UNIVERSIDADE FEDERAL DE SANTA CATARINA

PÓS-GRADUAÇÃO EM LETRAS/INGLÊS E LITERATURA

CORRESPONDENTE

THE PROBLEM OF CODIFYING LINGUISTIC KNOWLEDGE IN TWO TRANSLATIONS OF SHAKESPEARE'S SONNETS: A CORPUS-BASED STUDY

FLÁVIA AZEVEDO

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

DOUTOR EM LETRAS

FLORIANÓPOLIS

Novembro de 2012

Esta Tese de Flávia Azevedo, intitulada The problem of codifying linguistic knowledge in two translations of Shakespeare's Sonnets: a corpus-based study, 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

DOUTOR EM LETRAS

Área de concentração: Língua Inglesa e Lingüística Aplicada

_____
Profª Drª Susana Bornéo Funk
Coordenadora

BANCA EXAMINADORA: _____
Prof. Dr Marco Rocha
Orientador e Presidente

_____
Profª Drª Diva Camargo

_____
Prof. Dr Lincoln Fernandes

_____
Profª Drª Maria Alice Antunes

_____
Profª Drª Maria Lucia Vasconcellos

_____
Profª Drª. Viviane Herbele

Florianópolis, 29 de novembro de 2012.

The completion of my dissertation has been a long and rewarding process, and it is now time to thank some people who had helped me during this important period of my life.
First, I would like to thank CAPES and REUNI for the scholarship that I received during the course.
Second, I would like to thank all the professors from PPGI, who contributed to my work, especially professor José Roberto O'Shea and professor Lincoln Fernandes who examined and made important contributions to my Research Project.
I want to express my gratitude to Martha Thunes, who inspired this research and provided a copy of her doctoral dissertation when we first met in Oslo during the ICAME Conference.
I deeply thank my husband Guilherme Amaral who was always so full of understanding and who has been encouraging and supporting me in all projects of my life.
I also thank my dear friend Denize Nobre who read my dissertation and helped me with text formatting.
Finally, I deeply thank my friend and adviser professor Marco Rocha, for having carried the responsibility of supervising this project and for being my mentor since I started my graduation course at Universidade Federal de Santa Catarina.

ABSTRACT

THE PROBLEM OF CODIFYING LINGUISTIC KNOWLEDGE IN TWO TRANSLATIONS OF SHAKESPEARE'S SONNETS: A CORPUS-BASED STUDY

FLÁVIA AZEVEDO

UNIVERSIDADE FEDERAL DE SANTA CATARINA
2012

Supervising Professor: PhD. Marco Rocha

The present study deals with the problem of codifying linguistic knowledge in a parallel corpus, in other words, the process of corpus annotation. The purpose of the present study was to test the identification of four types of translational correspondence, as defined by Thunes (2011) in a parallel corpus made up of 45 Shakespeare's Sonnets and two distinct translations into Brazilian Portuguese. The obtained results show that Thunes' model can be considered effective when applied to classify alignment units in a parallel corpus of translated poetry, but it needs some adjustments in order to cope with some translational pairs which did not fit properly into any of the four categories. The advantage of Thunes' proposal is that it establishes criteria to analyse complexity involved in the translation process in a very clear way.

Keywords: Shakespeare's sonnets, parallel corpus, corpus annotation, alignment units, types of translational correspondence, poetry translation, translators' styles.

RESUMO

O PROBLEMA DE CODIFICAR CONHECIMENTO LINGUÍSTICO EM DUAS TRADUÇÕES DOS SONETOS DE SHAKEPEARE: UM ESTUDO BASEADO EM CORPUS

FLÁVIA AZEVEDO

UNIVERSIDADE FEDERAL DE SANTA CATARINA
2012

Professor Orientador: Dr. Marco Rocha

Este estudo aborda o problema de codificação do conhecimento linguístico em um corpus paralelo, em outras palavras, o processo de anotação de corpus. O objetivo deste estudo foi testar a identificação dos quatro tipos de correspondência tradutória descritos por Thunes (2011) em um corpus paralelo constituído por 45 sonetos de Shakespeare e duas traduções distintas em Português. Os resultados obtidos mostram que o modelo de Thunes pode ser considerado eficaz quando utilizado para classificar unidades de alinhamento em um corpus paralelo de poesia traduzida, mas precisa de algumas adaptações, a fim de lidar com alguns pares tradutórios que não se ajustaram adequadamente em nenhuma das quatro categorias propostas. O modelo proposto por Thunes pode ser considerado vantajoso por estabelecer critérios para analisar a complexidade envolvida no processo de tradução de uma forma muito clara.

Palavras-chave: sonetos de Shakespeare, corpus paralelo, anotação de corpus, unidades de alinhamento, tipos de correspondência tradutória, tradução de poesia, estilos de tradutores.

List of Figures

List of Tables

List of Abbreviations

ACL	Association for Computational Linguistics
API	Application Programming Interface
BNC	British National Corpus
EEC	European Economic Community
EModE	English in Early Modern English
JW	Jorge Wanderley
LPT	Linguistically Predictable Translations
MIME	Multipurpose Internet Mail Extensions
MT	Machine Translation
POS	Part of Speech
PS	Péricles Eugênio da Silva Ramos
SL	Source Language
URL	Uniform Resource Locator
USAS	UCREL Semantic Annotation Scheme
UG	Universal Grammar
TL	Target Language

Chapter 1

Introduction

1.1 The study of complexity in Poetry Translation: How it began

This study was inspired by Martha Thunes' work entitled "Classifying translational correspondences" (1998). My first contact with Thunes' model of translational complexity happened when I was taking a Masters Course at Universidade Federal de Santa Catarina (UFSC). My adviser, Dr. Marco Rocha offered a discipline of computational linguistics and he presented this model as an option to be used for corpus annotation. Two years later, while working on my Doctoral Proposal, I decided to combine this model with my interest on Shakespeare's Sonnets and Translation Studies. Initially, my objective was to analyse the complexity involved in two translations of Shakespeare's Sonnets into Brazilian Portuguese.

In 2011, I attended the conference ICAME (International Computer Archive of Modern and Medieval English) in Norway, where I had the opportunity to meet Martha Thunes in person. At that time, she had just defended her Doctoral Dissertation entitled "*Complexity in Transaltion, An-English-Norwegian Study of Two Text Types*" (2011). In this work, Thunes develops in greater depth the ideas initially presented in the article "Classifying translational correspondences", published in 1998. I received a copy of her work, which constitutes the basis of the present investigation.

Similarly to Thunes (2011), I have chosen an empirical approach to analyse verses extracted from parallel texts, considering that they are part of the extension of the translational relation. By definition, a parallel corpus contains source texts and their translations and it can be bilingual or multilingual (McEnery & Xiao, 2008). Differently from Thunes, who expected to verify to what extent the translation between the language pair English-Norwegian could be done automatically, my main objective was to show that her model on complexity is useful as a

means to reveal stylistic differences between two translators and shed light on some of their choices during the translation process.

There would be no interest in finding out if the translation of poetry could be done automatically because we already know that the results expected from this kind of translation so far can only be achieved by a "bilingually competent human translator"[1] (Thunes, 2011, p. 3). Therefore, this study attempts to answer these two primary research questions:

1. How can we adapt Thunes' model in order to be used to analyse any parallel corpus made up of Brazilian Portuguese translated poetry?
2. To what extent can the analysis of complexity in two translations of that pair of languages point to stylistic differences between the translators?

Following Thunes (2011), I recognize that the extent to which this study can answer the proposed research questions is limited to the scope of my empirical analysis. That means that my results apply only to that part of the translational relation between English and Brazilian Portuguese which is covered by the very specific selected parallel texts of translated poetry.

The study carried out by Thunes applied "a method where translationally corresponding text units are classified according to a measure of the complexity of the relation between source and target expression" (2011, p. 3-4). According to Thunes,

> the complexity measure is based on assumptions concerning a translator's need for information when producing the given target text, and this need for information is analysed in terms of how much information is needed, what types of information this involves, and the effort required in order to access and process them. We assume a scale of translational complexity, and on this scale we have identified four main types of translational competence. When a pair of translational unit is analysed, it is assigned one of these four types, as a classification of the

[1] The sophistication of automatic translation systems might lead to better results in terms of poetry translation.

> complexity of the translational relation between the two units. (Thunes, 2001, p. 4)

It is important to emphasise that the classification of correspondence proposed by Thunes involves no evaluation of translational quality, but in the present study translational quality will be discussed during the discussion of the results. Being aware of the difficulties in establishing criteria to evaluate translational quality (which tend to be very subjective), this work attempts to adapt Thunes' methodological framework in order to be used to investigate the question of style in literary translation, similarly to what is proposed by Baker in the article "*Towards a Methodology for Investigating the Style of a Literary Translator*" (2000).

Since this investigation was inspired by Thunes' work and uses her model and methodology for the analysis, both studies share the same "nature". In other words, the main objective is to analyse the product of translation, assuming that "an empirical investigation of parallel texts, as instantiations of the translational relation, may serve as a basis for studying translation competence" (Thunes, 2011, p. 5). Therefore, similarly to Thunes' study (2011), the present work cannot be considered a cognitive or psycholinguistic investigation of translation, the focus is not the procedure involved during human translation, but the external and objective result of it, the translational pairs.

The analysis of the information that is accessible through the competence of translators is an important topic in Thunes' investigation because she believes that the analysis of a translation (comparing it to its original) might reveal not only types of information related to translator's competence, but also other types of information which are accessed by him/her in order to produce a specific target text.

Although Thunes' study does not approach *translation competence*, she considered important to list a simple and intuitive conception of translation competence before explaining the nature of her study. These aspects, which are also related to the present investigation, are listed below:

> (i) Competence in the source language (SL) as well as in the target language (TL), and knowledge of how these two language systems are interrelated.
> (ii) Necessary background knowledge of various kinds.

> (iii) The ability to assign an interpretation to the SL text by merging the information encoded in the text itself with the information present in the textual context and in the utterance situation.
> (iv) The ability to construct a translation which will receive an interpretation in the TL context and utterance situation which is as close as possible to the interpretation of the original, given its purpose. (Thunes, 2011, p. 6)

As observed, the task of translation involves many competences, but when it comes to poetry translation it would be necessary to include a fourth item (iv): the ability to construct a translation that fits the metre and rhymes. Besides that, there are issues of beauty and musicality, subjective characteristics which are intrinsically related to poetry in general.

The great complexity involved on this type of translation led me to build and analyse a parallel corpus of poetry. When we first met, Thunes said that she was very surprised to know that her model was being used in such type of corpus. Thunes' parallel corpus consisted of fiction and law texts, but I realised that her model could be applied to analyse complexity in poetry translation as well, it would just need some modifications to cope with some issues that the original model could not predict, since it was not designed for this purpose.

