

UNIVERSIDADE FEDERAL DE SANTA CATARINA

**DEVELOPING ONLINE PARALLEL CORPUS-BASED  
PROCESSING TOOLS FOR TRANSLATION RESEARCH AND  
PEDAGOGY**

Carlos Eduardo da Silva  
2013



UNIVERSIDADE FEDERAL DE SANTA CATARINA  
PÓS-GRADUAÇÃO EM LETRAS/INGLÊS E LITERATURA  
CORRESPONDENTE

**DEVELOPING ONLINE PARALLEL CORPUS-BASED  
PROCESSING TOOLS FOR TRANSLATION RESEARCH AND  
PEDAGOGY**

CARLOS EDUARDO DA SILVA

Dissertação submetida à Universidade Federal de Santa Catarina em  
cumprimento parcial dos requisitos para a obtenção do grau de

MESTRE EM LETRAS

Florianópolis  
Março – 2013.

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

Silva, Carlos Eduardo da  
Developing Online Parallel Corpus-based Processing  
Tools for Translation Research and Pedagogy / Carlos  
Eduardo da Silva ; orientador, Lincoln Paulo Fernandes -  
Florianópolis, SC, 2013.  
142 p.

Dissertação (mestrado) - Universidade Federal de Santa  
Catarina, Centro de Comunicação e Expressão. Programa de Pós-  
Graduação em Letras/Inglês e Literatura Correspondente.

Inclui referências

1. Letras. 2. Translation. 3. Parallel Corpus. 4.  
Software Engineering. I. Fernandes, Lincoln Paulo. II.  
Universidade Federal de Santa Catarina. Programa de Pós-  
Graduação em Letras/Inglês e Literatura Correspondente.  
III. Título.

Esta dissertação de Carlos Eduardo da Silva, intitulada ‘Developing Online Parallel Corpus-Based Processing Tools for Translation Research and Pedagogy’, foi julgada adequada e aprovada em sua forma final, pelo programa de Pós-Graduação em Letras/Inglês e Literatura Correspondente, da Universidade Federal de Santa Catarina, para fins de obtenção do grau de

## MESTRE EM LETRAS

Área de concentração: Inglês e Literatura Correspondente  
Opção: Língua Inglesa e Linguística Aplicada

---

Dra. Susana Funck  
Coordenadora

---

Dr. Lincoln P. Fernandes  
Orientador e Presidente

---

Dr. Denilson Sell  
Examinador

BANCA EXAMINADORA:

---

Dra. Maria Lúcia Barbosa de Vasconcellos  
Examinadora

---

Dr. Markus J. Weininger  
Examinador

Florianópolis, 04 de março de 2013.



I dedicate this thesis to my mother, a brave woman, Eliane da Silva and my grandparents Seu Antônio e Dona Glorinha.





## ACKNOWLEDGEMENTS

תודה

I would like to express my sincere gratitude for the people who accompanied me during these two years of Master's and who have, somehow, contributed and supported me to finish the writing of this endeavor. I would like to mention the following people.

My wife, **Danielle Amanda**, for her love, support and encouragement to never give up on my dreams.

My supervisor, **Dr. Lincoln P. Fernandes** for the guidance throughout my M.A. course – in the development of COPA-TRAD and the writing of my dissertation. Without his support and friendship, this study would not have been possible.

My friend **David Arnan Smith**, for the friendship and support.

My former teacher at Instituto Federal de Educação (Unidade São José), **Vidomar Silva Filho**, who said during his classes in high school that one day I would study linguistics. At that time I did not believe him, but now, here I am.

My professors from PPGI and PGET whose classes have certainly contributed to my academic life.

My classmates – especially the under graduates, who have accompanied me from the first to the last year.

All the members of the research group **TraCor** (Tradução e Corpora) at UFSC.

The support team from **SeTIC** – UFSC for their patience and support provided to set up the server to receive COPA-TRAD.

A special thanks to **Mona Baker** who I had the chance to meet and present the embryo of this study and also for the insights she gave me.

Professor **Dr. William H. Fletcher** from the United States Naval Academy who gave me insightful information regarding the searching mechanism of COPA-TRAD and how to apply the Sphinx search engine in a linguistic corpora.

For all personnel at PPGI and PPGET whose assistance was invaluable. CAPES, for its financial support.

All my family.

February 20, 2013.



ABSTRACT  
**Developing Online Parallel Corpus-based Tools for Translation  
Research and Pedagogy**

Carlos Eduardo da Silva

UNIVERSIDADE FEDERAL DE SANTA CATARINA  
2013

Supervising Professor: Dr. Lincoln P. Fernandes

This study describes the key steps in developing online parallel corpus-based tools for processing COPA-TRAD ([copa-trad.ufsc.br](http://copa-trad.ufsc.br)), a parallel corpus compiled for translation research and pedagogy. The study draws on Fernandes's (2009) proposal for corpus compilation, which divides the compiling process into three main parts: corpus design, corpus building and corpus processing. This compiling process received contributions from the good development practices of Software Engineering, especially the ones advocated by Pressman (2011). The tools developed can, for example, assist in the investigation of certain types of texts and translational practices related to certain linguistic patterns such as collocations and semantic prosody. As a result of these applications, COPA-TRAD becomes a suitable tool for the investigation of empirical phenomena with a view to translation research and pedagogy.

Keywords: Translation, Parallel Corpus, Software Engineering.

Number of Pages: 142

Number of Words: 22,364



RESUMO

**Desenvolvendo Ferramentas Online com base em Corpus Paralelo  
para Pesquisa e Ensino de Tradução**

Carlos Eduardo da Silva

UNIVERSIDADE FEDERAL DE SANTA CATARINA  
2013

Orientador: Dr. Lincoln P. Fernandes

Este estudo descreve as principais etapas no desenvolvimento de ferramentas online com base em corpus para o processamento do COPA-TRAD (Corpus Paralelo de Tradução – [www.copa-trad.ufsc.br](http://www.copa-trad.ufsc.br)), um corpus paralelo compilado para a pesquisa e ensino de tradução. Para a compilação do corpus, o estudo utiliza a proposta de Fernandes (2009) que divide o processo de compilação em três etapas principais: desenho do corpus, construção do corpus e processamento do corpus. Este processo de compilação recebeu contribuições das boas práticas de desenvolvimento fornecidas pela Engenharia de Software, especialmente as que foram sugeridas por Pressman (2011). As ferramentas desenvolvidas podem, por exemplo, auxiliar na investigação de certos tipos de textos, bem como em práticas tradutórias relacionadas a certos padrões linguísticos tais como colocações e prosódia semântica. Como resultado dessas aplicações, o COPA-TRAD configura-se em uma ferramenta útil para a investigação empírica de fenômenos tradutórios com vistas à pesquisa e ao ensino de tradução.

Descritores: Corpus Paralelo, Corpus para Tradução, padrões linguísticos.

Número de páginas: 142

Número de palavras: 22.364



## TABLE OF CONTENTS

CHAPTER ONE: Introduction.....	23
1.1 The Context of Investigation.....	25
1.2 The Emergence of Corpus-based Translation Studies .....	30
1.3 The Purpose of the Study .....	32
1.4 Significance of the Study .....	33
1.5 Organization of the Study .....	34
CHAPTER TWO: Review of Literature .....	35
2. Initial Remarks .....	37
2.1 Corpus-Based Translation Studies.....	37
2.2 Corpus Typology.....	39
2.3 Fernandes' Parallel Corpus of Children's Fantasy Literature .....	45
2.4 Children's Literature, Collocation and Semantic Prosody .	45
2.5 Summing Up.....	47
CHAPTER THREE: Methodology .....	49
3. Opening remarks .....	51
3.1 Corpus Design .....	51
3.1.1 Purpose of Corpus Creation .....	54
3.1.2 Type of Corpus.....	55
3.1.3 Copyright.....	55
3.1.4 Selection of Texts .....	56
3.1.5 Corpus Creation: Shortcomings and Possible Solutions .....	57
3.1.6 Requirements Analysis .....	59
3.1.7 Database for Text Storage.....	62
3.2 Corpus Implementation Phase .....	68
3.2.1 COPA-TRAD Framework .....	70

3.2.2	<b>COPA-TRAD Tools</b> .....	72
3.2.3	<b>COPA-TRAD Textual Processing Modules</b> .....	74
3.3	<b>Corpus Processing</b> .....	75
3.3.1	<b>How COPA-TRAD’s Search Engine Works</b> .....	75
3.3.2	<b>How Texts are Processed in the COPA-TRAD System</b> .....	78
3.3.3	<b>COPA ALIGNER Processing and Alignment</b> .....	81
3.3.4	<b>COPA TOKENIZER Word Extraction</b> .....	82
3.4	<b>Chapter Closing Remarks</b> .....	83
	<b>CHAPTER FOUR: Analysis of COPA-TRAD tools</b> .....	85
4.	<b>Initial Remarks</b> .....	87
4.1	<b>Revisiting COPA-TRAD: A Brief Overview</b> .....	87
4.2	<b>COPA-TRAD External Tools</b> .....	89
4.2.1	<b>COPA-CONC – Parallel Concordance</b> .....	89
4.2.2	<b>MONO-CONC – Monolingual Concordance</b> .....	101
4.2.3	<b>WORDLIST – Frequency List</b> .....	104
4.2.4	<b>CORPUS-BUILDER – Creating a Disposable Corpus on the Fly</b> .....	107
4.2.5	<b>COPA-STATS – The COPA-TRAD Statistics Tool</b> .....	110
4.2.6	<b>Text Submission</b> .....	112
4.3	<b>Final Remarks</b> .....	116
	<b>CHAPTER FIVE: Concluding Remarks</b> .....	117
5.	<b>Summarizing the Study</b> .....	119
5.1	<b>The Research Questions</b> .....	120
5.2	<b>Limitation and Suggestion for Future Study</b> .....	123
	<b>BIBLIOGRAPHY</b> .....	125
	<b>APPENDIXES</b> .....	131
	<b>APPENDIX A</b> .....	132
	<b>(COPA-TRAD Patent application request form)</b> .....	132
	<b>APPENDIX B</b> .....	135



<b>(Legal Announcement of COPA-TRAD patent in <i>Revista da Propriedade Industrial</i> equivalent to “Official Gazette for Patents” in United States)</b> .....	135
APPENDIX C .....	136
<b>(COPA-TRAD running in a mobile browser)</b> .....	136
APPENDIX D .....	137
<b>(COPA-TRAD public website)</b> .....	137
APPENDIX E.....	138
<b>(Website of “University Research Program for Google Translate”)</b> 138	
APPENDIX F.....	139
<b>(Microsoft Translator Advertisement)</b> .....	139
APPENDIX G .....	140
<b>(Microsoft Translator API Management for COPA-TRAD)</b> .....	140
APPENDIX H .....	141
<b>(Me, Mona Baker, Lincoln Fernandes and Danielle Amanda at my first presentation of COPA-TRAD)</b> .....	141



## FIGURES

<i>Figure 1.</i> CORTRAD search form. ....	26
<i>Figure 2.</i> COMPARA simple search form. ....	27
<i>Figure 3.</i> Results from Opus Corpus for the word “take”. ....	28
<i>Figure 4.</i> Example extracted from COPACONC. ....	31
<i>Figure 5.</i> The stages on corpus compilation proposed by Fernandes (2004). ....	41
<i>Figure 6.</i> Baker’s (1995) Corpus Typology. ....	42
<i>Figure 7.</i> Print screen showing how the requirements were organized. ....	52
<i>Figure 8.</i> The completed task list. ....	53
<i>Figure 9.</i> Material organized on the Evernote Application. ....	54
<i>Figure 10.</i> Bilingual parallel concordance showing how a parallel corpus works. ....	55
<i>Figure 11.</i> Requirements and the overall design. ....	60
<i>Figure 12.</i> The relationship between source and target sentences. ....	64
<i>Figure 13.</i> The relationship between sentences and words. ....	65
<i>Figure 14.</i> Chart displaying the frequency of words. ....	66
<i>Figure 15.</i> The incremental development process (Sommerville, 2000). ....	69
<i>Figure 16.</i> Prototyping process (Sommerville, 2000). ....	69
<i>Figure 17.</i> A working prototype for COPA-CONC. ....	70
<i>Figure 18.</i> Example showing the diagram classes for the news management module. ....	71
<i>Figure 19.</i> Example showing the MONO-CONC’s KWIC display format. ....	73
<i>Figure 20.</i> Word cloud from WORDLIST tool. ....	74
<i>Figure 21.</i> System Topology showing how a user request is processed inside COPA-CONC. ....	77
<i>Figure 22.</i> System topology showing how texts are submitted to COPA-TRAD. ....	80
<i>Figure 23.</i> Architecture Diagram of COPA ALIGNER showing how texts are processed. ....	81
<i>Figure 24.</i> Architecture Diagram of COPA TOKENIZER showing how words are extracted. ....	83
<i>Figure 25.</i> The five subcorpora, which are part of COPA-TRAD. ....	88
<i>Figure 26.</i> COPA-CONC search input. ....	90
<i>Figure 27.</i> Results from COPA-CONC using the "AND" operator. ....	91
<i>Figure 28.</i> Results from COPA-CONC using the "OR" operator. ....	91
<i>Figure 29.</i> Results from COPA-CONC using the "NOT" operator. ....	92

<i>Figure 30.</i> Results from COPA-CONC using composition/grouping expression. ....	93
<i>Figure 31.</i> Results from COPA-CONC using the exact operator. ....	93
<i>Figure 32.</i> Results from COPA-CONC using the field start operator. .	94
<i>Figure 33.</i> Results from COPA-CONC using the field end operator. ..	94
<i>Figure 34.</i> COPA-CONC highlight in both sides when the term match across English – Portuguese languages.....	95
<i>Figure 35.</i> COPA-CONC highlight in both sides when the term match across English – Portuguese languages.....	95
<i>Figure 36.</i> Main table of COPA-CONC. ....	96
<i>Figure 37.</i> COPA-CONC auxiliary tools.....	97
<i>Figure 38.</i> COPA-CONC statistical information.....	98
<i>Figure 39.</i> COPA-CONC filter.....	99
<i>Figure 40.</i> COPA-CONC Source Text and Target Text visual aid.....	100
<i>Figure 41.</i> COPA-CONC auxiliary window to show the extra linguistic information related to a source/target sentence.....	101
<i>Figure 42.</i> MONO-CONC KWIC display format. ....	102
<i>Figure 43.</i> MONO-CONC filters options. ....	102
<i>Figure 44.</i> MONO-CONC results for "have a".....	103
<i>Figure 45.</i> Word cloud for English words from COPA-TRAD.....	104
<i>Figure 46.</i> Details for the word "spiro". ....	105
<i>Figure 47.</i> WORDLIST filter form.....	105
<i>Figure 48.</i> Word frequency list with words starting with "A". ....	106
<i>Figure 49.</i> CORPUS-BUILDER text alignment.....	108
<i>Figure 50.</i> COPA-BUILDER filters in use.....	109
<i>Figure 51.</i> COPA-BUILDER results. ....	110
<i>Figure 52.</i> COPA-STATS language statistics for type and token in Portuguese and English.....	111
<i>Figure 53.</i> Statistical information related to types and tokens for the book "Harry Potter and the Chamber of Secrets" and its Brazilian version "Harry Potter e a Câmara Secreta". ....	112
<i>Figure 54.</i> Text submission main screen. ....	113
<i>Figure 55.</i> Message sent from the moderator to user.....	113
<i>Figure 56.</i> Form text submission. Source Text section.....	115
<i>Figure 57.</i> COPA-CONC and MONO-CONC display formats.....	123

## ABBREVIATIONS AND ACRONYMS

CL – Children’s Literature

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

CTS – Corpus-based Translation Studies

DTS – Descriptive Translation Studies

SP – Semantic Prosody

ST – Source Text

TS – Translation Studies

TT – Target Text

TraCor – Grupo de Pesquisa Tradução e Corpora

XML – Extensible Markup Language



## **CHAPTER ONE: Introduction**





## 1.1 The Context of Investigation

In the context of Translation Studies, a parallel corpus is viewed as a computational tool used for translation research and pedagogy. This type of corpus provides useful tools to help the investigation of certain types of text, translation practices and also linguistic patterns such as collocation and semantic prosody (Baker, 1995; Zanettin, et al., 2003).

Basically, a parallel corpus consists of a collection of electronic texts in a source language aligned with their respective translation in a target language, and analyzable by means of specific computerized translational tools (Baker, 1995, p. 230-231). From a research perspective, parallel corpora have been used to investigate the practices of translating specific types of text, as they allow for the analyst to unveil the strategies or techniques employed by professional translators in dealing with certain elements within a particular type of text (Kenny, D. 2001; Fernandes, 2004). Now from a pedagogical perspective, the use of a parallel corpus in translator education supports students to find solutions proposed by professional translators for problems that characteristically arise in translation (Pearson, 2003, p. 18). In addition, the use of parallel corpus can help translator educators to explain to their trainee students why a particular decision can be considered suitable or not for a particular translational situation (Fernandes, 2007, p.142).

To my knowledge, currently there are three relevant on-line parallel corpora available for academic research and pedagogy in the linguistic pair Portuguese – English, namely CORTRAD – *Parallel Corpus for Translation* (2009), COMPARA – *Bidirectional Portuguese and English Parallel Corpus* (2011) and Opus Corpus – *An Open Source Parallel Corpus* (2012). CORTRAD<sup>1</sup>, for instance, is a parallel corpus in the Portuguese – English linguistic pair, which is one of the corpora developed by the COMET project at University of São Paulo. CORTRAD has at least two innovative functionalities: (i) the possibility of comparison between different versions of the same text (original text, revised versions and published translations), and (ii) it also has different searching mechanisms (Figure 1) for each text type investigated, for instance, it is possible to conduct an investigation on specific sections from different textual types.

---

<sup>1</sup> Available online at  
[http://www.fflch.usp.br/dlm/comet/consulta\\_cortrad.html](http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html)

Início	Jornalístico	Literário	Técnico-científico
--------	--------------	-----------	--------------------

**CorTrad jornalístico divulgação científica**

O CorTrad é um corpus aberto, sujeito a alterações. Veja dados quantitativos para informações atualizadas sobre o conteúdo do corpus.

A **parte jornalística do CorTrad** conta atualmente com textos das edições de 2001, 2002 e 2003 da Revista Pesquisa FAPESP, totalizando 20 números. As seções incluídas foram: Humanidades, Ciência, Tecnologia, Estratégias, Laboratório, Linha de Produção e Política de C & T. Veja uma tabela pormenorizada por assunto e gênero. A disponibilização do CorTrad na rede é um projeto conjunto entre o COMET, a Linguateca e o NILC, usando o sistema DISPARA.

**Pesquisar no corpus** ?

Original	<input type="radio"/> principal	<input type="text"/>	<input checked="" type="checkbox"/> ver
Tradução publicada	<input type="radio"/> principal	<input type="text"/>	<input checked="" type="checkbox"/> ver

Ignorar maiúsculas/minúsculas

**Resultado**

<input checked="" type="radio"/> Concordância	<input type="radio"/> Distribuição das formas
<input type="radio"/> Distribuição dos lemas	<input type="radio"/> Distribuição da categoria gramatical (PoS)
<input type="radio"/> Distribuição do tempo verbal e/ou do caso pronominal	<input type="radio"/> Distribuição de pessoa e/ou número
<input type="radio"/> Distribuição do gênero morfológico	<input type="radio"/> Distribuição da função sintática
<input type="radio"/> Distribuição por documento	<input type="radio"/> Distribuição por data de publicação
<input type="radio"/> Distribuição por gênero de texto	<input type="radio"/> Distribuição por tema
<input type="radio"/> Distribuição por campo semântico	<input type="radio"/> Distribuição por grupo (de cor, de vestuário, etc.)

**Opções**

Figure 1. CORTRAD search form.

CORTRAD has three different subcorpora; journalistic, scientific and literary texts. CORTRAD uses the DISPARA system (*DIStribuição de corpora PARAlelos na Web*, which provides an easy interface for parallel texts which was created at a first hand for COMPARA), which was provided through a partnership between COMET project and Linguateca project from Foundation for National Scientific Computing in Portugal (Tagnin, Teixeira, & Santos, 2009, p. 314-315).

The second parallel corpus to be mentioned here is COMPARA<sup>2</sup>, based on a project from *Linguateca*, which uses the DISPARA system (the same used in CORTRAD). COMPARA is an “open-ended” corpus, which “means that it can grow in whichever direction proves to become important to users, and that the texts incorporated in the corpus can be put to use as soon as they are processed” (Frankenberg-Garcia, A. & Santos, D., 2003, p. 71). This corpus consists of fictional texts in the Portuguese – English language pair, aligned at sentence level (originals with translations). According to Figure 2, it is possible to understand why COMPARA was developed aiming “people who are not necessarily corpus-literate as well as for experienced corpus users” (Frankenberg-Garcia, A. & Santos, D., 2003,

<sup>2</sup> Available online at <http://193.136.2.104/COMPARA/index.php>

p. 71). This is due to the fact COMPARA provides a “simple search” tool which requires a basic knowledge from the user in order to operate it.

### Simple search

A simple search enables you to search the whole of COMPARA either (1) from **Portuguese to English** or (2) from **English to Portuguese**. Your results will be presented in the form of parallel concordances.

1. From Portuguese to English 2. From English to Portuguese

help

Enter a word or a search string in **Portuguese**. Put quotation marks around each separate word (e.g. "Até" "logo")

Search (from Portuguese to English) Reset form

Case insensitive  
 Ignore accents and cedillas

Figure 2. COMPARA simple search form.

COMPARA, however, needs the insertion of quotation marks around each word in separate (i.e. fixed expression, phraseology, and idiom), for example, "*pay*" "*attention*". This search feature can be viewed as a limitation of the system since it can become a hassle for the user due to the need of inserting quotation marks in each word of the string when performing a search of long word-strings.

Finally, the Opus Corpus<sup>3</sup>, a multilingual parallel corpus that consists of online texts available on the Internet for free. Figure 3 displays Opus Corpus and its search tools at the right; the available subcorpora at the left and below the results for the node (search word or keyword) "take", using the subcorpus of PHP texts.

<sup>3</sup> Available at <http://opus.lingfil.uu.se/>

OPUS - Corpus query (CWB)

The image shows the OPUS Corpus query interface. On the left, a list of corpora includes ECB, EMEA, EUconst, EUROPARL, KDE4, KDEdoc, OpenOffice, OpenSubtitles, and PHP. The main area displays a query editor with a search term 'take' and a list of results. The results table has columns for 'id', 'text', and 'corpus'. The first result is '1186 Many functions take on multiple parameters, such as my\_array()' from the 'emea' corpus. Below the table, there is a detailed description of the 'take' function in both English and Portuguese, explaining its usage and parameters.

Figure 3. Results from Opus Corpus for the word “take”.

The available subcorpora that comprise Opus Corpus are “EMEA - European Medicines Agency documents”, “EUconst - The European constitution”, “EUROPARL - European Parliament Proceedings”, “Open Office - A collection of documents”, “Open Subtitles - A collection of documents”, “KDE4 - KDE4 localization files”, “KDEdoc - KDE manuals”, “PHP - A collection of PHP manuals”, “SETIMES - South East European Times” and “SPC - Stockholm Parallel Corpus”.

The aforementioned on-line corpora, however, lack a user-friendly interface providing an easy interaction with filters as well as how data is displayed. As a consequence, these corpora can present a challenge for novice researchers and TS students who are not acquainted to working with such technological tools. As an example, the textual alignment of Opus Corpus is carried out automatically, but “no manual corrections have been carried out” for the texts (Opus Corpus, 2012b). From a qualitative perspective, the automatic alignment without manual supervision can pose an obstacle, especially for trainee students, since the paired sentences sometimes do not exist or if they do exist, they are aligned with the wrong correspondent pair (Table 1).

Table 1. *Parallel Concordance Lines from the Opus Subcorpus of PHP Programming Language*

	<b>Original: English</b>	<b>Translation: Brazilian Portuguese</b>
27862	4 Many examples in this reference <b>require</b> an XML string.	Exemplos
31219	4 Most <code>fdf</code> functions <b>require</b> a <code>fdf</code> resource as their first parameter.	<code>fdf</code>
32908	4 Some functions may <b>require</b> more recent version of the GMP library.	Instalação
43142	4 <code>qmail-inject</code> does not <b>require</b> any option to process mail correctly .	Esta extensão não possui nenhum tipo resource.
49823	4 Configuration of the client will <b>require</b> installation of all the tools.	A extensão requer que as ferramentas de cliente do MS SQL sejam instaladas no sistema onde o PHP esta instalado.
55377	4 Note that it is possible to override the default path from the script using the <code>configargs</code> of the functions that <b>require</b> a configuration file.	Configuração durante execução Esta extensão [sic] não define nenhum parâmetro de configuração no <code>php.ini</code> .
60392	4 You will <b>require</b> the appropriate SDK for your platform, which may be downloaded from within the manager interface once you have registered.	Requisitos

*NB.* It is important to note that these parallel concordance lines exemplify the fact that sometimes what might seem as mistakes is in reality the technical language of PHP programming code (e.g. `fdf`, `configargs`).

The exception is COMPARA but, unfortunately, it consists mainly of fiction texts “extracts of around 30% of the total of each work” (Frankenberg-Garcia, A. & Santos, D., 2003, p. 74). The text types in the three corpora do not include “Children’s Fantasy Literature” and “Academic Texts”, just to name some examples. In addition, these corpora do not provide users a possibility for submitting their own text to be processed and included in the corpus. With the advent and popularization of machine translation such as Google Translate (Appendix E) and Microsoft Translator (Appendix F, and Appendix G), the mentioned three corpora do not provide an interface between texts (original and translation) from the corpus and such automatic translation engines. As a result of some of the limitations (i.e. lack of a user friendly interface and the fact that types of texts covered do not include children’s literature) listed here, the necessity of a parallel corpus aimed for translation research and pedagogy becomes paramount.

