Natural language processing for content analysis in social. networking. In *International Journal of Engineering Inventions*.

Scott, M., (2001). Mapping key words to problem and solution. In M. Scott & G. Thompson (eds.) *Patterns of Text: in honour of Michael Hoey* (pp. 109-127). Amsterdam: Benjamins.

Scott, M., & Tribble, C. (2006). *Textual patterns : key words and corpus analysis in language education*. Amsterdam: John Benjamins publishing company. ISBN: 9027222940 9789027222947

Santos, A. (2011). A survey on parallel corpora alignment. In *MI-STAR* (pp. 117-128).

Srihari, S. N., & Hull, J. J. (1992). Character recognition. In Shapiro, S. C., (ed.), *Encyclopedia of Artificial Intelligence* (2nd ed., pp. 138-150). John Wiley.

Silva, C. E. (2013). Developing online parallel corpus-based processing tools for translation research and pedagogy. Unpublished Master's Dissertation, Universidade Federal de Santa Catarina, Florianópolis.

Silva, C. E. (2014). Utilização de técnicas de mineração de textos em corpora paralelo para auxílio na pesquisa acadêmica em estudos da tradução: Um estudo de caso. Unpublished post-graduate monograph. Universidade do Sul de Santa Catarina, Florianópolis.

Simoudis, E. (1996). Reality check for data mining. In *IEEE Expert*, 11(5), 26-33. DOI: 10.1109/64.539014.

Sinclair, J. M. (1987). Corpus creation. In G. Sampson and D. McCarthy (eds.) *Corpus Linguistics: Readings in a Widening Discipline* (pp. 78-84). London: Continuum.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. M. (2004a). Introduction. In Sinclair, J. McH. (ed.), *How to Use Corpora in Language Teaching* (pp. 1–10). Amsterdam: John Benjamins.

Sinclair, J. M. (2004b). *Trust the Text*. London: Routledge.

Sucharita, V., Rao, P. V., Satya, K. A., & Rajarajeswari, P. (2017). Big Data Mining: A Forecast to the Future. In A.V. K. Prasad (Ed.), *Exploring the Convergence of Big Data and the Internet of Things* (pp. 36-42). IGI Global.

Sumathy, K. L. & Chidambaram, M. (2013). Text Mining: Concepts, Applications, Tools and Issues – An Overview. *International Journal of Computer Applications*, 80(4), 29-32.

Tan, A. (1999). Text Mining: The state of the art and the challenges. In Proceedings, *PAKDD'99 workshop on Knowledge Discovery from Advanced Databases* (pp. 65-70). Beijing, China.

TheFreeDictionary.com,. (2018). Connotation. Retrieved 10 June 2018, from https://www.thefreedictionary.com/connotation

Tiedemann, J. (2004). Word to word alignment strategies. In *Proceedings of COLING 2004*, pp. 212-218, Geneva, Switzerland.

Tiedemann, J. (2011). Bitext alignment. In Graeme Hirst (Ed.), *Synthesis Lectures on Human Language Technologies* (Vol. 14, pp. vii-153), San Rafael: Morgan and Claypool Publishers.

Tiedemann, J. (2012). parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)* (pp. 2214-2218).

Tóth, K., Farkas, R., & Kocsor, A. (2008). Sentence alignment of Hungarian-English parallel corpora using a hybrid algorithm. *Acta Cybern*, 18, 463-478.

Tsatsoulis, C. I. (2013). Unsupervised text mining methods for literature analysis: a case study for Thomas Pynchon's V.. In *A Journal of American Literature*. 1(2). DOI: https://doi.org/10.7766/orbit.v1.2.44

Tufiş, D., & Ion, R. (2007). Parallel corpora, alignment technologies and further prospects in multilingual resources and technology infrastructure. In C. Burileanu and H-N Teodorescu (Eds.), *Proceedings of the 4th Conference on Speech Technology and Human-Computer Dialogue SpeD 2007* (pp. 1-12), Iaşi, Romania.

trim. (n.d.). Retrieved from http://php.net/manual/en/function.trim.php

Tymoczko, M. (1998). Computerized corpora and the future of translation studies. *Meta*, 43(4), 652–660. DOI:10.7202/004515ar

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing (RANLP 2005)* (pp. 590-596).

Varantola, K. (2002). Disposable corpora as intelligent tools in translation. In S. E. O. Tagnin (Ed.), *Cadernos de Tradução IX - Tradução e Corpora*, 1(9), 171-189. Florianópolis: NUT.

Veni, R. M., Praveena, M., & GanaPriya, V. (2013). A Review on duo mining techniques. In *International Journal of Science and Research (IJSR)*, 2 (3), 124-128. Retrieved from https://www.ijsr.net/archive/v2i3/v2i3.php

Vieira, R., & Lima, V. L. S. (2001). Lingüística computacional: princípios e aplicações. In: *IX Escola de Informática da SBC-Sul*. Luciana Nedel (Ed.) Passo Fundo, Maringá, São José. SBC-Sul.

Zethsen, K. K. (2006). Semantic prosody: Creating awareness about a versatile tool. *Journal of Sprogforskning*, 4(1), 275-294.

Zipf, G. K. (1932). Selected studies of the principle of relative frequency in language. Cambridge: Harvard University Press.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, (pp. 29-39). London.

Witten, I. H. (2004). Text mining. In M.P. Singh (Ed.), *The Practical Handbook of Internet Computing*, (pp. 14-22). Boca Raton: Chapman/Hall/CRC Press.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., & Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation (pp. 1-23). *arXiv preprint arXiv:1609.08144*.

Yang, C. C., & Li, K. W. (2003). Automatic construction of English/Chinese parallel corpora. *JASIST*, 54 (8), 730-742.

Youzhi, Z. (2010). Research and application of hidden Markov model in data mining. In *2010 Second IITA International Conference on Geoscience and Remote Sensing, IITA-GRS 2010*, 1, 459-462.

APPENDIXES

APPENDIX A

(COPA-TRAD Grant of Patent - Legal Document)

REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DO DESENVOLVIMENTO, INDÚSTRIA E COMÉRCIO EXTERIOR
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL

CERTIFICADO DE REGISTRO
DE PROGRAMA DE COMPUTADOR

Processo: 13281-6

O INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL expede o presente Certificado de Registro de Programa de Computador, **válido por 50 anos** a partir de 1º de janeiro subsequente à data de criação indicada, em conformidade com o art. 3º da Lei Nº 9.609, de 19 de Fevereiro de 1998, e arts. 1º e 2º do Decreto 2.556 de 20 de Abril de 1998.

Título: **COPA-TRAD: CORPUS PARALELO DE TRADUÇÃO**

Data de Criação 01 de Junho de 2011

Titular: 83.899.526/0001-82 **UNIVERSIDADE FEDERAL DE SANTA CATARINA**

Criadores: 807.832.529-00 **LINCOLN PAULO FERNANDES**
039.255.359-77 **CARLOS EDUARDO DA SILVA**

Linguagens: **JAVASCRIPT, PHP**

Campo de Aplicação: **CO-03**

Tipos de Programa: **FA-01, GI-01, GI-08, UT-01**

Documentação Técnica em depósito **SOB SIGILO até 16/05/2022.**

A exclusividade de comercialização do programa de computador objeto deste Certificado não tem a abrangência relativa à exclusividade de fornecimento estatuída pelo art. 25, inciso I da Lei Nº 8.666, de 21 de Junho de 1993, para fins de inexigibilidade de licitação para compras pelo poder público.

Expedido em 26 de Março de 2013.

Rodrigo Moerbeck de Almeida Rego
Chefe da Divisão de Registro de Programas de Computador e Topografia de Circuitos Integrados

Breno Bello de Almeida Neves
Diretor de Contratos, Indicações Geográficas e Registros

APPENDIX B

(COPA-TRAD Patent application request form)

INPI

PEDIDO DE REGISTRO DE PROGRAMA DE COMPUTADOR

protocolo

INPI INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL PROTOCOLO GERAL

16/05/2012 01712000049 15:20 RESC

0000271202187537

IDENTIFICAÇÃO DO PEDIDO (Para uso do INPI)

Número do Pedido

Protocolo, Data e Hora

DADOS DO AUTOR DO PROGRAMA

Nº de Autores 2 Se mais de um, preencha a "Continuação", com todos os dados solicitados neste Quadro. Date e assine.

CPF* 807.832.529-00

Nome LINCOLN PAULO FERNANDES

Nome Abreviado, pseudônimo ou sinal convencional (se houver)

Data de Nascimento Nacionalidade BRASILEIRA

Endereço RODOVIA JOÃO PAULO, 710, T1, 301-B

Cidade FLORIANÓPOLIS UF SC País BRASIL

CEP 88.030-300 Telefone 4833048936 FAX

E-mail lincoln.fernandes@ufsc.br

DADOS DO TITULAR DOS DIREITOS PATRIMONIAIS

Nº de Titulares 1 Se mais de um, preencha a "Continuação", com todos os dados solicitados neste Quadro. Date e assine.