Thunes affirms that "the various kinds of information that are accessible through translation competence are part of the information needed to produce a specific translation from a given SL expression" (2011, p. 6). Since the objective of her study was to describe a typology of information sources for translation, she makes a distinction between three main types of information:

> (a) Purely linguistic information, some of which is encoded in the SL expression, and some of which is inherent in a translator's bilingual competence and knowledge of interrelations between source and target language systems.
> (b) Pragmatic information from the textual context and the utterance situation of the source expression.
> (c) Various kinds of extra-linguistic background information. (Thunes, 2011, p. 6)

In addition to the types of information presented above, Thunes also makes a distinction between general and task-specific information sources. The first one includes "information about source and target language systems and their interrelations" (2011, p. 7), while the second covers "information about a particular piece of source text and the concrete task of translating it into a given target language" (ibid). The general information is available prior to the translation process, while the task-specific task is related to process involved during the translation of a particular piece of text.

The distinction of these types of information is necessary because the translational complexity described by Thunes is based on the amount and types of information needed during the production of a target text, therefore the information typology is used to analyse "the degree of translational complexity in correspondences between expressions of two languages" (Thunes, 2011, p. 7).

1.2 The background for the correspondence type hierarchy

In Thunes' work, a scale of translational complexity is defined by a hierarchy of four types of translational correspondence. This hierarchy was originally developed by Helge Dyvik in a work on an experimental machine translation system entitled *The PONS Project: Features of a Translation System* (1990). This system contains information about source and target language systems and their respective interrelations, a model that was supposed to be similar to the translator's bilingual competence:

> The first step of the translation task is to analyse the input, a procedure which is comparable to a translator's reading and understanding of the source sentence. The analysis provides the system with information about the syntactic structure of the input text, which is then compared with information about source and target language interrelations. (Thunes, 2011, p.12)

The PONS machine translation system uses this comparison between source and target language to distinguish three distinct modes of translation. This distinction is done according to the complexity

involved during the translation task. The system basically works identifying the matching of syntactic structures between source and target languages (Thunes, 2011).

In order to explain and exemplify how the PONS machine translation system works, some examples of Portuguese sentences translated into English by Google Translator are presented below, pretending that they would be produced by PONS system, just to illustrate hypothetically how it functions. If we take the following sentence in Portuguese and use *Google Translate* to translate it, we would obtain the translation pair below:

(1) Durante esses vinte anos muitos sonetos foram escritos na Inglaterra. (SL)
During these twenty years many sonnets were written in England. (TL)

As observed, all the elements of the input text have a match in the TL grammar. This would be an example of type 1 correspondence[2] produced by the PONS machine translation system (Section 2.6), which would translate word by word during the process. In order to create the target sentence it would be only necessary to have information about word order and syntactic structure of the source sentence.

In other cases, the system would realize that the source sentence structure could be partially matched by the target grammar because there is at least one difference related to constituent sequence and/or the presence of grammatical form words (Thunes, 2011). The two examples below can hypothetically represent type 2 correspondences (Section 2.6) produced by the system. In (2), there is no correspondence for the pronoun *it* in the source string, characterizing an example which contains an extra grammatical word form in the TL. Example (3) represents a difference in relation to constituent sequence, since adjectives in English are usually placed in front of the noun[3].

(2) Está chovendo muito aqui. (SL)
It is raining a lot here. (TL)

(3) A primeira quadra do soneto é quase traduzível com a mesma **estrutura sintática**.

[2] This type of correspondence was also described by other authors, like Vinay and Dalbernet (1968), who called it obligatory transposition.
[3] In English, adjectives can come either before the noun (e.g. She is a beautiful girl) or after the verb (She is beautiful).

The first block of the sonnet is almost translatable with the same **syntactic structure**.

With respect to type 3 correspondences (Section 2.6), they represent cases "where the PONS system finds that with respect to the function and/or category of at least one lexical word, the syntactic structure of the source sentence cannot be matched by the target language" (Thunes, 2011, p. 12). Observe Example (4):

(4) **No que se refere** ao inglês, é muito difícil conservar o mesmo metro e **o** mesmo ritmo do original nas **traduções em português**.
With regard to English, **it** is very difficult **to keep** the same pace and metre of the original in **Portuguese translations**.

In this case, the system needs to produce "a full semantic analysis of the input, and use a semantic representation of the source sentence as the basis for the target text generation" (ibid). The task here requires "semantic information about the input text together with structural and lexical information about the target language" (Thunes, 2011, p. 13). The system would need to know that the expression *no que se refere* cannot be translated literally (it corresponds to *with regard*), that the personal pronoun *it* needs to be added in English to conform to its grammar, and finally, that the infinitive form *conservar* becomes *to keep* in English.

In summary, these are the three distinct modes of translation used by the system to produce the translations. The fourth type of correspondence (type 4) which is included in Thunes' model does not belong to the PONS system because it refers to

> (…) cases where purely linguistic information is insufficient, and the translation task requires additional information sources, such as extra-linguistic background information and discourse information derived from a wider linguistic context. (Thunes, 2011, p. 13)

The PONS system represents, therefore, the original background of Thunes' model. However, she emphasises that all the discussion related to automatisation in her investigation "is discussed without reference to the architecture of any particular machine translation system, although the analytical framework is inspired by the PONS design" (ibid). The same principle applies to the present investigation, I

expect to estimate the degree of translational complexity in the parallel corpus, but I do not intend to use this information to verify to what extent automatic translation is feasible within the investigated parallel texts.

From now on I will present and illustrate the four types of translational correspondence with reference to the verse, which was chosen as the basic unit of translation in this study. All the examples used to illustrate the correspondence types were extracted from the parallel corpus.

Type 1 is considered the least complex type of correspondence. It is described by Thunes as "cases of word-by-word translations where source and target string are identical with respect to the sequence of word forms" (2011, p. 8). Example (5) was the only occurrence of type 1 correspondence identified in the corpus.

(5a) And other strains of woe, which now seem woe,
(5b) E outras formas de dor, que ora parecem dor,

Type 2 correspondences are considered more complex because "source and target string are not matched word by word, but every lexical word in the source expression has a target correspondent of the same lexical category and with the same syntactic function as the source word" (Thunes, 2011, p. 8). Consider Example (6) which was identified in the parallel corpus and classified as type 2:

(6a) Kissing with golden face the meadows green,
(6b) Beijar com face de ouro o prado verdejante,

As for type 3 correspondences, translational complexity is considered higher if compared to type 2 because there are more structural differences between source and target strings. Although there is no mismatch on the semantic level, there is at least one structural difference in one of the strings that violates syntactic functional equivalence (Thunes, 2011). One example of violation in the syntactic function is shown in Example (7):

(7a) For thee and for myself no quiet find.
(7b) Não podem repousar, graças a ti e a mim.

There is no correspondence for the noun *quiet* in the target string. While in the original verse the syntactic structure is adverb + noun + verb (*no*

quiet find), in the target text the correspondent structure is adverb + verb + verb (*não podem repousar*).

In type 4 correspondences, "complexity is even higher: in such cases there are discrepancies between original and translation not only on the structural level, but also on the semantic [level]" (Thunes, 2011, p. 10).

(8a) Yet, do thy worst, old Time: <u>despite thy wrong</u>,
(8b) Faze o pior, porém: <u>malgrado o teu rigor</u>,

Here, the expression *despite thy wrong* was translated as *malgrado o teu rigor*, and this represents a mismatch on the semantic level, there is no correspondence between source and target string in relation to meaning.

Therefore, as shown in the examples above, "a central aspect of the correspondence type hierarchy is the increase in the degree of translational complexity from type 1 upwards" (Thunes, 2011 p. 10). As observed, the degree of translational complexity seems to gradually increase from Examples (1) to (4). The four types of translational correspondence described in this subsection represent the starting point of this investigation.

1.3 The value of corpus-based studies

After briefly presenting the framework chosen for this investigation, it is important to justify the value of corpus-based studies for the field of Translation Studies and eventually show the contributions that this study in particular might bring to this field.

The first issue related to the value of corpus-based studies is justified by McEnery and Xiao (2008). According to these authors, if one considers and observes the evolution of corpus-based approaches during the last years, he/she will visualise that the theoretical elaborations and empirical realisations related to them evolved into a coherent and rich paradigm which addresses diverse issues such as theory, description and also the practice of translation.

Corpus-based studies can be divided in two broad areas, theoretical and practical, which in turn offer many possibilities of investigations:

> In theoretical terms, corpora are used mainly to study the translation process by exploring how an idea in one language is conveyed in another language and by comparing the linguistic features and their frequencies in translated L2 texts and comparable L1 texts. In the practical approach, corpora provide a workbench for training translators and a basis for developing applications like MT [Machine Translation] and computer-assisted translation (CAT) systems. (McEnery & Xiao, 2008, p. 22)

Interestingly, the present study could be placed between these two broad areas because part of the analysis consists of exploring how an idea in English is conveyed in Portuguese according to the measure of complexity in the translation task. On the other hand, the results might also be used to create a workbench for training translators or, to offer basis for developing applications for MT systems, although this is not one of the specific objectives of this investigation.

With respect to the contributions that this study might bring to this field, I expect that the application of the model (that was initially used for MT translations) in a parallel corpus of translated poetry will shed light on some of the translators' choices and their styles in a practical way. This might be useful for literary translators, since most studies focus on translation quality, and there are no studies that focus on structural differences between distinct translations of poetry based on a complexity model to the best of my knowledge. The advantage here is that Thunes' model offers very clear criteria to explain translators' choices on the syntactic level.

1.4 Organisation

This dissertation consists of five parts, among which the present chapter constitutes the first one. The purpose of this chapter has been to state my research questions, to introduce the main framework, and to present some important topics related to the present study. Chapter 2 covers the theoretical and analytical foundations of this investigation. Chapter 3 describes the method applied during the empirical

investigation. Chapter 4 presents the results of the analysis and discusses them in relation to the initial research questions. Finally, Chapter 5 brings out the conclusions centred on the presented framework, the method, and the results obtained from the study. In addition, I present suggestions for future research that could be viewed as an extension of the analytical approach used in this investigation.

Chapter 2

Review of Literature

This review of literature is divided into eight main parts, which together present the theoretical basis of this investigation on translational complexity in two translations of Shakespeare's Sonnets into Brazilian Portuguese. In the first section, I introduce the issue of the use of the Web as a corpus, a new tendency in Computational Linguistics which deserves to be explored. In Section two, the role of corpora in Translation Studies describes major influences on the present study. In sequence, some studies that discuss the relation between the linguist and the translator are presented. Section three presents an overview about corpus annotation. In Section four, four articles related to some of the issues proposed by the present investigation are reviewed. In Section five, the concept of Universals of Translation is explained. Section six brings Thune's model to light, as it constitutes the main theoretical basis of this investigation. Section seven discusses Baker's suggestions presented in the article *Toward a Methodology for Investigating the Style of a Literary Translator*. In Section eight, suggestions towards a more objective evaluation of poetic translations by Paulo Henrique Britto is presented. Finally, section nine brings information about Shakespeare's Sonnets and their two respective translations into Brazilian Portuguese.