Focusing on these three parallel corpora, which cover the linguistic pair Portuguese and English, it is possible to verify that as far as I know, there is no Children’s Literature (CL) parallel corpus available in the linguistic pair Portuguese – English. Another important characteristic to mention is related to the mentioned machine translation (MT) such as Google Translate and Microsoft Translator, as observed, none of these tools support the comparison of official translated texts with their automatic translation from a machine-translation system.

## **1.2 The Emergence of Corpus-based Translation Studies**

Translation Studies (TS) was largely developed during the 80’s (Baker, 1998, p. 277) and with the advent of descriptive approaches, Corpus-based Translation Studies (CTS) emerges as an approach to describe translational phenomena. The idea of descriptive approaches comes from Gideon Toury’s concept of norms and the methodology he developed related to Descriptive Translation Studies (DTS), an area part of TS whose main objective is the analysis of translated texts. Moreover, DTS is engaged to understand the translator decisions for a particular problem rather than just criticize the final product based on a comparison between source text and target text. In other words, the descriptive analyst refuses to “make a priori statements about what translation is, what it should be, or what kinds of relationship a translated text should have with its original” (Baker, 2001, p. 163).

In this sense, Descriptive Translation Studies has paved the way for CTS to become one of the most influential methods used to describe translations (Baker, 1998). The merit of CTS in this case has to do with

the fact that all the drudgery associated with descriptive work has been reduced to the click of a button (FERNANDES, 2009, p. 17). The use of specialized software enables the analyst to investigate a large amount of data in a fast and precise way, thus making descriptive results more reliable and encompassing.

According to Baker (1995), a corpus (plural: corpora) for many years was associated with hard-copy texts. However, in last decades with the popularization of computer and related technology (i.e. scanners, specific software to deal with texts, etc.) corpus in a broad sense is a collection of texts put into electronic format, readable and analyzable automatically or semi-automatically rather than manually. Among the various types of corpora available for the study of translation, the parallel corpus has been chosen for this research due to the fact that it is the only type of corpus capable of showing how professional translators deal with specific linguistic patterns (p. 226).

One of the linguistic patterns that have been widely investigated with the use of corpus-based tools is that of collocation. Sinclair (1991) has highlighted the fact that without the use of electronic corpora, it would be extremely difficult to identify collocational patterns, because linguists would have to “rely on their intuitions, their limited capacity for thorough textual analysis, and whatever has caught their eye or ear” (p. 100). This is so due to the necessity of large amounts of textual evidence in order to extract such collocational patterns. Collocation here means the tendency certain words have to co-occur in a regular and frequent way in a given language (Baker, 1992, p. 47). The following example (Figure 4) offers an illustration of a collocational pattern extracted from COPA-LIJ, the children’s literature subcorpus of COPA-TRAD:

#### Concordância do COPA-TRAD

Língua 1: Inglês	Língua 2: Português
Total em Exibição: 1 Total Processado: 1 Total Encontrado: 1 Tempo Decorrido: 0.028s	
<i>Resultados por palavra:</i> <b>pay:</b> Entradas > 49, Total de Ocorrências > 51. <b>him:</b> Entradas > 2477, Total de Ocorrências > 3059. <b>a:</b> Entradas > 18788, Total de Ocorrências > 31903. <b>visit:</b> Entradas > 56, Total de Ocorrências > 56.	
<b>Traduzir</b> Oh, I get it. So I gotta <b>pay him a visit</b> . 	Ah, saquei. Então eu tenho de fazer uma visita. 
 Token: 12   Type: 10   Ratio: 90.9091%	Token: 9   Type: 9   Ratio: 100%
<a href="#">Imprimir</a>   <a href="#">Exportar CSV</a>   <a href="#">Exportar XML</a>   <a href="#">Exportar PDF</a>	

Figure 4. Example extracted from COPACONC.

The occurrence shows that English speakers typically use the expression *pay a visit* while Brazilian Portuguese speakers typically use *fazer uma visita*. Of course, less typically English speakers can also use *make a visit*, but this would sound more formal (Baker, 1992, p. 47). The example above shows that patterns of collocation are largely arbitrary and independent of semantic meaning (“pay” is not the same as “make”), therefore, the understanding of the notion of collocation can help translators and TS students to avoid potential shortcomings involved in the act of translating (Baker, 1992, p. 47).

After contextualizing this research project and defining briefly some terms within Translation Studies, I would like to present my research questions and objectives, as well as the significance of the project being conducted.

### 1.3 The Purpose of the Study

The current study draws upon the corpus of children’s literature (henceforth CL) proposed by Fernandes PhD work (2004) and aims to develop an online parallel corpus-based tool. This electronic tool is going to be used for investigating the practices of translating collocational patterns in CL. An important part of the design and development of the mentioned tool is related to its storage mechanisms, through which it is possible to organize, catalog, relate and attach meta-information (i.e. extra-linguistic information) to a large amount of data. The way data is stored and structured is undoubtedly important when dealing with large amounts of text. It is worth noting that the system overall response time can get compromised if the corpus is not properly planned (Chapter 3). Nevertheless there is a solution for the potential shortcoming and it is to combine a storage mechanism with a search engine for indexing the corpus textual content. Consequently, the search engine promotes an increase of processing speed, productivity and it can also perform statistics on the fly and make complex recoveries of data, just to mention a few of its possibilities. Apart from processing mechanisms, the visual elements displaying the information provided by the corpus to the user were developed to be in line with the current visual aspects related to on-line applications i.e., responsive web design (Marcotte, 2011).

As already mentioned, meta-information such as bibliographical references were added, providing the researcher/user a way to track the source of a specific entry and details about it such as name of the author, gender, relevant dates, publisher house, etc. The possibility of storing the complete digital version of the source and



target text is also present, providing a rich contextualization environment for investigation.

In order to achieve the objectives mentioned above which are related to a technological tool and possible contributions of it, this study aims to answer the following research questions:

- **What aspects should be taken into account when compiling a parallel corpus to be made available online?**
- **In what ways can a parallel corpus of children’s literature proposed and developed by the present study contribute to the analysis of translational patterns?**
- **What alternative display formats can be used for this specific kind of analysis and why?**

It is important to highlight the fact that the research questions above reflect the applied nature of the present study as it aims to build an online parallel corpus for translation research and pedagogy.

#### **1.4 Significance of the Study**

The significance and relevance of the study have to do with the practical perspective it offers; it aims to develop an online parallel corpus, which concentrates useful tools to help the investigation of certain types of text, translation practices and also linguistic elements and patterns (such as proper names, collocational patterns, semantic prosody). Additionally, this study can contribute from a methodological perspective as it offers a set of easy-to-use tools for the empirical investigation of translation practices associated with the specific linguistic elements and patterns aforementioned. As a result, COPA-TRAD can be viewed as a valuable set of tools for novice researchers learning how to carry out descriptive research in Translation Studies. All in all, the online parallel corpus developed in this study can be an excellent resource for professional translators, translation students and TS researchers focusing on the translation of children’s literature. In terms of implications, the applicability of this online parallel corpus consists, for example, in translation research and pedagogy. The present research is also expected to contribute with translators as well as under- and postgraduate TS students to improve and develop their awareness in relation to translational patterns by using tools provided by CTS.

## **1.5 Organization of the Study**

After introducing the context as well as the objectives, research questions and the significance of the study, Chapter Two brings the theoretical notions and concepts informing the type of analysis here envisaged, Chapter Three reports the methodology employed in the study. Next, Chapter Four is related to the data analysis and description 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.

## **CHAPTER TWO: Review of Literature**



## 2. Initial Remarks

In this Chapter, the theoretical foundation informing the present study is discussed. It starts by locating the study within the research area of Corpus-based Translation Studies. Then, it presents the types of corpora for translation research and pedagogy suggested by Baker (1995), with a special focus on parallel corpora. After that, Fernandes' parallel corpus of children's literature is briefly described, as the method used in the compilation of this particular corpus is the basis of the present study. Finally, some issues related to children's literature, collocation and semantic prosody are briefly described, as they had to be kept in mind when compiling COPA-LIJ, the subcorpus of children's literature in COPA-TRAD.

### 2.1 Corpus-Based Translation Studies

According to Olohan (2002), "Corpus-based Translation Studies is a relatively new area of research within translation studies" (p. 153), whose main interest is the investigation of translated texts in a real context. This new research area has come into being in order to fill in the gap left by the discipline of Corpus Linguistics, which tends to "exclude translated text" from their collection of texts due to the fact the translated text "might distort the view of 'real' language under investigation" (Baker, 1993, p. 234).

On this basis, Corpus-based Translation Studies aims to investigate translated texts combining quantitative and qualitative analysis to explore pragmatic factors related to discourse, specific text types and regularities of translational behavior (i.e. norms) (Baker, 1998, p. 189; Olohan, 2004, p. 22). This investigation is carried without losing track of the constraints and influences on the translator and the translation activity, which are not directly related to the differences between language systems, but the influences that arise from the translation brief, translation situation, cognitive factors, and so on (p. 14). In this vein, Laviosa-Braithwaite (as cited in Kruger, 2002, p. 79) explains that:

[t]he corpus-based approach in translation studies emerges as a composite, rich and coherent paradigm, covering many different aspects of the translational phenomenon and concerned with unveiling both the universal and the specific features of translation, through the interplay of theoretical constructs and hypotheses, variety of

data, novel descriptive categories and a rigorous, flexible methodology, which can be applied to inductive and deductive research, as well as product- and process-oriented studies.

In addition to Laviosa-Braithwaite's words, Tymoczko (1998) argues that CTS allow us to investigate vast quantities of data "more data than any single human being could ever manage to gather or examine in a productive lifetime without electronic assistance" (p. 652). According to Tymoczko (*ibid.*), the focus of CTS is twofold because it relies on the process of translation as well as the product of translation (p. 653). To clarify, the process of translation is the act of translating, including cognitive aspect related to the translation activity. This process involves "reading, text comprehension, semantic transfer between linguistic systems, and writing a text in the target language" (Shreve, 2009, p. 255). On the other hand, the product is the final result, that is, the translated text itself. As stated before these two aspects are the main focuses of CTS.

Olohan (2004) goes on to explain that the need to move beyond vague generalizations based on quantitative data is paramount in any corpus-based study and adds "purely quantitative studies of corpus data are regarded as limited in their usefulness" (p. 22). Similarly, Tymoczko (1998) also raises a caveat in relation to the "quantification" aspect of corpus-based research, particularly on its use to "prove the obvious" (p. 658). In other words, this researcher emphasizes that relying on quantitative data from corpora to confirm answers to questions previously known "merely to prove the obvious" is an empty exercise. Therefore, "researchers must take care to ask 'the right questions': to pose questions and construct research programs that have as their goal substantive investigations that are worthy" of the technological tools and solutions provided by CTS (*ibid.*). Tymoczko's view on the qualitative aspect related to CTS is the same shared by this study. The technological tool proposed here is an example of the qualitative aspect since it provides automatic search with easy-to-use display mechanisms to facilitate the manual investigation.

Despite these notes of caution, the potential of CTS cannot be denied as Malmkjaer (2003) puts it,

[t]he use in translation studies of methodologies inspired by corpus linguistics has proved to be one of the most important gate-openers to progress in

the discipline since Toury's re-thinking of the concept of equivalence (p. 119).

Malmkjaer recognizes the importance corpus linguistics methodologies applied to Translation Studies and she treats it with same level of Toury's concept. Indeed, the importance of corpus methodologies is attributed to the fact it allows "scholars to investigate all kinds of issues from a range of perspectives" and it "reflects the current concerns and trends of Translation Studies as a whole" (Olohan, 2004, p. 22). The following section explores some of the assumptions concerning the typology of corpus based on Baker (1995) and the advantages of a parallel corpus in relation to other types of corpus (p. 229-235).

## 2.2 Corpus Typology

Baker (1995) listed a number of criteria to take into account in the design of specific corpora (p. 229). These criteria are listed as follows.

1. General language vs. restricted domain;
2. Written vs. spoken language;
3. Synchronic vs. diachronic;
4. Typicality in terms of range sources (writers/speakers) and genres (e.g. newspaper editorials, radio interviews, fiction, journal articles, court hearings);
5. Geographical limits, e.g. British and American English;
6. Monolingual vs. Bilingual or multilingual.

Important to mention that Baker (1995) argues that the above criteria are important "but not sufficient" and "it is up to translation scholars now to refine these criteria and adapt them to our needs" (p. 230). Baker gives a clear clue that depending on the project we are carrying out the criteria can be extended in order to cover all the necessities and shortcomings involved in the selection of text, for instance, Fernandes (2009) suggests mode and medium as well as extralinguistic information to be included in the corpus (p. 24-26).

In this sense, the compilation method used in this study is based the three main stages of corpus compilation proposed by Fernandes (2004). These three main stages were adapted and extended in order to fit this study. Figure 5, displays the three main stages; the first stage is related to CORPUS DESIGN. Fernandes (ibid.) states that at this part "the general theoretical issues associated with corpus planning are discussed" (p. 74). The mentioned stage was extended here to cover not

only the theoretical aspects related to planning but also the practical aspects and how it was conducted in the present study. The extended aspects here envisaged are the following ones and they will be fully covered in Chapter 3.

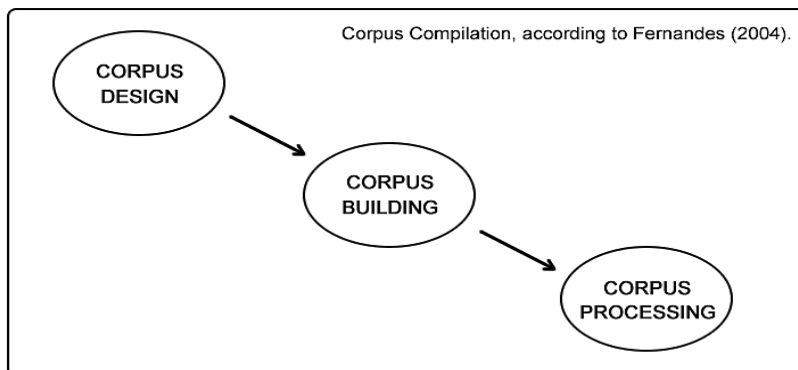
- Corpus creation: shortcomings and possible solutions –

This section aims to answer and make researchers aware in relation to important issues to be considered during the design and planning phase. The issues covered at this part impact directly upon the final product which is the corpus and the system that manipulates the data;

- Requirements Analysis – This section is based on the assumptions provided by Software Engineering area. In general lines, the researcher on requirements analysis phase gathers the needs desired for the software, establishes a plan, validate the needs and manages it by preparing technical documentation to guide the software development (Pressman, 2011, p. 127-133). The requirements analysis is the phase when the necessities for the creation of specific software are gathered and organized. The procedures to gather the requirements comprehends from the interview and discussions to raise the necessities of the people interested in the project, find similar projects in the area and reading of material published on the subject. The final step of the requirements analysis is the validation of each requirement to testify if the real need was understood between the personnel involved in the project.

- Database for Text Storage – This part complements the last one but it was decided to separate it, because the design of the database impact directly in the system, how data is stored and recovered, etc. This section lists a set of questions to be considered before starting the design of the database and after it shows a preview of how the tables were modeled using specific software.





*Figure 5.* The stages on corpus compilation proposed by Fernandes (2004).

Following the diagram presented in Figure 5, the next stage is **CORPUS BUILDING**. According to Fernandes (2004), this stage is aimed to cover the technical decisions “made throughout the corpus compilation” (p. 74). Here, this stage was named to “corpus implementation phase” in order to fit the nature of the technical and developmental stage of the corpus as well as the system which manipulates such corpus. For example, the creation of prototypes, the programmatic development of COPA-TRAD framework, processing tools and modules are described in this section.

The last stage proposed by Fernandes (2004) and displayed in Figure 5 is **CORPUS PROCESSING**. The author explains that in this stage “the hardware, software and set of computational tools used for processing the corpus are specified” (p. 74). Fernandes (ibid.) in this stage specifies the hardware and the software he utilizes to process his corpus. However the main objective of this study is to create a parallel corpus and the system which manipulates it. As a result, no third party software was used to process the corpus. In this sense, **CORPUS PROCESSING** here was used to describe how COPA-TRAD inner mechanisms process the textual material. The following subsections were added in order to better explain how COPA-TRAD system works in its core.

- **How COPA-TRAD Search Engine Works** – The objective of this subsection is to describe how the COPA-TRAD search engine works starting from the search performed by the user to the results returned by the server.

- How Texts are Processed in the COPA-TRAD System – at this point the focus is on two of COPA-TRAD internal tools aimed to process the unstructured textual information and store it in structured form into the database. The mentioned internal tools are COPA ALIGNER aimed to read the text files, extract the sentences, align it with its correlated translated sentence and store all these processed material in database. The second tool is COPA TOKENIZER, this tool retrieves from database all processed sentences in order to extract all tokens, proceed the statistics of these tokens finally COPA TOKENIZER stores the information processed in database also.

After discussing some of the criteria for corpus design, Baker (1995) goes on to propose three main types of corpora for translation research and pedagogy (p. 230): (i) multilingual corpus, (ii) comparable corpus and (iii) parallel corpus (Figure 6).

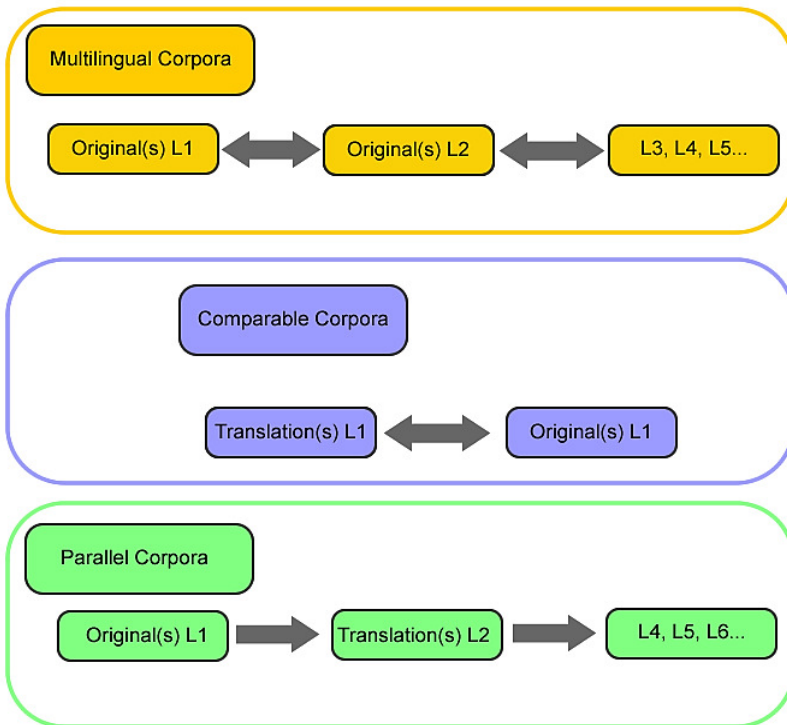


Figure 6. Baker's (1995) Corpus Typology.

(i) The multilingual corpus is a “set of two or more monolingual corpora in different languages, built up either in the same or different institutions on the basis of similar criteria” (p. 232). This type of corpus is especially used for contrastive linguistic work, especially for lexicography, and the Council of Europe Multilingual Lexicography Project is an example of a multilingual corpus.

(ii) According to Baker (1995), the comparable corpus, is the one that encompasses “two separate collections of texts in the same language: one corpus consists of original texts in the language in question and the other consists of translations in that language from a given source language or languages” (p. 234). This type of corpus aims to capture “patterns which are either restricted to translated text or which occur with a significantly higher or lower frequency in translated text” (p. 235). An example of this kind of corpus is the Translation English Corpus (TEC) developed mainly for the investigation of translation features such as simplification, normalization, explicitation, etc.

(iii) The parallel corpus - as defined in Chapter 1 (Section 1.1) – is a set of texts in one language and its translations in another language. This type of corpus provides an important contribution to TS area because it “supports a shift of emphasis, from prescription to description” (Baker, 1995, p. 231). Prescriptivism is a normative approach which stipulates how “translation should be performed in a particular culture” based on the original text (Shuttleworth & Cowie, 1997, p. 130-131). On the other hand, descriptivism, in Translation Studies, investigates what “translation was and is” and CTS “is clearly aligned with the descriptive perspective” (Olohan, 2004, p. 10). Thus descriptivism in Translation Studies tries to understand translator decision for a particular problem (Section 1.2). A parallel corpus helps researchers to understand how translators overcome translational challenges in the process of translation and all of this evidence is an excellent input for TS students (Baker, 1995, p. 231). The parallel corpus is not a monolithic database, instead it can be an unidirectional “source text in language A and target texts in language B” or even bidirectional “source texts in language A and translations in language B, and source texts in language B and their translations in language A” (Olohan, 2004, p. 24). Olohan (2004) observes an important point for Parallel Corpora, which is the alignment of source texts with target texts she says that “alignment means linking a unit of text in one language with a unit of text in another language” (p. 26). The alignment of texts at first sight can be a sort of thing trivial to do but this procedure is very important for a parallel corpus, without it the corpus is not parallel.

There should always exist a reference from a piece of text in language A to a translated piece of text in language B. There are projects, software specifically developed to align texts in two or more languages. These technologies use statistical data, dictionaries or both to predict which line in the original text corresponds to the correct line in the translated text. However depending on the size of the corpus and the team working on the project is also possible to align the texts manually with the aid of technological equipment, such as scanners and also specific software like text editors, optical character recognition systems, etc. Malmkjær (1998) in an article mentioning the advantages and disadvantages of a parallel corpus cites some applicability within Translation Studies for this type of corpora, the main ones are reproduced as follows (p. 535):

- i) Parallel corpus studies can reveal characteristics of translated texts, such as tendencies towards explicitness and avoidance of repetition.
- ii) Comparison between the translation part of the corpus and a corpus of texts of the same genre, written in the target language for the translation corpus, reveals a tendency towards what we might call the Eliza Doolittle phenomenon: the translated texts, more than the texts in the control corpus, tend to contain those TL phrases, structures, and so on, which, from a comparative point of view, seem particularly characteristic of the TL.
- iii) When the translation-part of the corpus is used in conjunction with a corpus containing the Source Texts (together constituting a parallel text corpus), the method can promote sense-disambiguation, and can help to identify translation norms and to create machine translation programs and bilingual dictionaries.
- iv) The method can be exploited for language learning/teaching purposes and for translator training.

This applicability of a parallel corpus was important in guiding the development of the tools proposed in this research. It is interesting to note how this applicability has been present in the study carried out by Fernandes (2004), which, as mentioned earlier, was the basis for the study here envisaged. In order to understand the characteristics of translated texts in relation to proper names in Children's Fantasy Literature, Fernandes (2004) proposed a corpus-based approach to

investigate the translator decision (process) and the translated text (product). This is discussed in the following section.

### **2.3 Fernandes' Parallel Corpus of Children's Fantasy Literature**

This study is based on Fernandes PhD research (2004). During his PhD, Fernandes compiled a corpus of children's fantasy literature aiming to investigate how children's fantasy books have been translated from English into Brazilian-Portuguese, with a specific focus on proper names. The corpus he compiled is bilingual and unidirectional with English source texts and Brazilian-Portuguese translations, using complete texts. Fernandes research data relied on texts from the period between 2000 and 2003, since this period witnessed a revival in popularity of fantasy books for children, notably after the successful launch of the Harry Potter series all over the world (Olohan, 2004, p. 59). Thus, the twenty-four texts, twelve in each language, chosen for inclusion in his corpus were published during this period.

Fernandes (2004) describes the procedures used to collect extralinguistic information on the translation product and process. Some of this information is obtained from the books themselves and paratexts. Additionally, a questionnaire to collect information about the translator, the translation, the source text, and the translation process was prepared. This questionnaire is used as a guide in the development of the module responsible for insertion of new texts into the corpus (Chapter 4).

It is important to note that Fernandes (ibid.) corpus was compiled to be processed using off-the-shelf software. COPA-TRAD, on the other hand, is a web-based corpus that needs its own storage and processing tools. Therefore, this study contributes to Fernandes method of corpus compilation by adding new elements that must be taken into account when compiling a web-based parallel corpus.

### **2.4 Children's Literature, Collocation and Semantic Prosody**

The Routledge Encyclopedia of Translation Studies (2009) defines Children's Literature as "texts intentionally written for children by adults, texts addressed to adults but read by children, texts read by both children and adult" (p. 31). In this study, this definition covers the texts comprising the subcorpus COPA-LIJ part of COPA-TRAD. According to Knowles & Malmkjaer (1996), two patterns in the translation of children's literature that have benefit from corpus-based tools are those of collocation and semantic prosody.