CPF/CNPJ* 83899526000182

Nome/Razão Social UNIVERSIDADE FEDERAL DE SANTA CATARINA

Nome abreviado, pseudônimo ou sinal convencional (se houver) UFSC

Data de Nascimento Nacionalidade/Origem

Endereço CAMPUS UNIVERSITÁRIO, SN, CP 476, TRINDADE

Cidade FLORIANÓPOLIS UF SC País BRASIL

CEP 88.040-900 Telefone 4837219628 FAX

E-mail dit@reitoria.ufsc.br

☐ SIM, este Titular é Pessoa Jurídica. Caso afirmativo, assinale a melhor classificação:

☐ Órgão Público ☐ Sociedade co Intuito não Econômico ☐ Microempresa ☐ Software House

☒ Instituição Pública de Ensino ou Pesquisa ☐ Instituição Privada de Ensino ou Pesquisa ☐ Outras

ENDEREÇO PARA CORRESPONDÊNCIA E CONTATO (Preencha apenas o necessário)

Toda correspondência será enviada para: ☐ O Procurador ou ☒ O Titular acima ou

☐ Escaninho nº ☐ Representação INPI em: ☐ O Endereço abaixo:

Nome

Endereço

Cidade UF País

CEP Telefone FAX

E-mail

Modelo I (folha 1/2) E

DADOS DO PROGRAMA

Título	COPA-TRAD: Corpus Paralelo de Tradução
Data de Criação do Programa	1/6/2011
Regime de Guarda	☒ COM SIGILO ☐ SEM SIGILO
Linguagens	PHP / JAVASCRIPT
Classificação do Campo de Aplicação	CO - 03
Classificação do Tipo de Programa	FA - 01 / GI - 01 / UT - 01 / GI - 08

☐ **SIM**, este Programa é Modificação Tecnológica ou Derivação. Caso afirmativo, informe Título do Programa Original e (se houver) Número de Registro:

Título do Programa Original:

☐ **SIM**, este Registro é composto por obra(s) de outra(s) natureza(s) de ordem intelectual. Caso afirmativo assinale-a(s) abaixo:

☐ Literária ☐ Musical ☐ Artes Plásticas ☐ Áudio-Visual ☐ Arquitetura ☐ Engenharia

DOCUMENTOS ANEXADOS (Informe as quantidades de documentos, não o número de páginas)

Quant	Nome	Quant	Nome
1	Guia de Recolhimento		Contrato de Trabalho/Prestação de Serviço
	Procuração		Invólucros/mídia eletrônica Utilizados
	Termo de Cessão		Contrato/Estatuto Social e Alterações (ou equivalente)
	Termo de Autorização para Modificações Tecnológicas ou Derivações	1	Autorização para Cópia do CD
		6	Outros(especificar) RESOLUÇÃO Nº 14 UFSC; COMPR. TE CNPJ UFSC; DECLARAÇÃO AUTENTICIDADE DOCS; POU NOMEAÇÃO REITOR; CÓPIA REITOR

DECLARAÇÕES

DECLARO, PARA TODOS OS FINS DE DIREITO:

A) que estou ciente de **TODAS AS RECOMENDAÇÕES** constantes do "Manual do Usuário de Registro de Programas de Computador", **ESPECIALMENTE NO QUE TANGE AO TÍTULO E AOS DOCUMENTOS DO PROGRAMA**, bem como da legislação pertinente ao assunto, constante dos anexos "A" ; "B" ; "C" ; "E" e "F", do referido Manual;

B) que se deixar de solicitar a prorrogação do sigilo, nos casos necessários, estarei desistindo desse caráter de guarda dos documentos de programa do presente depósito, na forma do art. 3º, § 2º, da Lei 9.609, de 12 de fevereiro de 1998;

C) que, se devido à qualidade do papel ou à qualidade gráfica dos documentos sigilosos anexos ao presente, houver deterioração ou perda de seu conteúdo, nenhuma responsabilidade caberá ao INPI, desde que mantida a inviolabilidade dos invólucros (ressalvadas as hipóteses de serem abertos por ordem judicial ou motivo de força maior);

D) que em caso de perda do SIGILO ou dos documentos, por culpa exclusiva do INPI, a indenização por perdas e danos, porventura cabível, estará limitada a 20 (vinte) salários mínimos;

E) que devo manter guardado, em segurança e inviolado, o COMPARTIMENTO "3" do invólucro especial para depósito, que é restituído pelo INPI, para fins de recomposição do arquivo do Instituto, no caso de sua destruição total ou parcial por algum tipo de sinistro;

F) que deverei manter endereço atualizado junto à Divisão de Registro de Programa de Computador, a fim de garantir o recebimento das comunicações relativas ao andamento do meu pedido/registro, ressalvando o INPI de qualquer responsabilidade decorrente da não observação deste preceito.

DADOS DO PROCURADOR

CPF/CNPJ*	Código do Procurador (se houver)
Nome	
Endereço	
Cidade	UF / País
CEP	Telefone / FAX
E-mail	

DECLARO, SOB AS PENAS DA LEI, SEREM VERDADEIRAS AS INFORMAÇÕES PRESTADAS

Florianópolis, 14 / 2012

Local/Data

Assinatura/Carimbo

Prof. Alvaro Toubes Prata
Universidade Federal de Santa Catarina
Reitor

Modelo I (folha 2/2) E

REGISTRO DE PROGRAMA DE COMPUTADOR - CONTINUAÇÃO

Utilize este ANEXO, em quantas folhas forem necessárias, para complementar as informações dos formulários "Pedido de Registro de Programa de Computador" e "Folha de Petição" (DIRTEC).

TÍTULO

"COPA-TRAD: Corpus Paralelo de Tradução"

Dados dos demais autores:

Tem outro(s) programa(s) registrado(s) no INPI?
Sim ( ) Não (X)
CPF: 039.255.359-77
Nome civil completo: Carlos Eduardo da Silva
Nome abreviado:
Nacionalidade: Brasileiro
Data de nascimento: 21/07/1984
Endereço: Rua Sandro Pain, 380, Praia Cumprida
Cidade: Florianópolis
UF: SC
CEP: 88103770
Cód. Pais: 55 Telefone: 48 3247 2528/ 48 9915 4018
E-mail: carlosedasilva@gmail.com

APPENDIX C

(Legal Announcement of COPA-TRAD patent in Revista da Propriedade Industrial equivalent to "Official Gazette for Patents" in United States)

Regime de Guarda: Sigilo Até 23/04/2022
Procurador: PAULO AUGUSTO MALTA MOREIRA - CPF:66320844604

Processo: 13279-5 **080**
Título: ALFA
Titular: UNIVERSIDADE FEDERAL DE VIÇOSA - CPF/CNPJ:25944455000196
Criador: ANDRÉ FERNANDO DE OLIVEIRA
Linguagem: VBA, VISUAL BASIC
Campo de Aplicação: FQ-16
Tipo de Programa: FA-01
Data da Criação: 28/02/2010
Regime de Guarda: Sigilo Até 18/05/2022
Procurador: PAULO AUGUSTO MALTA MOREIRA - CPF:66320844604

Processo: 13280-4 **080**
Título: THOTAU/THOTAU - SISTEMA INTEGRADO DE GESTÃO EMPRESARIAL
Titular: ORION SISTEMAS LTDA. - CPF/CNPJ:03005347000115
Criador: EDSON TEIXEIRA MARQUES , EDUARDO BARBOSA DE SOUZA
Linguagem: OBJECT PASCAL
Campo de Aplicação: AD-05, AD-08, AD-09, FN-05, FN-06
Tipo de Programa: AT-03
Data da Criação: 07/11/2008
Regime de Guarda: Sigilo Até 18/05/2022
Procurador: Não informado ou inexistente

Processo: 13281-6 **080**
Título: COPA-TRAD: CORPUS PARALELO DE TRADUÇÃO
Titular: UNIVERSIDADE FEDERAL DE SANTA CATARINA - CPF/CNPJ:83899526000182
Criador: CARLOS EDUARDO DA SILVA , LINCOLN PAULO FERNANDES
Linguagem: JAVASCRIPT, PHP
Campo de Aplicação: CO-03
Tipo de Programa: FA-01, GI-01, GI-08, UT-01
Data da Criação: 01/06/2011
Regime de Guarda: Sigilo Até 16/05/2022
Procurador: Não informado ou inexistente

Processo: 13282-1 **080**
Título: MATA ATLÂNTICA, O BIOMA ONDE EU MORO
Titular: UNIVERSIDADE FEDERAL DE SANTA CATARINA - CPF/CNPJ:83899526000182
Criador: ANA BEATRIZ BAHIA SPINOLA BITTENCOURT, CRISTINA VALÉRIA SANTOS, EMÍLIO TAKASE , MATHEUS BASSI BLANK GONÇALVES
Linguagem: FLASH
Campo de Aplicação: ED-01
Tipo de Programa: ET-01, ET-02
Data da Criação: 20/01/2012
Regime de Guarda: Sigilo Até 16/05/2022
Procurador: Não informado ou inexistente

Processo: 13283-3 **080**
Título: D-1 DISTRIBUTION ONE
Titular: ALCIRO MARCOS ORLAMÜNDER - CPF/CNPJ:68401361915
Criador: ALCIRO MARCOS ORLAMÜNDER
Linguagem: 4GL, JAVA, PROGRESS, VISUAL BASIC
Campo de Aplicação: AD-05, AD-08, AD-10, AD-11, FN-06
Tipo de Programa: AP-01, AP-02, AP-03, IA-01, IA-02
Data da Criação: 01/07/2001
Regime de Guarda: Sigilo Até 17/05/2022
Procurador: ALCIRO MARCOS ORLAMÜNDER - CPF:68401361915

Processo: 13329-1 **080**
Título: DOMINIO ESCRITA FISCAL VERSÃO 04
Titular: DOMÍNIO SISTEMAS LTDA. - CPF/CNPJ:02825945000178
Criador: ADRIANO DIAS, ADRIANO FRANCISCO, ALESSANDRA TEREZINHA DA SILVA, ALEXANDRE DE ALMEIDA, ALEXANDRE NIERO, ALEXANDRE ROBERTO LEMES MARTINS, ALINE CORREA RAMOS, ALISSON DE VILLA GERONIMO, ALISSON DOS SANTOS SILVA, ANDERSON FELISBERTO MANOEL, ANDERSON RICARDO DOS SANTOS RODRIGUES, ANDERSON SILVESTRI FERRO, ANTONIO JOSÉ VIEIRA JUNIOR, ANTONIO MARCOS DE OLIVEIRA, BRUNO BRISTOT LOLI, CAMILA MOTTA WOSNIESKI, CARLA EYNG, CESAR EDUARDO FRANCO ISE COLONETTI, CIRILO PINTER COLOMBO, CLEVERSON REINERT, DANGELO ROSSO ZANETTE, DANIEL DE MEDEIROS BOFF, DAVI GONÇALVES, DIEGO GOMES ANTONELI, DIEGO MACHADO MEDEIROS, DIEGO MARIANI DE MELO, DIEGO MARTINS DA ROCHA, EDGAR SOUZA DA CRUZ, EDIVALDO LUCIO, EVERSON NERI FRANCELINO, FAGNER LEANDRO DE SOUZA, FELIPE CORAL SASSO, FELIPE ORTMEYER HENRIQUE DA SILVA, FERNANDA D AGOSTIN, FERNANDO NAZARIO PIZZETTI, FLARIS BARRETO MARTINHAGO, GABRIEL GUADANHIM GENEROSO, GUILHERME FRANCISCO DE SOUZA, GUILHERME TEODORO DE OLIVEIRA, GUSTAVO GRIGGIO DE SOUZA, HEMERSON BEZ BIROLO, HENRIQUE COLOMBO GUINZAIN, HENRIQUE PIAZZA LUCIANO, HERLON HILBERT, HERON POTRIKUS CRESTANI, IURI SONEGO CARDOSO, JAISSON RODRIGUES DEMBOSKI, JEFERSON LUIZ BATISTI, JESSICA RONCONI DONDOSSOLA, JULIANA GUADANHIM GENEROSO, JULIANO MARQUES, LEONARDO BENEDET, LUANA GASPAR SOARES, LUCAS VITORINO GONÇALVES, MARCELO DEHON BATISTA DE PRA, MARCIO DAGOSTIM DE CASTRO, MARCONDES DE BORBA, MARIANA ANTONIO SARTORI, MARIANI COLONETTI, MARIANNA SANTOS SAGGIORATO, MARILIA TEIXEIRA PIRES, MARINA KURTZ SCHMIDT, MARITY EYNG NUERNBERG, MATHEUS MEDEIROS ANACLETO, MELISSA DA PAZ TEIXEIRA, MICHAEL CELSO BITENCOURT, PAULA CRISTINA VIEIRA RONSANI, PAULO HENRIQUE ELI, PAULO ROBERTO DABOIT MILANEZ, RAFAEL CECHINEL SILVESTRI, REGINALDO DAROLT, RENAN ROSSO DA SILVA, RICHARDSON PICININI CORREIA, ROBERTO MENDES GARCIA, ROBERTO VEFAGO CAROLLI, ROGERIO BRUM HERMANY, ROGERIO DAMACENO DE FÁRIAS, SAMUEL LODETTI GHELLERE, SIMONE PEREIRA DA CUNHA, SUELEN JUVENCIO DAMAZIO, TALINE FELTRIN DE SOUZA, TAMARA JOSEPHINO FERNANDES, TAMIRES JUSTI ROCHA, THALES MENDES MILANESE, THIAGO APOLINARIO BILLIERI, TIAGO BITENCORT MARQUES, TULIO DAMINELLI BORGES, VANESSA CRISTINA CARPES DA SILVA, VANESSA FELISBERTO BILESIMO, WAGNER JOSÉ DENONI FREITAS, WELLINGTON ZOMER NUNES
Linguagem: POWERBUILDER, SQL
Campo de Aplicação: IF-10
Tipo de Programa: AT-02
Data da Criação: 01/01/1999
Regime de Guarda: Sigilo Até 29/05/2022
Procurador: DMARK REGISTROS DE MARCAS E PATENTES LTDA - CPF:03389474000165