2.1 Corpus Linguistics: the special issue of the Web as a Corpus

The advent of computers was followed by the creation and use of corpora in linguistic studies. The Brown Corpus was the first one, compiled in the 1960's. After the Brown Corpus, Sinclair and Atkins developed the COBUILD Corpus, which in 1980 already contained eight million words. Ten years later, Atkins got involved in the

development of the British National Corpus (BNC) which was even bigger, containing 100 million words (Kilgarriff, 2003). Nowadays, one of the issues involving corpus linguistics is related to the use of the Web as a corpus.

The Web is an immense and free source of information available by the simple action of a mouse click. The 'hundreds of billions of words of texts' provided by it can be used in many areas of research. One of the most common uses of this tool is spelling check. If you are in doubt in relation to the spelling of a word, Google will easily answer the question: the most frequent word is the correct one (Kilgarriff, 2003).

According to Kilgarriff (ibid), language scientists and technologists are appealing to the Web for three main reasons: First, the Web provides a large amount of data; second, sometimes it is the only available source for a specific type of language which they are interested in; and finally, the Web is free and immediately available.

The first subject discussed by Kilgarriff is related to the primary question "Is the Web a corpus?". He prefers to avoid the misconceptions and controversial discussions involving corpus-hood and defines corpus simply as "a collection of texts". For those who consider this definition too broad, he proposes the following: "a corpus is a collection of texts when considered as an object of language or literary study" (2003, p. 2). By considering this definition and the fact that the Web represents a collection of texts, his answer to the question "Is the Web a corpus?" is affirmative.

After affirming that the Web is a corpus, the second issue approached by Kilgarriff relates to the main tool used to process all the information provided by the Web: the computer. Computers are used by distinct areas of study in different ways. Researchers of Chemistry and Biology, for example, use the computer basically as a place to store and process information related to their fields of study. Linguists, on the other hand, find in computers their object of study represented on its two primary forms: written and acoustic. In this context, the text is considered the information object and the computer's hard disc is the valid place to look for it, similar to a printed page or any other location (Kilgarriff, 2003).

The starting point for computer-based language study happened in 1960 with the creation of the Brown corpus, containing one million words. In 1970, Sinclair and Atkins noticed the necessity of developing lexicography for the vast amount of data available and they started the COBUILD project, which was responsible for the growth of corpus size, reaching the extent of eight million words in the 1980s. Confirming the

geometric progression growth of corpora, the creation of 100-million-word BNC, for which Atkins was also responsible, consolidated the concept of vast data available at low cost (Kilgarriff, 2003).

The incorporation of corpora into computational linguistics was formalized in 1989, during the Association for Computational Linguistics (ACL) meeting which took place in Vancouver. At that time, they were considered "large, messy, ugly objects clearly lacking in theoretical integrity in all sorts of ways, and many people were skeptical regarding their role in the discipline" (Kilgarriff, 2003, p. 334). There was not an agreement in relation to corpus work as part of the field of computational linguistics, and the consummation between the corpora and this field of study only happened in 1993, with the publication of the special issue on "Using Large Corpora" of the journal *Computational Linguistics 19* (Kilgarriff, 2003).

Based on this "messy view" of corpora left by the ACL meeting, it is not difficult to trace a parallel with Web corpus work. One of the problems related to the Web is its anarchic nature, and the fact that its use is still placed out of the territory of computation linguistics. But given the vast amount of information offered by the Web that sometimes is not available in traditional corpus and the fact that it is easily accessible, free from the copyrights that limit corpus development, Web-based work will probably continue to grow (Kilgarriff, 2003).

Another important aspect in relation to the use of the Web as a corpus is that large corpora containing about one hundred million words provide enough data for lexicographers or linguists who want to investigate empirical strategies for learning about language, or for technologies that demand quantitative information about the behaviour of words. But depending on the purpose of the investigation, one million words might not be enough (Kilgarriff, 2003).

Kilgarriff illustrates this by mentioning the issue of word frequencies, according to which the amount of data provided by a 100-million-word corpus might not guarantee enough information, depending on the object of study. In those cases it is better to get out of the comfort zone of traditional corpora and take the risk of using the Web. Using the Web as an option of larger data source, therefore, represents an advance in terms of computational linguistics studies, as performance tends to improve with the increase of the amount of data.

According to Kilgarriff, besides being an option of larger data source, the Web is multilingual and it continuously grows. In July 1999, the Web contained 56 million registered network addresses. In January 2003, the number of identified pages was 172 million. Considering the

significant growth in several languages, the Web can definitely be considered a multilingual corpus. At that time, "71% of the Web pages were written in English, followed by Japanese (6.8%), German (5.1%), French (1.8 %), Chinese (1.5%), Spanish (1.1%), Italian (0.9%), and Swedish (0.7%)" (Xu, 2000, qtd. in Kilgarriff, 2003, p. 337).

In one of his investigations, Kilgarriff (2003) has measured the occurrence of some English phrases using several search engines available in the Web and compared them with the counts identified in the BNC. The phrase *deep breath*, for example, appeared 732 times in BNC, while in 1998 the Web search engine Alta Vista indexed it 54,550 times. In 2001, the number increased to 170,921, finally reaching the number of 868,631 Web pages in 2003, data provided by AlltheWeb[4] search engine. At that point, the search engines were able to identify numbers significantly higher when compared to BNC counts and this type of information is a clear indication of the size of the English corpus provided by the Web.

Another way to estimate the number of words available in the Web through the use of a search engine is by counting function words. Function words, like articles and prepositions, have little semantic content; in fact, they usually indicate a grammatical relationship. According to Kilgarriff (2003), they can be used as 'predictors of corpus size' because their frequency of occurrence over different types of text seems to be stable. By knowing the corpus size, it is possible to calculate the frequency of these words. For example, the article *the* occurs 5,776,487 times in BNC, specifically about seven times for every 100 words of the corpus. The same word appears 84 times in the U.S. Declaration of Independency. Considering this information, he predicts that the Declarion is about 84 x 100/7, that is, it probably contains 1,200 words. Interestingly, the text contains in fact approximately 1,500 words. The conclusion is that the frequency of one word might indicate the first approximation of the size of the whole text. A better result can be obtained through the use of more data points.

The author used the method described above to calculate frequencies of function words and short common words of the German text extracted from the European Corpus Initiative (ECI) Multilingual Corpus[5]. The first procedure was to remove from the list those words

[4] An Internet search engine created in 1999 which once rivaled Google in size and technology, but never reached its popularity. More information available at http://en.wikipedia.org/wiki/AlltheWeb.

[5] http://www.elsnet.org/resources/eciCorpus.html.

which were also common in other languages. In order to obtain the occurrence of each query word on the Web, Alta Vista search engine was used. Table 1 presents the estimate frequency of words from the European Corpus Initiative Multilingual Corpus, along with the frequencies provided by Alta Vista in February 2000, and the estimative of the German-language Web size according to the data provided by the search engine.

Table 1. Short German words in the ECI corpus and via Alta Vista, giving German Web estimates (Kilgarriff, 2003, p. 338).

Word	Known-Size-Corpus Relative Frequency	AltaVista Frequency	Prediction for German-Language Web
oder	0.00561180	13,566,463	2,417,488,684
sind	0.00477555	11,944,284	2,501,132,644
auch	0.00581108	15,504,327	2,668,062,907
wird	0.00400690	11,286,438	2,816,750,605
nicht	0.00646585	18,294,174	2,829,353,294
eine	0.00691066	19,739,540	2,856,389,983
sich	0.00604594	17,547,518	2,902,363,900
ist	0.00886430	26,429,327	2,981,546,991
auf	0.00744444	24,852,802	3,338,438,082
und	0.02892370	101,250,806	3,500,617,348
Average			3,068,760,356

In conclusion, the average of the remaining predictions gave an estimate of three billion words of German that could be accessed through Alta Vista on the day that Kilgarriff's team conducted the test (February 2000).

The same technique has been applied on other controlled data and reliable results were obtained. Kilgarriff estimated "the number of words that were available in 30 different Latin-script languages through Alta Vista in March 2001" (2003, p. 339). As observed in Table 2, English easily surpassed the other languages with 76 billion words. Seven of the other languages already contained more than a billion.

Table 2. Estimates of Web size in words, as indexed by Alta Vista, for various languages (Kilgarriff, 2003, p. 339).

Language	Web Size	Language	Web Size
Albanian	10,332,000	Catalan	203,592,000
Breton	12,705,000	Slovakian	216,595,000
Welsh	14,993,000	Polish	322,283,000
Lithuanian	35,426,000	Finnish	326,379,000
Latvian	39,679,000	Danish	346,945,000
Icelandic	53,941,000	Hungarian	457,522,000
Basque	55,340,000	Czech	520,181,000
Latin	55,943,000	Norwegian	609,934,000
Esperanto	57,154,000	Swedish	1,003,075,000
Roumanian	86,392,000	Dutch	1,063,012,000
Irish	88,283,000	Portuguese	1,333,664,000
Estonian	98,066,000	Italian	1,845,026,000
Slovenian	119,153,000	Spanish	2,658,631,000
Croatian	136,073,000	French	3,836,874,000
Malay	157,241,000	German	7,035,850,000
Turkish	187,356,000	English	76,598,718,000

Table 2 shows that even languages like Malay, Croatian, and Slovenian, which can be considered 'smaller' (when compared to English) have the considerable number of one hundred million words on the Web. Kilgarriff (2003) suggests that much research that exploits this type of scale could be applied to other languages, instead of just being undertaken on the traditional BNC, a corpus of English language.

In order to justify the numbers presented in Table 2, Kilgarriff presents three reasons: First, Alta Vista does not cover all the indexable pages available at the Web, covering only a fraction of them, which is estimated at just 15% by Lawrence and Giles (1999, qtd. in Kilgarriff, 2003). Second, Alta Vista tends to be biased toward North American English pages by the strategy used to crawl the Web. Maybe this happens because North American English is considered the main English variant. Finally, Alta Vista does not consider and index texts that are accessible through dialogue windows on Web pages, texts which are considered as part of the 'hidden Web'. This search engine indexes only pages that can be directly called by a Uniform Resource Locator (URL). That means that a large amount of data is missed,

because the hidden Web is vast. Just to illustrate, a database like MedLine[6] contains more than five billion words, an example of a significant number not considered at all in Alta Vista estimates (Kilgarriff, 2003).

The procedure performed by Kilgarriff's team was repeated after a period of time and they found that the amount of non-English text when compared to English is growing significantly. Alta Vista indexed 38 German words for every 1,000 words of English in 1996, and this number increased respectively to 71 in 1999, and 92 in 2001 (Kilgarriff, 2003). This fact can be interpreted in a positive way, since the increase in the amount of non-English texts represents more data for linguists interested in studying other languages besides English.