The relevance of investigating collocational patterns is an important subject in CL because children use specific collocational patterns when they are explicitly told about it, the child "is more likely to gain this impression through exposure to use, and to store it as a part of a developing base of implicit knowledge about the language system" (Knowles & Malmkjaer, 1996, p. 70) this has to do with the idea children learn language by words in composition not in isolation. The following citation provides a glimpse explaining how certain linguistic constructions can be accompanied by a positive or negative meaning and shows the author responsibility when writing and using collocations in CL books:

For example, if the semantic prosody of black is negative, as its tendency to occur in expressions like black magic, black Wednesday, the black sheep of the family and so on suggests, then this prosody may spill over onto the person referred to in expressions like black man/woman/boy/girl. This particular phenomenon was actively combated through the 'Black is Beautiful' slogan used by campaigners for racial equality in the 1970's and 1980's - a slogan which itself manipulates collocation and whose success testifies to the effectiveness of such manipulation. Through frequent collocation with the positively loaded term beautiful, any negative connotations of black are overridden (Knowles & Malmkjaer, 1996, p. 70).

In other words, collocational patterns can unveil how the author in order to transfer to the child the specific meaning intended and thus avoid misunderstanding constructs these linguistic structures. In addition how collocational patterns are represented by means of the choices made by translators and the solution used to translate these structures can provide to TS students a good resource to overcome translational problems when dealing with collocations. Additionally, collocational patterns are "semantically arbitrary restrictions which do not follow logically from the prepositional meaning of a word" (Baker, 1992, p. 47). Therefore, these patterns, as observed, can be seen as potential problematic areas for TS students and as such should be included in any syllabus design for translator education course and corpus has a part to play here.

Semantic prosody (SP), for example, has shown to be a grammatical phenomenon possible to be studied through parallel corpus within the area of TS. According to Stubbs (1996) SP is “a particular collocational phenomenon” (p. 176) that can be categorized in negative prosody, positive prosody and neutral prosody; for instance, the verb *cause* is mainly related to something negative such as problems, death or damage, while the verb *provide* is related to positive things such as support and the verb *deal* is related to something neutral such as technical matters (Zethsen, 2006, p. 279-285). As shown before, “the semantic prosody of black is negative, as its tendency to occur in expressions like black magic, black Wednesday, the black sheep of the family and so on” (Knowles & Malmkjaer, 1996, p. 70) this particular example show us some of the most common meanings related to a specific word and how it can be understood. As a consequence, translators have to be aware of such semantic patterns while translating text for children. Teachers from Translation Studies have to be aware of such phenomena while teaching for their students the aforementioned subject.

## **2.5 Summing Up**

This Chapter has addressed the theoretical framework supporting this research. Firstly, Corpus-Based Translation Studies was used to locate the area this study is inserted in. Next, to present the array of variables involved in a corpus study, the corpus typology based on Baker (1995) was delineated to contextualize parallel corpus and its underpinnings. In addition, the stages in corpus compilation proposed by Fernandes (2004) were introduced. Next, an overview on Fernandes (2004) research was given to point out COPA-TRAD beginning. Finally, collocational patterns and semantic prosody were discussed to show an applicability of a parallel corpus. The next chapter describes in detail the methods used for compiling COPA-TRAD with a special focus on the steps added to the design, building and processing stages of a web-based parallel corpus.





## **CHAPTER THREE: Methodology**



### 3. Opening remarks

The study being carried out here is based on the development of an online parallel corpus for translation research and pedagogy. This chapter is organized following Fernandes (2004) proposal for corpus compilation and it was extended to cover up the technical development of COPA-TRAD:

- The first stage elicited by Fernandes (2004) is; “corpus design, where general theoretical issues associated with corpus planning are discussed” (p. 74), at this stage a layer for problem analysis was added (i.e. the development of a parallel corpus);
- Next, Fernandes (2004) gives the second stage “corpus building, where the technical decisions made throughout the corpus compilation are described” (p. 74), here a second layer related to the implementation phase of the corpus system tools was included;
- The last stage raised by Fernandes (2004) is related to “corpus processing, where the hardware, software and set of computational tools used for processing of the corpus are specified” (p. 74), at this part the display formats internal functionality are described—to show how these search and retrieval tools process information in the COPA-TRAD system.

The procedures related to corpus design, development and the technical apparatus are discussed below. In general, this chapter brings:

- An internal and external view of COPA-TRAD;
- The main phases involved in its development - from the initial planning until the final product, which is a working version of the corpus;
- The display format tools used to present the information retrieved from the corpus.

#### 3.1 Corpus Design

COPA-TRAD is a web-based corpus developed using technology to run on a web browser. This web-based approach contrasts with software developed for computers where it runs on Operational Systems such as Windows, Linux or MacOS. The first procedures to discuss in this chapter are related to the initial process involved in the development of COPA-TRAD. In order to develop a solid and reliable system, a set of steps preceding the building/creation stage of a corpus and its system processing tools must be taken into consideration.

The technical procedures adopted in the development of COPA-TRAD were borrowed from Software Engineering which is “the

application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software; that is, the application of engineering to software” (Software Engineering, 1990, p. 67).

From a Corpus-based Translation Studies perspective, the researches carried out by Fernandes (2009) and Baker (1993/1995) guide this project because the information elicited from these studies provide a theoretical and practical framework contributing for the requirements analysis and leading to the final product (i.e. COPA-TRAD).

Though, it is necessary emphasizing that not all techniques and methods from Software Engineering were applied for this study due to the small size of the research team, the overall size of the project itself and its particular nature. The lists of tasks (software requirements) to be accomplished were organized and shared by the team on an on-line application called “do.com”, as can be seen in Figure 7 below.

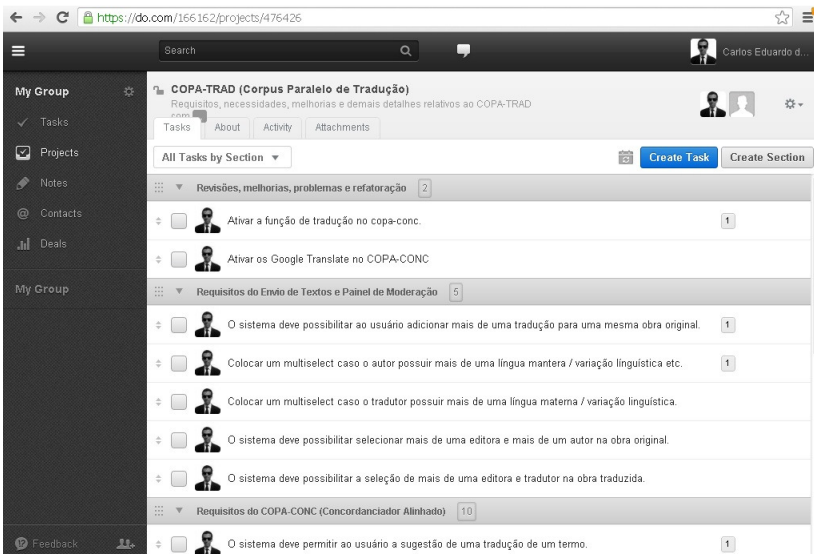


Figure 7. Print screen showing how the requirements were organized.

Upon the completion of the requirements listed in the aforementioned task management tool, a working version of the system was made available. The major requirements were imported from Enterprise Architect version 8.0. In the “do.com”, a section was also

allocated to receive possible new features and improvements to be analyzed and added to the work list. Completed tasks can be easily accessed in as shown in Figure 8.

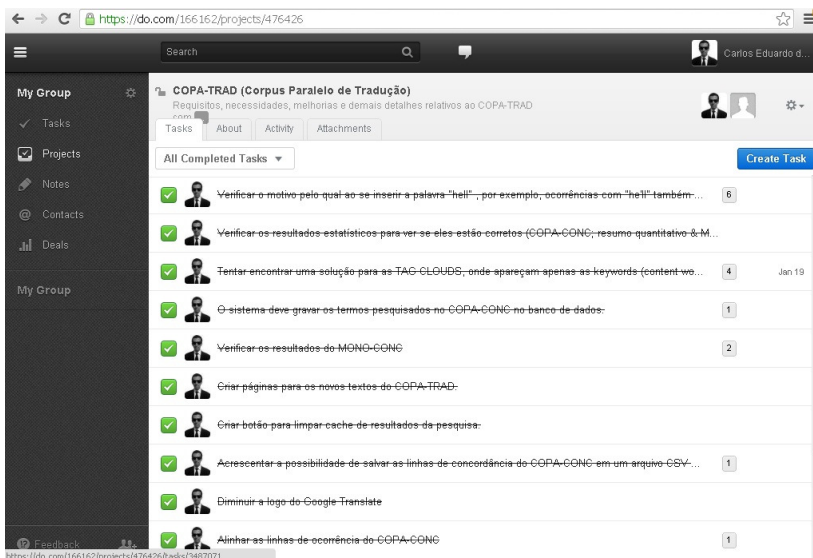


Figure 8. The completed task list.

The textual and media information collected in the requirements analysis phase were documented and organized in Evernote, an on-line software (Figure 9) available on <http://evernote.com>. This application was a key tool for the organization of technical information as well as providing easy access across multiple computers.

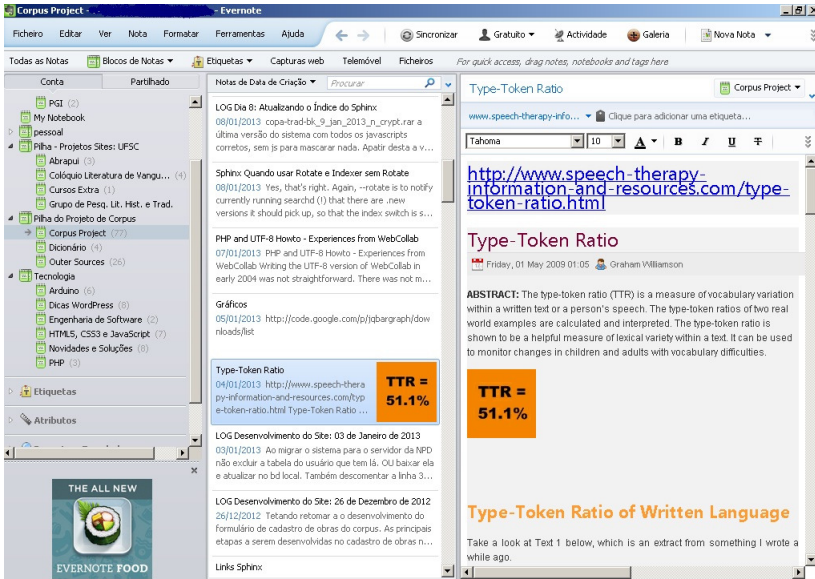


Figure 9. Material organized on the Evernote Application.

In addition, written notes, memos and forms resulted from the meetings discussions were kept in those electronic logs for further reference. Having discussed these managerial aspects of the corpus compilation, the next section goes deeper into corpus design by establishing the purpose for creating COPA-TRAD.

### 3.1.1 Purpose of Corpus Creation

One of the main purposes of this study was the creation of COPA-TRAD and its five subcorpora for translation research and pedagogy. In other words, COPA-TRAD aims at providing tools that allow users to investigate the practices involved in the translation of five specific text types (i.e. children's literature, academic abstracts, literary texts, multimodal texts and translation theory texts).

Other applications for COPA-TRAD are also possible, such as its use by professional translators as a Computer-Assisted Tool to support their decision while translating. However, the scope of the present study focuses only on the development of an online system capable of storing, processing and retrieving the texts comprising a parallel corpus. It is important observing that, even though the system developed is ready to include five subcorpora, due to time constraints,

only children’s literature texts have been used to feed the system, at the moment.

### 3.1.2 Type of Corpus

For the research envisaged, the best suitable type of corpus is the parallel corpus. As already mentioned, a parallel corpus consists of a set of texts in one language and their translations in another language (Olohan, 2004). Figure 10 illustrates it and also brings a small example extracted from the current version of COPA-TRAD.

Texto Fonte	Texto Alvo
You've forgotten the <b>magic</b> word,' said Harry irritably.	Você esqueceu a palavra mágica – "disse Harry irritado.
Improper Use of <b>Magic</b> Office	ESCRITÓRIO DE CONTROLE DO USO INDEVIDO DE MÁGICA

Figure 10. Bilingual parallel concordance showing how a parallel corpus works.

Baker (1995) discussion on parallel corpus was important to establish the theoretical and methodological framework for this study, she suggests that parallel corpus “allow us to establish, objectively, how translators overcome difficulties of translation in practice, and to use this evidence to provide realistic models for trainee translators” (p. 230-231). Based on Baker’s observation, COPA-TRAD can support translation students and researchers to investigate the practices professional translators use in order to translate collocational patterns, just to name a possibility.

Nevertheless, building a corpus is not only a matter of will; the following subsection discusses the issue of copyright in corpus design.

### 3.1.3 Copyright

A problem that can set the boundaries for the size of corpus is the issue of copyright. The data used by the corpus, in order to be available

for translators and researchers on the Internet, need special permission from the copyright holders. For overcoming this initial matter, it was decided to make COPA-TRAD available in restricted mode just for researchers from UFSC (Universidade Federal de Santa Catarina). Therefore, to access COPA-TRAD in its full access mode, users will need a username and password. Moreover, the second step is allowing the general public access to COPA-TRAD, but with a limited access (e.g. the corpus provides only texts without copyright).

In relation to the issue of copyright and to safeguard the present study, Fernandes (2004) motioned in his PhD thesis the following subject:

I have looked into the UK and Brazilian copyright laws and learned that the degree of academic privilege in copyright matters is often uncertain as it is difficult to say for sure how far scholars can go using other people's intellectual property as a corpus, especially if the corpus being created is for research purposes only and the research is not going to allow anyone to make use of it. Thus based on these three safeguards, I believe to be able to continue carrying out my research until I manage to get the permissions needed to make the corpus available to the academic community (p. 96).

The arguments provided by Fernandes (*ibid.*) seem to be sufficient for the present study because most of the texts in the system's database have come from his own research and some have been collected from the Gutenberg Project (see Section 3.1.4 below).

### **3.1.4 Selection of Texts**

The text selection phase for COPA-TRAD was based on the assumptions discussed by Baker (1995) and Fernandes (2004). Fernandes (*ibid.*) paraphrasing Kennedy (1998) argues that corpus-based work involves a great deal of planning before establishing explicit and rigorous criteria in the selection of hardware, software and texts. It is this careful planning that enables a corpus to provide accurate and reliable descriptions and ensures it can be used or referred to by other researchers (p. 90).

The selection of complete texts for COPA-TRAD comes from three different main sources. The first source of texts is Fernandes (2004) PhD thesis in which he has compiled a parallel corpus consisting



of 12 children's fantasy books and their respective 12 Brazilian-Portuguese translations in order to investigate the practices involved in the translation of names (see Chapter 2 - Literature Review).

COPA-TRAD, nevertheless, is an open-ended corpus and since the beginning of its conception, it was designed taking into account this important characteristic. As a result, the insertion of new books becomes a permanent process.

During the development of COPA-TRAD, Bruna Coletti, from an Undergraduate Research Mentorship Program, was responsible for selecting texts from Gutenberg Project<sup>4</sup> (a project which digitizes and provides public domain books in many languages) to be included in the COPA-LIJ subcorpus. This student organized and adapted default textual files and pre-aligned them at paragraph level. She also collected extralinguistic information about each of those texts. These efforts combined with pre-existing aligned texts provided by Fernandes (2004) have contributed for COPA-TRAD start its operations with 18 books in English as well as 18 books in Portuguese, which was a reasonable number of books at the beginning.

Finally, the users themselves will be the third source of texts. They can submit their own texts following a specific set of criteria established and made available on the COPA-TRAD website. The system has a specific module for text submission. In this module, users can check the texts they have submitted as well as select the option to submit new text. In the submission panel there is a form to fill in with extralinguistic information of the new book to make available on COPA-TRAD. The user has two options to upload the source text and the target text. A moderator verifies the content submitted in order to check if the extralinguistic information and texts were provided correctly. The moderator has a special module for the verification of the texts submitted. For this reason COPA-TRAD started with a reasonable quantity of books. However more books can be added because COPA-TRAD is an open-ended corpus.

### **3.1.5 Corpus Creation: Shortcomings and Possible Solutions**

Before discussing each stage carried out during the project it is necessary to discuss the possible shortcomings and solutions related to the creation of an on-line Parallel Corpus. The building of a parallel corpus may appear to be a simple task for a researcher who has just

---

<sup>4</sup> More information on: <http://www.gutenberg.org/>

begun studying CTS. The creation of a parallel corpus and its processing system involves a set of technical procedures that will be discussed here. The basic functionality (e.g. find, retrieve and display) of a corpus might give the impression that is a simple task. Therefore lay people, or maybe beginners to CTS, could possibly think about corpus building as a simple task because it involves “text compilation only”; there would be no multimedia, no 3D complex animation or any other computational high-tech product whatsoever.

For beginners who have a misconception in relation to corpus creation problems could emerge and become a pitfall. After starting to work with the development of an electronic corpus without a solid knowledge on the subject and taking into account the steps involved at the design level, the neophyte researcher would face a set of problems to solve. As a result, the neophyte researcher would perceive that:

- Texts in electronic format are difficult to handle in computer programming, various types of textual patterning have to be used such as regular expression (i.e. a pattern to recognize strings/texts automatically);
- Problems such as lexical items, punctuation and different characters for a specific language can consume many hours to find a suitable solution for the intrinsic features of the languages and its conversion to digital format;
- Dealing with a substantial textual volume in electronic format is time consuming and demands high levels of computer processing power and large amount of memory;
- The malformed (e.g. logical errors, bugs) software algorithm utilized to parse and process the text can lead to a computer crash, causing the operational system to become unresponsive. (i.e. Windows, Linux or MacOS);
- Texts in a corpus have to be indexed for data retrieving; a badly organized index could give for the researcher inaccurate results leading the investigation to false results jeopardizing the project’s reliability;
- The decision to build the corpus in a specific computer programming language (such as PHP, Java, C, C++ and others) can affect the whole project in a positive or negative way. It all depends on the developer’s expertise;
- The researcher will have to determine what specific tools are necessary for future development. For example, is the information stored in the corpus going to be used for a future mechanism not developed yet? The database internal structure

without prior planning predicting any future mechanism could cause problems in the future.

Other problems or insights could be listed but the most important ones were mentioned to contextualize the reader about the importance of a detailed corpus design. Designing a parallel corpus is a challenging task because the lack of technical information available and the complexity involved. An instance of this is the number of relevant parallel corpora available and developed in Portuguese speaking countries: COMPARA, from Portugal, is a bidirectional parallel corpus in the linguistic pair English/Portuguese. In Brazil the project COMET (*Corpus Multilíngue para Ensino e Tradução* hosted at Universidade de São Paulo) is a multilingual corpus for teaching and translation. However, as observed in the introduction, there is a gap (e.g. no user friendly interface, in some cases the use automatic alignment without human supervision or the availability of Children's Literature texts) to be filled in relation to parallel corpus in the linguistic pair Portuguese – English (Section 1.1).

This section shed light on the possible shortcomings involved in the creation of a parallel corpus and its complexities. In addition, this chapter will encourage researchers to reflect on the issues covered in this section. There are three phases of development of a parallel corpus. The first phase of development the requirements analysis is discussed below. Afterwards, two other phases (building and implementation) are going to be discussed.

### **3.1.6 Requirements Analysis**

The applied objective of this study is the creation of a web based bi-directional (English / Portuguese) parallel corpus and five subcorpora for Translation Studies Discipline (Section 3.1). According to Pressman (2001) “requirements engineering provides the appropriate mechanism for understanding what the customer wants, analyzing need, assessing feasibility, negotiating a reasonable solution, specifying the solution unambiguously, validating the specification, and managing the requirements (...)” (p. 256). The requirements were gathered in the initial phase using interviews, forms and a comprehensive research related to corpus linguistics, corpus-based translation studies and parallel corpus. The requirements, use-cases, amongst other elements (e.g. workflows and diagrams) were stored, modeled and managed in Enterprise Architect version 8.0 (Figure 11).

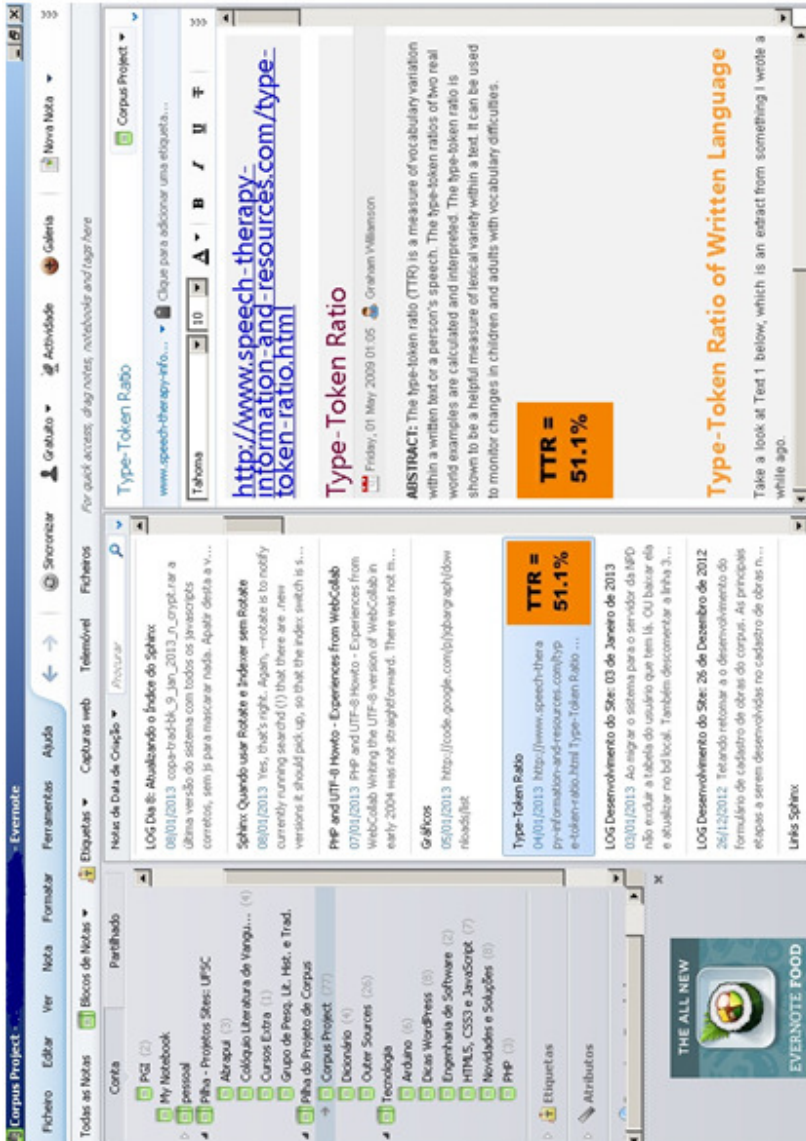


Figure 11. Requirements and the overall design.

The technology adopted for COPA-TRAD was defined in the requirement analysis phase also. The major technologies defined at this

part are listed below. It was given preference to Open Source (i.e. available for free) technological alternatives:

- The Operational System (OS) for the server is OpenSUSE Linux (<http://www.opensuse.org/pt-br/>), this is the default choice server's OS at UFSC;
- MySQL (<http://www.mysql.com/>) database was chosen because it is reliable and it supports a large quantity of data. According to Oracle Corporation (n.d.-a), the size amount is 256 Terabytes;
- The NGINX HTTP Server was chosen as the web server, (<http://www.nginx.org/>) which is the default server available for UFSC hosted sites and systems;
- PHP (<http://php.net/>), JavaScript and HTML were chosen as programming languages mainly because of the author's expertise in these languages and also its popularity and reliable technology for Web Applications (i.e. software developed for the web);
- CodeIgniter was chosen as the framework for development (<http://ellislab.com/codeigniter>), it was built for PHP programming language and it complements the project conducted providing a set of useful libraries and a fast learning curve;
- As a search engine, Sphinx - an Open Source search server - was chosen. Sphinx is used for COPA-TRAD because MySQL search support is not as accurate for the specific kind of search in COPA-TRAD. As an example, MySQL ignores words with less than 4 characters such as common words (i.e. also known as stopwords) like "the", "again", "be", "is" and "will", among many others. For a linguistic corpus, such characteristic is not desirable. The support documentation for MySQL dictates that such features can be changed in the configuration file. However, on a shared server that hosts other websites, such changes can present a problem of search relevance. Weak use of complex search/retrieval, amongst other factors, contributed to the decision of not using MySQL's search support for COPA-TRAD tools (i.e. COPA-CONC and MONO-CONC). As a result, another solution had to be implemented. This solution was the Sphinx Search server, which provides deep indexing, searching and fast retrieval of data. In addition Sphinx provides an API (i.e. Application Programming

Interface – a set of programming codes to integrate different computer software) to integrate with systems developed in PHP, which runs on a server. Sphinx is robust and reliable, for example, big websites such as Infegy - <http://infegy.com/> - which uses Sphinx has indexed 22 billion documents from Twitter, Facebook and blog posts “the speed and precision of the Sphinx engine enables Infegy's customers to efficiently measure online sentiments and trends”, another example is the website Craigslist - <http://craigslist.org/> which executes 250,000,000 million queries each day (Sphinx Technologies Inc., n.d.). Such technology was chosen for COPA-TRAD because the positive points mentioned above and in the short and long terms, there are plans to expand COPA-TRAD database.

An essential point to mention is the Sphinx technology applicability for Corpus-based Translation Studies. To attest the reliability of COPA-TRAD, another researcher who also used Sphinx for corpus linguistics was contacted by e-mail. Professor William H. Fletcher, from the US Naval Academy, shared his experiences in dealing with Sphinx, PHP and Corpus Linguistics, which helped to further the development of COPA-TRAD. Professor Fletcher is the developer of the “Phrases in English (PIE)” corpus, which is available online at <http://phrasesinenglish.org/searchBNC.html>.