Processo: 13332-4 **080**
Título: DOMÍNIO ATENDIMENTO VERSÃO 02
Titular: DOMÍNIO SISTEMAS LTDA. - CPF/CNPJ:02825945000178
Criador: ADRIANO DIAS, ADRIANO FRANCISCO, ALESSANDRA TEREZINHA DA SILVA, ALEXANDRE DE ALMEIDA, ALEXANDRE NIERO, ALEXANDRE ROBERTO LEMES MARTINS, ALINE CORREA RAMOS, ALISSON DE VILLA GERONIMO, ALISSON DOS SANTOS SILVA, ANDERSON FELISBERTO MANOEL, ANDERSON RICARDO DOS SANTOS RODRIGUES, ANDERSON SILVESTRI FERRO, ANTONIO JOSÉ VIEIRA JUNIOR, ANTONIO MARCOS DE OLIVEIRA, BRUNO BRISTOT LOLI, CAMILA MOTTA WOSNIESKI, CARLA EYNG, CESAR EDUARDO FRANCO ISE COLONETTI, CIRILO PINTER COLOMBO, CLEVERSON REINERT, DANGELO ROSSO ZANETTE, DANIEL DE MEDEIROS BOFF, DAVI GONÇALVES, DIEGO GOMES ANTONELI, DIEGO MACHADO MEDEIROS, DIEGO MARIANI DE MELO, DIEGO MARTINS DA ROCHA, EDGAR SOUZA DA CRUZ, EDIVALDO LUCIO, EVERSON NERI FRANCELINO, FAGNER LEANDRO DE SOUZA, FELIPE CORAL SASSO, FELIPE ORTMEYER HENRIQUE DA SILVA, FERNANDA D AGOSTIN, FERNANDO NAZARIO PIZZETTI, FLARIS BARRETO MARTINHAGO, GABRIEL GUADANHIM GENEROSO, GUILHERME FRANCISCO DE SOUZA, GUILHERME TEODORO DE OLIVEIRA, GUSTAVO GRIGGIO DE SOUZA, HEMERSON BEZ BIROLO, HENRIQUE COLOMBO GUINZAIN, HENRIQUE PIAZZA LUCIANO, HERLON HILBERT, HERON POTRIKUS CRESTANI, IURI SONEGO CARDOSO, JAISSON RODRIGUES DEMBOSKI, JEFERSON LUIZ BATISTI, JESSICA RONCONI DONDOSSOLA, JULIANA GUADANHIM GENEROSO, JULIANO MARQUES, LEONARDO BENEDET, LUANA GASPAR SOARES, LUCAS VITORINO GONÇALVES, MARCELO DEHON BATISTA DE PRA, MARCIO DAGOSTIM DE CASTRO, MARCONDES DE BORBA, MARIANA ANTONIO SARTORI, MARIANI COLONETTI, MARIANNA SANTOS SAGGIORATO, MARILIA TEIXEIRA PIRES, MARINA KURTZ SCHMIDT, MARITY EYNG NUERNBERG, MATHEUS MEDEIROS ANACLETO, MELISSA DA PAZ TEIXEIRA, MICHAEL CELSO BITENCOURT, PAULA CRISTINA VIEIRA RONSANI, PAULO HENRIQUE ELI, PAULO ROBERTO DABOIT MILANEZ, RAFAEL CECHINEL SILVESTRI, REGINALDO DAROLT, RENAN ROSSO DA SILVA, RICHARDSON PICININI CORREIA, ROBERTO MENDES GARCIA, ROBERTO VEFAGO CAROLLI, ROGERIO BRUM HERMANY, ROGERIO DAMACENO DE FÁRIAS, SAMUEL LODETTI GHELLERE, SIMONE PEREIRA DA CUNHA, SUELEN JUVENCIO DAMAZIO, TALINE FELTRIN DE SOUZA, TAMARA JOSEPHINO FERNANDES, TAMIRES JUSTI ROCHA, THALES MENDES MILANESE, THIAGO APOLINARIO BILLIERI, TIAGO BITENCORT MARQUES, TULIO DAMINELLI BORGES, VANESSA CRISTINA CARPES DA SILVA, VANESSA FELISBERTO BILESIMO, WAGNER JOSÉ DENONI FREITAS, WELLINGTON ZOMER NUNES
Linguagem: POWER BUILDER, SQL
Campo de Aplicação: IF-10
Tipo de Programa: AT-02
Data da Criação: 15/09/2009
Regime de Guarda: Sigilo Até 29/05/2022
Procurador: DMARK REGISTROS DE MARCAS E PATENTES LTDA - CPF:03389474000165

Processo: 13333-6 **080**
Título: CAPTA CLIENTE
Titular: NOVOCLIENTE TECNOLOGIA LTDA. - CPF/CNPJ:11496248700133
Criador: GUILHERME LEMOS SANTOS
Linguagem: PHP
Campo de Aplicação: AD-01, AD-02, AD-03, AD-05, AD-10
Tipo de Programa: GI-01, GI-02, GI-04, GI-05, GI-07
Data da Criação: 14/06/2012
Regime de Guarda: Sigilo Até 29/05/2022
Procurador: RICARDO PREIS DE FREITAS VALLE CORREA - CPF:63149591015

Processo: 13450-3 **080**
Título: PLATAFORMA E COMMERCE
Titular: EDUARDO MALVEIRO PEREIRA LEITE - CPF/CNPJ:06542126864
Criador: EDUARDO MALVEIRO PEREIRA LEITE
Linguagem: PHP
Campo de Aplicação: IF-09, SV-03, TC-02
Tipo de Programa: GI-01, GI-02, GI-04, GI-07, SO-07
Data da Criação: 10/06/2010
Regime de Guarda: Sigilo Até 21/06/2022
Procurador: SUL AMÉRICA MARCAS E PATENTES LTDA. - CPF:60848983000142

Processo: 13451-5 **080**
Título: CLIENTE TR - 69
Titular: EDUARDO MALVEIRO PEREIRA LEITE - CPF/CNPJ:06542126864
Criador: EDUARDO MALVEIRO PEREIRA LEITE
Linguagem: PASCAL
Campo de Aplicação: TC-02
Tipo de Programa: CD-04, GI-01, SO-06, SO-08
Data da Criação: 13/08/2012
Regime de Guarda: Sigilo Até 21/06/2022
Procurador: SUL AMÉRICA MARCAS E PATENTES LTDA. - CPF:60848983000142

Processo: 13452-0 **080**
Título: IP TOUCH
Titular: EDUARDO MALVEIRO PEREIRA LEITE - CPF/CNPJ:06542126864
Criador: EDUARDO MALVEIRO PEREIRA LEITE
Linguagem: C, PHP, PYTHON
Campo de Aplicação: TC-02
Tipo de Programa: CD-01, CT-01, SO-07, TI-01, TI-04
Data da Criação: 10/06/2010
Regime de Guarda: Sigilo Até 21/06/2022
Procurador: SUL AMÉRICA MARCAS E PATENTES LTDA. - CPF:60848983000142

APPENDIX D

(COPA-TRAD Version 2.0 Patent application request form)

INPI

protocolo

INPI 01716000093
04/10/2016 16:06 RESC
BR 51 2016 001294 3

PEDIDO DE REGISTRO DE PROGRAMA DE COMPUTADOR

IDENTIFICAÇÃO DO PEDIDO (Para uso do INPI)

Número do Pedido | Protocolo, Data e Hora

DADOS DO AUTOR DO PROGRAMA

Nº de Autores: 2 — Se mais de um, preencha a "Continuação", com todos os dados solicitados neste Quadro. Date e assine.

CPF*: 807.832.529-00
Nome: LINCOLN PAULO FERNANDES
Nome Abreviado, pseudônimo ou sinal convencional (se houver):
Data de Nascimento: 30/09/1971 — Nacionalidade: BRASILEIRA
Endereço: ROD. JOÃO PAULO, 710
Cidade: FLORIANÓPOLIS — UF: SC — País: BRASIL
CEP: 88.030-300 — Telefone: 4833048936 — FAX:
E-mail: lincoln.fernandes@ufsc.br

DADOS DO TITULAR DOS DIREITOS PATRIMONIAIS

Nº de Titulares: 1 — Se mais de um, preencha a "Continuação", com todos os dados solicitados neste Quadro. Date e assine.