Kilgarriff also discusses the issue of representativeness. The term itself leads to the question 'representative of what?'. As a matter of fact, it is difficult to say precisely what the available corpora can be representative of. When someone decides to build a corpus of general English, the obvious expectation is because he/she wants the corpus to be representative of this variety of English. As a consequence, it is necessary to establish the 'general English events' that the corpus will represent (Kilgarriff, 2003). The author discusses the issue of representativeness based on six main topics: theory, technology, language modelling, language errors, sublanguages and general-language-corpus composition, and literature.

In relation to *theory*, four issues should be considered. First, when it comes to production and reception, it is necessary to decide whether the language event is an event of speaking or writing, or reading or hearing. A standard conversation usually has one speaker and one hearer for each utterance produced. An article published in a newspaper like *Times*, for example, is written by at least one author but has several readers. The second issue concerns speech and text. In that case, deciding whether speech and written events share the same status is necessary. So far, most corpus research has tended to focus and work with written material, probably because they are easier to compile and manipulate. Third, the issue of background knowledge, which involves deciding if 'muttering under one's breath' or 'talking in one's sleep' should be considered a speech event. Other examples are deciding if 'doodling with words' can be considered a writing event, or if observing a roadside advertisement can be characterized as a reading event. Finally, the author presents the issue of copying. The example given was

[6] http://www4.ncbi.nlm.gov/PubMed/.

a successful singer or group (like Michael Jackson or the Spice Girls) which attempts to make many people sing their songs. In that case, it is necessary to decide whether each individual singing would be considered a unique language production event (Kilgarriff, 2003).

The topic related to *technology* concerns the decisions that linguists who use the Web need to make in relation to sublanguages[7]. In sublanguages, there are not many ambiguous words and the amount of grammatical structures used is limited. Given this characteristics, 'sublanguage's-specific application development' can be considered simpler when compared to 'general-language application development'. But many of the available resources that developers might use (e.g. WordNet or BNC) are examples of general-language resources. The problem emerged by this fact is that nobody knows yet if these resources can be considered relevant for the creation of sublanguages' applications, if they can be actually be used, or if it is better "to use a language model based on a large general-language corpus or a relatively tiny corpus of the right kind of text" (Kilgarriff, 2003, p. 341). According to the author, the field lacks theory, discussions and mathematical models related to sublanguages.

The third topic approached by Kilgarriff is the use of *language modelling*. Similarly to the matter of sublanguages, the lack of theory related to text types constrains the assessment of the usefulness of language-modelling work, although there are a lot of works focusing on this subject. The fact that statistics of different text types tend to be different implicates in the limitation of language model's application, in general (Kilgarriff, 2003). A language model

> (…) predicts the behavior of language samples of the same text type as the training-data text type (and we can be entirely confident only if training and test samples are random samples from the same source). (Kilgarriff, 2003, p. 339)

Therefore, the problem involving language modelling is that when a language technology application is used in a new text, it is not possible to predict the characteristics of this text type. It is necessary to investigate the efficiency of language models when those are applied to text types which are distinct from the training corpus originally used (Kilgarriff, 2003).

---

[7] A language of restricted domain, used in a particular field, which usually contains distinctive vocabulary.

The fourth topic concerns *language errors*, which are recurrent on the Web. Differently from 'paper-based, copy-edited published texts', the texts published on the Web are written by several and distinct authors who are not always concerned with correctness. In principle, this might be considered a problem, but although the erroneous forms exist, they occur less frequently than the 'correct' forms. A search for 'I beleave' on Google results in 3,910 hits, while the 'correct' form 'I believe' occurs 70,900 times. Thus, despite the fact of being a dirty corpus[8], the Web can still be considered trustable because the correct usage is much more frequent than the undesirable incorrect forms (Kilgarriff, 2003).

Kilgarriff, then, approaches the issue of *sublanguages* from a different perspective. Considering that a language can be defined "as a modest core of lexis, grammar, and constructions, plus a wide array of sublanguages" (2003, p. 342), he questions whether sublanguages should be included in corpus composition. He establishes three possible positions to answer this question. The first possible answer would be "No, none should" (ibid), which is problematic for resulting in an "impoverished view of language" (ibid). The second answer would be "Some, but not all should" (ibid), which is also problematic given its arbitrariness. One example that illustrates this is BNC, which includes cake recipes and research papers on diseases on its data, but sets aside astronomy texts and car manuals. One could easily argue why. Finally, the last possible answer, "Yes, all should" (ibid) seems not to be a viable option, according to the author, although he does not justify this point of view.

The last topic concerns the matter of *representativeness* and the available *literature* on text classification. First, the use of the term itself is questioned. According to Kilgarriff, "the word *representative* has tended to fall out of discussions, to be replaced by the meeker *balanced*" (2003, p. 342). Second, he criticises the extensive literature on text classification. Although he considers this material relevant, the problem is that "it most often starts from a given set of categories and cannot readily be applied to the situation in which the categories are not known in advance" (ibid).

Finally, the author concludes that

[8] The Web can be considered a dirty corpus because it is a place of public domain where texts can be published without proper revision.

> the Web is not representative of anything else. But neither are other corpora, in any well-understood sense. Picking away at the question merely exposes how primitive our understanding of the topic is and leads inexorably to larger and altogether more interesting questions about the nature of language, and how it might be modeled. (Kilgarriff, 2003, p. 343)

He also emphasises the fact that 'text type' is still a very limited area that needs to be investigated. It is one of the issues intrinsically related to the Web, and its use as a corpus requires a better understanding of it.

After presenting a wide prospect of the use of search engines, and showing in practice how efficient they might be at the task that they were originally designed for, Kilgarriff also explains why they are still frustrating when used by linguists. First, "the search engine results do not present enough instances (1,000 or 5,000 maximum)" (2003, p. 345). These numbers might be considered insufficient, depending on the purpose of the research. Second, "they do not present enough context for each instance (Google provides a fragment of around ten words)" (ibid). The lack of context might me problematic for some types of research. If the study involves anaphora, for example, the analysis of the context is crucial. Third, "they are selected according to criteria that are, from a linguistic perspective, distorting (with uses of the search term in titles and headings going to the top of the list and often occupying all the top slots" (ibid). Fourth, "they do not allow searches to be specified according to linguistic criteria such as the citation form for a word, or word class" (ibid). If you are looking for the adjective 'talking', the verb form will also appear on the results. In a corpus annotated with grammatical tags, for instance, this would not be a problem. Finally, "the statistics are unreliable, with frequencies given for 'pages containing x' varying according to search engine load and many other factors" (ibid), which might distort results based on this type of data.

The removal of the constraints listed above could transform the search engines into a wonderful tool for researchers who investigate language phenomena. Search engine designers could easily solve each of these 'limitations', but there is no interest from engine companies in meeting the needs of linguists, since those are not considered a 'powerful lobby'. Thus, the task of solving these limitations remains only in the hands of linguists. The advantage is that the kinds of

querying and processing designed would attend explicitly their needs. There are many possibilities to be explored. Kilgarriff believes that all successful processes which were already applied to smaller corpora could be used in the Web. Instead of focusing only on the search of strings, the searches could be organised by lemmas, noun phrases, or according to grammatical relations. The production of Thesaurus and lexicons of several languages then could be created through the Web (Kilgarriff, 2003).

According to this view, the large amount of text in a huge variety of languages makes the Web 'a fabulous playground' for linguists and deserves to be explored (Kilgarriff, 2003). By linking the existence of this 'fabulous playground' containing large amount of information and the fact that for many languages there are no corpora available yet, in the article *A Corpus Factory for many languages* Kilgarriff describes a method that can be used to build large[9] corpora from the Web. The importance of this can be justified considering that there are many large corpora available for the major languages of the world, while other languages still lack this kind of database.

While traditional corpus collection are known for being long, slow and expensive, with the advent of the internet, enormous quantities of text became available by simply clicking a mouse. The first attempts of using the Web as a corpus happened in the 1990s and the results were promising when compared to the ones obtained from traditional corpora (Kilgarriff, 2010).

The method to build corpora through the Web described by Kilgarriff (2010) is based on the functioning of current commercial search engines which search and index the Web, recognize text-rich pages and deal with character-encoding subjects. According to the author, it is advantageous to utilize the tools offered by search engines because they usually work well and help researchers solve many tasks.

Kilgarriff lists 6 steps involved in corpora collection:

> 1. Gather a 'seed word' list of several hundred mid-frequency words of the language
> 2. Repeat several thousand times (until the corpus is large enough);
>
> - Randomly select three (typically) of these words to create a query

[9] By 'large', he means a corpus containing the minimum of 50 million words (Kilgarriff, 2010).

> - Send the query to a commercial search engine (…) which returns a 'search hits' page.
> - Retrieve pages identified in the search hits page. Store them.
>
> 3. 'Clean' the text, to remove navigation bars, advertisements and other recurring material.
> 4. Remove duplicates
> 5. Tokenise, and, where tools are available, lemmatize and part-of-speech tag
> 6. Load into a corpus query tool. (Kilgarriff, Method Section, ¶ 2)

The first step (seed word selection) is necessary to start every process of corpus collection. Some authors use common words extracted from lists gathered from preexisting corpora (e.g. BNC). Nevertheless, for languages that do not have corpora available, these lists do not exist yet, requiring the researcher to build his/her own seed word list (Kilgarriff, 2010).

If a preexisting corpus is not available, one of the possibilities is to use Wikipedia (a large resource of language containing articles from several domains) to generate the word list. The advantage is that the whole database is available for download. The idea of using Wikipedia as a corpus itself does not seem appropriate given its limited size and diversity; it would be better to use it to generate frequency lists and identify the seed words. The next step is to use the data obtained from the Web through these seed words as the corpus. Another advantage of using Wikipedia is that it contains texts written in 265 languages, allowing researches to apply the same method several times and producing corpora that are likely to be 'comparable' or at least similar to each other (Kilgarriff, 2010).

Wiki corpora are extracted from a Wiki dump of language, which consists of "a single large XML file containing all the articles of the Wikipedia" (2010, Method Section, ¶ 6). Because most of the articles do not contain connected text (they are basically concise definitions or a group of links), the process of filtering is necessary. In order to decide whether a file has connected text or not during this compilation process, Kilgarriff established that the word count for each selected text needed to be over 500.

After extracting Wiki corpora, the next step is to build the frequency lists. In order to do that, the corpus needs to be tokenized, using space and punctuation marks as criteria[10]:

> Once the Wiki corpus is tokenized, term frequency and document frequency are calculated and a frequency list is build. Words are sorted in the frequency list based on document frequency. (Kilgarriff, Method Section, ¶ 4)

The top 1,000 words obtained are treated as the high frequency words of the language, while the next 5,000 are considered the midfrequency ones. The midfrequency words from the frequency word list are the ones used as the seed words (Kilgarriff, 2010).