### **3.1.7 Database for Text Storage**

The storage mechanism for a corpus may vary in its nature; each of them has advantages and disadvantages. For the present study, the storage mechanism chosen was of database type. In short, some issues that lead to choose this kind of storage mechanism were because the capacity to store large amounts of data (some databases have millions or even billions of words, e.g. some companies/websites such as Wikipedia and YouTube (Oracle Corporation, n.d.-b) use the same database adopted for this project), the speed of data retrieval, processing and delivering the results the researcher wants in a fast but accurate way.

The common problem related to a database that uses this kind of mechanism is the complexities involved in the design of its internal structure that must be well designed to avoid potential problems in the present moment or future. For the present study, the first step in the design of the database was to verify which corpora use the same type of storage mechanism. After identifying the corpora that used such storage mechanism, the researchers involved in such projects were contacted to

collect more information, as well as, the technical approach they used to create the database infrastructure. Another crucial part was related to finding literature that covers the use of databases in the corpus area.

After acquiring the necessary expertise in dealing with databases to adopt it for linguistic corpora, the next step was designing the corpus database itself. In the requirements analysis phase, questions concerning the structure of the database, for example, how data will be organized, managed and stored, were raised in order to meet the necessities of the project. Some of these questions are suggested below:

- Should the texts be stored at word level, sentence level or paragraph level?
- What information from the texts should be made available?
- Will any reporting of statistics such as word frequencies or any other mathematical operation be performed?
- What would be the future plans of the project? Is there at least any prediction?
- What are the source languages for the texts provided? Will any of the text have special characters from languages such as Russian, Mandarin, Greek, Hebrew, Japanese or Arabic?
- What are the requirements of the corpus under development?
- Which meta-textual information will be stored? Can the author/translator nationality, sexual orientation, and biography amongst other factors have an influence on the final product?
- How texts will be processed and used in the corpus?
- How many languages the corpus will cover?
- What type of corpus is it? Is it a parallel, multilingual or comparable corpus?

Such questions have a direct impact on the database design phase. One important aspect during the design of COPA-TRAD relies on how the database was constructed. The database is the core of the corpus because textual information is stored in it. Focusing on the main aspect of the database, which is the storage of parallel data, careful attention was given to the design task.

The database tables were designed in Enterprise Architect 8.0. Knowing that texts are unstructured information, some techniques from the text mining area were adopted because it “deals with the machine supported analysis of text” (Hotho, Nürnberger, & Paaß, 2005, p. 4). For instance, in text processing procedure which extracts and align the text sentences as well as the words which are extracted through a tokenization process i.e. “a text document is split into a stream of words

by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces” (Hotho et al., 2005, p. 6) after the application of the mentioned text mining technique the data obtained was stored in the database (i.e. structured information).

The database tables (i.e. a set of elements organized using columns and rows to store structured information) were designed in a way to store parallel texts and the relationship between the corpus data. In database context, a relationship is understood to be the association between  $x$  and  $y$  (e.g. the table which stores a specific information on a book such as name or year of publication has a relationship with the table which stores the sentences and words). The parallel sentences were stored in the same table and an auxiliary table was used to provide the relationship between a source and target sentence as Figure 12 illustrates.

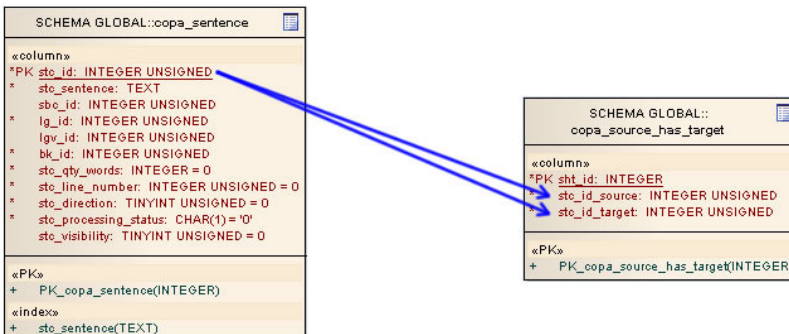


Figure 12. The relationship between source and target sentences.

Figure 12, represents the schema of two tables, the first table stores the sentence in “stc\_sentence” and each sentence has an ID (i.e. unique identification) named “stc\_id”. The second table “copa\_source\_has\_target” is responsible for providing the relationship between a source sentence “stc\_id\_source” and a target sentence “stc\_id\_target”. The relationship between sentences and words were also designed as we can observe in Figure 13.



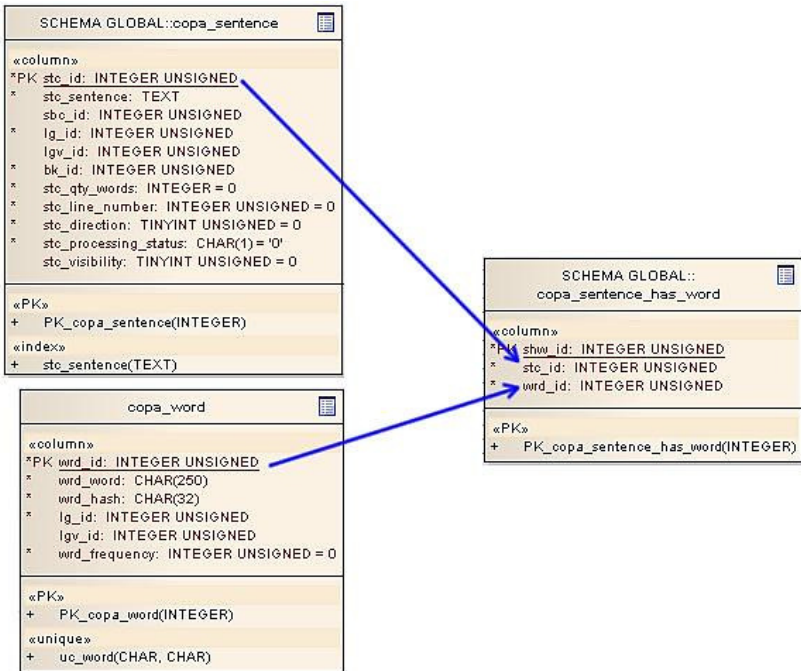


Figure 13. The relationship between sentences and words.

Figure 13 shows the schema representing the relationship between sentences and words. Each sentence ID “stc\_id” stored in the table “copa\_sentence” and each word id “wrд\_id” stored in “copa\_word” are associated in the auxiliary table “copa\_sentence\_has\_word”. These relationships in the table are countable and can be used for reporting or statistics. Following the same idea, we can compare the relationship between books and sentences and between books and words. These relationships are useful for counting the word frequency in a book and make it possible to extract that information by counting the number of relationships between books and words. Figure 14 illustrates a practical example of how the frequency data is displayed to the final user.

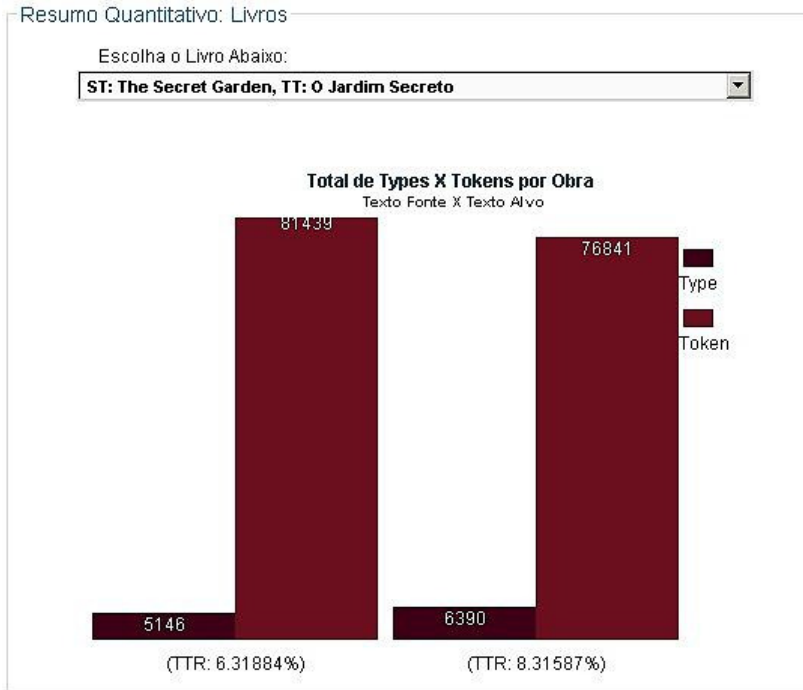


Figure 14. Chart displaying the frequency of words.

Baker (1995) explains the importance of type, token and ratio in investigating translation using a corpus. The author says that “any sequence of letters with an orthographic space” is considered a word or more technically a “token”, according to the author, “each occurrence of the word *day* is counted as an individual token and we can say that there are x tokens of *day* in a given corpus” (p. 236).

In relation to types Baker says that “*day* itself is a type, no matter how often it occurs”, finally she concludes “we can say that there are x tokens of the type *day* in a corpus”. This kind of information could be calculated while the data is delivered for the user, but for a better performance the type, token and ratio were stored in database so there is no performance shortcoming in COPA-TRAD whatsoever.

The type, token and ratio were calculated and stored during the sentence processing operation by COPA-ALIGNER module. Similarly, the word extraction and processing was carried out by COPA TOKENIZER module.

Finally other relevant data found during the requirements analysis were stored in the database. The list below provides an overview of other relevant information collected and stored in database:

- **Books** – all the extralinguistic information listed by Fernandes (2009, p. 25-26) including language, language variation, genre, type, token and ratio from the whole book;
- **Author** – Name, biography/profile, photograph, gender, nationality at birth, current nationality, domicile, language, language variation;
- **Translator** – all the extralinguistic information listed by Fernandes (2009, 25-26) and also language variation, photo and biography/profile.

A key point in the design of the tables which constitutes the table database *per se* is the decision about what kind of information should be stored or not. The data provided was based on a comprehensive theoretical framework. Common extralinguistic information for the corpus (below), as suggested by Fernandes (2009, p. 25-26), were used:

**Translator:** name, gender, age, employment status, translation workload, nationality at birth, current nationality, mother tongue, domicile, membership of translator associations.

**Translation:** text category, collection, text title, mode, word-count, special features, date of publication, place of publication, publisher, publication of the name of the translator(s), copyright, reviews and appraisals, prizes (not only who received them, but also who awards them and under which circumstances), distribution data (e.g. if the text has been reprinted, perhaps with alterations).

**Translation process:** relation between translation and source text, direction of translation, written translating mode, commissioner, subcommissioner, editing, time lag (i.e. time elapsing between commissioning and publication).

**Source text:** language, status, name of the author(s), gender of the author(s), date of publication, place of publication, publisher and prizes.

With the tables modeled, a routine to generate and export the table's schema from Enterprise Architect version 8.0 was used to transmit the tables to the database. With the importation procedures finished a crucial stage was completed. The next stage was the implementation of COPA-TRAD system to process the texts in electronic format and store these texts in the database. The following section is going to discuss the main snippets for COPA-TRAD implementation.

### **3.2 Corpus Implementation Phase**

After discussing the design issues informing COPA-TRAD compilation and project documentation, the focus in this section relies on the application and technical issues related to the system development, tools and support modules (or internal tools). One point to have in mind is that this project was carried out during a two-year period and covers only the main practical applicability of the corpus building.

The development process adopted for COPA-TRAD was the incremental one, that means: with each increment of the software (or addition of new functionalities) a new version is delivered and ready to be used (Pressman, 2011, p. 41). Figure 15 illustrates the steps in this process: Firstly, the developer defines what will be delivered after s/he prepares the design of system architecture. Next the requirements and specifications are defined. Following the workflow, the next step, is the development of the system. Subsequently, comes the developed part (increment) is validated. Then the part developed is integrated in the whole system and another validation comes again to check if the system is working properly. Finally there is a decision box to check if the system is complete or not to determine if another cycle start or not.

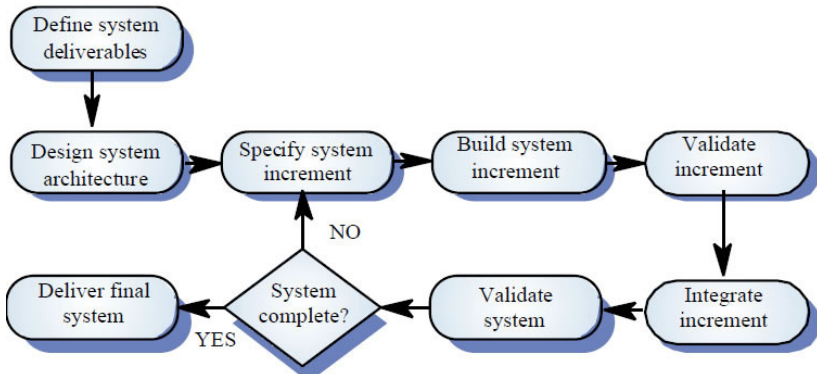


Figure 15. The incremental development process (Sommerville, 2000).

In addition, it was adopted the creation of prototypes to verify the functionalities that would be part of COPA-TRAD tools. According to Somerville (2000), a prototype eliminates misunderstandings, provides a way to find missing functionalities and may serve as a “basis” for system analysis. The workflow presented in Figure 16 gives a glimpse about the steps involved in the development of a prototype system. The workflow starts with the establishment of the main objectives of the prototype after a specific functionality is defined; the next step is then the development of the prototype and finally its evaluation.

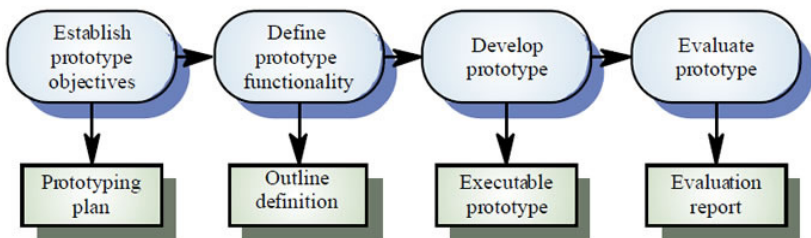


Figure 16. Prototyping process (Sommerville, 2000).

The development of prototypes for COPA-TRAD tools unveiled the real necessity of such technology in the Corpus-based Translation Studies area. Figure 17 presents one of the first COPA-CONC prototype: on the left, there is a filter panel to narrow down search on the corpus; on the top, there are meta-data about the text and, under it, a table displays the corpus results: on the left side column the original text

(i.e. source text) and on the right column, the translated text (i.e. target text).

**Termo:**  
hiccoughing

**Direção:**  
Inglês para Português

**Sub-Corpus:**  
COPA-LIJ

**Gênero:**  
Infantil

**Coleção:**  
Harry Potter

Val »

\* Opções de Filtro Avançadas \*

Foram encontrados 2 entradas para o termo "hiccoughing" na coleção de textos do Harry Potter no sub-corpus: COPA-LIJ.

Texto Fonte: harry potter philosopher stone.

Texto Alvo: harry potter e a pedra filosofal.

Funções	Texto Fonte	Texto Alvo
✖	Harry, trying to say 'Shhh!' and look comforting at the same time, ushered Dobby back onto the bed where he sat <b>hiccoughing</b> , looking like a large and very ugly doll. At last he managed to control himself, and sat with his great eyes fixed on Harry in an expression of watery adoration.	Harry, tentando ao mesmo tempo fazer O elfo se calar e dar a impressão de consolá-lo, levou Dobby de volta à cama, onde O elfo se sentou entre soluços, parecendo uma boneca enorme e muito feia. Por fim ele conseguiu se controlar e se sentou, os grandes olhos fixos em Harry com uma expressão de aquosa admiração.
✖	It was nearly lunchtime and as Harry had only had one bit of treacle fudge since dawn, he was keen to go back to school to eat. They said goodbye to Hagrid and walked back up to the castle, Ron <b>hiccoughing</b> occasionally, but only bringing up two, very small slugs.	Era quase hora do almoço e como Harry só comera uns quadradinhos de chocolate desde o amanhecer, estava doído para voltar à escola e almoçar. Eles se despediram de Hagrid e regressaram ao castelo. Ronny tossia de vez em quando, mas só vomitou duas lesminhas.

Figure 17. A working prototype for COPA-CONC.

The development of prototypes for each system module demonstrated as useful and capable of processing large amount of texts even during the design phase. These prototypes helped to find and determine specific features for the system. Due to the nature of the project, newer TS researcher's needs will be prototyped in future design phases of the system.

### 3.2.1 COPA-TRAD Framework

The internal structure of COPA-TRAD system relies on *CodeIgniter* framework. The mechanisms for the system were developed on this framework, following the requirements previously planned. Due to the requirements and the problem of copyright, COPA-TRAD offers two different types of access to its text collection: a public one, aimed to provide access to texts without copyright and a secured area for texts that are copyrighted or not.

COPA-TRAD users are divided in three different access roles. In each user role, the system adapts the corpus content (i.e. provide text with copyright or not). This security measure was adopted to protect copyrighted content. The CodeIgniter framework already provided other parts that constitute the system, such as database support, security environment, among other basic and complex modules.

The development of COPA-TRAD focused on its real requirements and problems *without reinventing the wheel*. Firstly, the Interface Design and its events (i.e. reactions such as message alerts and color changing) planned according to possible user actions were implemented. The visual elements that comprise the interface received special attention to make it as *user friendly* as possible. To do so, Ajax<sup>5</sup> technology was used to implement a comparable interface like in RIA (Rich Internet Application). This technique provides the best user experience. In addition, the use of Ajax improves the performance because it enables the system to send and receive data asynchronously to a server i.e. without reloading the entire web page multiple times. The HTML, CSS and JavaScript code were carefully implemented in order to run COPA-TRAD on smartphones and tablets (Appendix C). Secondly, the user management and authentication modules were implemented to achieve the security requirements. Thirdly, the auxiliary web pages as well as the news management module were developed. Figure 18 illustrates an example of a class diagram created in the design phase; the figure shows two diagrams related to the system news management module.

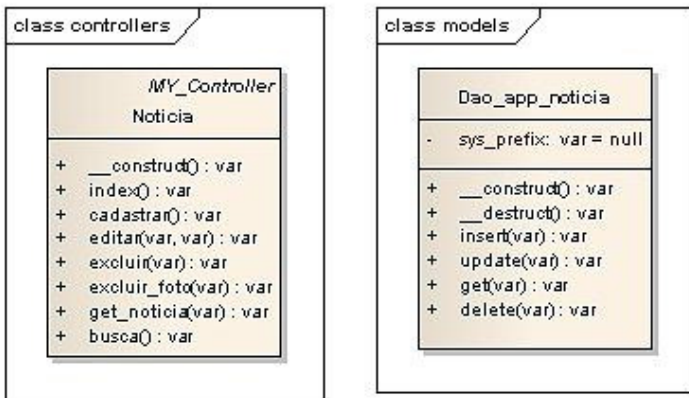


Figure 18. Example showing the diagram classes for the news management module.

<sup>5</sup> Ajax stands for Asynchronous JavaScript and XML – a kind of data exchange between a client/user and the server, the data exchange runs in background avoiding the refresh of the whole page on an Internet browser.

It is necessary to emphasize that each requirement implemented on COPA-TRAD, class diagrams were designed in order to assist the development process and were also used for the system documentation. Figure 18 was an example extracted from the design documentation.

### 3.2.2 COPA-TRAD Tools

Currently, COPA-TRAD has the following tools to assist researchers and translators: COPA-CONC, MONO-CONC, CORPUS-BUILDER, WORDLIST and COPA-STATS. The implementation of these tools followed the three stages mentioned in the previous section. Naturally, there are small differences due to the complexity of some tools. In order to COPA-CONC work as previously planned, it has to access three different APIs<sup>6</sup>. The first and most important API is the one, which communicates with Sphinx Search. There are also two APIs, one to connect and request content from Microsoft Translator, and another to communicate with Google Translate. The search algorithm employed to find the user requested keyword was provided by Sphinx. In addition, customized filters are also present to narrow down the search to a specific subcorpus, a book or the desired language pair.

In the near future, new filters will be implemented. MONO-CONC uses just Sphinx API and, at present, has the KWIC (Key Word in Context) display format to show the results. Figure 19 gives a glimpse of the KWIC display format for the keyword “replied”, with its co-texts, in each side of the keyword, limited in one line. Having in mind that COPA-TRAD stores full sentences, not just a part of it, for MONO-CONC, a special algorithm was developed in order to split sentences in co-texts for the left and right side of a keyword. The same algorithm works for non-English words with Latin characters such as the ones from Portuguese language.

---

<sup>6</sup> Application Programming Interface (API) – technology, which facilitates the interface between two different systems. For example, system “X” needs to access the functionality of system “Y”, the communication between them occurs through an API.



Lingua Ingles		
He had a look of trollish cunning on his face as he	replied	„Plenty of room for all of us, Wood.“
made up his mind to make friends with thee,”	replied	Ben. “Dang me if he hasn’t took a fancy to thee.”
“No, tha’ hasn’t,”	replied	Martha.
“Yes, I think so,” he	replied	.
“No,” he	replied	after waiting a moment or so. “I am Colin.”
“I dont when I am by myself,”	replied	the Rajah, “but my cousin is going out with me.”
“Very good, sir,”	replied	Mr. Roach, much relieved to hear that the oaks might
“Nothing disagrees with me now,”	replied	Colin, and then seeing the nurse looking at him
was something between a sneeze and a cough,” she	replied	with reproachful dignity, “and it got into my
“Dickon can sing it for thee, I’ll warrant,”	replied	Ben Weatherstaff.

Figure 19. Example showing the MONO-CONC’s KWIC display format.

COPA-BUILDER incorporates a tool developed by Yasu Imao (2011) and it was customized for COPA-TRAD needs and visual elements. The changes implemented in the existing tool were an improvement to achieve a friendly interface. Such implements were: the possibility for choosing source and target languages, a better error handling system to report messages to the user, and the database interaction to save the data provided by the user for further research.

The WORDLIST module implementation did not require any kind of API; this tool was developed to retrieve and present the words and their frequency from the database table “copa\_word”. Particular attention was given to the implementation of the word cloud view (Figure 20), depending on the frequency of a word, its size and color tone changes. Due to the quantity of database requests and processing time to create such word cloud, a special feature from *CodeIgniter* framework was used, the cache support. Once a day, by the first time the WORDLIST home is accessed *CodeIgniter* generates a cache in HTML of the page including the word cloud. After that, while accessing the WORDLIST home, the page displayed is the cached version, which brings the content in a fast way without overloading the server.

## Keywords: Inglês

As palavras aqui listadas apareceram pelo menos 600 vezes. As palavras mais comuns foram excluídas desta lista.

potter head black fowl holly gwendolen colin **harry** spiro door  
 dumbledore time malfoy ron janet eyes mary lupin hermione butler  
 weasley artemis hagrid cat looked professor sname

Figure 20. Word cloud from WORDLIST tool.

COPA-STATS is the statistical tool for COPA-TRAD. Currently, only the basic functionality of this tool was developed. The generation of the graph was implemented in JavaScript language; as a result there is no processing overload on the server side. The user's computer is responsible to receive the statistical data and render the graph on screen.

### 3.2.3 COPA-TRAD Textual Processing Modules

COPA-TRAD has individual processing modules for statistics generation, among others. However, at this part, attention is focused in two main processing modules: The first one is COPA ALIGNER and the second one is COPA TOKENIZER; both modules were built from scratch and they are responsible for a vital part of the system, which is the application of text mining techniques.

COPA ALIGNER is the module used for the source text and target text alignment. An important feature developed for this module is a code from PHP language that guarantees the execution of this module. Even if the system administrator refreshes the page accidentally or loses the connection with the server, the entire text file will be processed in any case. The basic functionality from the previous module is also present in COPA TOKENIZER, which is responsible for word extraction and the frequency list update.

COPA TOKENIZER was developed with subroutines responsible for word normalization of non-English words existent in English texts, non-alphanumeric character removal and token and type counter. Regarding word normalization for non-English words within English texts the necessity of this subroutine was mainly because English texts are in UTF-8 encoding (i.e. an encoding which represents a set of characters in computers mainly non Latin characters). Non-English words that appear in English texts might cause a problem: the COPA

TOKENIZER module would consider 2 words rather than one. Consider the following sentence from the book *The Secret Garden* extracted from COPA-TRAD:

Her black dress made her look yellower than ever, and her limp light hair straggled from under her black crêpe hat.

During the word extraction at this sentence COPA TOKENIZER would split the word in red (i.e. crêpe) in two chunks because the algorithm for English language does not know the character “ê”. This is the reason why normalization was necessary at this point. Considering Table 2, it is possible to observe how the code developed converts the characters in line A by the ones in line B.

Table 2. *Characters normalized when appeared in English texts*

	ŠŒŽšœžŸÿµÀÁÃÄÅĂÆÇÈÉÊËÏÎĴŃŇŌÓÔÕÖØÙÚÛÜÝ ßàáâãäåæçèéêëìíîïðñóôõöøùúûýÿ
	SOZsozYYuAAAAAAAAACEEEEEIIIDNOOOOOOUUUUYsa aaaaaaceeeeeeiiionooooouuuuyy

Consequently, the word “crepê” would be converted to “crepe” during word extraction for English language texts.

### 3.3 Corpus Processing

The objective of this section is the description of COPA-TRAD processing mechanisms for textual analysis, extraction and searching. Depending on the complexity and technicality some parts will be omitted if necessary. To do so, it will be described how the modules from COPA-TRAD process the textual information. According to Kenny (as cited in Fernandes, 2009) “a corpus on its own is of little practical use if there are no techniques to search, catalogue and extract potential data” (p. 30). Based on Kenny observation, the following subsection explains how the searching, extraction and data presentation works in the COPA-TRAD system.