CPF/CNPJ*: 83899526000182
Nome/Razão Social: UNIVERSIDADE FEDERAL DE SANTA CATARINA
Nome abreviado, pseudônimo ou sinal convencional (se houver): UFSC
Data de Nascimento: — Nacionalidade/Origem:
Endereço: CAMPUS UNIVERSITÁRIO, SN, CP 476, TRINDADE
Cidade: FLORIANÓPOLIS — UF: SC — País: BRASIL
CEP: 88.040-900 — Telefone: 4837219628 — FAX:
E-mail: sinova@contato.ufsc.br

☒ SIM, este Titular é Pessoa Jurídica. Caso afirmativo, assinale a melhor classificação:

☐ Órgão Público ☐ Sociedade co Intuito não Econômico ☐ Microempresa ☐ Software House
☒ Instituição Pública de Ensino ou Pesquisa ☐ Instituição Privada de Ensino ou Pesquisa ☐ Outras

ENDEREÇO PARA CORRESPONDÊNCIA E CONTATO (Preencha apenas o necessário)

Toda correspondência será enviada para: ☐ O Procurador ou ☒ O Titular acima ou
☐ Escaninho nº ☐ Representação INPI em: ☐ O Endereço abaixo:

Nome:
Endereço:
Cidade: — UF: — País:
CEP: — Telefone: — FAX:
E-mail:

Modelo I (COPA-TRAD) E

DADOS DO PROGRAMA

Título	COPA-TRAD: Corpus Paralelo de Tradução - Versão 2.0				
Data de Criação do Programa	01/06/2011	Regime de Guarda	☒ COM SIGILO	☐ SEM SIGILO	
Linguagens	PHP	Javascript			
Classificação do Campo de Aplicação	CO - 03	-	-	-	-
Classificação do Tipo de Programa	FA - 01	GI - 01	UT - 01	GI - 08	-

☒ **SIM**, este Programa é Modificação Tecnológica ou Derivação. Caso afirmativo, informe Título do Programa Original e (se houver) Número de Registro:

Título do Programa Original: COPA-TRAD: Corpus Paralelo de Tradução (13281-6)

☐ **SIM**, este Registro é composto por obra(s) de outra(s) natureza(s) de ordem intelectual. Caso afirmativo assinale-a(s) abaixo:

☐ Literária ☐ Musical ☐ Artes Plásticas ☐ Áudio-Visual ☐ Arquitetura ☐ Engenharia

DOCUMENTOS ANEXADOS (Informe as quantidades de documentos, não o número de páginas)

Quant	Nome	Quant	Nome
1	Guia de Recolhimento		Contrato de Trabalho/Prestação de Serviço
	Procuração		Invólucros/mídia eletrônica Utilizados
	Termo de Cessão	1	Contrato/Estatuto Social e Alterações (ou equivalente)
	Termo de Autorização para Modificações Tecnológicas ou Derivações	1	Autorização para Cópia do CD
		3	Outros(especificar)

DECLARAÇÕES

DOU Nomeação Vice-Reitora UFSC; Compr. Sit. Fiscal UFSC; Resolução 14 UFSC;

DECLARO, PARA TODOS OS FINS DE DIREITO:

A) que estou ciente de **TODAS AS RECOMENDAÇÕES** constantes do "Manual do Usuário de Registro de Programas de Computador", **ESPECIALMENTE NO QUE TANGE AO TÍTULO E AOS DOCUMENTOS DO PROGRAMA**, bem como da legislação pertinente ao assunto, constante dos anexos "A" ; "B" ; "C" ; "E" e "F", do referido Manual;

B) que se deixar de solicitar a prorrogação do sigilo, nos casos necessários, estarei desistindo desse caráter de guarda dos documentos de programa do presente depósito, na forma do art. 3º, § 2º, da Lei 9.609, de 12 de fevereiro de 1998;

C) que, se devido à qualidade do papel ou à qualidade gráfica dos documentos sigilosos anexos ao presente, houver deterioração ou perda de seu conteúdo, nenhuma responsabilidade caberá ao **INPI**, desde que mantida a inviolabilidade dos invólucros (ressalvadas as hipóteses de serem abertos por ordem judicial ou motivo de força maior);

D) que em caso de perda do SIGILO ou dos documentos, por culpa exclusiva do INPI, a indenização por perdas e danos, porventura cabível, estará limitada a 20 (vinte) salários mínimos;

E) que devo manter guardado, em segurança e inviolado, o COMPARTIMENTO "3" do invólucro especial para depósito, que é restituído pelo INPI, para fins de recomposição do arquivo do Instituto, no caso de sua destruição total ou parcial por algum tipo de sinistro;

F) que deverei manter endereço atualizado junto à Divisão de Registro de Programa de Computador, a fim de garantir o recebimento das comunicações relativas ao andamento do meu pedido/registro, ressalvando o **INPI** de qualquer responsabilidade decorrente da não observação deste preceito.

DADOS DO PROCURADOR

CPF/CNPJ*		Código do Procurador (se houver)	
Nome			
Endereço			
Cidade		UF	País
CEP	Telefone	FAX	
E-mail			

DECLARO, SOB AS PENAS DA LEI, SEREM VERDADEIRAS AS INFORMAÇÕES PRESTADAS

Florianópolis, 15 de setembro de 2016

Local/Data

Assinatura/Carimbo

Prof. Ubaldo Cesar Balthazar
Vice-Reitora / UFSC
Port. 955/2016/GR

Modelo I (folha 2/2) E

REGISTRO DE PROGRAMA DE COMPUTADOR - CONTINUAÇÃO

Utilize este ANEXO, em quantas folhas forem necessárias, para complementar as informações dos formulários "Pedido de Registro de Programa de Computador" e "Folha de Petição" (DIRTEC).

Título:

"COPA-TRAD: Corpus Paralelo de Tradução - Versão 2.0"

Dados dos demais autores:

Tem outro(s) programa(s) registrado(s) no INPI?
Sim (X) Não ( )
CPF: 039.255.359-77
Nome civil completo: Carlos Eduardo da Silva
Nacionalidade: Brasileiro
Data de nascimento: 21/07/1984
Endereço: Rua Sandro Pain, 380, Praia Cumprida
Cidade: São José
UF: SC
CEP: 88103770
Cód. Pais: 55 Telefone: 48 3247 2528/ 48 9915 4018
E-mail: carlosedasilva@gmail.com

APPENDIX E

(COPA-TRAD Version 2.0 Grant of Patent - Legal Document)

REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DA INDÚSTRIA, COMÉRCIO EXTERIOR E SERVIÇOS
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIA DE CIRCUITOS INTEGRADOS

CERTIFICADO DE REGISTRO DE PROGRAMA DE COMPUTADOR

Processo: BR 51 2016 001294-3

O INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL expede o presente Certificado de Registro de Programa de Computador, **válido por 50 anos** a partir de 1º de janeiro subsequente à data de criação indicada, em conformidade com o parágrafo 2º, artigo 2º da Lei Nº 9.609, de 19 de Fevereiro de 1998, e arts. 1º e 2º do Decreto 2.556 de 20 de Abril de 1998.

Título:	**COPA-TRAD: CORPUS PARALELO DE TRADUÇÃO - VERSÃO 2.0**
Criação:	01 de junho de 2011
Titular(es):	UNIVERSIDADE FEDERAL DE SANTA CATARINA (83.899.526/0001-82)
Autor(es):	CARLOS EDUARDO DA SILVA (039.255.359-77) LINCOLN PAULO FERNANDES (807.832.529-00)
Linguagem:	JAVASCRIPT, PHP
Aplicação:	CO-03
Tipo Prog.:	FA-01, GI-01, GI-08, UT-01

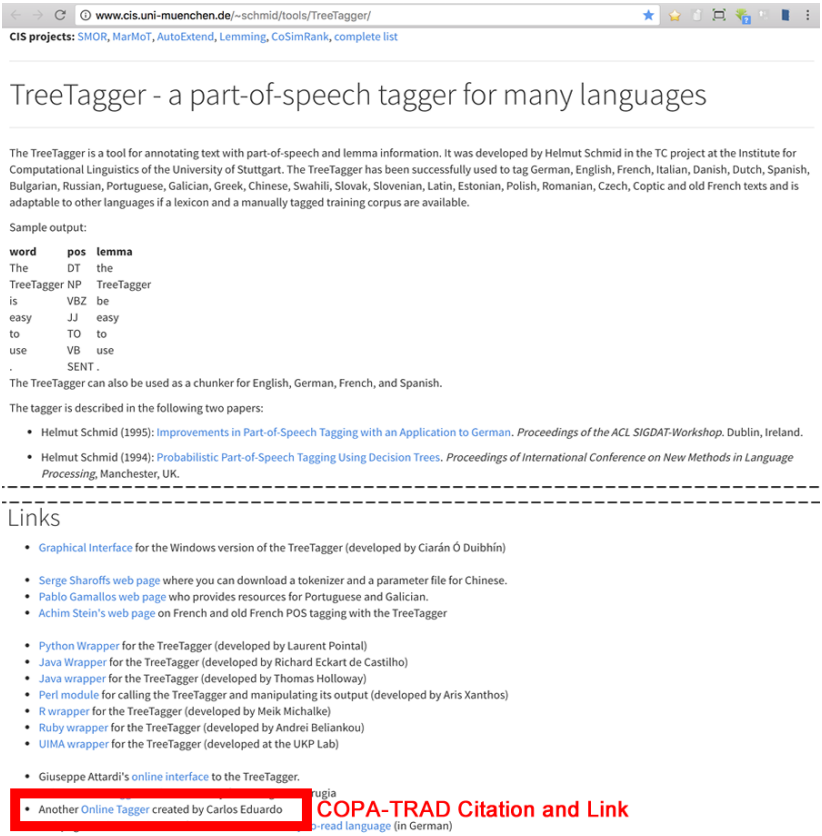
DOCUMENTAÇÃO TÉCNICA EM DEPÓSITO SOB SIGILO ATÉ 05/10/2026.

A exclusividade de comercialização deste programa de computador não tem a abrangência relativa à exclusividade de fornecimento estatuída pelo art.25, I, da Lei nº8.666, de 21 de Junho de 1993, para fins de inexigibilidade de licitação para compras pelo poder público.