The next step is query generation. The seeds are used to generate Web queries through a query generation module called BootCaT[11]. This module generates ordered lists "of length *n* by random selection (without replacement) of *n* words" (Kilgarriff, 2010, Method Section, ¶ 10). Once query length was defined, Kilgarriff generated about 30,000 queries for each language (Dutch, Hindi, Indonesian, Norwegian, Swedish, Telugu, Thai, and Vietnamese).

The URL collection was obtained using Yahoo's or Bing's Application Programming Interface (API) in Kilgarriff's corpora compilation, but other search engines (such as Google) can be used depending on the purposes of the corpus. During the process of collection of the URLs, the query, page size and Multipurpose Internet Mail Extensions (MIME) type were stored, just as they were provided by the search engine output (Kilgarriff, 2010).

All the URLs were downloaded using free software called *unix wget*. Only files containing more than 5KB were downloaded, to increase the probability of finding connected texts. In order to filter the html markup and 'boilerplate' text (navigation bars, advertisements, etc.) and only obtain the connected text, Kilgarriff (2010) used the Body Text Extraction algorithm (BTE):

---

[10] For some languages (like Thai and Vietnamese) this is not possible because word delimiters are not present. In those cases it is necessary to use other language-specific tools (Kilgarriff, 2010).

[11] http://bootcat.sslmit.unibo.it/

> BTE starts from the observation that Web pages typically have material at the beginning and end which is rich in boilerplate and which tends to be heavily light in markup. It calculates the ratio of text to markup for different parts of the page, divides the page into three sections on the basis of this ratio, and retains only the middle one. (Kilgarriff, Method Section, ¶ 10)

The pages obtained are further filtered again to guarantee the presence of connected text. One of the characteristics of a connected text is the presence of a high proportion of function words, and pages that do not have this characteristic should be discarded. The author assumed that the top 500 words obtained from the frequency list will probably include most function words (Kilgarriff, 2010).

The final step of Web corpus compilation is the detection of duplicates. Kilgarriff used DeDuper module to detect similar documents based on the text. Duplicate detection can be described as a memory intensive task. The module works as the following:

> N-grams (n=5) for each document are generated and similarity is measured between two documents based on the number of overlaps in their n-grams. Since main memory size is limited and can hold only a limited number of files, duplicate detection is done using a sliding window. At each interaction a fixed number of non-duplicate files, say 500, whose n-grams can fit in memory, are identified using DeDuper module. All other files are taken one file at a time and compared with the n-grams of these non-duplicate files to identify if they are duplicates or not. This process is repeated until all files are covered. (Kilgarriff, Method Section, ¶ 11)

After compilation, all the corpora obtained were loaded into the Sketch Engine and became accessible at the address http://www.sketchengine.co.uk.

When reflecting about the evaluation of the developed corpora, Kilgarriff affirms that in order to state if a corpus can be considered good, first it is necessary to know what one wants to use the corpus for. His straightforward answer to the question “What does it mean for a corpus to be good” is “if it supports us in doing what we want to do”

(2010, Evaluation Section, ¶ 1). He states that this is how the corpora produced should be evaluated by researchers who decide to use them. By using the corpora, they will eventually discover the most appropriate corpus for a specific research purpose. He also reinforces that this kind of evaluation cannot be done immediately, since it is time-consuming and there is no alternative but to wait. The corpora will only be assessed as they are used by linguists in their researches.

Final remarks

Kilgarriff shows that corpus size does not always guarantee that the researcher is going to find the necessary information that he/she is looking for. Therefore, all the discussion involving corpus size and representativeness can be replaced by the simple question proposed by Sardinha (2004): The corpus is "representative of what and for whom?". The corpus used in the present study is definitely small; it contains less than 80,000 words. In terms of representativeness, the small amount of data could be considered insufficient, especially if compared to a corpus containing 10 million words. A large corpus can be considered representative of a language, but it does not mean that it is adequate to any kind of linguistic investigation. If someone is interested in Shakespearean language, "the collection of all the works written by Shakespeare would be a representative corpus of this author" (Sardinha, 2004, p. 27). Therefore, the corpus compiled for this study can be considered representative, but just like any other corpora it has its limits. It can help answer few types of questions, but still, it is appropriate to the interests of the researcher and the issues related to the investigation, which demanded a specific subcorpus of Shakespeare's work.

Kilgarriff's initiative of building a corpus factory probably represents a new chapter in the history of corpus linguistics. Not only because corpora of language that were not available so far will be available to linguists who want to investigate them, but also because it allows autonomy to those who want to build their own datasets. The corpus used in this investigation was not compiled by the Web, but since this practice represents a new tendency in corpus linguistics and it has been widely discussed, it seemed relevant to present it in this review of literature.

2.2 Corpora in Translation Studies

Translation Studies is defined by Baker as "the academic discipline concerned with the study of translation at large, including literary and nonliterary translation, various forms of oral interpreting, as well as dubbing and subtitling" (1998, p. 277). The first attempt to structure the field of Translation Studies in a map is attributed to Holmes (as cited in Baker, 1998). According to Holmes' definition, the present investigation can be characterised as a pure, descriptive product-oriented study. It is pure because it "has the objective of describing translation phenomena as they occur, and developing principles for describing and explaining such phenomena" and descriptive product-oriented because it is a text-focused study which attempts to describe existing translations (ibid).

According to Baker (1998), the area of Translation Studies is still trying to establish itself as a discipline, notwithstanding it also maintains an interchange with other disciplines and with distinct theoretical perspectives from which translation can be approached. Baker affirms that "the study of translation has gone far beyond the confines of any other discipline and it has become clear that research requirements in this area cannot be catered for by any existing field of study" (ibid, p. 279). For this reason, "various methodologies and theoretical frameworks borrowed from different disciplines are increasingly being adapted and reassessed to meet the specific needs of translation scholars" (ibid).

Baker (1995, p. 224) also points out that "the potential for using corpora is beginning to take shape in translation studies". Historically speaking, "Corpus Linguistics was originally centered on monolingual corpora" (Anderman & Rogers, 2008, p. 14), but the several publications on corpora and translations increased the interest for parallel corpora[12] in the area of Descriptive Translation Studies. Some of the advantages related to the use of this type of corpora, especially the multilingual ones, are presented by Johansson:

> (…), through corpora we can observe patterns in language which we were unaware of before or only vaguely glimpsed. My claim is that this applies particularly to multilingual corpora. We

[12] "'Parallel corpora'- source texts and their translations" (Anderman & Rogers, 2008, p. 14).

> can see more clearly what individual languages are alike, what they share and – perhaps eventually – what characterizes language in general. Seeing through corpora we can see through language. (2007, p. 1)

According to Johansson (ibid) the development and use of multilingual or parallel corpora occurred in the last 10-15 years. He defines those corpora as "collections of texts in two or more languages which are parallel in some way, either by being in a translation relationship or by being comparable in other respects, such as genre, time of publication, intended readership, and so on" (ibid).

The first example of parallel text is the Rosetta Stone, which was discovered in a small town near Alexandria. The stone contains inscriptions in Egyptian hieroglyphs, demotic script[13] and Greek, and a comparison of these texts contributed to the deciphering of the hieroglyphs (Johansson, 2007). Nowadays it would be unacceptable to conceive a parallel corpus which was not digitalized. The actual corpora used in corpus linguistics are always machine-readable, since the advent of computers allowed researchers to compile increased volume of texts which are used both in translation studies and in comparative language studies.

Johansson (2007) claims that there are two types of multilingual corpora: translation corpora and comparable corpora. The former consist of "original texts and their translations into two or more other languages" (ibid, p. 52); and the latter consist of "original texts in two or more languages matched by criteria such as genre, time of publication" (ibid, p. 53). In current researches, parallel texts in diverse languages have been used in translation studies and in comparative language studies. Through analytical comparisons it is possible to determine characteristics of languages and also gain a deeper insight into their specific features (ibid).

According to McEnery and Xiao (2008, p. 21), "parallel and comparable corpora are used primarily for translation and contrastive studies". Furthermore, both types present advantages and disadvantages, and thus serve for different purposes in the field of Translation Studies. Source and translated texts in a parallel corpus are considered useful for

[13] Egyptian hieroglyphic writing of cursive form that was used in handwritten texts from the early 7th century bce until the 5th century ce.
(http://www.britannica.com/EBchecked/topic/157464/demotic-script).

exploring "how the same content is expressed in two languages" (ibid), but alone they are considered poor to be used as basis for cross-linguistic contrasts because of the effect of translationese[14]. Comparable corpus, on the other hand, overcome the issue of translationese but are less useful for the study of how a message is conveyed from one language to another. Johansson states that parallel and comparable corpora "can be combined within the same overall framework, as has been done with the *English-Norwegian Parallel Corpus (ENPC)*" (2007, p. 53). In summary, "comparable corpora are useful resource for contrastive studies and translation studies when used in combination with parallel corpora". McEnery and Xiao also reinforce that "comparable corpora can be a poor basis for contrastive studies if the sampling frames for the comparable corpora are not fully comparable" (ibid, p. 22).

The corpus compiled for this investigation is a parallel corpus. Although its observation might show how the sonnets are expressed in two languages, this is not the main objective of the research. The objective is to evaluate Thune's model of complexity when applied to translated poetry. A comparable corpus was not used because the investigation is not characterized as a contrastive study between the two languages.

2.2.1 The linguist and the translator

In this section, I present how corpora were introduced into the area of Translation Studies. It is divided into two distinct parts. In the first part there is an overview of how the relation between the linguist and the translator was built throughout history. In the second section I discuss the differences between parallel and comparable corpora in more detail.

2.2.2 Incorporating Corpora: the linguist and the translator

[14] "Translations may reflect features of the source language, a phenomenon which has been given the label translationese" (Johansson, 2007, p. 11).

In 1956, after reading a paper entitled 'Linguistics and Translation' to an audience at Birkbeck College, J. R. Firth concluded that the need for mutual translation from and into languages such as English, Russian, Arabic, and Chinese was emerging as an immediate consequence of the spread of these languages across the world. Furthermore, there was also the translation of other languages which were part of what was called 'a common world civilization' at that time (as cited in Anderman & Rogers, 2008).

By saying these words, Firth clearly predicted the need for translation that was about to arise as a consequence of the European Economic Community's creation (EEC) in 1957, shortly after his lecture. He also anticipated the consolidation of English as a global language followed by Chinese, Spanish, and Hindi, which later would become the most frequently spoken languages in the world (Anderman & Rogers, 2008).