#### 3.3.1 How COPA-TRAD’s Search Engine Works

Section 3.3.3 and 3.3.4 presents the procedures followed in the implementation phase of the support tools named COPA ALIGNER and COPA TOKENIZER. Before going deeper in the internal description of such tools, it is necessary to explain the overall mechanics of COPA-TRAD starting with the user requested information on COPA-CONC.

Figure 21 presents the system topology explaining how a search for a term triggered by the user is processed inside COPA-TRAD. To begin, the first moment signaled in the figure, COPA-TRAD (running on the “NGINX / PHP” server) do a request to Sphinx Search Server (through an API) asking to search for the term “XYZ”, Sphinx checks its own keyword index and returns all sentences ID’s (through an API) to COPA-TRAD. Next, at the second moment signaled in the figure, COPA-TRAD in possession of the sentence’s ID provided by Sphinx; requests from MySQL all the sentences based on the available ID’s. Hence, MySQL returns all sentences from the database, COPA-TRAD process all the sentences and highlight in each sentence the term “XYZ”, after that it sends all processed data to user’s computer. An operation not covered in the Figure 21 is the storing in database of search terms (or keywords) requested by the user. It is necessary to emphasize that this searching mechanism works in similar way for both, COPA-CONC and MONO-CONC. The difference between the two modules relies on the content displayed to the user. COPA-CONC delivers the aligned sentences (source and target) while MONO-CONC delivers just sentences of one language previously chosen by the user.

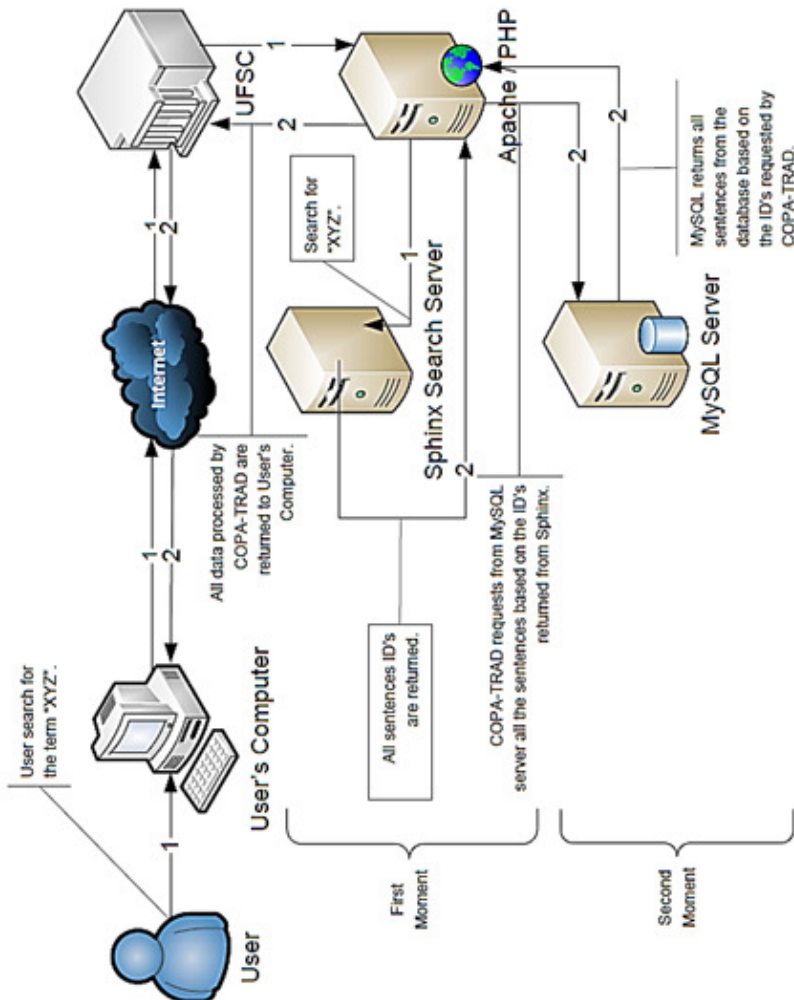


Figure 21. System Topology showing how a user request is processed inside COPA-CONC.

After all, a lay person could think why get just the sentence ID's from Sphinx and then do a request to MySQL to return the sentences and its related extra textual information. Sphinx does the hardest and most complex part, which is the search and ranking operations without burden the developer with the implementation of complex matching

algorithms. Some advantages in returning the IDs, for example, are related to the fact the IDs are unique and easy to handle; with the IDs, it is possible to retrieve many other pieces of information associated with the sentence's table and not just the textual sentence itself.

In addition, a factor to be considered is the availability of just IDs provided by Sphinx's is the default system mode. Under these circumstances the system topology presented in Figure 21 is the most convenient for this research.

### **3.3.2 How Texts are Processed in the COPA-TRAD System**

This subsection focalizes on the internal tools: COPA ALIGNER and COPA TOKENIZER by extending the discussion to the textual processing techniques applied in both tools. These tools extract and manipulate the raw text files sent by the user. COPA ALIGNER is focused on the creation of bilingual concordances and COPA TOKENIZER is focused on the creation of wordlists.

Nevertheless, before going into details and the internal functionality of the mentioned tools, it is necessary to understand how texts are delivered for COPA-TRAD. First of all, COPA-TRAD does not have a limited quantity of texts; the system has a growing database because each COPA-TRAD user has the option, inside the system, to send his/her texts. Certainly, these texts must have some relation with the COPA-TRAD subcorpus and its genres. Figure 22 brings the system topology showing how texts are sent and also the interaction between users and moderators.

The first step, as pointed in the top left of the figure, user logs in into his/her account and access the text submission panel. Then the user accesses the form submission and fills in the required information such as book title, author name, language, etc. (Section 3.2). At this part the submission of the source and target texts files are mandatory. Finally the user clicks of the button "save" to save the provided information and send to COPA-TRAD.

The system receives the information and returns a success message if all information provided is correct, or rather, an error message is shown. At this moment, COPA-TRAD sends an email to the available moderators (i.e. users with privileges to check all texts sent by other users) informing that a new text was sent and needs the inspection to confirm if the data sent is correct. The moderator checks the data including the source and target text and writes a small feedback which can be read by the user who sent the text, and after, the moderator chooses if the text is approved or not to be part of COPA-TRAD. In case

the text is not accepted due to some incorrect information provided the user has the option to change that information and send it again for moderation.

However, if the text is approved, the user receives by email a notification about it. The notification message tells that the submitted text has been accepted and will be available in 48 hours or less (i.e. this time is necessary because the new content has to be indexed by Sphinx which is scheduled to run every 24 hours). After this point, the user editing action in the text submission panel is blocked because no adjustment is permitted after the text is approved; then the texts sent are ready to be processed (i.e. the tools COPA ALIGNER and COPA TOKENIZER can be started to process the approved texts). Finally, the texts are processed and the administrator can activate two auxiliary mechanisms, which will check the new information in the corpus to update the available statistics.

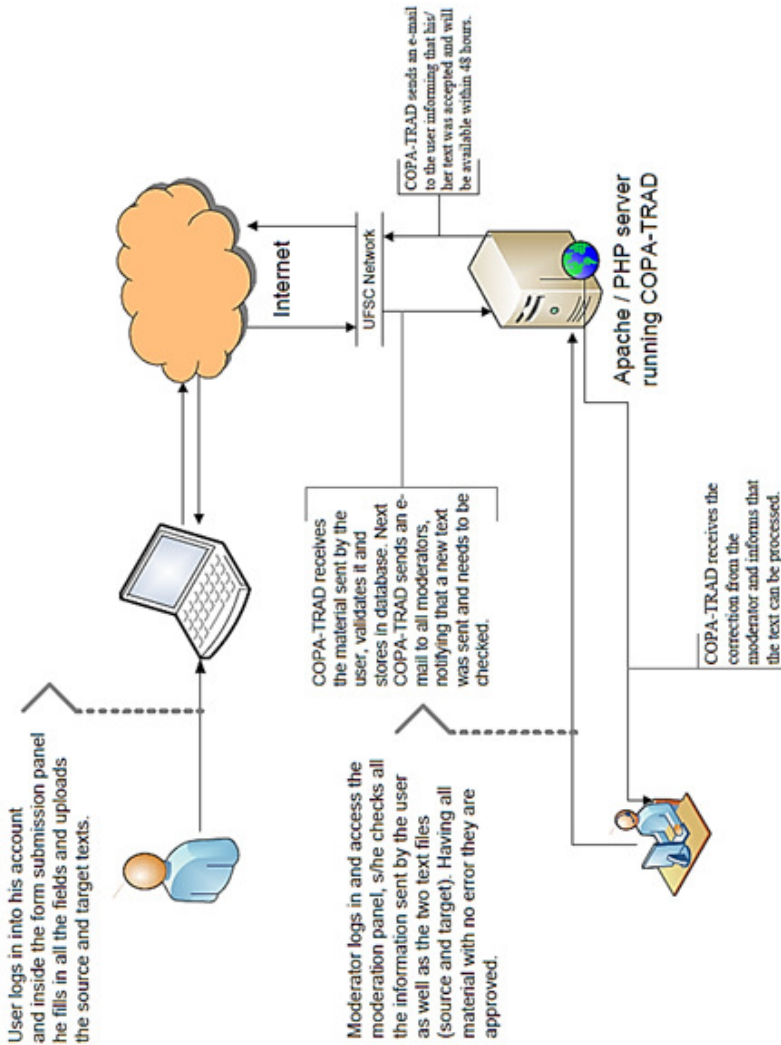


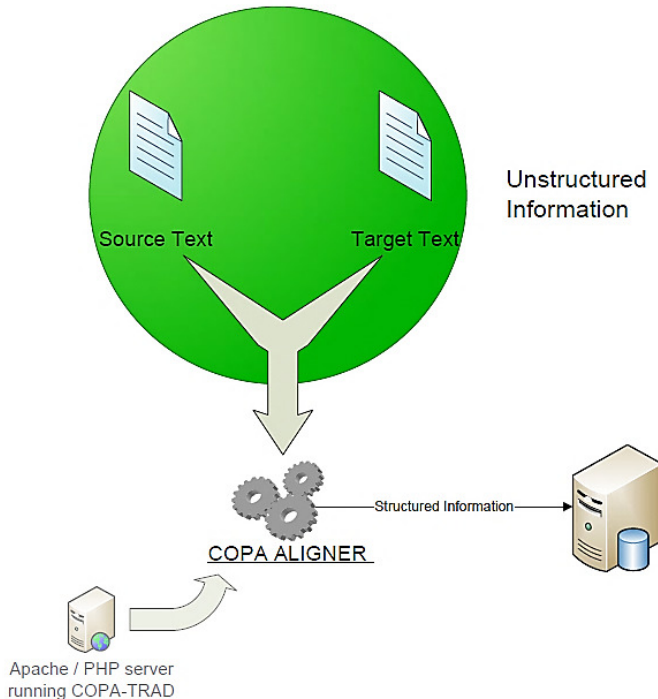
Figure 22. System topology showing how texts are submitted to COPA-TRAD.

Having explained how texts are sent to COPA-TRAD now I will explain the processing text routine for COPA-ALIGNER and COPA-TOKENIZER.



### 3.3.3 COPA ALIGNER Processing and Alignment

The mentioned parallel corpora available on Internet for the Portuguese-English language pair lack update in its major tools to make the information processing easier and robust, helping users to get their tasks done quickly and accurately. Facing the challenges in text processing, COPA ALIGNER was developed from scratch. According to Figure 23, when COPA ALIGNER is triggered, it checks and fetches the text to be processed. These texts (source and target) are in txt file format. Basically, COPA ALIGNER opens both texts and each sentence from the text is extracted and stored as structured information into the database. Before the sentences storage, COPA ALIGNER performs statistics for each sentence and these quantitative data are stored together with each sentence.



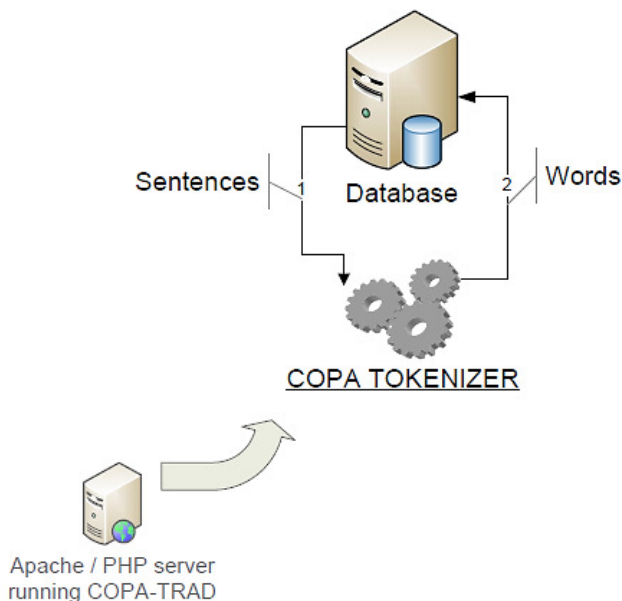
*Figure 23.* Architecture Diagram of COPA ALIGNER showing how texts are processed.

During the previous process, the sentence is aligned. For each sentence pair (source and target), after stored in database, COPA ALIGNER recovers both sentence ID's and creates the relationship between these two sentences (i.e. the parallel alignment). COPA ALIGNER creates a relationship between each sentence stored for the texts under processing as well as a relationship between sentences and books. Hence, it is possible to count, for example, how many sentences a text has or when erasing a text from the corpus these relationships support in this operation.

### **3.3.4 COPA TOKENIZER Word Extraction**

Besides sentence alignment, a second tool named COPA TOKENIZER was developed from scratch for the system. 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.

The development of a tool to extract words that work for more than one language is not an easy task because for each language there is a specific character pattern to identify what is a word and what is not. According to Fernandes (2009), the creation of “wordlists allows the researcher to obtain statistical information about the number of types (different words) and tokens (total number of words) for individual texts in a corpus and for the corpus as a whole” (p. 30). Wordlists, as well as statistical information related to it, are possible just through a list of words created by COPA TOKENIZER. Another interesting point is observed by Fernandes (2009) who explains, “the ratio of types to tokens in a corpus displays the range and diversity of vocabulary used by a writer or translator represented in that corpus”; as observed, the necessity of wordlists is an important part of a corpus.



*Figure 24.* Architecture Diagram of COPA TOKENIZER showing how words are extracted.

Figure 24 is the basic architecture diagram of COPA TOKENIZER. After aligning all sentences, COPA ALIGNER stops its execution and COPA TOKENIZER is triggered. Firstly, the routine checks and requests all new sentences aligned. Secondly, all sentences are returned from the database and, one by one, these sentences are tokenized, statistics are performed and the word is stored in the table “copa\_word”. If the words already exist, the frequency of each specific word is incremented. The final step is the creation of relationships between words and their corresponding book and sentence. Depending on the text size, the execution of COPA TOKENIZER can take from 10 minutes to up to 2 hours.

### 3.4 Chapter Closing Remarks

Following the presentation of the phases related to the creation of COPA-TRAD, which showed the three stages of corpus compilation, the next chapter sets out to describe the usability and applicability of COPA-TRAD tools and how these tools can be used, how a user query for a term can be narrowed down in order to bring accurate results, etc.

In addition, it will be showed how the user can export the data in three different file formats as well as to print it.

The next section explains in detail the tools offered by the COPA-TRAD system for translation research and pedagogy.

## **CHAPTER FOUR: Analysis of COPA-TRAD tools**



#### **4. Initial Remarks**

This chapter discusses the analysis of COPA-TRAD display tools namely COPA-CONC, MONO-CONC, WORDLIST and COPA-STATS. The referred discussion has the main objective to provide hints on how these tools can be used for translation research and pedagogy. This chapter is organized in the following fashion. Firstly, an overview of COPA-TRAD and its system is presented in order to give a better understanding of the system. Following this subsection, the next one deals with the description and analysis of COPA-TRAD tools. The analysis conveyed uses practical examples in order to illustrate how researchers and students can take advantage of COPA-TRAD tools.

##### **4.1 Revisiting COPA-TRAD: A Brief Overview**

The primary characteristic of COPA-TRAD (Corpus Paralelo de Tradução) an online bidirectional Portuguese/English parallel corpus is the online availability. COPA-TRAD provides useful tools to help in the investigation of translated texts. This in turn can be used for translation research and pedagogy, in the sense that these tools provide the decisions made by professional translators when dealing with specific linguistic patterns (such as collocation, semantic prosody, idioms, among others).

The online availability of COPA-TRAD (<http://copa-trad.ufsc.br>) is made one of its main assets (Appendix D). Because of its online nature, COPA-TRAD can be accessed from anywhere around the world no matter what device accessing it, whether a laptop, a desktop computer, tablets or smartphones (Appendix C).

COPA-TRAD and its five subcorpora will provide the data necessary for the analysis driven here. Figure 25 presents the five mentioned subcorpora.



*Figure 25.* The five subcorpora, which are part of COPA-TRAD.

The first one, COPA-LIJ consists of classical texts from Children’s Literature ranging from various subgenres such as “Picture Story Books”, “Fantasy”, “Fairy Tales”, “Modern Fantasy”, “Fables”, etc.

The second subcorpora is COPA-TEL which consists of classical text from general Literature (e.g. “Poetry and Drama”, “Biography”, “Historical Fiction”, etc.) in a public domain (i.e. no copyright). The following subcorpora are COPA-MDT which includes theoretical texts from Translation Studies discipline in order to investigate issues such as metadiscourse in translation. The next is COPA-RAC which comprehends academic abstracts from various different areas.

Finally COPA-MUM, which consists of multimodal material such as text and image (e.g. comics), video (e.g. movies) among others. The multimodal subcorpus will be developed and investigated in a future study (Section 5.2).

The main tools available in COPA-TRAD at the present moment are listed as follows.

- COPA-CONC – Parallel Concordancer.
- MONO-CONC – Monolingual Concordancer.
- WORDLIST – Keyword Frequency.
- COPA BUILDER – Disposable Corpus Creator and Parallel Concordancer.
- COPA STATS – Corpus Statistical Data.

The listed tools were designed and implemented to help translation researchers to conduct their investigations.



The following section sets out in detail the descriptions and analysis of each aforementioned tool. Finally, an explanation on how to use the text submission panel is conducted in order to give a detailed view of COPA-TRAD system.

## 4.2 COPA-TRAD External Tools

For “external” tools, it is meant an opposition to “internal” tools (Chapter 3), as only the former are directly available for the final user. In the next section, the details and the practical use of such external tools are presented.

### 4.2.1 COPA-CONC – Parallel Concordance

COPA-CONC is a bilingual parallel concordancer. It is understood the fact that a corpus consists of original texts in English and its respective translated texts in Portuguese as well as original texts in Portuguese and its respective translations in English.

COPA-CONC is capable of showing the most relevant<sup>7</sup> results based on the user’s query, which can be a word or a sentence (even part of it). The use of wildcards (i.e. special characters – operators and modifiers – to narrow down a specific query) is also available. As observed in Chapter 3, COPA-TRAD uses as a search engine, the Sphinx Search Server which provides a set of solutions ready to be used such as the indexer and search tool. The Sphinx’s search tool is a very powerful engine that can search from keywords ranging from the simpler to the most complex ones with grouped wildcards. Sphinx provides a well-documented<sup>8</sup> manual related to wildcards for a deeper study. The tested and assessed wildcards to be used in COPA-TRAD are discussed below. In addition, it is necessary to emphasize that the wildcards are used in combination with the investigated keyword; that is the use of wildcards without a keyword is not possible. The keyword to be queried, according to Figure 26, has to be specified in the textual form field (highlighted in red). Simple and complex query expressions

---

<sup>7</sup> The relevance is calculated by Sphinx using the ranker *SPH\_RANK\_PROXIMITY\_BM25*, for more information visit: <http://sphinxsearch.com/docs/2.0.4/weighting.html>

<sup>8</sup> The whole documentation in relation to wildcards is available on Sphinx on-line manual at these following links: <http://sphinxsearch.com/docs/2.0.5/boolean-syntax.html> and <http://sphinxsearch.com/docs/2.0.5/extended-syntax.html>

can be requested through the same input, making the use of COPA-CONC easier for advanced and novice users.

Figure 26. COPA-CONC search input.

**Boolean Operators** – A set of special wildcards aimed to do logical operations with the queried keywords. The three Boolean operators are covered below.

- Boolean Operator AND – The schematic structure is listed as follows: “*term X*” & “*term Y*”. It is possible to observe that the wildcard is the “&”. A query with the AND operator brings sentences which contains the “keyword X” and the “keyword Y” no matter which is the position of both keywords. To illustrate, when using the expression “*after & him*” COPA-CONC searches and returns the following results (Figure 27).

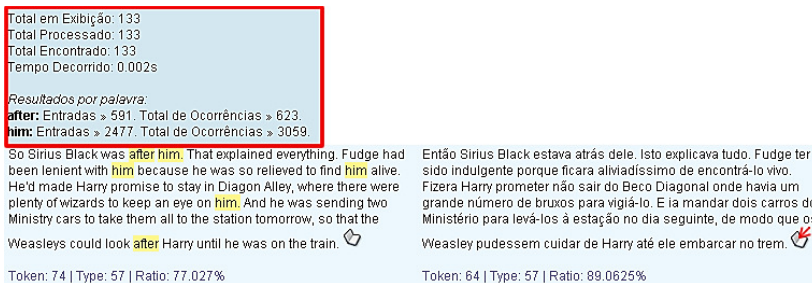


Figure 27. Results from COPA-CONC using the "AND" operator.

- Boolean Operator OR – “*term X*” | “*term Y*” the wildcard for that is the “|”. A query with the OR operator will return sentences which contains the “keyword X” or the “keyword Y” no matter which position both keywords occur in a sentence. For example, when searching using the expression “*after* | *him*” COPA-CONC displays the results exemplified in Figure 28. In brief, it is possible to observe in Figure 28 that using the “OR” operator the results shown contains sentences with the keywords “after” and “him” as well as sentences with just the keyword “after” or just the keyword “him”.

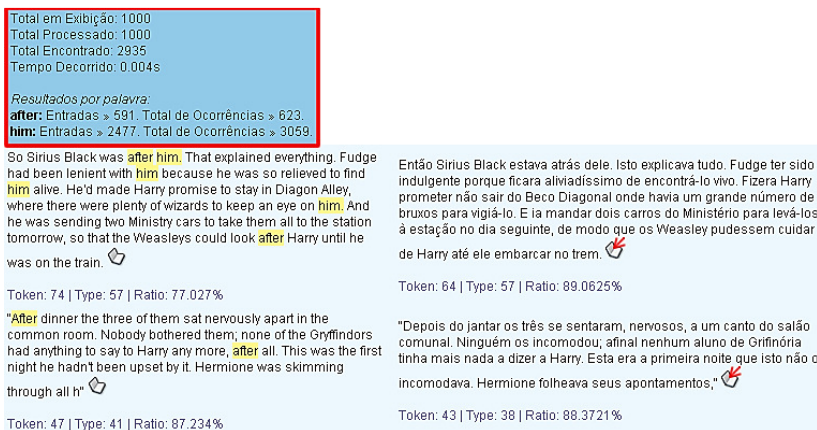


Figure 28. Results from COPA-CONC using the "OR" operator.

- Operator NOT – “*term X -term Y*”. The wildcard for the NOT operator is the “-” (dash). A query with the NOT operator locates and shows sentences with the “keyword X” but not “keyword Y” no matter

which position these keywords occur in a sentence. For example, when querying using the expression “*after -him*” the provided results is illustrated in Figure 29. It is necessary to emphasize that when using the NOT operator the results shown contains sentences just with the term “after” all sentences that contains “after and/or him” are ignored.

Total em Exibição: 458 Total Processado: 458 Total Encontrado: 458 Tempo Decorrido: 0.002s	
<b>Resultados por palavra:</b> <b>after:</b> Entradas > 591. Total de Ocorrências > 623. <b>him:</b> Entradas > 2477. Total de Ocorrências > 3059.	
"My mother died just <b>after</b> I was born, sir. They told me at the orphanage she lived just long enough to name me. Tom <b>after</b> my father, Marvolo <b>after</b> my grandfather."	Minha mãe morreu logo depois que eu nasci. Me disseram no orfanato que ela só viveu o tempo suficiente para me dar um nome... Tom, em homenagem ao meu pai, Servolo, ao meu avô.
Token: 31   Type: 27   Ratio: 87.0968%	Token: 34   Type: 31   Ratio: 91.1765%
"I don't want this <b>afternoon</b> to go," he said; "but I shall come back tomorrow, and the day <b>after</b> , and the day <b>after</b> , and the day <b>after</b> ."	- Queria que esta tarde não terminasse nunca - disse. - Mas vou voltar amanhã, e depois, e depois, e depois.
Token: 27   Type: 20   Ratio: 74.0741%	Token: 21   Type: 17   Ratio: 80.9524%

Figure 29. Results from COPA-CONC using the "NOT" operator.

- Composite Boolean Operation – Composite queries by grouping all the three mentioned Boolean operators is also possible. To conduct this kind of search, it is necessary to use the grouping wildcard which is the “()” (parenthesis). For example, when searching using the expression “(*harry & jumped*) | (*harry & ran*)” the results are shown as in Figure 30. In other words, when using query composition/grouping the returned results contains sentences with the keywords “harry and jumped” or “harry and ran”.