Expedido em 04 de abril de 2017

Assinado digitalmente por:
Julio Cesar Castelo Branco Reis Moreira
Diretor de Patentes, Programas de Computador e Topografia de Circuitos Integrados

APPENDIX F
(COPA-TRAD citation on TreeTagger Official Website)



www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

CIS projects: SMOR, MarMoT, AutoExtend, Lemming, CoSimRank, complete list

TreeTagger - a part-of-speech tagger for many languages

The TreeTagger is a tool for annotating text with part-of-speech and lemma information. It was developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart. The TreeTagger has been successfully used to tag German, English, French, Italian, Danish, Dutch, Spanish, Bulgarian, Russian, Portuguese, Galician, Greek, Chinese, Swahili, Slovak, Slovenian, Latin, Estonian, Polish, Romanian, Czech, Coptic and old French texts and is adaptable to other languages if a lexicon and a manually tagged training corpus are available.

Sample output:

word	pos	lemma
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

The TreeTagger can also be used as a chunker for English, German, French, and Spanish.

The tagger is described in the following two papers:

- Helmut Schmid (1995): Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.
- Helmut Schmid (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Links

- Graphical Interface for the Windows version of the TreeTagger (developed by Ciarán Ó Duibhín)
- Serge Sharoffs web page where you can download a tokenizer and a parameter file for Chinese.
- Pablo Gamallos web page who provides resources for Portuguese and Galician.
- Achim Stein's web page on French and old French POS tagging with the TreeTagger
- Python Wrapper for the TreeTagger (developed by Laurent Pointal)
- Java Wrapper for the TreeTagger (developed by Richard Eckart de Castilho)
- Java wrapper for the TreeTagger (developed by Thomas Holloway)
- Perl module for calling the TreeTagger and manipulating its output (developed by Aris Xanthos)
- R wrapper for the TreeTagger (developed by Meik Michalke)
- Ruby wrapper for the TreeTagger (developed by Andrei Beliankou)
- UIMA wrapper for the TreeTagger (developed at the UKP Lab)
- Giuseppe Attardi's online interface to the TreeTagger.
- Another Online Tagger created by Carlos Eduardo

COPA-TRAD Citation and Link

-read language (in German)

APPENDIX G

(List of institutions with registered users on COPA-TRAD platform)

AIX-MARSEILLE UNIVERSITÉ
BEIJING NORMAL UNIVERSITY
DEBRECENI EGYETEM
ECOLE D'INGÉNIEURS AÉRONAUTIQUE ET SPATIALE TOULOUSE
FACULDADE INTEGRADA BRASIL AMAZÔNIA - FIBRA
FACULDADES METROPOLITANAS UNIDAS
INSTITUTO FEDERAL DO MARANHÃO
INSTITUTO NACIONAL DE EDUCAÇÃO DE SURDOS
KATHOLIEKE UNIVERSITEIT LEUVEN
NORTH SOUTH UNIVERSITY
PONTIFÍCIA UNIVERSIDADE CATÓLICA SAO PAULO
SHANGHAI JIAO TONG UNIVERSITY
SS. CYRIL AND METHODIUS UNIVERSITY IN SKOPJE
TECHNISCHE UNIVERSITÄT DRESDEN
UNIVERSIDAD DE SANTIAGO DE COMPOSTELA
UNIVERSIDAD DE VALLADOLID
UNIVERSIDADE DE BRASÍLIA
UNIVERSIDADE DE COIMBRA
UNIVERSIDADE DE SÃO PAULO
UNIVERSIDADE DO ESTADO DE SANTA CATARINA
UNIVERSIDADE DO ESTADO DO RIO DE JANEIRO
UNIVERSIDADE ESTADUAL DE LONDRINA
UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
UNIVERSIDADE FEDERAL DA GRANDE DOURADOS
UNIVERSIDADE FEDERAL DA PARAÍBA
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
UNIVERSIDADE FEDERAL DE MINAS GERAIS
UNIVERSIDADE FEDERAL DE SANTA CATARINA
UNIVERSIDADE FEDERAL DE SÃO CARLOS
UNIVERSIDADE FEDERAL DO CEARÁ
UNIVERSIDADE FEDERAL DO PARÁ
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
UNIVERSIDADE FEDERAL DO TRIÂNGULO MINEIRO

UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI
UNIVERSITÀ DI PERUGIA
UNIVERSITÄT ZÜRICH
UNIVERSITE DE LORRAINE
UNIVERSITÉ DE NANTES
UNIVERSITÉ RENNES
UNIVERSITY OF ADELAIDE
UNIVERSITY OF BIRMINGHAM
UNIVERSITY OF GENEVA
UNIVERSITY OF PESHAWAR
UNIVERSITY OF REGENSBURG
UNIVERSITY OF TAMPERE

APPENDIX H

(COPA-TRAD quantitative information for each text, during the period of this thesis. TTR: Type/Token Ratio | SD: Standard Deviation | STTR: Standardized Type/Token Ratio)

Title	Type	Token	TTR	TTR(%)	SD (σ)	STTR (/1000)
Harry Potter and the Chamber of Secrets	7057	85875	0.0821776	8.21776%	88.8374	0.484
Harry Potter e a Câmara Secreta	9617	89543	0.107401	10.7401%	79.3679	0.507
Harry Potter and the Philosopher's Stone	5885	78374	0.0750887	7.50887%	88.0948	0.433
Harry Potter e a Pedra Filosofal	8618	78645	0.109581	10.9581%	71.9721	0.482
Harry Potter and the Prisoner of Azkaban	7876	108498	0.0725912	7.25912%	115.9521	0.476
Harry Potter e o Prisioneiro de Azkaban	10613	116786	0.0908756	9.08756%	107.6824	0.4712

Artemis Fowl	8062	56658	0.142292	14.2292%	56.3244	0.5021
Artemis Fowl o Menino Prodígio do Crime	8502	58282	0.145877	14.5877%	51.3908	0.548
Artemis Fowl: The Arctic Incident	7715	58550	0.131768	13.1768%	59.4926	0.506
Artemis Fowl Uma Aventura no Ártico	8439	61379	0.13749	13.749%	55.501	0.532
Artemis Fowl The Eternity Code	8242	66398	0.12413	12.413%	63.4854	0.4941
Artemis Fowl O Código Eterno	8928	68603	0.13014	13.014%	60.0477	0.539
Revolting Rhymes	1221	4137	0.295141	29.5141%	10.4368	0.4383
Historinhas em versos perversos	1389	4270	0.325293	32.5293%	9.623	0.5228
Cuentos en verso para niños perversos	1562	4070	0.383784	38.3784%	9.1235	0.5419
Carta das Nações Unidas	1481	8491	0.17442	17.442%	22.3194	0.407
Charte des Nations Unies	1499	8752	0.171275	17.1275%	25.4846	0.3789

Charter of the United Nations	1185	8766	0.135181	13.5181%	37.621	0.3213
Protocolo de genebra sobre proibição do emprego na guerra de gases asfixiantes, tóxicos ou similares e de meios bacteriológicos de guerra	171	336	0.508929	50.8929%	2.0996	0
Protocole concernant la prohibition d'emploi à la guerre de gaz asphyxiants, toxiques ou similaires et de moyens bactériologiques	172	340	0.505882	50.5882%	2.2974	0
Protocol for the prohibition of the use in war of asphyxiating, poisonous or other gases, and of bacteriological methods of warfare	151	368	0.410326	41.0326%	4.529	0
Convenção sobre a Escravatura	431	1182	0.364636	36.4636%	5.0425	0.403
Convention relative à l'esclavage	449	1238	0.362682	36.2682%	5.4376	0.405
Declaração Universal dos Direitos Humanos	585	1717	0.340711	34.0711%	7.3242	0.404

Déclaration Universelle des Droits de l'Homme	650	1956	0.332311	33.2311%	8.5267	0.41
Slavery Convention	397	1303	0.304682	30.4682%	8.7534	0.361
Universal Declaration of Human Rights	537	1787	0.300504	30.0504%	9.6109	0.376
Convenção das Nações Unidas Contra a Corrupção	2333	18530	0.125904	12.5904%	39.4255	0.3548
Convention des Nations Unies contre la corruption	2413	18746	0.128721	12.8721%	38.1591	0.3567
Convenção sobre os Direitos Políticos da Mulher	275	777	0.353925	35.3925%	4.5076	0
Convention sur les droits politiques de la femme	284	816	0.348039	34.8039%	5.0862	0
Aligned Hansards of the 36th Parliament of Canada	15042	491590	0.0305987	3.05987%	427.7396	0.394

Débats (Hansard) Aligné - 36e Parlement du Canada	22392	521027	0.0429767	4.29767%	322.3643	0.463
Convention on the Political Rights of Women	257	780	0.329487	32.9487%	6.6557	0
United Nations Convention against Corruption	1837	18112	0.101424	10.1424%	50.1252	0.3128
Convenção relativa ao Tratamento dos Prisioneiros de Guerra	1880	7894	0.238156	23.8156%	17.1812	0.4551
Convention relative au traitement des prisonniers de guerre	1811	8007	0.226177	22.6177%	20.5021	0.4257
Convenção de Genebra Relativa ao Estatuto dos Refugiados	1184	5309	0.223018	22.3018%	14.294	0.3774
Convention de Genève relative au statut des réfugiés	1210	5614	0.215533	21.5533%	14.3269	0.3832
Refugee Convention relating to the Status of Refugees	968	5388	0.179659	17.9659%	21.266	0.3324

Convention relative to the Treatment of Prisoners of War	1508	8211	0.183656	18.3656%	28.7564	0.3829
Protocolo de Quioto à Convenção-Quadro das Nações Unidas sobre Mudança do Clima	1235	8511	0.145106	14.5106%	24.9668	0.327
Protocole de kyoto à la convention-cadre des nations unies sur les changements climatiques	1448	9236	0.156778	15.6778%	26.2422	0.3355
Convenção Internacional sobre a Supressão de Atentados Terroristas com Bombas	956	3678	0.259924	25.9924%	11.1633	0.386
Convention internationale pour la répression des attentats terroristes à l’explosif	993	3975	0.249811	24.9811%	11.904	0.3697
Convenção sobre a Proibição do Desenvolvimento, Produção, Estocagem e Uso de Armas Químicas	1776	13998	0.126875	12.6875%	31.8843	0.3997

e sobre a Destruição das Armas Químicas Existentes no Mundo						
Convention sur l'interdiction de la mise au point, de la fabrication, du stockage et de l'emploi des armes chimiques et sur leur destruction	1864	15161	0.122947	12.2947%	36.0579	0.3483
Convenção-Quadro das Nações Unidas Sobre Mudança do Clima	1480	8305	0.178206	17.8206%	21.6268	0.38
Convention-cadre des nations unies sur les changements climatiques	1688	8873	0.19024	19.024%	23.7529	0.3884
Convenção sobre Diversidade Biológica	1444	9056	0.159452	15.9452%	23.3727	0.3606
Convention sur la diversité biologique	1657	9877	0.167764	16.7764%	25.456	0.369
Convenção sobre os direitos da criança	1538	7489	0.205368	20.5368%	19.1048	0.396
Convention internationale des droits de l'enfant	1558	7968	0.195532	19.5532%	21.9493	0.3893