> As a pioneer of the new discipline of linguistics in the UK, Firth's insight into the nature of language not only led him to predict an increased need for translation, it also made him an early advocate of the study of meaning in linguistics. At a time when American structuralist linguists were attempting to exclude meaning from linguistic analysis along with all psychological, or as Bloomfield called it 'mentalistic references', Firth clearly realized the importance of the task of incorporating linguistic meaning into the science of language. And as his definition of meaning as 'function in context' suggests, he was well aware of the importance of running text of the kind that computers are now able to process. In looking at words in their context, he was not, however, the first linguist to understand that – in isolation – separate lexical items are less likely to reveal to us their actual meaning. (Anderman & Rogers, 2008, p. 5)

The concept of context is also related to the earlier developments in foreign language teaching, anticipating what would be later called the 'communicative turn' at the end of the 20th century. Long before Firth, Henry Sweet pointed out the importance of the use of connected texts rather than isolated sentences in the study of spoken English because "it is only in connected texts that the language itself can be given with each

word in a natural and adequate context" (qtd. in Anderman & Rogers, 2008, p. 6).

The linguist Otto Jespersen was also aware of "the importance of not viewing words and constructions in isolation" (Anderman & Rogers, 2008, p.6). In his work entitled *A Modern English Grammar on Historical Principles*, he explains facts related to English usage during different periods of its history. In order to support his discussion, Jespersen used examples extracted from a data source[15] made up of the English Canon and other sources to "place grammatical phenomena in a true light (qtd. in Anderman & Rogers, 2008, p.6). Nowadays, this difficult and exhaustive task performed by Jespersen can be achieved in machine-readable corpus studies by using automatic grammatical tagging of words. If Jespersen had access to an annotated corpus, he would be able to observe nouns, adjectives, adverbs from a different perspective and to have a more realistic view of how they 'behave' in context, besides using statistic tools that would probably help him to explain some grammatical phenomena.

The notion of context was also an important object of study for the social anthropologist Bronislaw Malinowski:

> For Malinowski, the notion of translation into English was crucial in his anthropological studies and was extended to include the definition of a term by ethnographic analysis, that is, by placing it within its context of situation and its context of culture, 'putting it within the set of kindred and cognate expressions, by contrasting it with its opposites, by grammatical analysis and above all by a number of well chosen examples […] the only correct way of defining the linguistic and cultural character of a word'. (Anderman & Rogers, 2008, p. 6)

Malinowski was not only aware of the importance of context in the translation process, but also anticipated the problems usually expressed by 21st-century translators in relation to the dictionaries. He realised that the growing of the fast-moving knowledge society demanded from translators contextual solutions to problems involving

[15] We call it "data source" because the term corpus did not exist at that time. But his procedure of collecting data can be considered the first attempt of corpus compilation.

terminology and phraseology. As a consequence of this current age of fast moving information, traditional dictionaries and electronic term bases, which usually do not contextualize meaning and use, became obsolete. As an alternative, translators started to consult on-line documentation or customized electronic corpora when dealing with the mentioned problems (Anderman & Rogers, 2008).

John Firth was one of the first linguists to recognise the importance of translation in the 20th century. He pointed out to the existence of four distinct types of translation. The first type was defined as 'creative translation', which includes the translation of literature. The second type is defined as 'official translation' and refers to language transfer used in treaties and documents, also known as 'controlled' or 'restricted languages', usually treated as specialist translation. The third type would be the translations used by linguists to describe a particular language. The fourth type is known as 'mechanical translation'[16] (Anderman & Rogers, 2008).

The translations analysed in this study fit into the first category, creative translation, since they are "intended primarily as literature in the language into which it is rendered by the translator" (Anderman & Rogers, 2008, p.7).

2.2.3 Parallel and Comparable Corpora: What is Happening?

After explaining how the relation between the linguist and the translator was built in Section 2.1, I will now explore the differences between parallel and comparable corpora. As a starting point, the increase of international exchange and the process of globalisation led to the popularisation of translation and contrastive studies. As a consequence, the use of corpora and multilingual corpora started to play an important role in this field of study (McEnery & Xiao, 2008).

There has been a considerable acceleration on the development of corpus linguistics from the 1980 onwards. It is true that the construction and exploitation of corpora in English still predominate in the area of corpus linguistics, but corpora of several other languages like Chinese, Japanese, and Korean have been developed, contributing significantly to

[16] Mechanical translation is a term that refers to the first attempt of developing devices for mechanizing translation, which started with the creation of mechanical dictionaries.

the diversification of corpus-based language studies. Besides monolingual corpora, parallel and comparable corpora have been occupying the core of non-English corpus linguistics, as a consequence of the importance of these two last types for translation and contrastive studies (McEnery & Xiao, 2008).

Among the specific uses and possibilities enabled by parallel and comparable corpora, McEnery and Xiao (2008) present four aspects related to contrastive and translation studies: First, "they give new insights into the languages compared – insights that are not likely to be gained via the study of monolingual corpora". Second, "they can be used for a range of comparative purposes and increase our knowledge of language-specific, typological and cultural differences, as well as universal features". Third, "they illuminate differences between source texts and translations, and between native and nonnative texts". Finally, "they can be used for a number of practical applications, e.g. in lexicography, language teaching and translation" (Aijmer & Altenberg, 1996, qtd. in McEnery & Xiao, 2008, p.18).

A corpus that contains more than one language is usually characterised as a multilingual corpus, if we use the term in a broad sense. However, by definition, a multilingual corpus should contain at least three distinct languages, while those which contain two languages should be called *bilingual corpora*. This problem of terminology can be explained by the fact that using corpora containing more than one language in research is something that started in the early 1990s; it is a recent phenomenon (McEnery & Xiao, 2008, p.19). McEnery and Xiao point out three types of corpora involving more than one language:

> (1) Type A: Source text plus translations, e.g. *Canadian Hansard* (cf. Brown *et al.*, 1991), CRATER (cf. McEnery & Oakes, 1995).
> (2) Type B: Monolingual subcorpora designed using the same sampling frame, e.g. *The Aarhus corpus of contract law* (cf. Faber & Lauridsen, 1991).
> (3) A combination of A and B, e.g. the *ENPC* (cf. Johansson & Hofland, 1994), the *EMILLE.* (2008, p. 19)

According to McEnery and Xiao (ibid), there is a variety of terms used to describe different types of corpora:

> For Aijmer and Altenberg (1996) and Granger (1996: 38), type A is a translation corpus whereas type B is a parallel corpus; for McEnery & Wilson (1996: 57), Baker (1993: 248; 1995; 1999) and Hunstin (2002: 15), type A is a parallel corpus whereas type B is a comparable corpus; and for Johansson & Hofland (1994) and Johansson (1998: 4) the term parallel corpus applies to both types A and B. Barlow (1995; 2000: 110) certainly interpreted a parallel corpus as type A when he developed the *ParaConc* corpus tool. It is clear that some confusion centers on the term *parallel*.

The main point stated by the authors is that different criteria such as content, form, or number of languages can be used during the process of defining distinct types of corpora. Once a criterion is chosen, it should be used in a consistent way. If one's criterion for definition is the number of languages involved, then one can call his/her corpus monolingual, bilingual, or multilingual. If one decides for content as a criterion, the corpus can be called a translation (L2) or a nontranslation (L1) corpus. But when the criterion chosen is corpus form, it is necessary to do it very consciously. Therefore, a corpus can be considered parallel "if it "contains source texts and translations in parallel", or comparable "if its subcorpora are comparable by applying the same sampling frame" (ibid, p. 19). It would be illogical to consider type A a translation corpora by the criterion of content, as well as type B a comparable if the criterion used to define it was form.

By definition, a parallel corpus contains source texts and their translations and it can be bilingual or multilingual. They can also be classified as unidirection or bidirection corpora. An example of unidirection corpus could be one that contains texts from English to Italian, or from Italian to English alone, while a bidirection would contain both English source texts and their English translations. A third type would be a multidirection, where the same piece of work is translated into English, French and German, for example. On the other hand, a comparable corpus necessarily contains components collected according to the same sampling frame and representativeness (McEnery & Xiao, 2008). This would include "the *same proportions* of texts of the *same genres* in the *same domains* in a range of *different languages* in the *same sampling period*" (ibid, p. 20).

Another important aspect regarding the use of parallel and comparable corpora is that they are supposed to be used for different purposes. Parallel corpora should be used in translation studies, while comparable corpora fit contrastive studies. Besides this difference in relation to use, the two types of corpora are also designed with different purposes. Many aspects need to be considered during the compilation of a comparable corpus, starting by the use of an appropriate sampling frame. The components which are supposed to represent the languages involved must be correspondent in relation to genre, domain, proportion, and sampling period. It is not an easy task to fulfill all these requirements. For a parallel corpus, on the other hand, the sampling frame should not be a problem, since the corpus components already are precise translations of each other. In this specific case, the sampling frame is used only to select the source texts. But it does not mean that building a parallel corpus is a simple task. In order to be useful, the parallel text need to be aligned, the source texts and their respective translations need to be put together. In other words, there might be a link between them at some level. So far, the automatic alignment of parallel corpora is still a complex task for some language pairs, and many alignment processes need to be done manually (Piao, 2000, 2002, qtd. in McEnery & Xiao, 2008).

In addition, McEnery and Xiao (2008) point out that depending on the specific research question, the researcher will need to choose a *specialized* or a *general* corpus. The first type would contain texts of a particular type, like Shakespeare's Sonnets, for example. The second type of corpus should be balanced, containing as many text types as possible. This kind of corpus would be used to build school dictionary entries, for example. According to McEnery and Xiao (ibid), these corpora can be of either type: In terms of terminology extraction, for example, both corpora can be considered useful, but if the interest focuses on contrast of certain general linguistic features, like tense and aspect, balanced corpora would be more recommended because in general they are supposed to be more representative of the language chosen. Parallel corpora tend to be specialized (McEnery & Xiao, 2008). Thunes' parallel corpus consists of law and fiction texts, while the corpus used in this investigation comprises Shakespeare's lyrical poetry. They are examples of specialized corpora. According to McEnery and Xiao, this level of specialization

> (…) is quite natural, considering the availability of translated texts by genre (in machine-readable

> form) in different languages, and indeed, specialized parallel corpora can be especially useful in domain-specific translation research. While most of the existing comparable corpora are also specialized, it is relatively easier to find comparable text types in different languages. Therefore, in relation to parallel corpora, it is more likely for comparable corpora to be designed as general balanced corpora. (McEnery & Xiao, 2008, p. 21)

In sum, the authors emphasise that both comparable and parallel corpora "have their own advantages and disadvantages, and thus serve for different purposes" (2008, p. 21). In order to illustrate that, source and translated texts in a parallel corpus could be used by someone willing to explore "how the same content is expressed in two languages" (Aijmer & Altenberg, 1996, qtd. in McEnery & Xiao, 2008, p. 21). On the other hand, "comparable corpora are a useful resource for contrastive studies and translation studies when used in combination with parallel corpora" (McEnery & Xiao, 2008, p. 22).