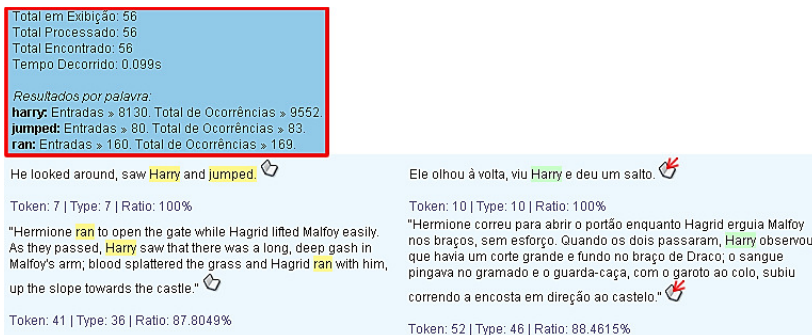


Figure 30. Results from COPA-CONC using composition/grouping expression.

**Exact phrase operator:** The quotation marks (“ ”) are a special wildcard aimed to match an exact keyword or phrases. The exact wildcard requests from Sphinx to find and deliver results with all the keywords in the same order. The wildcard mentioned is double quotes with the keyword in middle of it. For example, consider the following structure “*keyword X Y Z*” (in this case including the double quotes). Consider the expression “*by the book*”, for this query the results are shown in Figure 31. Observing Figure 31, it is possible to note that the result contains sentence only and with the whole expression queried.

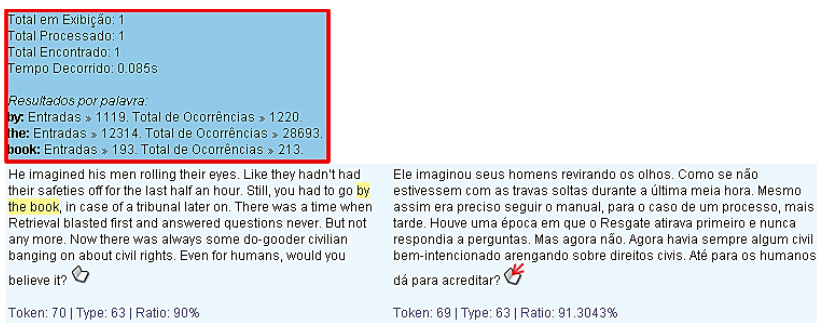


Figure 31. Results from COPA-CONC using the exact operator.

**Field start operator:** This wildcard is aimed to match a term when it is present at the start of a sentence. The wildcard for that is the “^” (caret). When using the caret wildcard in front of a specific

keyword, just sentences starting with this specific keyword are searched and provided. For example, when searching for the collocational pattern “**^take your time**” the results are like in Figure 32. To make it clear, it is possible to observe that just the sentence starting with the queried collocational pattern is listed.

Total em Exibição: 1  
 Total Processado: 1  
 Total Encontrado: 1  
 Tempo Decorrido: 0.235s

*Resultados por palavra:*  
**take:** Entradas > 425. Total de Ocorrências > 435.  
**your:** Entradas > 1078. Total de Ocorrências > 1223.  
**time:** Entradas > 1089. Total de Ocorrências > 1172.

<p><b>Take your time.</b> We have eight hours... excuse me, seven and a half hours, then <b>time's</b> up for everybody! </p>	<p>Demore o quanto quiser. Nós temos oito horas... desculpe, sete e meia, depois acaba o tempo para todo mundo. </p>
Token: 19   Type: 19   Ratio: 100%	Token: 19   Type: 18   Ratio: 94.7368%

Figure 32. Results from COPA-CONC using the field start operator.

**Field end operator:** Likewise the *field start* operator and the *field end* operator works in the same mode but with a clear difference which is the fact that the wildcard is used to match a keyword at the end of sentence. The wildcard for that is the “**\$**” dollar sign. Using the dollar sign at the end of a keyword, just sentences finishing with the matched keyword will be searched and delivered. To illustrate, when searching for “**leave off\$**”, the results are like in Figure 33 snippet. Note that only sentences ending with the queried keyword are listed.

Total em Exibição: 9  
 Total Processado: 9  
 Total Encontrado: 9  
 Tempo Decorrido: 0.002s

*Resultados por palavra:*  
**leave:** Entradas > 156. Total de Ocorrências > 157.  
**off:** Entradas > 887. Total de Ocorrências > 937.

<p>No, Michael," said Chrestomanci, "is the right answer. It's quite clear elementary magic isn't going to mean much to Cat. I'll have to teach you myself, Cat, and we'll be starting on Advanced Theory, I think, by the look of it. You seem to start where most people <b>leave off.</b>" </p>	<p>" A resposta certa é "não", Michael – Chrestomanci interveio. – Está bem claro que Magia Elementar não vai adiantar grande coisa para Gato. Eu mesmo terei que lhe ensinar, Gato, e começaremos com Teoria Avançada, eu acho. Pelo que estou vendo, parece que você começa onde a maioria das pessoas termina." </p>
Token: 50   Type: 46   Ratio: 92%	Token: 52   Type: 48   Ratio: 92.3077%

Figure 33. Results from COPA-CONC using the field end operator.

**Find non-translated words across languages:** COPA-CONC helps the researcher to find words that were not translated. COPA-CONC does that by highlighting these words in left column in yellow

and in the right column the word is highlighted in green (Figure 34). For example, personal names, object names, places, among others, depending on the translator’s decision sometimes are not translated. For instance, when investigating the name of the most modern broomstick in Harry Potter books which is “nimbus” it is to recognize that this name was not translated.

The screenshot shows the COPA-CONC interface. On the left, a blue box contains statistics: 'Total em Exibição: 36', 'Total Processado: 36', 'Total Encontrado: 36', and 'Tempo Decorrido: 0.001s'. Below this, it says 'Resultados por palavra:' and 'nimbus: Entradas > 72, Total de Ocorrências > 76'. The main area shows an English sentence: "It's not any old broomstick," he said, "it's a Nimbus Two Thousand. What did you say you've got at home, Malfoy, a Comet Two Sixty?" Ron grinned at Harry. "Comets look flashy, but they're not in the same league as the Nimbus." The word "Nimbus" is highlighted in green. On the right, the Portuguese translation is: "Não é uma vassoura velha qualquer, é uma Nimbus 2000. Que foi que você disse que tem em casa, Draco, uma Comet 260? – Rony riu para Harry. – A Comet enche os olhos, mas não tem a mesma classe da Nimbus." The word "Nimbus" is also highlighted in green. At the bottom, statistics for both sides are shown: English (Token: 42 | Type: 37 | Ratio: 88.0952%) and Portuguese (Token: 42 | Type: 35 | Ratio: 83.3333%).

Figure 34. COPA-CONC highlight in both sides when the term match across English – Portuguese languages.

Latin words or French words, among other languages, sometimes are not translated as well. For example, in Harry Potter texts Latin expressions are commonly used for words of wizardry, investigating, for example, the term “**aparecium**” when translated to Portuguese the translator decided to maintain the same Latin word.

The screenshot shows the COPA-CONC interface. On the left, a blue box contains statistics: 'Total em Exibição: 1', 'Total Processado: 1', 'Total Encontrado: 1', and 'Tempo Decorrido: 0.027s'. Below this, it says 'Resultados por palavra:' and 'aparecium: Entradas > 2, Total de Ocorrências > 2'. The main area shows an English sentence: "She tapped the diary three times and said, 'Aparecium!'" The word "Aparecium!" is highlighted in green. On the right, the Portuguese translation is: "A garota deu três toques no diário e disse: 'Aparecium!'" The word "Aparecium!" is also highlighted in green. At the bottom, statistics for both sides are shown: English (Token: 9 | Type: 9 | Ratio: 100%) and Portuguese (Token: 10 | Type: 10 | Ratio: 100%).

Figure 35. COPA-CONC highlight in both sides when the term match across English – Portuguese languages.

Having covered the query expressions for COPA-CONC’s filter, attention now will be given to the table that displays the results for the user. The results (Figure 36) provided by COPA-CONC are shown in an interactive table. Here, the term “interactive” is being used because the table has two main characteristics; the first characteristic is in relation to

the highlighted feature of the entire table row. For instance, when the cursor is on a specific row, this row is highlighted giving the user a better visual experience while reading a specific sentence. All sentences displayed in the table are complete, there is no omission. The sentences are not displayed alone, below each of them are basic statistical information such as type, token and ratio.

Concordância do COPA-TRAD	
Lingua 1: Inglês	Lingua 2: Portugues
Total em Exibição: 36 Total Processado: 36 Total Encontrado: 36 Tempo Decorrido: 0.019s	
Resultados por palavra: <b>nimbus:</b> Entradas > 72, Total de Ocorrências > 76.	
It's not any old broomstick', he said, 'It's a Nimbus Two Thousand. What did you say you've got at home, Malfoy, a Comet Two Soxy?' Ron grinned at Harry. 'Comets look flashy, but they're not in the same league as the Nimbus.'	Não é uma vassoura velha qualquer, é uma Nimbus 2000. Que foi que você disse que tem em casa, Dragão, uma Comet 260? – Rony riu para Harry. – A Comet enche os olhos, mas não tem a mesma classe da Nimbus.
Traduzir 	Traduzir 
Token: 42   Type: 37   Ratio: 88.0952%	Token: 42   Type: 35   Ratio: 83.3333%
"Madam Pomfrey insisted on keeping Harry in the hospital wing for the rest of the weekend. He didn't argue or complain, but he wouldn't let her throw away the shattered remnants of his Nimbus Two Thousand. He knew he was being stupid, knew that the Nimbus was beyond repair, but Harry couldn't help it, he felt as though he'd lost one of his best friends."	"Madame Pomfrey insistiu em manter Harry na ala hospitalar pelo resto do fim de semana. Ele não discutiu nem se queixou, mas não deixou jogarem no lixo os estilhaços de sua Nimbus 2000. Sabia que era uma atitude burra, sabia que a vassoura não tinha consento, mas o sentimento era mais forte que ele, era como se tivesse perdido um dos seus melhores amigos."
Traduzir 	Traduzir 
Token: 65   Type: 50   Ratio: 76.9231%	Token: 64   Type: 55   Ratio: 85.9375%

Figure 36. Main table of COPA-CONC.



The second characteristic is in relation to the removal of each table row. Along the table there are red buttons at the beginning of each row and when the user clicks on a specific red button, the entire row is removed from the table. As a result, the user can keep just the results that interest him/her. In addition when using some auxiliary tools, part of COPA-CONC, such as data exportation in CSV, XML and PDF (Figure 37) just the results in the present table are available (i.e. the removed rows do not appear).



Figure 37. COPA-CONC auxiliary tools.

As outlined here, COPA-CONC results table is not used just to display information on the screen. Another aspect is in relation to table's header which has a particular aspect; there are statistical information related to the keyword being researched. Figure 38 is an example extracted from COPA-CONC and it shows the results for the requested term "as a matter of fact". The first three statistical results are related to the total amount for the exact match "as a matter of fact" how many results are being presented on screen, how many were processed and found. The last statistical information is the quantity of entries and tokens for each word that are part of the keyword/expression.

## Concordância do COPA-TRAD

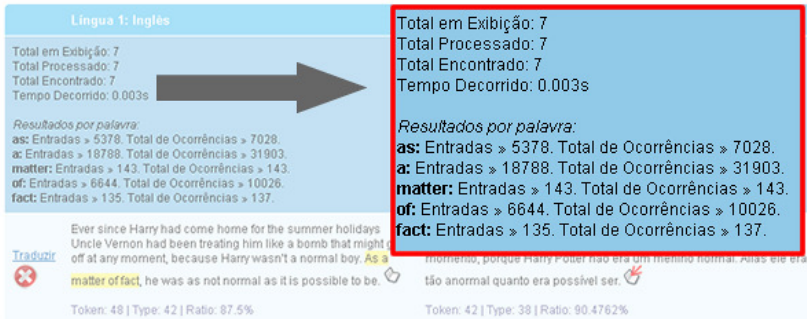


Figure 38. COPA-CONC statistical information.

Finally, COPA-CONC has a basic but powerful filter (Figure 39) to retrieve and narrow down the data presented. The filter can be used by novice users without any kind of experience in a translational corpus and also by advanced users who can explore the whole functionality of the filter using wildcards, as discussed in the aforementioned paragraphs.

According to Figure 39, the first form text field is on the top and it is responsible to receive the keywords/expressions from the user. Then the second field is a list of options giving the possibility for the user to choose one specific subcorpus or the whole corpus (which is composed of all texts in the database). The numerical information at the side of each option is the current quantity of texts indexed and available to be investigated in the corpus or each subcorpus. For example, Figure 39 shows that there are 20 texts available. The last two inputs are related to the first language (i.e. the language on the left column of the table) and the second language (i.e. the language on the right column of the table).

:: Busca Simples ::  
 Termo:  
**"as a matter of fact"**  
 Subcorpus:  
 [Todo corpus (20 texto(s) indexad]

Língua 1:  
 [Inglês (10 texto(s) indexados)]

Língua 2:  
 [Português (10 texto(s) indexados)]

Realizar Busca no COPA-CONC »

---

Concordância do COPA-TRAD  
 Língua 1: Inglês  
 Língua 2: Português  
 Total em Exibição: 7  
 Total Processado: 7

Termo:  
**"as a matter of fact"**  
 Subcorpus:  
 [Todo corpus (20 texto(s) indexad]

Língua 1:  
 [Inglês (10 texto(s) indexados)]

Língua 2:  
 [Português (10 texto(s) indexados)]

Realizar Busca no COPA-CONC »

Figure 39. COPA-CONC filter.

It is necessary to emphasize that the structural organization of the presented sentences in COPA-CONC are ordered by language A on the left and language B on the right “ignoring” the fact that it is a source or translated text. However, source and target texts are visibly signaled because such information is extremely relevant. In view of this problem, a special feature was designed in order for the user to determine whether a text is translated or not. This feature is a visual aid as shown in Figure 40. For source texts at the end of each sentence there is an opened book and for the target texts there is an opened book with an arrow signaling that it is translated. In addition when the mouse is over each image a small hint appears informing whether that sentence comes from a source or a target text.

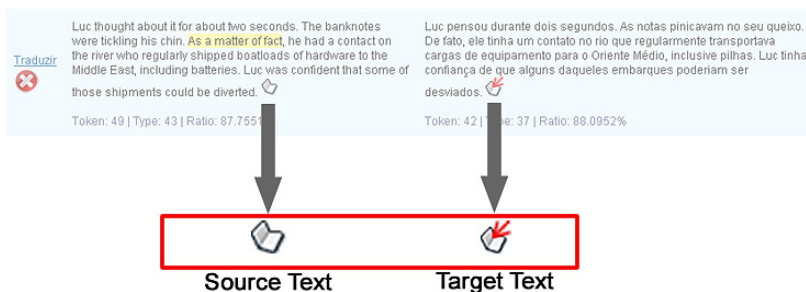


Figure 40. COPA-CONC Source Text and Target Text visual aid.

The images used as a visual aid to indicate if a text is a translation or not has a second functionality when the user clicks on them. After the click, a window is opened bringing extra linguistic information about that specific entry (Figure 41). This extralinguistic information involves, for example, the information raised in Section 3.1.7 as well as the statistical information related to the book of which this sentence is part. The difference when the user clicks on the source text image or the target text image is in relation to which kind of data in the new window is presented. The source text image opens a window with the information about the author while the target opens a window with the information about the translator. The same characteristic is used for the publishing house of the source text and the target text.

**Detalhes sobre a entrada**

**Título:** Harry Potter and the Chamber of Secrets  
**Ano:** 1998  
**Subcorpus:** COPA-LU  
**Língua:** Inglês  
**Variante da Língua:** English - Great Britain  
**Gênero:** Fantasy

**Autor(es):**

Nome	Foto	Sexo	Nacionalidade
J. K. Rowling		female	Britânica

**Biografia:**

Joanne "Jo" Rowling pen name J. K. Rowling, is a British novelist, best known as the author of the Harry Potter fantasy series. The Potter books have gained worldwide attention, won multiple awards, and sold more than 400 million copies. They have become the best-selling book series in history, and been the basis for a series of films which has become the highest-grossing film series in history. Rowling had overall approval on the scripts as well as maintaining creative control by serving as a producer on the final instalment.

Born in Yate, Gloucestershire, Rowling was working as a researcher and bilingual secretary for Amnesty International when she conceived the idea for the Harry Potter series on a delayed train from Manchester to London in 1990. The seven year period that followed entailed the death of her mother, divorce from her first husband and poverty until Rowling finished the first novel in the series, Harry Potter and the Philosopher's Stone (1997).

FONTE: [http://en.wikipedia.org/wiki/J.\\_K.\\_Rowling](http://en.wikipedia.org/wiki/J._K._Rowling)

**Editora(s):**

Nome	Local

Figure 41. COPA-CONC auxiliary window to show the extra linguistic information related to a source/target sentence.

In short, all the features in COPA-CONC were presented; the next section will deal with the monolingual concordance.

#### 4.2.2 MONO-CONC – Monolingual Concordance

MONO-CONC is a monolingual concordancer and it is capable of showing the most relevant results (based on Sphinx ranker) for a user query composed by a keyword or expression. The display format used to present and structure the information in MONO-CONC is the KWIC (Key Word in Context), which is the most suitable display format for monolingual concordances.

Olohan (2004) description of the KWIC display format fits well for MONO-CONC; she says that the keyword requested by the user is shown highlighted and centered with its co-text on each side of the keyword to give an overview of the context. The co-text is “most conveniently viewed if it is limited to one line” (p. 63). Figure 42

illustrates how the sentences from the corpus are organized and limited to one line in the table.

Lingua Inglês		
He had a look of trollich cunning on his face as he	replied	, 'Plenty of room for all of us, Wood.'
made up his mind to make friends with thee,"	replied	Ben. "Dang me if he hasn't took a fancy to thee."
"No, tha' hasn't,"	replied	Martha.
"Yes, I think so," he	replied	-
"No," he	replied	after waiting a moment or so. "I am Colin."
"I don't when I am by myself,"	replied	the Rajah;"but my cousin is going out with me."
"Very good, sir,"	replied	Mr. Roach, much relieved to hear that the oaks might
"Nothing disagrees with me now,"	replied	Colin, and then seeing the nurse looking at him
was something between a sneeze and a cough," she	replied	with reproachful dignity, "and it got into my
"Dickon can sing it for thee, I'll warrant,"	replied	Ben Weatherstaff.

Figure 42. MONO-CONC KWIC display format.

Likewise COPA-CONC parallel concordancer, MONO-CONC uses the same search engine which is Sphinx and because of that all the technology behind the searching mechanism is also available here. As a result, there is no reason to discuss all the available wildcards again, but it is important to point out the options available in COPA-CONC filter.

Figure 43 shows the form filter for COPA-CONC. All the options available are numbered and are covered by this paragraph. Number 1 is the text field in which a keyword or expression can be informed. Next, number 2 is an option list to choose the preferred language. Then, in number 3 the possibility to choose the display format to show the concordances is available, but at this part there is just one display format which is the KWIC. Finally, number 4 is an option list to choose the whole corpus or any of its subcorpus.

**:: Concordância ::**

**1**

Termo:

**2** Língua:

**3** Formato:

**4** Subcorpus:

Figure 43. MONO-CONC filters options.

As mentioned before, MONO-CONC uses the same search engine used for COPA-CONC. Figure 44 is an example extracted from MONO-CONC for the keyword *“have a”*. Currently the production version of MONO-CONC does not have statistical information or any other auxiliary tool. This subsection discussed all the features available in MONO-CONC. The next subsection covers the WORDLIST tool.

Lingua Ingles	
on the kitchen table. "If a person has a gift, they	<b>have a</b> right to have it developed - and so I told him! But
"I may be obliged to	<b>have a</b> tantrum," said Colin regretfully. "I don't want to
again - and I may get worse this very night. I might	<b>have a</b> raging fever. I feel as if I might be beginning to
elf, but my cat's climbed a stalactite. Or, if you	<b>have a</b> minute, Captain, could you tell me how to get to the
Mary. "I may have it where I like! I am not going to	<b>have a</b> governess for a long time! Your mother is coming to
the way Gwendolen would have done it. "We'd better	<b>have a</b> hunt round," she said, "in case dear Gwendolen has
Well, I didn't think much of it, but I thought I'd	<b>have a</b> go. And I did what he said, as far as I could
Sig Sauer. His own skills aside, it would be nice to	<b>have a</b> weapon. Something with a bit of weight to it. His
I still have to finish mesmerizing these goons. We	<b>have a</b> fifty-five-minute window here. Let's not waste it
pocket and unravelling it for Mr Borgin to read. I	<b>have a</b> few - ah - items at home that might embarrass me, if
curling his mouth as though he doubted it, "but we do	<b>have a</b> set of suspicious circumstances here. Why were they
Eane thundered. "What are you doing? You	<b>have a</b> human on your back! Have you no shame? Are you a
But, it'll have to wait until next term, I'm afraid. I	<b>have a</b> lot to do before the holidays. I chose a very

Figure 44. MONO-CONC results for "have a".

### 4.2.3 WORDLIST – Frequency List

WORDLIST was developed to provide auxiliary tools to manipulate and retrieve statistical information related to the lexical composition of texts. However, the current version of WORDLIST tool has two features that are covered here. Differently from the last two tools already discussed, the WORDLIST home screen has the form filter as well as a word cloud (Figure 45), which was designed to show the most frequent words from the corpus. The words to be part of the word cloud should have a frequency equal to or greater than 600 occurrences inside the corpus. In addition, the frequency of each keyword implies in a distinguished font size and color tone. The number of 600 occurrences was defined in the system solely because the visual aspect because less than 600 used to show a big amount of words.

#### Keywords: Inglês

As palavras aqui listadas apareceram pelo menos 600 vezes. As palavras mais comuns foram excluídas desta lista.

dumbledore snape artemis door butler malfoy hermione black  
 spiro **harry** cat mary weasley gwendolen looked potter eyes  
 time janet colin head lupin professor holly ron hagrid fowl

Figure 45. Word cloud for English words from COPA-TRAD.

A hidden feature in word cloud can be noted when the user clicks on any of the words listed. As a result, a new window is opened bringing the frequency of the selected word as well as an example extracted from the corpus containing the chosen word (Figure 46).



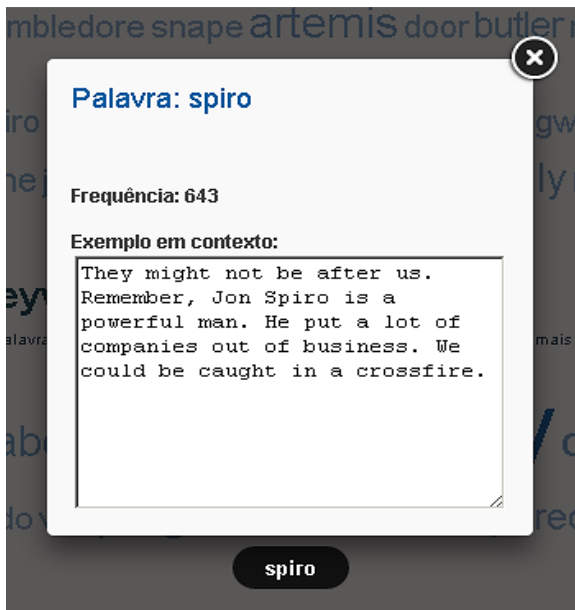


Figure 46. Details for the word "spiro".

WORDLIST has a small filter form to facilitate the creation of the word frequency list. The first field is a list containing the letters from the alphabet. The second field is a list to select the language of the frequency list (Figure 47).

 A screenshot of a filter form titled "Busca Simples". The form has a white background and a light grey border. It contains the following elements:
 

- The title "Busca Simples" in blue.
- A label "Letra:" followed by a dropdown menu showing the letter "A".
- A label "Língua:" followed by a dropdown menu showing "Escolher...".
- A button labeled "Buscar Entradas »" at the bottom.

Figure 47. WORDLIST filter form.

Figure 48 displays the word frequency list with words starting with the letter “A”.

Palavra	Frequência
a	30193
and	14247
as	6706
at	3770
artemis	2614
all	2267
about	1500
an	1409
are	1163
again	814
around	721
any	700
after	619
away	607
ah	590

Figure 48. Word frequency list with words starting with "A".

The main features of WORDLIST tool was covered in this section. The next one is going to discuss a tool named CORPUS-BUILDER.

#### 4.2.4 CORPUS-BUILDER – Creating a Disposable Corpus on the Fly

COPUS-BUILDER is tool capable of creating a disposable corpus (i.e. generally a small corpus used for a specific research which the user chooses the texts to be processed as well as the filters to be used to narrow down the search). This kind of corpus is disposable because it is created in real time and is not part of COPA-TRAD or any of its subcorpora. In addition, after the investigation the information is lost.

However the information provided by the user is stored in the database for further research. The first step to start using CORPUS-BUILDER, according to Figure 49, is the sentence alignment (in the example the alignment is in paragraph level) after that the source language of the text as well as the target language has to be chosen. For the sake of this example, a text from a bilingual newspaper<sup>9</sup> (English/Spanish) was extracted and aligned at paragraph level in COPA-BUILDER. With both texts aligned, the second step is to select the language for the source text as well as the language for the target text as Figure 49 illustrates.