Convenção de viena para a proteção da camada de ozônio	1255	6036	0.207919	20.7919%	16.9532	0.3103
Convention de Vienne pour la protection de la couche d'ozone	1282	6402	0.20025	20.025%	19.5006	0.362
Convenção Contra a Tortura e Outros Tratamentos ou Penas Cruéis, Desumanos ou Degradantes	1073	5047	0.212602	21.2602%	14.1661	0.3695
Convention contre la torture et autres peines ou traitements cruels, inhumains ou dégradants	1076	5272	0.204097	20.4097%	14.966	0.3596
Tratado Sobre a Não-Proliferação de Armas Nucleares	602	2130	0.282629	28.2629%	8.8147	0.356
Traité sur la non-prolifération des armes nucléaires	626	2324	0.269363	26.9363%	9.5284	0.3575
Convenção Internacional sobre a Eliminação de Todas as Formas de Discriminação Racial	1102	4612	0.238942	23.8942%	12.8316	0.3923

Convention internationale sur l'élimination de toutes les formes de discrimination raciale	1118	4987	0.224183	22.4183%	15.7405	0.381
Kyoto protocol to the united nations framework convention on climate change	1066	8530	0.124971	12.4971%	36.1413	0.2785
Measures to eliminate international terrorism	818	3810	0.214698	21.4698%	15.6712	0.3267
Convention on the prohibition of the development, production, stockpiling and use of chemical weapons and on their destruction	1447	13909	0.104033	10.4033%	44.9875	0.3085
Convention on Climate Change	1269	8192	0.154907	15.4907%	30.1886	0.3484
convention on biological diversity	1245	9134	0.136304	13.6304%	32.2396	0.3291
Convention on the Rights of the Child	1244	7473	0.166466	16.6466%	28.622	0.3172

The Vienna Convention for the Protection of the Ozone Layer	1008	5860	0.172014	17.2014%	22.8833	0.3146
Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment	862	5101	0.168986	16.8986%	22.3221	0.3095
Treaty on the non-proliferation of nuclear weapons	534	2130	0.250704	25.0704%	12.7885	0.3325
International Convention on the Elimination of All Forms of Racial Discrimination	913	4683	0.19496	19.496%	21.024	0.336
Artemis Fowl: El Cubo B	11154	74245	0.150232	15.0232%	66.6212	0.524

APPENDIX I
(COPA-TRAD stopwords lists)

English	Portuguese	French	German	Italian	Spanish
Zero	vos	Vu	zur	vostro	yo
z	vós	vous	zum	volte	y
yourselves	vocês	votre	zu	voi	voy
yourself	você	vont	wohin	vai	vosotros
yours	vir	voient	woher	va	vosotras
youre	vindo	tu	wo	uno	verdadero
your	vez	trop	wirst	una	verdadera
youngest	verdadeiro	très	wird	un	verdad
younger	verdade	tout	wir	ultimo	vaya
young	ver	tous	wieso	triplo	van
youd	vendo	ton	wieder	tre	vamos
you've	veja	tes	wie	tra	valor
you're	valor	tels	weshalb	terzo	vais
you'll	usar	tellement	werdet	tempo	va
you'd	usa	tandis	werden	te	uso
you	uns	ta	werde	tanto	usas
yet	umas	sur	wer	sulla	usar
yes	uma	sujet	wenn	sul	usan
years	um	soyez	weitere	subito	usamos
year	últimos	sous	weiter	su	usais
y	último	sont	was	stesso	usa

x	últimas	son	warum	stato	unos
www	última	sien	wann	stati	uno
wouldn't	tudo	si	vor	sotto	unas
would've	tuas	seulement	von	soprattutto	una
would	tua	ses	vom	sopra	un
world	tu	sans	unter	sono	ultimo
works	trabalho	sa	unsere	solo	tuyo
working	trabalhar	qui	unser	siete	tras
worked	todos	quels	und	siamo	trabajo
work	todo	quelles	über	sia	trabajas
words	todavia	quelle	sowie	sette	trabajar
wonder	todas	quel	soweit	senza	trabajan
won't	toda	que	sonst	sembrava	trabajamos
without	tivéssemos	quand	sollt	sembra	trabajais
within	tivessem	pourquoi	sollst	sei	trabaja
with	tivesse	pour	sollen	secondo	todo
wish	tivermos	plupart	soll	sara	tienen
willing	tiverem	peut	sind	rispetto	tiene
will	tivéramos	peu	sie	quinto	tiempo
widely	tiveram	pas	sich	quindi	tengo
why's	tivera	parce	seine	qui	tener
why'll	tiver	par	sein	questo	tenemos
why'd	tivemos	ou	seid	quello	teneis
why	tive	où	oder	quattro	también
whose	tipo	nous	nun	quasi	sus

whos	tínhamos	notre	nicht	quarto	su
whomever	tinham	nommés	nein	qua	soy
whom	tinha	ni	nachdem	promesso	somos
whole	tido	mot	nach	primo	solo
whoever	ti	mon	mußt	poco	solamente
whod	teve	moins	müßt	piu	sois
who's	teus	mine	musst	persone	sobre
who'll	teu	mes	müssen	pero	sin
who'd	teríamos	même	muß	peggio	siendo
who	teriam	mais	mit	otto	si
whither	teria	maintenant	meine	ora	ser
whim	teremos	ma	mein	oltre	sabes
while	terei	leur	machen	o	saber
which	terão	les	könnt	nuovo	saben
whether	terá	le	können	nuovi	sabemos
wherever	ter	la	kannst	nove	sabeis
whereupon	tentei	là	kann	nostro	sabe
wheres	tente	juste	jetzt	nome	quien
wherein	tentaram	je	jenes	noi	que
whereby	tentar	ils	jener	no	puedo
whereas	tenho	il	jedes	nella	pueden
whereafter	tenhamos	ici	jeder	nei	puede
where's	tenham	hors	jeden	molto	primero
where'll	tenha	font	jedem	molti	porque
where'd	tendo	fois	jede	molta	por qué

where	tempo	faites	ja	meglio	por
whenever	temos	fait	ist	me	podrias
whence	tem	eu	in	ma	podrian
when's	tém	être	im	lungo	podriamos
when'll	têm	étions	ihre	lui	podriais
when'd	te	été	ihr	loro	podria
when	tampouco	état	ich	lo	poder
whats	também	étaient	hinter	lei	podemos
whatever	talvez	et	hier	le	podeis
what's	tal	est	hattet	lavoro	pero
what'll	suas	essai	hattest	la	para
what'd	sua	encore	hatten	io	otro
what	sr	en	hatte	invece	nosotros
weren't	sou	elles	für	indietro	nos
were	somos	elle	eure	il	muy
went	somente	du	euer	ho	muchos
wells	sobre	dos	es	hanno	modo
well	sob	donc	er	hai	mio
welcome	só	doit	eines	ha	mientras
wed	sido	devrait	einer	giu	los
we've	si	des	einen	gente	lo
we're	seus	depuis	einem	fra	las
we'll	seu	dehors	eine	fino	largo
we'd	seríamos	dedans	ein	fine	la
we	seriam	début	durch	fare	ir

ways	seria	dans	du	ecco	intento
way	seremos	comment	dort	e	intentas
wasn't	serei	comme	doch	due	intentar
was	serão	ci	dieses	doppio	intentan
wants	será	chaque	dieser	di	intentamos
wanting	ser	ceux	dies	devo	intentais
wanted	sendo	ces	die	deve	intenta
want	sempre	cela	dessen	dentro	incluso
w	sem	ce	deshalb	dello	hago
vs	sejamos	car	des	della	haces
vols	sejam	ça	der	del	hacer
vol	seja	bon	den	da	hacen
viz	se	avoir	dem	cui	hacemos
via	são	avec	deine	cosa	haceis
very	saber	avant	dein	consecutivo	hace
various	quieto	autre	dass	consecutivi	ha
value	quem	aussi	das	con	gueno
v	queda	aucuns	daß	comprare	fuimos
usually	que	au	darum	cinque	fui
using	quê	alors	daher	chi	fueron
uses	quantos		dadurch	che	fue
usefulness	quanto		da	buono	fin
usefully	quando		bist	ben	estoy
useful	qualquer		bis	avevano	estan
used	qual		bin	aveva	estamos

use	quais		bei	avere	estais
us	pude		aus	ancora	estado
ups	próprios		auf	anche	estaba
upon	próprio		auch	altro	esta
up	próprias		an	altri	es
unto	própria		am	altre	eres
until	primeiros		als	allora	eras
unlikely	primeiro		aber	allo	eran
unlike	primeiro			alla	eramos
unless	povo			al	era
unfortunately	poucos			ai	entre
under	pouco			adesso	entonces
un	poucas			a	encima
u	pouca				en
two	posso				empleo
twice	porque				empleas
twas	porém				emplear
turns	por				emplean
turning	pois				empleais
turned	podiam				ellos
turn	podia				ellas
ts	poderiam				el
trying	poderia				dos
try	poderá				donde
truly	poder				desde

tries	podendo				dentro
tried	pode				de
towards	pôde				cuando
toward	pessoas				cual
took	perante				consigues
too	per				consiguen
together	pequenos				consigue
today	pequeno				consigo
to	pequenas				conseguir
tis	pequena				conseguimos
tip	pelos				con
til	pelo				como
thus	pelas				ciertos
thru	pela				cierto
throughout	pegar				ciertas
through	parte				cierta
throug	para				cada
three	outros				bien
thousand	outro				bastante
thoughts	outras				bajo
thought	outra				atras
thoughh	ou				arriba
though	os				aqui
thou	onde				aquellos
those	o				aquellas

thoroughly	nunca				aquel
thorough	numa				antes
this	num				ante
third	novo				ampleamos
thinks	nossos				ambos
think	nosso				algunos
things	nossas				alguno
thing	nossa				algunas
theyre	nos				alguna
theyd	nós				algún
they've	nome				a
they're	no				
they'll	ninguém				
they'd	nestas				
they	nesta				
these	nessas				
thereupon	nessa				
thereto	nenhum				
theres	nem				
therere	nas				
thereof	não				
therein	na				
therefore	muitos				
thered	muito				
thereby	muitas				