But what would be the real value of parallel and comparable corpora to translation and contrastive studies? Authors like Laviosa state that theoretical elaborations and empirical studies involving corpus based approaches are transforming the field into "a coherent, composite and rich paradigm that addresses a variety of issues pertaining to theory, description, and the practice of translation" (1998, qtd. McEnery & Xiao, 2008, p. 22). Corpus-based translation studies can be divided into two broad areas, theoretical and practical:

> Corpus-based translation studies come in two broad areas: theoretical and practical (Hunston, 2002: 123). In theoretical terms, corpora are used mainly to study the translation process by exploring how an idea in one language is conveyed in another language and by comparing the linguistic features and their frequencies in translated L2 texts and comparable L1 texts. In the practical approach, corpora provide a workbench for training translators and a basis for developing applications like MT and computer-assisted translation (CAT) systems. (McEnery & Xiao, 2008, p. 22)

The relevance of this discussion proposed by McEnery and Xiao (2008) is that it clarifies the misunderstanding involving the terminology related to multilingual corpora and reinforces the importance of choosing consistent criteria to define these types of corpora. Thus, parallel corpora "refer to those that contain collections of L1 texts and their translations while comparable corpora refer to those that contain matched L1 samples from different languages" (ibid, p. 27).

It also states that "while parallel corpora are well suited to research and teaching in translation studies, they provide a poor basis for cross-linguistic contrast if used as the sole source of data" (McEnery & Xiao, 2008, p. 27). On the other hand, "comparable corpora used alone are less useful for translation studies" but "they certainly serve as a reliable basis for contrastive studies" (ibid).

In this study, I will follow McEnery and Xiao's and Baker's terminology, which means that the corpus used in the investigation is classified as type A, a bilingual parallel corpora. The parallel corpus was aligned manually and sometimes the alignment process was not easy, reinforcing McEnery and Xiao's statement about the difficulties involved during this task. The link between the source text and their translations was done at verse level, which seemed to be the best option as a starting point. This issue will be discussed further in the Method Chapter. As mentioned in the previous section, the goal of the study is not to establish a contrastive analysis between Portuguese and English, since the type of corpus chosen suits the area of translation studies, which is consistent with McEnery and Xiao's point of view.

2.3 Corpus Annotation

After defining what a corpus is, explaining the different types of corpora and tracing an overview of their applications in Translation Studies, I will now explore the process of annotation, which constitutes the basis of the methodology developed in this study, based on the ideas proposed by Leech (1997).

The year of 1961, when the first manned space flight was launched was also the date of birth of computer corpus linguistics. The first electronic corpus (Brown Corpus) was developed by Brown University and contained just over one million words. About thirty years

later, the Brown Corpus would be considered small if compared to the 100-million-word British National Corpus (BNC) (Leech, 1997).

However, people noticed that the value of a corpus should not be measured only in terms of size, and other aspects such as the diversity of the corpus and annotation were important criteria which added value to it. Researchers soon realized that corpus annotation would bring crucial contributions and also add value to the existing corpora. The annotation process "enriches the corpus as a source of linguistic information for future research and development" (Leech, 1997, p. 2).

But what is corpus annotation at all? According to Leech, annotation is

> (…) the practice of adding **interpretative**, **linguistic** information to an electronic corpus of spoken and / or written language data. Annotation' can also refer to the end-product of this process: the linguistic symbols which are attached to, linked with, or interspersed with the electronic **representation** of the language material itself. (1997, p.2)

Grammatical tagging[17] is pointed out by the author as a typical example of corpus annotation. In this type of annotation, a label or tag is attached to a word to indicate its grammatical class. "For example, in *taken*_VVN, the grammatical tag VVN indicates that *taken* is a past participle" (Leech, 1997, p.2).

According to the author, interpretive annotation of this kind necessarily involves "the product of the human mind's understanding of the text" (Leech, 1997, p. 2). In that case it is difficult to establish clear and objective criteria in order to analyse a certain linguistic phenomenon. On the other hand, someone would not disagree with the label VVN (past participle) attached to *taken*, since this is a grammatical class which belongs to English and it is not a controversial category in terms of grammar.

However, Leech affirms that in some cases the classification would not be so simple. If we take the lexical item *future* from the expression *future bride*, it is necessary to decide whether the word is a noun or an adjective. If you consider *future* an adjective, should it be

[17] Also called word-class tagging, part-of-speech tagging or POS tagging (Leech, 1997, p.2).

considered an adjective or a particular subclass that must occur in a pre-nominal position? These are decisions that have to be taken during the process of corpus annotation.

But why should we annotate a corpus? The first justification presented by Leech is that "corpora are useful only if we can extract knowledge from them" (1997, p. 4). In order to extract information from them, first we need to add information to it, and one way of doing this is through annotations:

> The 'raw corpus' in its orthographic form contains no direct information, for example, about grammar – and this can hinder many of the applications to which a corpus can be put. Consider the word spelt *left.* As a word meaning the opposite of *right*, it can be an adjective ('my *left* hand'), an adverb ('turn *left*') or a noun ('on your *left*'). As the past tense or past participle of *leave*, it is a verb ('I *left* early'). *Left* is therefore a very versatile piece of language – but its various meanings and uses cannot be detected from its orthographic form. (Leech, 1997, p. 4)

The example above illustrates how the use of annotated corpora can be useful for making dictionaries. If the corpus is successfully grammatically tagged, it will be possible to know the word-class of all the occurrences of the lexical item *left*.

Leech's second justification is the idea of reusability. One might argue that in order to extract the information presented above it would not be necessary to go through the exhaustive process of annotation. A simple program that could recognize that *left* following a verb is an adverb, while *left* preceding a noun is an adjective would solve the problem. Yet such program would have to recognize the word-class of neighbouring words to complete the task. Therefore, the identification of word-classes is an essential tool in this kind of processing. Finally, once a corpus is annotated, it becomes a more valuable resource (if compared to the raw corpus) because it becomes available to other users that can use it for different purposes. The idea of reusability somehow justifies the annotation of any corpora, since it is an expensive and time-consuming activity.

The third justification presented by Leech refers to the multifunctionality of an annotated corpus. "(…) annotation gives 'added value' to a corpus in the general sense: it adds overt linguistic

information, which can then be used for a multitude of purposes" (ibid, p. 5).

Leech describes six important issues related to annotation that need to be considered. First, the reversion to the raw corpus needs to be possible and easy, that is, the annotations should be easily removed if the user desires it. Recovering the raw corpus should be a simple task. Second, the possibility of storing the annotations independently should be available if there is a need for that. Third, the user needs to have access to specific documentation containing information about the *annotation scheme*, *where*, *how*, and *by whom* the annotations were applied and also information about the quality of the annotation process. Some examples of annotation quality include to what extent the corpus was checked, information about the percentage of annotations that were considered correct, and the consistency of its application. Fourth, one should bear in mind that there is no annotation scheme that will represent 'God's truth'. Leech states that there is no guarantee in relation to its application, they are only offered for practical reasons. Many users might prefer to use a corpus which is already annotated instead of doing the entire job from scratch, a task which could last years. Fifth, when possible, annotation schemes should be based on theory-neutral or on consensual analyses of the data. This would be necessary to avoid possible misunderstandings and misapplications. For example, it would be safer to use structural or classificatory information provided by dictionaries. Since they offer information based on general descriptive traditions, it does not need to be justified. Besides the need to cope with sensitive decisions, annotators should adopt widely accepted and understandable annotation schemes. Even within this 'accepted' annotations the annotator will face problems to be solved. Finally, annotation schemes cannot require authority as an absolute standard. There are good reasons for the existing variation between them. One example is related to the size of corpus to be annotated, which might be incompatible with too much detail. Depending on the purposes of the annotation, specific types of information might be prioritized. Leech affirms that other aspects that might also lead to differences in the choice of annotations are the corpus type or the identity of the language chosen.

The first three standards for corpus annotation proposed by Leech are applied to this study. The annotations were placed by each verse's side, and it would not be difficult to remove them in order to recover the raw corpus. The review of Thunes' model offers all the information necessary to the understanding of the annotation scheme. Chapter 3

explains how the annotations were applied. It will not be necessary to observe to what extent the corpus has been checked because the process of annotation itself was manual. Finally, the discussion of results approaches the quality of the annotation process and all the difficulties faced during it.

In relation to the other items, it is expected that the final product can be used by a translation research community to bring light into some decisions related to poetry translation processes. The annotation scheme has been developed recently and is not based on consensual or theory-neutral analyses if compared to the information offered by dictionaries, but this might not be a problem because Thunes' model is based on clear and specific criteria, which might avoid disagreements in relation to its application.

2.4 Studies on Corpora Annotation

Previous studies which inspired the approach used in this research include five works that discuss some of the issues proposed by the present investigation and which are intrinsically related to the process of adding linguistic information to electronic corpora.

The first article related to the issue proposed by this investigation is entitled "Tagging the Bard: Evaluating the Accuracy of a Modern POS Tagger on Early Modern English Corpora" (Rayson et al., 2007). This article focuses on automatic part-of-speech (POS) annotation of historical English texts. The authors applied some techniques which were originally created to the annotation of modern English in Early Modern English (EModE) datasets. The corpus used in the experiment contained files from the Lampeter Corpus (tracts and pamphlets from the 17th and 18th centuries) and 5 comedies by Shakespeare: *The Taming of the Shrew*, *Love's Labour's lost*, *The Merry Wives of Windsor*, *Twelfth Night* and *The Tempest* (all the plays were taken from the First Folio printed in 1623).

The researchers wanted to measure the accuracy of the CLAWS Part-of-Speech (POS) tagger[18]. Their objective was to check how a

[18] CLAWS is an Automatic Word-Tagging System. "Part-of-speech (POS) tagging is the most common form of corpus annotation: grammatical annotation can be useful in situations where you want to distinguish the grammatical functions of particular word forms or identify all the

tagger trained on modern English would work when applied to an EModE corpus. In other words, they wanted to check if the grammatical tagger would be able to distinguish the grammatical functions of particular word forms in the EModE corpus. This study is a typical example of grammatical tagging. They concluded that in Shakespeare's data, there was a significant reduction in POS tagging accuracy from 96% to 82%, as a consequence of spelling variants in Early Modern English.

In the second article, "Features in Translated Brazilian Portuguese Texts: a Corpus-Based Research" (2002) written by Célia Magalhães and Maria da Conceição Batista (UFMG)[19], the authors analysed the presence of two types of universals of translation (simplification and explicitation) in two translations of the novel *Frankenstein* by Mary Shelley into Brazilian Portuguese. The authors concluded that there was a tendency to simplification in the translated texts, which could be justified by the average sentence length in both translations: The sentence length in the translations was shorter than in the original texts. In relation to explicitation, the authors concluded that both translations were shorter than the original, contradicting the common idea that translations are usually longer than their originals.

Madan M. Sarma also investigated the features of universals of translation in the article "Translating Shakespeare: Intervention and Universals in Translation" (2008). The corpus used in this study included three Assamese translations of Shakespeare's play *Hamlet* by three different translators, an Assamese translation of *Macbeth* by the author (Sarma) himself and other four translations of contemporary authors. According to the author, in total the corpus contained around a hundred thousand words. Part of the investigation consisted in the identification of some features which are considered to be universal in translated texts. His analysis pointed towards the presence of simplification and explicitation in the translated texts, but he concludes the article recognizing that "it is not easy to avoid some kind of overlapping in the use of terms like normalization, explicitation and simplification" (ibid, p. 85).