---

<sup>9</sup> The journalistic text in English was extracted from <http://tudecidesmedia.com/the-second-generation-a-portrait-of-the-adult-children-of-immigrants-p4442-128.htm> and the journalistic text in Spanish was extracted from <http://tudecidesmedia.com/la-segunda-generacin-un-retrato-de-los-hijos-adultos-de-inmigrantes-p4441-128.htm>

## CORPUS-BUILDER

**\* Preenchimento obrigatório.**

**\* Texto Fonte**

1 The Second Generation: A Portrait of t)  
 2 (Pew Hispanic Center) – Second-generat:  
 3 Hispanics and Asian Americans make up  
 4 The Pew Research surveys also find that  
 5 As the U.S. Congress gears up to consi  
 6 Given current immigration trends and b:  
 7 By then, the nation's "immigrant stock"  
 8 The focus of this report is on the 20  
 9 This is a heterogeneous group that inc.  
 10  
 11  
 12  
 13  
 14  
 15

**\* Língua Fonte:**  
 English - United States

**\* Texto Alvo**

1 La Segunda Generación: Un retrato de l  
 2 (Centro Hispánico Pew) – Los americano:  
 3 Los hispanos y los asiáticos-americanos:  
 4 Las encuestas del Centro Pew también h  
 5 A manera que el Congreso de los Estado:  
 6 Dadas las tendencias actuales de inmigr  
 7 Para entonces, el grupo inmigrante (un  
 8 El foco principal de este reporte son  
 9 Éste es un grupo heterogéneo que inclu  
 10  
 11  
 12  
 13  
 14  
 15

**\* Língua Alvo:**  
 Spanish - Mexico

Figure 49. CORPUS-BUILDER text alignment.

The next step is the use of filters. CORPUS-BUILDER has various kinds of filters to narrow down the search, as well as display formats. For the example here consider Figure 50, each filter in use is numbered. Number 1, the term under investigation is informed, in this case “Hispanics”. Next, in number 2 was selected to match the whole word. Then in number 3 the keyword for the parallel display format is informed. Finally, number 4 the parallel display format was chosen.

Tipo de Entrada:  2 Textos Alinhados  Textos com Tabulação

\* Termo: **Hispanics** **1**

Procurar em Texto Fonte:  Texto Alvo:

**2**  
Tipo de procura Por palavra:  Palavra começando com  Palavra terminando com  Palavras incluindo

Contexto:

Consultar no Alinhamento:  **3**

Ordenar:

Saída:  Paralelo  KWIC (Key Word in Context) **4**

Figure 50. COPA-BUILDER filters in use.

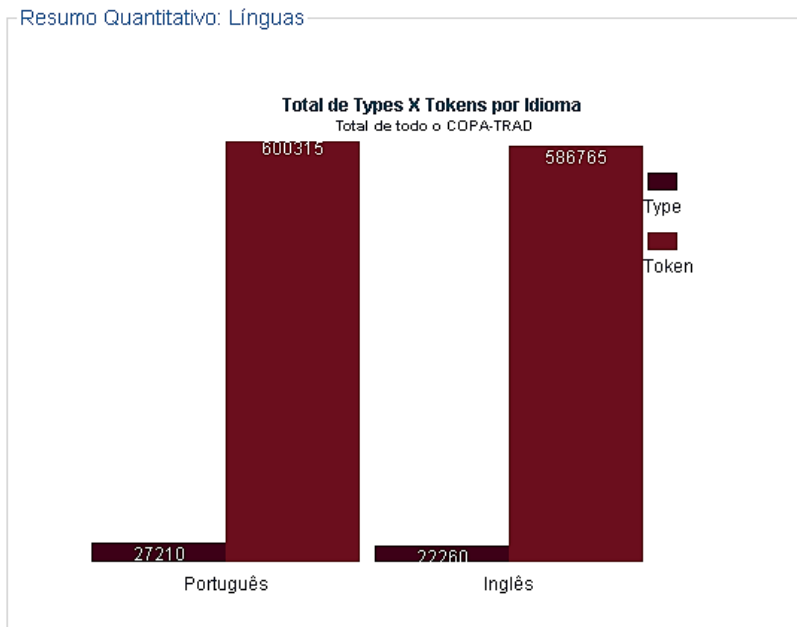
When the user clicks on the button “Buscar Termos”, the aligned texts are processed and in real time the table with sentences are displayed, as Figure 51 illustrates. The word highlighted in “red” is the keyword “Hispanic” specified by the user. It is necessary to emphasize that due to the nature of CORPUS-BUILDER, Sphinx Search engine is not used for this tool.

Texto Fonte	Texto Alvo
<p>The Second Generation: A Portrait of the Adult Children of Immigrants</p> <p>(Pew Hispanic Center) — Second-generation Americans—the 20 million adult U.S.-born children of immigrants—are substantially better off than immigrants themselves on key measures of socioeconomic attainment, according to a new Pew Research Center analysis of U.S. Census Bureau data. They have higher incomes; more are college graduates and homeowners; and fewer live in poverty. In all of these measures, their characteristics resemble those of the full U.S. adult population.</p>	<p>La Segunda Generación: Un retrato de los hijos adultos de inmigrantes</p> <p>(Centro Hispánico Pew) — Los americanos de segunda generación—esos 20 millones de adultos nacidos de inmigrantes en Estados Unidos—están viviendo substancialmente mejor que sus predecesores en cuanto a medidas clave de logros socioeconómicos, según indica un análisis de la información del Buró de Censos de los Estados Unidos realizado por el Centro de Investigación Pew. Tienen mejores ingresos, más graduados universitarios y un mayor número de ellos son dueños de casas, además de que menos de ellos viven en la pobreza. En todas estas medidas, sus características se asemejan a las de la población estadounidense adulta completa.</p>
<p><b>Hispanics</b> and Asian Americans make up about seven-in-ten of today's adult immigrants and about half of today's adult second generation. Pew Research surveys find that the second generations of both groups are much more likely than the immigrants to speak English; to have friends and spouses outside their ethnic or racial group, to say their group gets along well with others, and to think of themselves as a "typical American."</p>	<p>Los hispanos y los asiáticos-americanos conforman casi siete de cada diez de los inmigrantes adultos del país, así como alrededor de la mitad de los actuales adultos de segunda generación. Las encuestas del Centro de Investigación Pew han encontrado que los miembros de la segunda generación de ambos grupos son mucho más propensos a hablar inglés, a tener amigos y cónyuges fuera de su grupo étnico o racial, a decir que su grupo se lleva bien con otros grupos y a referirse a sí mismos como un "típico americano".</p>

Figure 51. COPA-BUILDER results.

#### 4.2.5 COPA-STATS – The COPA-TRAD Statistics Tool

The main objective of COPA-STATS tool is to bring quantitative information about the corpus, subcorpora and the texts. For future versions of COPA-TRAD, information such as the most used terms requested by the users will be part of this tool. The home screen of COPA-STATS displays the number of tokens and types by language as well as a graphical representation of it (Figure 52).



*Figure 52.* COPA-STATS language statistics for type and token in Portuguese and English.

COPA-STATS also include an option to select specific texts (source and target) and bring its statistical information represented in a graph as Figure 53 presents. The filter to select the books is highlighted in red.

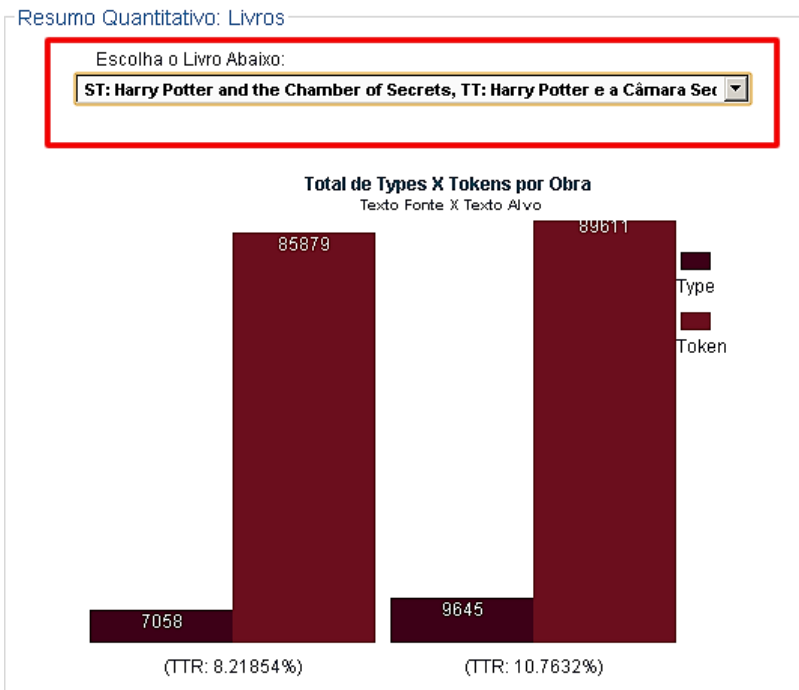


Figure 53. Statistical information related to types and tokens for the book "Harry Potter and the Chamber of Secrets" and its Brazilian version "Harry Potter e a Câmara Secreta".

#### 4.2.6 Text Submission

COPA-TRAD offers a simple and easy way for all users to submit their texts to be part of the corpus. The system has a special tool that users can use to submit how many texts they want (according to the subcorpora and text types defined for COPA-TRAD) and the use of this panel does not require any technical knowledge to use it. Chapter 3 covered in two section parts of the text submission panel. Section 3.1.4 dealt with the aspects of selection of texts and pointed out the three main sources of textual data in COPA-TRAD, one of them is the text submitted by the user. Finally, Section 3.3.2, shows how texts are processed in the internal part of COPA-TRAD. After covering the aforementioned issue now the focus of this subsection is on the practical use of text submission panel.



COPA-TRAD text submission panel consists of two main screens and various windows that can be opened or not, depending on information that is already available or not. When accessing the submission panel the first screen displayed, has two buttons left and in the middle it has a table listing all text that the user has already sent. Figure 54 presents this first screen.



Envio	Títulos	Staus	Mensagem	Editar
10/01/2013	Charmed... & Vida...	Aprovado		
10/01/2013	Artemis Fowl The... & Artemis Fowl O Código...	Aprovado		
10/01/2013	Artemis Fowl: The Arctic... & Artemis Fowl Uma...	Aprovado		
10/01/2013	Artemis... & Artemis Fowl o Menino...	Aprovado		
09/01/2013	Harry Potter and the... & Harry Potter e o...	Aprovado		
09/01/2013	Harry Potter and the... & Harry Potter e a Pedra...	Aprovado		
09/01/2013	Harry Potter and the... & Harry Potter e a Câmara...	Aprovado		
09/01/2013	The Secret... & O Jardim...	Aprovado		

Foram encontrados: 8 registros.

Figure 54. Text submission main screen.

The table does not list the submitted texts only, but it has a status showing if the text was approved to be part of the corpus or not. In edition, it has a message button, when clicked shows the message the moderator sent in relation to the assessment of the text (Figure 55). Finally it has an edit button that has two variants: The first is when the text is not approved yet, which gives the user the chance to edit the information sent. The second is when the text is approved, at this very moment the edit button is blocked and the user cannot edit the information.

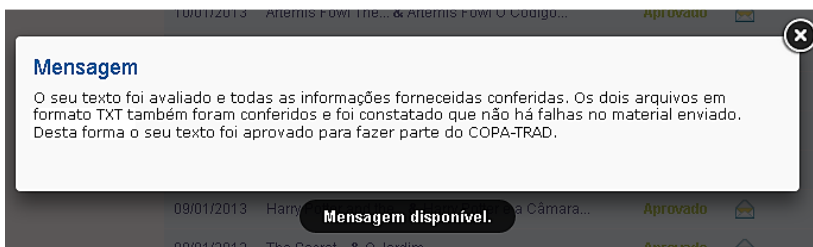


Figure 55. Message sent from the moderator to user.

In order to submit new text, it is necessary to click on the green button named “Enviar Texto” at the left of the screen (Figure 30). A new window is opened with a form to be filled with the extra linguistic and translational information of the text. This form also provides the possibility to send the source and target text in TXT file format. Figure 56 displays the first part of the form which focuses on the source text. All fields are self-explanatory, dismissing the discussion of each one here. However, two fields are necessary to discuss, both of them are highlighted in red in Figure 56. The field labeled number 1 has an option with the list of all authors already submitted, thus the user has to select the name of the author in the available list. However, if the name of the author is not present yet, he has to click on the button on the side of the field named “Incluir Autor” or “Include Author”. When the user clicks on this button a new window is presented with a form for the user to provide information related to the author. After the user clicks on the save button the name of this author is automatically include in the option list of the main form. The same functionality is also present for the name of Publishing House.

Dados do Texto Fonte (ST)

Você deve informar abaixo as informações paratextuais referente ao texto fonte ou obra original. Estas informações são importantes, pois garantem uma maior confiabilidade dos dados além de servirem de insumo para pesquisas científicas. Consulte abaixo, os gêneros e subcorpus disponíveis antes de realizar o cadastro completo! Caso seu gênero ou subcorpus não estiver disponível entre em contato com o desenvolvedor através do e-mail [carloosedasilva@gmail.com](mailto:carloosedasilva@gmail.com) para verificar a viabilidade de incluí-los no COPA-TRAD.

Título:

Ano:

Edição:

Características Especiais:

Copyright:

Prêmios:

Língua:

Variação Linguística:

Gênero:








Autor:  [\[Incluir Autor\]](#) **1**

Editora (original):  [\[Incluir Editora\]](#) **2**

**O arquivo a ser enviado precisa estar em formato TXT com codificação UTF-8 sem BOM. Dúvidas? Favor clicar aqui para consultar a documentação explicativa de como preparar o texto a ser submetido ao corpus.**

Texto Fonte (.TXT até 8Mb):  Nenhum arqui... selecionado

Resumo do texto:

**B I S**       

body p

Figure 56. Form text submission. Source Text section.

The second part of the submission form is related to target texts. The figure and details about this part of the form is not covered here because all the fields are the same with the exception of the author name, translator name and the upload field.

The last part of the form is to gather information related to the translation process (i.e. the editor, commissioner, etc.). Also the user has to choose which subcorpus the text will be part of. As a last procedure the user has to click on the “save” button to save all the information and send a message to all moderators informing that new text was submitted and needs to be assessed.

### **4.3 Final Remarks**

In this Chapter, the tools comprising the COPA-TRAD system were described in order to show the possible applications of these tools for translation research and pedagogy. The possibilities are endless, but only the use of the tools here developed can really reveal them. In the next section, some conclusions are drawn in relation to the present study.

## **CHAPTER FIVE: Concluding Remarks**



This chapter closes the present study by presenting a global view of the issues discussed here and is organized as follows; firstly a recapitulation of each chapter that encompasses this study is covered. Next, the objectives and research questions raised in the introductory chapter are summarized. These research questions are going to be answered in light of the investigation carried out here. Finally, the limitations of this study are presented as well as the suggestions for future research.

## **5. Summarizing the Study**

The first chapter brought a contextualization related to parallel corpus and its application for Translation Studies more specifically Corpus-based Translation Studies. This area from Translation Studies is inserted on the applied branch and took advantage on the shift from prescriptivism to descriptivism. As a result, the use of parallel corpus contributes for the investigation based on descriptive methodologies. Afterwards, three relevant parallel corpora for the linguistic pair Portuguese – English were described in order to survey what have been proposed in the field paving the way for this research by presenting the limitations (i.e. lack of a user-friendly interface and children's literature texts) of those three corpora. The last point presented was in relation to machine translation engines such as Google Translate and Microsoft Translator. Up until now, there is no translational tool available that is capable of comparing human translation with non-human translation and at the same time providing vital information for the improvement of translation systems.

The second chapter introduced the concept and theory that form this study. The discussion about the ideas presented by several researchers is carried out in order to provide a theoretical framework to help the researcher think about COPA-TRAD compilation. In order to establish this study, a comparison between corpus-based translation studies and corpus linguistics shows the main differences between these two research areas, without losing track of the fact that several methods and theoretical information from Corpus Linguistics are used in CTS.

Chapter three presented the method used to answer the research questions this study seeks to answer. This method is based on the three main stages for corpus compilation proposed by Fernandes (2004) and some additions to this proposal were provided in order to cover the practical and theoretical aspects of an online corpus compilation. The first stage is corpus design including the management aspects of the project. Secondly, the corpus building stage specifies the system

implementation phase, the practical and technical work conducted to create COPA-TRAD. The internal functioning of display formats were added to the corpus processing stage in order to show how COPA-TRAD processes data and presents its results to the user.

The fourth chapter showed possible ways of how researchers, translators and students can take advantage of COPA-TRAD. Firstly, it reviewed the purpose of corpus creation and brings an overall COPA-TRAD description in order to familiarize the reader with this specific technology. Finally, it presented the tools available in COPA-TRAD and how to use them to conduct investigations.

## **5.1 The Research Questions**

COPA-TRAD was designed to be a bidirectional parallel corpus for the investigation of translational phenomena. It serves as a tool for translation research and pedagogy by showing solutions taken by professionals when tackling specific translation challenges. The three questions brought by this study are revisited and answered as follows.

### **1. What aspects should be taken into account when compiling a parallel corpus to be made available online?**

Chapter three used the three stages involved in corpus compilation proposed by Fernandes (2004). Among these three stages, the one that covers the aspects to take into consideration when compiling a parallel corpus is in Section 3.1, Corpus Design. At this stage the aspects were listed that had to be considered to create and make available on the Internet the parallel corpus. Additionally, Section 3.1 explained how the information was collected and organized. After that, the aspects to be considered in a corpus design were discussed. These aspects are cited in the next paragraphs.

The first aspect considered was in relation to the purpose of the corpus and the goal, which was the creation of COPA-TRAD to provide tools for the investigation of translational practices. This kind of investigation can be carried out by user researchers, translators and students.

The second aspect considered was in relation to what kind of corpus satisfies the necessities for translational investigation, here the parallel corpus was proposed.

The third aspect highlights the copyright issues, which have to be considered when compiling a corpus because it can reduce the textual material to be stored; at this part a solution for the copyright aspect was presented. Any corpus can be jeopardized, if no criteria and planning is established in the selection of texts. At this part the criteria and



planning are justified as well the three main sources of text selection for COPA-TRAD are described.

Another aspect raised was in relation to shortcomings and possible solutions, when creating a corpus. As a result, a set of issues should be considered in order to have a successful outcome, these issues were provided in this section.

The next aspect discussed was in relation to requirements analysis and how it was conducted. In lieu of starting the development of the system in an unplanned way, for COPA-TRAD a careful analysis of the problems was carried out as well as documented, the development of COPA-TRAD started. During this phase the technological aspects to make COPA-TRAD available on the Internet was discovered.

Finally the aspect related to textual storage in database was discussed. In addition, important questions had to be taken into consideration before the modeling of the database. The answers for such questions have a direct impact on the final structure of the database. Next the organization of the tables in database was discussed to show how the textual material is organized and aligned in a structured level.

As observed, Corpus Design section listed the theoretical and practical aspects involved in the corpus compilation as well as the availability of it on the Internet.

## **2. In what ways can a parallel corpus of children's literature proposed and developed by the present study contribute to the analysis of translational patterns?**

The parallel corpus built during this study has five tools in order to support the investigation of translational patterns as showed in Chapter Four. From these five tools three of them (COPA-CONC, MONO-CONC and CORPUS BUILDER) provide the possibility of a more qualitative research and the last two tools (WORDLIST and COPA STATS) provide a more quantitative research. The mentioned tools use the same corpus in order to operate and extract data. One of the subcorpus that compose the main corpus is COPA-LIJ the Children's Literature parallel corpus. The data provided from COPA-LIJ through the 5 tools already mentioned, contributes to the analysis of translational patterns. For instance, the monolingual concordancer provides the possibility to investigate collocations through the KWIC display format, which is part of MONO-CONC.








## **3. What alternative display formats can be used for this specific kind of analysis and why?**

This question complements the last one in that COPA-TRAD has three specific tools, which has display formats useful to support the investigation of translational patterns.

The first tool is COPA-CONC (Figure 57) a parallel concordancer, which shows original, sentences aligned with its respective translation. The terms requested by the user is highlighted and these terms are part of sentences which gives not just specific words but the whole sentence and in this way provides context for better understanding of translational patterns. Translators and students can, for example, investigate how idioms are translated by professional translators. For such investigation COPA-CONC can display a format that suits this kind of investigation.

The second tool is MONO-CONC (Figure 57) which is a monolingual concordance which presents the results in KWIC (Key Word in Context) mode facilitating the investigation of patterns such as collocations. COPA-CONC and MONO-CONC tools provide autonomy for researchers by allowing them to select what kind of data to be retrieved and in what format to present the results; through a bilingual (COPA-CONC) or a monolingual (MONO-CONC) concordancer. The third tool to indicate here is CORPUS BUILDER, which has the same display format as COPA-CONC.

## COPA-CONC

Língua 1: Inglês	Língua 2: Português
Total em Exibição: 7 Total Processado: 7 Total Encontrado: 7 Tempo Decorrido: 0.003s	
<i>Resultados por palavra:</i> <b>as:</b> Entradas > 5378. Total de Ocorrências > 7028. <b>a:</b> Entradas > 18789. Total de Ocorrências > 31903. <b>matter:</b> Entradas > 143. Total de Ocorrências > 143. <b>of:</b> Entradas > 6644. Total de Ocorrências > 10026. <b>fact:</b> Entradas > 135. Total de Ocorrências > 137.	
Ever since Harry had come home for the summer holidays Uncle Vernon had been treating him like a bomb that might go off at any moment, because Harry wasn't a normal boy. <b>As a</b>  <b>matter of fact</b> , he was as not normal as it is possible to be.  	Desde que Harry voltara para passar as férias de verão em casa, tio Váiter o tratava como uma bomba que fosse explodir a qualquer momento, porque Harry Potter não era um menino normal. Alias ele era tão anormal quanto era possível ser. 
Token: 48   Type: 42   Ratio: 87.5%	Token: 42   Type: 38   Ratio: 90.4762%
"Precisely," said the Headmaster. "My dear boy, you must see how foolish it would be of me to allow you to remain at the castle when term ends. Particularly in the light of the recent tragedy... the death of that poor little girl... You will be safer by far at your orphanage. <b>As a matter of fact</b> , the Ministry of Magic is even now talking about closing the school. We are no nearer locating the – er – source of all this unpleasantness..."  	Precisamente – disse o diretor. – Meu rapaz, você deve entender que seria muito insensato de minha parte permitir que você permaneça no castelo quando terminar o ano letivo. Principalmente à luz da recente tragédia... a morte daquela pobre menininha ... Você estará muito mais seguro no seu orfanato. Aliás, o Ministério da Magia está neste momento falando em fechar a escola. Não estamos nem perto de identificar a... hum... fonte de todos esses contratempos... 
Token: 83   Type: 66   Ratio: 79.5181%	Token: 75   Type: 64   Ratio: 85.3333%

## MONO-CONC

Língua Inglês		
on the kitchen table. "If a person has a gift, they	<b>have a</b>	right to have it developed - and so I told him! But
"I may be obliged to	<b>have a</b>	tantrum," said Colin regretfully. "I don't want to
again - and I may get worse this very night. I might	<b>have a</b>	raging fever. I feel as if I might be beginning to
elf, but my cat's climbed a stalactite.' Or, 'If you	<b>have a</b>	minute, Captain, could you tell me how to get to the
Mary. "I may have it where I like! I am not going to	<b>have a</b>	governess for a long time! Your mother is coming to
the way Gwendolen would have done it. "We'd better	<b>have a</b>	hunt round," she said, "in case dear Gwendolen has
Well, I didn't think much of it, but I thought I'd	<b>have a</b>	go. And I did what he said, as far as I could
Sig Sauer. His own skills aside, it would be nice to	<b>have a</b>	weapon. Something with a bit of weight to it. His boot
I still have to finish mesmerizing these goons. We	<b>have a</b>	fifty-five-minute window here. Let's not waste it

Figure 57. COPA-CONC and MONO-CONC display formats.

## 5.2 Limitation and Suggestion for Future Study

This study covered the theoretical and practical issues related to the creation and availability of a parallel corpus on the Internet. Due to the size of the project and extensive technical procedures taken during this study, some limitations were discovered.

The first limitation was in relation to COPA-MUM (Corpus Paralelo de Multimodalidade), this subcorpus is not present in the

working version of COPA-TRAD because at the present moment, just the prototype version was developed and the development of this subcorpus as well as the tools to handle images, sounds, videos and texts were planned to be conducted during my doctoral research. The second limitation of this study is related to COPA STATS which just two kinds of quantitative data (type and token) can be extracted through the tool. However, the mentioned limitations will be solved in future versions of COPA-TRAD.

A suggestion for future research is to create the 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 alignment of the texts need some refinement, especially as regards the automatization of the whole aligning 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 translational phenomena.

Another suggestion for future research would that of translational activities for students. Such activities would demonstrate how to use and take advantage of a parallel corpus. Due to time constraints and the size of the whole project, this research effort could not be developed. Based upon student feedback, further development requirement can be gathered to provide resourceful information in relation to the use of COPA-TRAD for translator education. To conduct this kind of investigation a book edited by Zanettin, Bernardini and Stewart (2003) entitled *Corpora in Translator Education* can serve as a basis for the research suggested. Therefore, the proposal of the activities will show the practical applicability of COPA-TRAD in translator education, and also assist in helping educators make students aware of translational patterns (such as collocations and semantic prosody) showing them how to deal with these patterns while translating text. This will expose students to translational decisions taken by professional translators.

## BIBLIOGRAPHY

Baker, M. (1992). *In Other Words - A coursebook on translation*. London: Routledge.

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/Philadelphia: John Benjamins Publishing Company.

Baker, M. (1995). *Corpora in Translation Studies. An Overview and Suggestions for Future Research*. *Target*, 7(2), 223-243.