thereafter	muita				
there've	minhas				
there's	minha				
there'll	meus				
there	meu				
thence	mesmos				
then	mesmo				
themselves	mesmas				
them	mesma				
theirs	me				
their	mas				
the	mais				
thats	maiorias				
that've	maioria				
that's	lo				
that'll	ligado				
that	lhes				
thanx	lhe				
thanks	lá				
thank	lá				
than	já				
th	isto				
tends	isso				
tell	irá				
taking	ir				

taken	inicio				
take	iniciar				
t's	houvéssemos				
t	houvessem				
sure	houvesse				
sup	houvermos				
suggest	houvermos				
sufficiently	houveriam				
such	houveres				
successfully	houverem				
substantially	houvéramos				
sub	houvéramos				
strongly	houveram				
stop	houveram				
still	houvera				
states	houvera				
state	houvera				
specifying	houver				
specify	houvemos				
specified	houve				
specifically	horas				
sorry	hei				
soon	havia				
somewhere	havemos				
somewhat	hão				

sometimes	hajamos				
sometime	hajam				
something	haja				
somethan	há				
someone	grandes				
somehow	grande				
somebody	fui				
some	fôssemos				
so	fossem				
smallest	fosse				
smaller	formos				
small	forem				
slightly	fôramos				
six	foram				
since	fora				
similarly	for				
similar	fomos				
significantly	foi				
significant	fim				
sides	fez				
side	feitos				
shows	feito				
showns	feitas				
shown	feita				
showing	fazia				

showed	fazer				
show	fazendo				
shouldn't	faz				
should've	fará				
should	eu				
shes	estou				
shed	estivéssemos				
she's	estivessem				
she'll	estivesse				
she'd	estivermos				
she	estiverem				
shan't	estivéramos				
shall	estiveram				
several	estivera				
seven	estiver				
seriously	estivemos				
serious	estive				
sent	esteve				
sensible	estes				
selves	estejamos				
self	estejam				
sees	esteja				
seen	este				
seems	este				
seeming	estávamos				

seemed	estavam				
seem	estava				
seeing	estas				
see	estará				
section	estar				
seconds	estão				
secondly	estamos				
second	estado				
sec	esta				
says	está				
saying	esta				
say	esses				
saw	esse				
same	essas				
said	essa				
s	éramos				
run	eram				
rooms	era				
room	entre				
right	então				
results	enquanto				
resulting	em				
resulted	eles				
respectively	ele				
research	elas				

relatively	ela				
related	e				
regards	é				
regardless	é				
regarding	dos				
refs	dois				
ref	do				
recently	dizer				
recent	dizem				
reasonably	diz				
really	dito				
readily	disto				
re	disso				
rd	disse				
rather	direita				
ran	deviam				
r	devia				
qv	deveriam				
quite	deveria				
quickly	deverão				
que	deverá				
q	dever				
puts	devendo				
put	devem				
provides	deve				

proud	destes				
promptly	deste				
problems	destas				
problem	desta				
probably	desses				
primarily	desse				
previously	dessas				
presumably	dessa				
presents	desligado				
presenting	desde				
presented	depois				
present	dentro				
predominantly	deles				
pp	dele				
potentially	delas				
possibly	dela				
possible	debaixo				
poorly	de				
points	das				
pointing	daqueles				
pointed	daquele				
point	da				
plus	corrente				
please	contudo				
places	contra				

placed	conhecido				
place	comprido				
perhaps	como				
per	com				
past	coisas				
parts	coisa				
parting	cima				
particularly	caminho				
particular	cada				
parted	bom				
part	bem				
pages	através				
page	atrás				
p	até				
own	as				
owing	às				
overall	aquilo				
over	aqui				
outside	aqueles				
out	aquele				
ourselves	aquelas				
ours	aquela				
our	após				
ought	apontar				
otherwise	aos				

others other orders ordering ordered order ord or opens opening opened open onto only ones one once on omitted oldest older old okay ok oh	ao antes ante amplos amplo amplas ampla ambos ali alguns algumas algumas alguma algum alguém ainda agora acerca a à -				

often off of obviously obtained obtain o numbers number nowhere now novel nothing noted not nos normally nor noone nonetheless none non nobody no ninety					

nine next newest newer new nevertheless never neither needs needing needed need necessary necessarily nearly near nd nay namely name na n myself my mustn't					

must've must mug much mrs mr mostly most moreover more ml miss million mightn't might've might mg merely men members member meanwhile meantime means mean					

me maybe may many man making makes make mainly made m ltd looks looking look longest longer long little line likely liked like lets let's					

let lest less least latterly latter latest later lately last largely large l knows known know knew km kind kg kept keeps keep k just					

j itself its itd it's it'll it'd it isn't is inward invention into interests interesting interested interest instead insofar inner information indicates indicated indicate index					

indeed inc inasmuch in important importance immediately immediate im ignored if ie id i've i'm i'll i'd i hundred however howbeit how's how'll how'd how					

hopefully home hither his himself him highest higher high hid hi hes herself hers hereupon heres herein hereby hereafter here's here her hence help hello					

hed he's he'll he'd he having haven't have hasn't has hardly happens hadn't had h groups grouping grouped group greetings greatest greater great gotten got					

goods good gone going goes go giving gives given give getting gets get generally general gave g furthers furthermore furthering furthered further fully full from					

four found forth formerly former for follows following followed fix five first finds find fifth ff few felt far facts fact faces face f except					

example exactly ex everywhere everything everyone everybody every ever evenly even etc et-al et especially entirely enough ends ending ended end elsewhere else either eighty					

eight eg effect edu ed early each e during due downwards downs downing downed down done don't doing doesn't does do differently different differ didn't					

did despite described definitely dear date d currently course couldnt couldn't could've could corresponding contains containing contain considering consider consequently concerning comes come com co					

clearly clear changes certainly certain causes cause cases case cant cannot can't can came ca c's c'mon c by but briefly brief both biol big					

beyond between better best besides beside below believe beings being behind begins beginnings beginning begin began beforehand before been becoming becomes become because became be					

backs backing backed back b awfully away available auth at associated asks asking asked ask aside as around arise arent aren't aren areas area are					

approximately appropriate appreciate appear apparently apart anywhere anyways anyway anything anyone anymore anyhow anybody any another announce and an amongst among am always although also					

already along alone almost allows allow all ain't ah against again afterwards after affects affecting affected adj added actually act across accordingly according accordance abst					

above about able a's a - 've 'twas 'tis 'll					

APPENDIX J

(Monitor Corpus of Literary Brazilian-Portuguese Books compiled to support Portuguese translations in forthcoming versions of COPA-TRAD)

File/Group/Title	Type	Token	TTR	STTR	Ave w Lgth	1 L word	2 L word	3 L word	4 L word
TOTAL	108462	378230	2.87	48.71	4.55	444354	602508	596803	385340
1808 - Laurentino Gomes.txt	14444	106165	13.61	51.6	4.81	12123	19844	12793	9714
1822 - Laurentino Gomes.txt	14156	102474	13.81	52.2	4.88	11906	19564	11139	8785
50 Anos a Mil - Lobao.txt	21934	186951	11.73	51.3	4.57	20091	31230	29623	19358
A CamanaVaranda - Regina Navarro Lins.txt	15936	139066	11.46	49.47	4.79	16965	22330	20114	13408
A Carteira - Machado de Assis.txt	582	1425	40.84	44.4	4.25	243	187	244	134
A Igreja do Diabo - Machado de Assis.txt	1225	2876	42.59	51.4	4.48	342	446	490	283
A Logica da Emocao - Manoelita Dias dos Santos.txt	8794	64938	13.54	49.44	5.07	7278	9618	9045	5966

A Mao e a Luva - Machado de Assis.txt	5703	35329	16.14	46.78	4.37	4680	5277	6461	3442
A Querela do Estatismo - Machado de Assis.txt	3294	16209	20.32	45.89	3.17	1724	5299	4355	2059
A Vida Como Ela E_. - Saraiva de Bolso - Rodrigues_ Nelson.txt	12964	143612	9.03	48.87	4.4	20366	21323	23404	14647
A_Hora_da_Estrela_[Clarice_Lispector].txt	5317	27787	19.13	45.91	4.4	3249	4325	5238	3054
Adao e Eva - Machado de Assis.txt	816	2052	39.77	46.75	4.2	306	324	357	250
Americanas - Machado de Assis.txt	4773	18608	25.65	54.69	4.32	2504	2986	2500	2035
Amor e Prosa, Sexo e Poesia - Arnaldo Jabor.txt	7712	34485	22.36	50.62	4.57	3548	5677	5486	4072
As Bodas de Luiz Duarte - Machado de Assis.txt	1805	5833	30.94	48.46	4.51	765	936	840	548
As Cariocas - Sergio Porto.txt	7397	50703	14.59	45.48	4.39	7350	7351	8421	5227
As Esganadas - Soares_ Jo.txt	10768	51027	21.1	53.9	4.74	6530	8034	6464	4656
As MentirasQueosHomensCont	5178	26509	19.53	44.15	4.41	2969	3871	4774	3389

am - Luis Fernando Verissimo.txt									
As Valkirias - Paulo Coelho.txt	5473	40534	13.5	45.48	4.48	5426	5163	6517	4311
Assassinatos Na Academia Brasileira De Letras - Jo Soares.txt	10857	50541	21.48	55.22	4.91	5317	7755	7223	4291
Auto da Compadecida - Ariano Suassuna.txt	3508	23609	14.86	38.48	4.31	2623	3332	3951	3459
BanheiroFeminino - Andrea Cals e Ricardo Grynszpan.txt	3670	12052	30.45	50.87	4.46	1245	1971	2123	1431
Bau de Espantos - Mario Quintana.txt	4024	14706	27.36	50.3	4.65	1445	2562	1953	1525
Biografia - Machado de Assis.txt	2950	9159	32.21	51.26	4.72	1190	1709	1102	847
Brida - Paulo Coelho.txt	6419	53057	12.1	45.24	4.52	6261	7275	9080	5243
Capitaes Do Brasil - Eduardo Bueno.txt	12412	92051	13.48	50.23	4.67	9142	17887	11770	9629
Capitaes de Areia - Jorge Amado.txt	8119	76307	10.64	42.99	4.35	7994	11684	13488	9254
Casa Velha - Machado de Assis.txt	4037	22671	17.81	46.51	4.35	2608	3596	4314	2376