Baker, M. (1998). *Routledge Encyclopedia of Translation Studies*. London and New York: Routledge.

Baker, M. (2001). *Routledge Encyclopedia of Translation Studies*. London and New York: Routledge.

Brants, T., Popat, A. C., Xu, P., Och, F. J., & Dean, J. (2007). *Large language models in machine translation*. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867.

COMPARA (2011). Retrieved November 04, 2011, from <http://linguateca.pt/COMPARA/>

Fernandes, L. (2007). *On the Use of Portuguese-English Parallel Corpus of Children's Fantasy Literature in Translator Education*. *Cadernos de Tradução*, 20, 141 - 163.

Fernandes, L. (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. (2009). A Portal into the unknown: designing, building and Processing a Parallel Corpus. CTIS Occasional Papers, 4, 16 - 36.

Frankenberg-Garcia, A. & Santos, D. (2003). Introducing COMPARA: the Portuguese-English Parallel Corpus. In F. Zanettin, S. Bernardini & D. Stewart (Eds.), *Corpora in Translator Education* (71-87). Manchester: St Jerome.

Gerber, R. M. & Vasilévski, V. (Ed.). (2007). *Um percurso para pesquisas com base em corpus*. Florianópolis: Editora da UFSC.

Hartley, T. (2009). Technology and Translation. In J. Munday (Eds.), *The Routledge Companion to Translation Studies* (106-127). New York: Routledge.

Holmes, James S. (1972/1988). The Name and Nature of Translation Studies. In: James S. Holmes, *Translated! Papers on Literary Translation and Translation Studies*, Amsterdam: Rodopi.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. In: *Zeitschrift für Computerlinguistik und Sprachtechnologie*. 20, 19 - 62.

Imao, Y. (2011). Web Concordancer. Retrieved June 05, 2012, from <https://sites.google.com/site/casualconc/web-concordancer>.

Johansson, S., Leech, G. N., & Goodluck, H. (1978). *Manual of Information to Accompany the Lancaster-Oldo/Bergen Corpus of British English, for Use with Digital Computers*. Oslo: Unpublished document.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London/New York: Lognman.

Kenny, D. (2009). Corpora. In M. Baker & G. Saldanha (Eds.), *Routledge Encyclopedia of Translation Studies* (2nd ed.) (59-63). London and New York: Routledge.

Knowles, M. & Malmkjaer, K. (1996). *Language and Control in Children's Literature*. London and New York: Routledge.

- Kruger, A. (2002). Corpus-based translation research: its development and implications for general, literary and Bible translation. In Naudé, J.A. & van der Merwe, C.H.J. (Eds.), (*ActaTheologicaSupplementum*, 2.) *Contemporary TranslationStudies and Bible Translation* (70-106). Bloemfontein: University of the Free State.
- Laviosa-Braithwaite, S. (1996). *The English Comparable Corpus (ECC): a resource and a methodology for the empirical study of translation*. Unpublished Doctoral Dissertation, Dept of Language Engineering - UMIST, Manchester.
- Laviosa, S. (2002). *Corpus-based translation studies: theory, findings, applications*. Amsterdam and New York: Rodopi.
- Luz, S. (forthcoming) 'Web-based Corpus Software' in A. Kruger, K. Wallmachand J. Munday (eds) *Corpus-Based Translation Studies: Research and Applications*, London: Continuum.
- Malmkjær, K. (1998). Love thy neighbour: will parallel corpora endear linguists to translators?. *Meta: Translators' Journal*, 43, 534 - 541.
- Marcotte, E. (2011). *Responsive Web Design*. New York: Jeffrey Zeldman.
- O'Dell, F., & McCarthy, M. (2008). *English Collocations in Use: Advanced*. Cambridge: Cambridge University Press.
- Olohan, M. (2002). Leave it out! using a comparable corpus to investigate aspects of explicitation in translation. *Cadernos de Tradução*, 19, 153 - 169.
- Olohan, M. (2004). *Introducing Corpora in Translation Studies*. London and New York: Routledge.
- Opus Corpus (2012a). Retrieved November 04, 2012, from <http://opus.lingfil.uu.se/>
- Opus Corpus (2012b). Retrieved November 04, 2012, from <http://opus.lingfil.uu.se/trac>

Oracle Corporation. (n.d.-a). MySQL 5.5 Reference Manual: The MyISAM Storage Engine. Retrieved from <https://dev.mysql.com/doc/refman/5.5/en/myisam-storage-engine.html>

Oracle Corporation. (n.d.-b). About MySQL. Retrieved from <http://www.mysql.com/about/>

Pearson, J. (2003). Using Parallel Texts in the Translator Training Environment. In F. Zanettin, S. Bernardini & D. Stewart (Eds.), *Corpora in Translator Education* (15-24). Manchester: St Jerome.

Projeto COMET (2009). Retrieved November 04, 2011, from <http://www.fflch.usp.br/dlm/comet/>

Pressman, R. S. (2001). *Software Engineering: A Practitioner's Approach* (5th ed.). New York: McGraw-Hill.

Pressman, R. S. (2011). *Engenharia de Software: Uma abordagem profissional* (7th ed.). Porto Alegre: McGraw-Hill-Bookman.

Quasthoff, U., Richter, M. & Biemann, C. (n.d.). *Corpus Portal for Search in Monolingual Corpora*. Leipzig.

Rocha, M. (Ed.). (2007). *Ilha do Desterro: Corpus Linguistics*. Florianópolis: Editora da UFSC.

Rocha, M. (2010). Translating anaphoric this into Portuguese: a corpus-based study. In Xiao, R. (Ed.), *Using Corpora in Contrastive and Translation Studies* (11-48). Newcastle: Cambridge Scholars Publishing.

Shreve, G. M. (2009). Recipient-Oriented and Metacognition in the Translation Process. In Dimitriu, R. & Schlesinger, M. (Eds.), *Translators and Their Readers: In Homage to Eugene A. Nida* (255-270). Brussels: Éditions du Hazard.

Shuttleworth, M. & Cowie, M. (1997). *Dictionary of Translation Studies*. Manchester: St. Jerome.

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



Sinclair, J. M. (2001). Preface. In: Ghadessy, M. et al (Ed.). *Small corpus studies and ELT – theory and practice*, (pp. vii-xv). Amsterdam: John Benjamins.

"Software Engineering." IEEE Standard Glossary of Software Engineering Terminology. 1990. Retrived from <http://www.idi.ntnu.no/grupper/su/publ/ese/ieee-se-glossary-610.12-1990.pdf> (05 Sep 2012).

Sommerville, I. (2000). *Software Engineering*. New Jersey: Pearson Higher Education.

Sphinx Technologies Inc. (n.d.). *Powered-By Sphinx*. Retrieved from <http://sphinxsearch.com/info/powered/>

Stubbs, M. (1996). *Text and Corpus Linguistics*. Oxford: Blackwell.

Tagnin, S. E. O., Teixeira, E. D., & Santos, D. (2009, September). CorTrad: a multiversion translation corpus for the Portuguese-English pair. *Arena Romanistica*. Paper presented at The 28th International Conference on lexis and grammar, Bergen, Norway, 314-323. Retrieved from [http://www.linguateca.pt/Diana/download/Tagnin-Teixeira-Santos\\_final.pdf](http://www.linguateca.pt/Diana/download/Tagnin-Teixeira-Santos_final.pdf)

Tymoczko, M. (1998). *Computerized Corpora and the Future of Translation Studies*. *Meta: Translators' Journal*, 43, 652 - 660.

Zanettin, F., Bernardini, S., & Stewart, D. (Eds.).(2003). *Corpora in Translator Education*. Manchester: St Jerome.

Zethsen, K.K. (2006). *Semantic prosody: Creating awareness about a versatile tool*. In *Journal of Sprogforskning, Årgang*, 4 (1-2), 275 - 294.

Zuniga, G. R. F. (2006). *Construing the Translator: A Meta-Reflection Grounded in Corpus-Based Translation Studies and Systemic Functional Linguistics*. Unpublished Masters Dissertation, Universidade Federal de Santa Catarina, Florianópolis.

Williams, J. & Chesterman, A. (2010). *The Map: A Beginner's guide to Doing Research in Translation Studies*. Manchester: St Jerome.

**APPENDIXES**

## APPENDIX A

## (COPA-TRAD Patent application request form)


**PEDIDO DE REGISTRO DE  
PROGRAMA DE COMPUTADOR**

protocolo

**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 assin.

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 assin.

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

**DADOS DO PROGRAMA**

Título <b>COPA-TRAD: Corpus Paralelo de Tradução</b>				
Data de Criação do Programa	<b>1/6/2011</b>	Regime de Guarda	<input checked="" type="checkbox"/> COM SIGILO	<input type="checkbox"/> SEM SIGILO
Linguagens	<b>PHP</b>	<b>JAVASCRIPT</b>		
Classificação do Campo de Aplicação	<b>CO - 03</b>	-	-	-
Classificação do Tipo de Programa	<b>FA - 01</b>	<b>GI - 01</b>	<b>UT - 01</b>	<b>GI - 08</b>

**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 assinala-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
<input type="checkbox"/> 1	Guia de Recolhimento	<input type="checkbox"/>	Contrato de Trabalho/Prestação de Serviço
<input type="checkbox"/>	Procuração	<input type="checkbox"/>	Involúcos/mídia eletrônica Utilizados
<input type="checkbox"/>	Termo de Cessão	<input type="checkbox"/>	Contrato/Estatuto Social e Alterações (ou equivalente)
<input type="checkbox"/>	Termo de Autorização para Modificações Tecnológicas ou Derivações	<input type="checkbox"/> 1	Autorização para Cópia do CD
		<input checked="" type="checkbox"/>	Outros(especificar) <b>RESOLUÇÃO Nº 14 UFSC, COMPRE ENDESCRIÇÃO DE DOCUMENTOS; DECLARAÇÃO AUTENTICIDADE DOCS; DOU NOMEAÇÃO REITOR; Cópia REITOR</b>

**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 involúcos (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 involúco 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 / 12 / 2012

Local/Data

Assinatura/Carimbo

Prof. Alvaro Toubes Prata  
 Universidade Federal de  
 Santa Catarina  
 Reitor

Modelo I (folha 02) 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. País: 55 Telefone: 48 3247 2528/ 48 9915 4018

E-mail: carlosedasilva@gmail.com

## APPENDIX B

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

288 DICIG - Diretoria de Contratos, Indicações Geográficas e Registros

RPI 2182 de 30/10/2012

Regime de Guarda: Sigilo Até  
23/04/2022  
Procurador: PAULO AUGUSTO MALTA  
MOREIRA - CPF:68320844604

Processo: 13279-5 **080**  
Título: ALFA  
Titular: UNIVERSIDADE FEDERAL DE  
VIÇOSA - CPF/CNPJ:25944550001196  
Criador: ANDRÉ FERNANDO DE  
OLIVEIRA  
Linguagem: VBA, VISUAL BASIC  
Campo de Aplicação: FO-18  
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:68320844604

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: 13291-6 **080**  
Título: COPA-TRAD: CORPUS  
PARALELO DE TRADUÇÃO  
Titular: UNIVERSIDADE FEDERAL DE  
SANTA CATARINA  
CPF/CNPJ:83899526000182  
Criador: CARLOS EDUARDO DA SILVA  
L. 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: 01/09/2011  
Regime de Guarda: Sigilo Até  
18/05/2022  
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  
VALERIA SANTOS, EMILIO TAKASE,  
MATEUS 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  
18/05/2022  
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  
Titular: ALCIRO MARCOS  
ORLAMÜNDER - CPF:68401361915

Processo: 1329-1 **080**  
Título: DOMÍNIO ESCRITA FISCAL  
VERSÃO 04  
Titular: DOMÍNIO SISTEMAS LTDA. -  
CPF/CNPJ:0225945000178  
Criador: ADRIANO DIAS, ADRIANO  
FRANCISCO, ALESSANDRA  
TEREZINHA DA SILVA, ALEXANDRE  
DE ALMEIDA, ALEXANDRE NIERO,  
ALEXANDRE ROBERTO LEMES  
MARTINS, ALINE CORREIA RAMOS,  
ALISSON DE VILHA GERONIMO,  
ALISSON DOS SANTOS SILVA,  
ANDERSON FELISBERTO MANOEL,  
ANDERSON RICARDO DOS SANTOS  
RODRIGUES, ANDERSON SILVESTRI  
FERRO, ANTONIO JOSE VIEIRA  
JUNIOR, ANTONIO MARCOS DE  
OLIVEIRA, BRUNO BRISTOT LOULI,  
CAMILA MOTTA WOSNIESKI, CARLA  
EYNG, CESAR EDUARDO FRANCO  
ISE COLONETTI, CIRILO PINTER  
COLUMBO, CLEVERSON REINERT,  
DANGELO ROSSO ZANETTE, DANIEL  
DE MEDEIROS BOFF, DAVI  
GONÇALVES, DIEGO GOMES  
ANTONELLI, DIEGO MACHADO  
MEDEIROS, DIEGO MARIANI DE  
MELO, DIEGO MARTINS DA ROCHA,  
EDGAR SOUZA DA CRUZ, EDVALDO  
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 BIRELO,  
HENRIQUE COLOMBO GUINZAIN,  
HENRIQUE PIAZZA LUCIANO,  
HERLON HILBERT HERON  
POTRIKUS CRESTANI, IURI SONEGO  
CARDOSO, JAISSON RODRIGUES  
DEMBOSKI, JEFFERSON LUIZ BATISTI,  
JESSICA RONCONI DONDOSSOLA,  
JULIANA GUADANHIM GENEROSO,  
JULIANO MARQUES, LEONARDO  
BENEDETTI, LUANA GASPAR SOARES,  
LUCAS VITORINO GONÇALVES,  
MARCIO DEHON BATISTA DE PRA,  
MARCIO DAGOSTIM DE CASTRO,  
MARCONDES DE BORBA, MARIANA  
ANTONIO SARTORI, MARIANA  
COLONETTI, MARIANNA SANTOS  
SAGGIORATO, MARILIA TEIXEIRA  
PIRES, MARINA RUTZ SCHMIDT,  
MARTY EYNG NUERNBERG,  
MATEUS MEDEIROS ANACLETO,  
MELISSA DA PAZ TEIXEIRA,  
MICHAEL CELSO BITENCOURT,  
PAULA CRISTINA VIEIRA RONSANI,  
PAULO HENRIQUE ELI, PAULO  
ROBERTO DABOIT MILANEZ, RAFAEL  
CECHINI SILVESTRI, REGINALDO  
DAROLT, RENAN SASSO DA SILVA,  
RICHARDSON PICININI CORREIA,  
ROBERTO MENDES GARCIA,  
ROBERTO VEFAGO CAROLLI,  
ROGERIO BRUM HERMANY,  
ROGERIO DAMACENO DE FARIAS,  
SAMUEL LODETTI GHELLERE,  
SAMPLE PEREIRA DA CUNHA,  
SULEN JUVENCO DAMAZO,  
TALINE FELTRIN DE SOUZA,  
TAMARA JOSEPHINO FERNANDES,  
TAMARA CRISTINA VIEIRA RONSANI,  
THALES MENDES MILANESE, THIAGO  
APOLINARIO BILLIERI, THIAGO  
BITENCORT MARQUES, TULIO  
DAMINELLI BORGES, VANESSA  
CRISTINA CARPES DA SILVA,  
VANESSA FELISBERTO BILESIMO,

WAGNER JOSÉ DENONI FREITAS,  
WELLINGTON ZOMER NUNES  
Linguagem: POWERBUILDER, SOL  
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:0225945000178  
Criador: ADRIANO DIAS, ADRIANO  
FRANCISCO, ALESSANDRA  
TEREZINHA DA SILVA, ALEXANDRE  
DE ALMEIDA, ALEXANDRE NIERO,  
ALEXANDRE ROBERTO LEMES  
MARTINS, ALINE CORREIA RAMOS,  
ALISSON DE VILHA GERONIMO,  
ALISSON DOS SANTOS SILVA,  
ANDERSON FELISBERTO MANOEL,  
ANDERSON RICARDO DOS SANTOS  
RODRIGUES, ANDERSON SILVESTRI  
FERRO, ANTONIO JOSE VIEIRA  
JUNIOR, ANTONIO MARCOS DE  
OLIVEIRA, BRUNO BRISTOT LOULI,  
CAMILA MOTTA WOSNIESKI, CARLA  
EYNG, CESAR EDUARDO FRANCO  
ISE COLONETTI, CIRILO PINTER  
COLUMBO, CLEVERSON REINERT,  
DANGELO ROSSO ZANETTE, DANIEL  
DE MEDEIROS BOFF, DAVI  
GONÇALVES, DIEGO GOMES  
ANTONELLI, DIEGO MACHADO  
MEDEIROS, DIEGO MARIANI DE  
MELO, DIEGO MARTINS DA ROCHA,  
EDGAR SOUZA DA CRUZ, EDVALDO  
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 BIRELO,  
HENRIQUE COLOMBO GUINZAIN,  
HENRIQUE PIAZZA LUCIANO,  
HERLON HILBERT HERON  
POTRIKUS CRESTANI, IURI SONEGO  
CARDOSO, JAISSON RODRIGUES  
DEMBOSKI, JEFFERSON LUIZ BATISTI,  
JESSICA RONCONI DONDOSSOLA,  
JULIANA GUADANHIM GENEROSO,  
JULIANO MARQUES, LEONARDO  
BENEDETTI, LUANA GASPAR SOARES,  
LUCAS VITORINO GONÇALVES,  
MARCIO DEHON BATISTA DE PRA,  
MARCIO DAGOSTIM DE CASTRO,  
MARCONDES DE BORBA, MARIANA  
ANTONIO SARTORI, MARIANA  
COLONETTI, MARIANNA SANTOS  
SAGGIORATO, MARILIA TEIXEIRA  
PIRES, MARINA RUTZ SCHMIDT,  
MARTY EYNG NUERNBERG,  
MATEUS MEDEIROS ANACLETO,  
MELISSA DA PAZ TEIXEIRA,  
MICHAEL CELSO BITENCOURT,  
PAULA CRISTINA VIEIRA RONSANI,  
PAULO HENRIQUE ELI, PAULO  
ROBERTO DABOIT MILANEZ, RAFAEL  
CECHINI SILVESTRI, REGINALDO  
DAROLT, RENAN SASSO DA SILVA,  
RICHARDSON PICININI CORREIA,  
ROBERTO MENDES GARCIA,  
ROBERTO VEFAGO CAROLLI,  
ROGERIO BRUM HERMANY,  
ROGERIO DAMACENO DE FARIAS,  
SAMUEL LODETTI GHELLERE,  
SAMPLE PEREIRA DA CUNHA,  
SULEN JUVENCO DAMAZO,  
TALINE FELTRIN DE SOUZA,  
TAMARA JOSEPHINO FERNANDES,

TAMIRIS JUSTI ROCHA, THALES  
MENDES MILANESE, THIAGO  
APOLINARIO BILLIERI, THIAGO  
BITENCORT MARQUES, TULIO  
DAMINELLI BORGES, VANESSA  
CRISTINA CARPES DA SILVA,  
VANESSA FELISBERTO BILESIMO,  
WAGNER JOSÉ DENONI FREITAS,  
WELLINGTON ZOMER NUNES  
Linguagem: POWER BUILDER, SOL  
Campo de Aplicação: IF-10  
Tipo de Programa: AT-02  
Data da Criação: 16/09/2009  
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: NOVOCIENTE TECNOLOGIA  
LTD. - CPF/CNPJ:14962497000133  
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 CORREIA -  
CPF:63149591015

Processo: 13450-3 **080**  
Título: PLATAFORMA E COMMERCE  
Titular: 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:6084983000142

Processo: 13451-5 **080**  
Título: CLIENTE TR - 69  
Titular: EDUARDO MALVEIRO  
PEREIRA LEITE  
Linguagem: PASCAL  
Campo de Aplicação: TC-02  
Tipo de Programa: CD-04, GI-01, SO-  
05, SO-08  
Data da Criação: 13/08/2012  
Regime de Guarda: Sigilo Até  
21/06/2022  
Procurador: SUL AMÉRICA MARCAS E  
PATENTES LTDA. -  
CPF:6084983000142

Processo: 13452-0 **080**  
Título: IP TOUCH  
Titular: EDUARDO MALVEIRO  
PEREIRA LEITE  
Linguagem: C, PHP, PYTHON  
Campo de Aplicação: TC-02  
Tipo de Programa: CD-01, CI-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:6084983000142

## APPENDIX C

### (COPA-TRAD running in a mobile browser)

#### Opera Mini Simulator

---

Below is a live demo of Opera Mini 7.1 that functions as it would when installed on a handset.





## APPENDIX D

### (COPA-TRAD public website)

# COPA-TRAD (Corpus Paralelo de Tradução)

UFSC / TRACOR / PPGI

[Home](#)
[CORPUS](#)
[Cadastro](#)
[Equipe](#)
[Sobre](#)
[Tipos De Texto](#)

## COPA-TRAD – Corpus Paralelo de Tradução



O COPA-TRAD (Corpus Paralelo de Tradução) é uma ferramenta computacional para a pesquisa, ensino/aprendizagem e prática da tradução.

A princípio, o **COPA-TRAD** prevê 5 (cinco) subcorpora que logo estarão disponíveis a toda comunidade acadêmica:

- COPA-LIJ (Corpus Paralelo de Literatura Infantil e Juvenil);
- COPA-TEL (Corpus Paralelo de Textos Literários);
- COPA-MDT (Corpus Paralelo de Meta-Discursos em Tradução);
- COPA-RAC (Corpus Paralelo de Resumos Acadêmicos);
- COPA-MUM (Corpus Paralelo de Multimodalidade)

O acesso a ferramenta é feito somente através de usuário e senha no link [copa-trad.ufsc.br/corpus](http://copa-trad.ufsc.br/corpus).

### Sobre o COPA-TRAD

O COPA-TRAD é um corpus paralelo que tem como objetivo oferecer ferramentas computacionais disponíveis online para a pesquisa, ensino e prática da tradução.

### Traduzir esta página

Select Language

Powered by [Google Translate](#)

### Lista de Links

- [Acesso ao Corpus](#)
- [PPGI](#)
- [TracOR](#)

## APPENDIX E

## (Website of “University Research Program for Google Translate”)



Research at Google




---

Home   Research Areas & Publications   People   **Research Programs**   Work at Google

Overview

Award Programs

Student Support

Tools and Resources

Workshops &  
Conferences

## University Research Program for Google Translate

### Overview

The University Research Program for Google Translate provides researchers, in the field of automatic machine translation, tools to help compare and contrast with, and build on top of, Google's statistical machine translation system.

Participation in the program will allow researchers programmatic access to Google's translation service.

A translation request returns either:

- A single translation, the highest scoring output of Google Translate
- As above but with detailed word alignment information
- A list of the n-best translations with detailed scoring information

These options can support research into n-best reranking (machine learning), additional translation components, system combinations, and who knows what else researchers may come up with.

The program supports all languages available publicly at [translate.google.com](http://translate.google.com).

### Restrictions

This research program is being made available to members of the academic community. Each applicant must review and adhere to the full terms, which include the following restrictions:

- The research program is for research purposes only. Commercial use is strictly forbidden.
- Participants may submit no more than 1000 translation requests per day. If your project requires more daily throughput, let us know in your proposal and we'll do our best to accommodate you.
- Publications presenting research that was done using resources provided by this program must include attribution to Google for providing those resources.
- The program may be used only by registered researchers and their teams, and access may not be shared with others.

Before registering for the Google Translate research program, you must read and agree to the full Terms of Use. Please take the time to familiarize yourself with the terms of the program before submitting your application.

### Registration

## APPENDIX F

## (Microsoft Translator Advertisement)

Microsoft®  
**Translator**

**Online translation service for a truly worldwide web**

Microsoft Translator technology brings the power of instant translations to any destination, helping to break the language barrier for users, developers, webmasters and businesses alike.

**Microsoft Translator:**

- Destination for your translation needs
- Powerful widget for your website
- Always ready in Microsoft Office
- Rich API for your custom application
- Collaboration framework for human-quality translation

**Translating in:**




**bing™** 

**Microsoft Translator:**  
[www.bing.com/translator](http://www.bing.com/translator)


**Microsoft Translator Tools:**  
[www.microsofttranslator.com/tools/](http://www.microsofttranslator.com/tools/)

**Microsoft Translator API:**  
[api.microsofttranslator.com](http://api.microsofttranslator.com)



## APPENDIX G

### (Microsoft Translator API Management for COPA-TRAD)



**Microsoft Translator**


Microsoft Translator delivers automatic translation (Machine Translation) of a text into a specified language. It is a state-of-the-art statistical machine translation system translating between any of the supported languages, and powering millions of translations every day.

[1,959,695 Characters remaining](#)

[Primary Account Key Show](#)

Download Options:

- [Excel \(CSV\)](#)
- [PowerPivot 2010](#)
- [PowerPivot 2013](#)



URL for current expressed query:

<https://api.datamarket.azure.com/Bing/MicrosoftTranslator/v1/Translate>

Displaying 0 rows

**Required parameters:**

Text

Value like: : hello

To

**Optional parameters:**

From

---

[Translate](#)   [GetLanguagesForTranslation](#)   [Detect](#)

## APPENDIX H

**(Me, Mona Baker, Lincoln Fernandes and Danielle Amanda  
at my first presentation of COPA-TRAD)**



Advance, and never halt, for advancing is perfection. Advance and do not fear the thorns in the path, for they draw only corrupt blood.

**Kahlil Gibran**