Casada e Viuva - Machado de Assis.txt	2812	9416	29.86	48.82	4.69	1153	1536	1266	1004
CincoMinutos - Jose de Alencar.txt	3539	15097	23.44	48.57	4.26	1596	2751	2833	1443
Clube dos Anjos - Luis Fernando Verissimo.txt	5210	28018	18.6	46.01	4.54	3046	4348	4699	2849
Comedias da Vida Privada - Luis Fernando Verissimo.txt	989	2500	39.56	48.1	4.52	259	360	437	290
Comediaspara se lernaescola - Luis Fernando Verissimo.txt	4724	20028	23.59	46.46	4.49	2283	2945	3502	2470
Contos - Machado de Assis.txt	4601	20600	22.33	49.7	4.46	2358	3279	3464	2064
ContosFluminenses - Machado de Assis.txt	7851	56959	13.78	46.57	4.44	6448	9049	10369	5577
Crisalidas - Machado de Assis.txt	4148	15537	26.7	50.76	4.25	2262	2586	2178	1742
CriticasTeatrais - Machado de Assis.txt	4668	20980	22.25	48.24	4.63	2701	3579	3106	1909
Dom Casmurro - Machado de Assis.txt	8915	67894	13.13	47.26	4.28	8372	11007	12605	6833
Era no Tempo do Rei - Ruy Castro.txt	10834	63505	17.06	49.4	4.54	7355	10428	9845	6475

Esau e Jaco - Machado de Assis.txt	9755	73768	13.22	48.69	4.38	9525	11279	12990	6905
EspumasFlutuantes - Castro Alves.txt	5146	20258	25.4	53.92	4.33	2434	3514	2756	2218
Falaserio, amor! - Thalita Reboucas.txt	5388	37096	14.52	43.15	4.13	5308	5382	6976	4694
Falaserio, mae! - Thalita Reboucas.txt	5893	38671	15.24	45.33	4.31	3702	5715	7613	5424
Falaserio, professor! - Thalita Reboucas.txt	6210	35780	17.36	47.07	4.49	3500	5196	6312	4641
Falenas - Machado de Assis.txt	4865	18790	25.89	53.59	4.3	2419	3088	2561	2109
FazendaModelo - Chico Buarque de Holanda.txt	7358	25389	28.98	54.03	4.76	2801	3797	3942	2400
Helena - Machado de Assis.txt	8248	59385	13.89	48.39	4.49	8063	9242	9540	5034
Historias da meia - Machado de Assis.txt	6889	42899	16.06	47.57	4.42	5619	6715	7218	4001
Historias do Analista de Bage - Luis Fernando Verissimo.txt	2899	11512	25.18	47.95	4.28	1371	2022	2122	1338
Historiassem data - Machado de Assis.txt	8570	52037	16.47	48.81	4.45	6430	8253	8875	5021

Ideias do Canario - Machado de Assis.txt	798	1820	43.85	50.3	4.53	216	273	306	179
Infancia - Graciliano Ramos.txt	12458	57580	21.64	58.44	4.93	6297	10652	6882	3938
LeiteDerramado - Chico Buarque de Holanda.txt	8099	38681	20.94	51.27	4.51	3489	6964	6717	3796
MaisComedias Para LernaEscola - Luis Fernando Verissimo.txt	4653	20415	22.79	46.14	4.33	2992	3036	3451	2289
MarchaFunebre - Machado de Assis.txt	1114	2781	40.06	51.4	4.35	345	417	499	264
Mare Vermelha - Carlos Rocha.txt	14261	158938	8.97	49.1	4.71	19190	21338	23384	15115
Memorial de Aires - Machado de Assis.txt	6384	51878	12.31	47.04	4.23	6724	8479	9487	5300
MemoriasPostumas de Bras Cubas - Machado de Assis.txt	10196	63139	16.15	49.57	4.41	7338	10709	10959	5928
MentesPerigosas - O Psicopata Mora aoLado - Ana Beatriz Barbosa Silva.txt	8445	42922	19.68	51.44	5.07	4194	6714	6065	4124
Mil diasemVeneza - Marlena de Blasi.txt	11164	75343	14.82	50.37	4.5	8226	12591	11384	8274

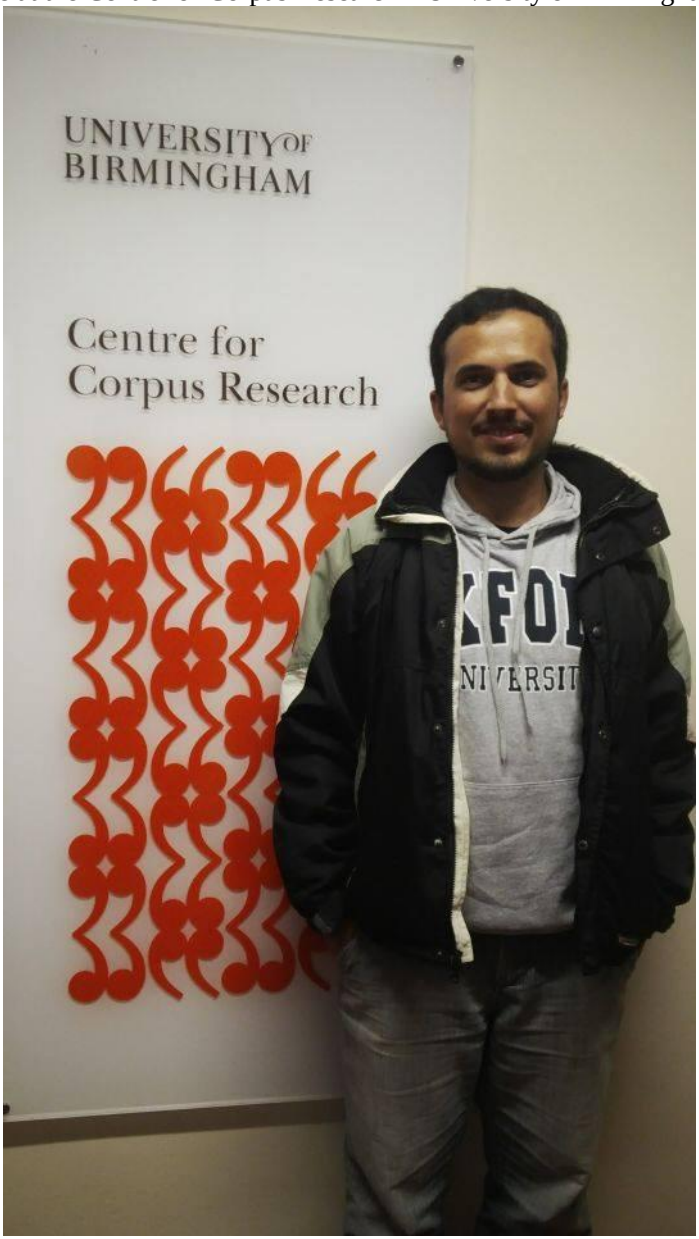
MorangosMofados - Caio Fernando Abreu.txt	8114	43123	18.82	49.61	4.59	3859	6695	6957	4839
Na Margem do Rio Piedra EuSentei e Chor - Paulo Coelho.txt	5077	37987	13.37	45.9	4.35	5336	5252	6209	4114
Naufragos, Traficantes e Degredados - Eduardo Bueno.txt	7933	50241	15.79	48.93	4.7	5064	9714	6291	5089
Noite de Almirante - Machado de Assis.txt	920	2570	35.8	46.3	4.37	289	378	500	289
O Alienista - Machado de Assis.txt	4190	16964	24.7	50.16	4.65	2150	2567	2626	1505
O Alquimista - Paulo Coelho.txt	4619	37752	12.24	42.7	4.42	5242	5151	5752	3670
O Demonio e a Srta. Prym - Paulo Coelho.txt	6258	44131	14.18	47.57	4.53	5558	5919	7426	4781
O Diario de Anne Frank - Laurentino Gomes.txt	9015	80401	11.21	47.04	4.33	9220	13314	14235	8816
O Diario de um Mago - Paulo Coelho.txt	7380	62346	11.84	45.83	4.48	7536	9522	9944	6265
O Diplomatico - Machado de Assis.txt	1361	3645	37.34	50.1	4.41	425	576	588	392
O Emprestimo - Machado de Assis.txt	1052	2715	38.75	48.6	4.49	280	433	519	263

O EncontroMarcado - Fernando Sabino.txt	9827	86553	11.35	46.01	4.36	12339	12112	14266	9963
O Enfermeiro - Machado de Assis.txt	1285	3380	38.02	49.17	4.38	348	628	563	358
O Espelho - Machado de Assis.txt	1248	3344	37.32	48.3	4.53	362	533	549	366
O Grande Mentecapto - Fernando Sabino.txt	10007	55172	18.14	50.55	4.66	5754	9211	8473	5656
O Manual do Guerreiro da Luz - Paulo Coelho.txt	2965	15254	19.44	44.91	4.56	1736	2248	2767	1665
O Monte Cinco - Paulo Coelho.txt	6321	41641	15.18	49.11	4.57	5337	5658	6336	4303
O Novo Mundo Digital - Ricardo Oliveira Neves.txt	10582	74108	14.28	49.48	5.05	7503	12528	9611	7618
O Xango de Baker Street - Jo Soares.txt	11503	66420	17.32	53.05	4.81	6895	10193	9283	6011
Ocidentais - Machado de Assis.txt	2680	7542	35.53	53.04	4.29	999	1180	1089	915
OnzeMinutos - Paulo Coelho.txt	7364	62471	11.79	47.46	4.58	5952	8906	11379	6907
OsCanibaisEstaonaSala de Jantar - Arnaldo Jabor.txt	11806	64727	18.24	49.24	4.57	7422	10785	9362	7232
PaginasRecolhidas - Machado de Assis.txt	7577	40247	18.83	49.3	4.44	4825	6575	6887	3827

PapeisAvulsos - Machado de Assis.txt	9469	56142	16.87	48.68	4.48	7598	8774	9077	5254
QuincasBorba - Machado de Assis.txt	10260	77668	13.21	48.75	4.44	8544	12208	13956	7273
Reliquias de Casa Velha - Machado de Assis.txt	6612	37255	17.75	48.26	4.41	4305	5975	6686	3605
Veronika Decide Morrer - Paulo Coelho.txt	6742	47879	14.08	48.38	4.64	5270	6673	7860	5323

APPENDIX K

(Me at the Centre for Corpus Research – University of Birmingham)



For a quarter of a century, corpus evidence was ignored, spurned and talked out of relevance, until its importance became just too obvious for it to be kept out in the cold.
Sinclair, 2004a