Another example of investigation involving annotation in a corpus consisting of Shakespeare's texts is presented in the article "Love, - 'a familiar or a devil'? An Exploration of Key Domains in

---

words performing a particular grammatical function" (Rayson et al., 2007, p. 3). More information can be found at: http://ucrel.lancs.ac.uk/claws/.

[19] Published in the periodical *Cadernos de Tradução n° IX* (2002).

Shakespeare's Comedies and Tragedies" (Archer et al., 2006). Differently from Rayson and colleague's article (2007), which involved grammatical tagging, this study is an example of semantic annotation. Semantic annotation requires the attachment of specific labels which indicate the semantic fields of all the words in a text. This paper shows how the UCREL Semantic Annotation Scheme (USAS)[20] can be used to determine "the semantic relationships between keywords via an investigation of key domains" and to provide "empirical support for some of the love-related conceptual metaphors put forward by cognitive metaphor theorists" (Archer et al., 2006, p. 1).

Since *love* is a common theme in Shakespeare's work, the authors decided to explore the examples of this concept taken from the *Nameless Shakespeare Corpus* (hosted by Northwestern University). The authors describe their approach as top-down, which means that "the categories are predefined and applied automatically by the USAS system" (Archer et al., 2006, p. 2). During the study they reported "on an exploration of key domains within three Shakespearean *love*-comedies and three Shakespearean *love*-tragedies" (ibid, p.13) and concluded that there were marked differences in the occurrence of *love* in the two datasets. The semantic fields of *intimate/sexual relationship* and *liking* are more common in the *love*-comedies, while the *love*-tragedies focus on issues like *war, lack of life/living things, religion and the supernatural.*

Finally, I would like to mention the investigation entitled "Desambiguação do Item Lexical Correto Através de Etiquetadores Semânticos: uma Abordagem Baseada em Corpus" (Azevedo, 2007). The objective of this study was to investigate the possible senses of the polysemic lexical item "*correto*" in a corpus of written Portuguese. In total, 956 occurrences of the lexical item were extracted from corpus NILC[21] (Núcleo Institucional de Linguística Computacional) by using the software WordSmith Tools (Scott, 1995). Each occurrence of this polysemic word was classified according to a previous set of seven senses taken from Portuguese Thesaurus called *Base de Dados TeP*. When the occurrence assumed a sense that would not fit any of the senses provided by the Thesaurus, a new category was created and

[20] UCREL Semantic Annotation Scheme: a software program for automatic, dictionary-based content analysis. The software traces the semantic relationships between keywords via an investigation of key domains.

[21] A corpus which contains about 35 million words.

considered during the analysis from that point on. At the end of the analysis, other five senses were created to cope with the annotation of all the occurrences. The author expected that the results could be used as a subsidy to linguistically support a system that would be able to carry the automatic disambiguation of the lexical item. The distinct senses identified in the corpus and the respective descriptions of co-occurrence patterns of "*correto*" with other words could be transformed into a set of tags to be used in a semantic tagger within a system which uses linguistic-analysis technology (Azevedo, 2007).

As observed, it is not unusual for researchers to include Shakespeare's plays in studies on corpora annotation, but the sonnets are not included in such corpora, to the best of my knowledge. Most studies involving poetry are centred on the field of Literary Translation. Since poetry translation is generally considered the most difficult and demanding form of translation (Baker, 1998), the analysis of complexity (as suggested by Thunes) on this type of text will probably bring contributions to scholars who desire to approach it from a different point of view.

2.5 Universals of Translation

Linguistic studies have demonstrated that there are certain rules which pertain to particular languages, but there are also rules which seem to be common to all human languages. The set of rules which represent the universal properties of all languages constitute a Universal Grammar (UG), a concept which emerged as a consequence of linguists' attempts to discover the "laws" of particular languages (Fromkin et. al, 2007). Although the concept of UG is generally attributed to Noam Chomsky, there is clear evidence that the idea of a Universal Grammar was considered by other scholars before him. The term *general grammar* was used by a German philosopher called Alsted in about 1630 to make reference to features which were common to all languages. Even earlier, approximately three centuries before Alsted, the scholar Robert Kilwardbly highlighted that "linguists should be concerned with discovering the nature of language in general" (ibid, p. 17). According to Chomskyian view, "there is a Universal Grammar (UG) that is part of the human biologically endowed language faculty. We can think of UG as a system of rules and principles that characterize

all grammars. The rules of UG provide the basic blue print that all languages follow" (ibid, p. 18).

The search for structural features which are common to all languages, therefore, culminates in the determination of the Universals of Language, a research paradigm developed by Chomsky in the 1950s. His generative theory of language "proposes a single set of rules from which all the grammatical sentences in a language can be derived" (Crystal, 1995, p. 84). According to Chomsky's view, linguistics should not be restricted to the study of individual languages, they should go beyond the investigation of individual languages and attempt to describe a 'universal grammar' that could be able to account for all kinds of linguistic variation that are "humanly possible" (Crystal, 1995).

Possibly as a consequence of the concept of Universals of Language systematised by Chomsky, the concept of universals of translation emerged within the area of translation studies. By definition, "universals of translation are linguistic features which typically occur in translated rather than original texts and are thought to be independent of the influence of the specific language pairs involved in the process of translation" (Baker, as cited in Laviosa, 1998, p. 288). According to Laviosa (1998), contrastive analysis of translations and their respective source texts pointed to the existence of some features that are considered to be present in all kinds of translated texts, independent of the languages involved in the translation process. These features are called universals of translation and "concern simplification, avoidance of repetitions present in the source text, explicitation, normalization, discourse transfer, and distinctive distribution of lexical items" (ibid, p. 288).

Simplification is "the tendency to simplify the language used in translation" (Baker, 1996, p. 176). Laviosa (1998) presents three different types of simplification which have been identified in translated text: lexical, syntactic and stylistic. Lexical simplification is "the process and/or result of making do with *less* words" (ibid, p. 288). Syntactic simplification occurs when "complex syntax is simplified by replacing nonfinite clauses with finite ones and by suppressing suspended periods" (ibid). With regard to stylistic simplification, evidence of this feature could be the "tendency to break up long sequences and sentences, replacing elaborate phraseology with shorter collocations, reducing or omitting repetitions and redundant information, shortening overlong circumlocutions and leaving out modifying phrases and words" (ibid, p. 289).

The term explicitation is defined by Baker as "(…) the overall tendency to spell things out rather than leave them implicit in translation" (1996, p. 176). That means that the language used in translations tends to make the information explicit, even if in the source text this information was originally implicit. This can be observed through the use or overuse of explanatory vocabulary and conjunctions. According to Laviosa, this feature was initially observed by Blum-Kulka (1986), who noticed that "shifts occur in the types of cohesion markers used in the target texts and records instances where the translator expands the target text by inserting additional words" (1998, p. 289).

Normalisation is characterised by Vanderauwera (1985) as a 'general tendency towards textual conventionality' (as cited in Laviosa, 1998, p. 289). Some examples are the standardisations of unusual punctuation, to complete sentences that are left unfinished in the source text, and to substitute idiosyncratic sentence structures by simpler structures. Besides that, structures like chapters, narrative sequences, paragraphs, and sentences are organized in a logical way. Adjustments are usually made in order to represent spoken language according to the rules of written prose; contrastively, formal dialogues are represented as intimate and colloquial conversations. The replacement of old-fashioned expressions by modern ones and the process of rewriting experimental narratives in a more familiar mode are also examples of normalisation (ibid).

Discourse transfer can be defined as the translator's tendency "to produce a translated utterance not by retrieving the target language via their own linguistic knowledge, but directly from the source utterance itself" (Baker, 1998, p. 290-291). Distinctive distribution of lexical items refers to the fact that some words are more frequent in source texts than in their respective translations (Baker, 1998).

The search for universals of translation has been object of study of many scholars, but this concept is considered by many authors a controversial issue. In the article "Beyond Intervention: Universals in Translation" (2008), House reflects on several suggestions of universals in language and universals of translation. According to her, translation universals are "universal tendencies of the translation process, laws of translation and norms of translation" (ibid, p. 10) which have been proposed by authors like Blum-Kulka (1986), Baker (1993), Laviosa (1998) and Toury (2001). House (ibid) mentions that while Blum-Kulka and Toury grounded their research on case studies and exhausting qualitative work, combining informed intuition and paper analysis,

many other researchers, however, "relied on, and copiously praised the methodological advantages of corpus-based qualitative and quantitative work" (ibid).

House (2008) suggests that "the quest for specific translation universals is in essence futile, i.e. that there are no, and there can be no, translation universals" (ibid, p. 11). In order to support her point of view, she highlights five reasons. First, she believes that there are not universals of translation *per se*. According to her, the concept of universals of language developed by Chomsky is enough to explain phenomena involved in the translation process, but she does not explain how this concept would be used in translation studies. Second, she considers translation a practical activity and it can be considered an act of performance, not of langue or competence. According to her, terms like "explicitness", "explicitation", "simplification", and so on are too general and "they should not be used unless one is perfectly clear about how they can be precisely defined and operationalised" (ibid, p. 11). Third, she talks about the issue of language-pair specificity in translation, affirming that "candidates of universality suggested for one particular translation direction need not necessarily be candidates for universality in the opposite direction" (ibid). Fourth, in her opinion universals of translation seem to be closely related to genre-specificity. Finally, she claims that it "is necessary to take the diachronic development of texts into account which belong to a certain genre" because "translations develop critically and they may be critically influenced by the status of the language of the source text genre which in turn may influence the nature of the translation text genre and also the nature of comparable texts in the same genre" (ibid). In summary, House is strongly skeptical in relation to the concept of translations universals because she believes that the postulation of universals of language (which necessarily includes translation) is quite sufficient to explain certain features.

I am not totally skeptical in relation to the existence of features of translation, but I agree with House when she mentions that the concept of universals is too general and should be carefully used. The first version of my Resarch Project also included the search for some features of translation in the parallel corpus (explicitation and simplification), but as soon as I started the analysis I realised that it would not be possible to annotate the corpus using these concepts as criteria because the available bibliography on universals usually focus on definitions, they do not provide clear information about the methods used to identify these features in a corpus.

2.6 Thunes' model

Thunes reports on an empirical study of complexity in translation which consisted in the extraction and classification of translationally corresponding text units according to the complexity involved in the translation task. The unit of translation was the finite clause and behind the method laid "the idea that studying the product of translation may reveal what information is needed in order to produce a specific translation from a given source text" (Thunes, 1998, p. 25).

In this case, the need for information during the translation process defined the criterion of classification: The string pairs were classified based on how much information and what kinds of information a translator would need, and the accessibility of the information. In order to measure this accessibility, the researcher created a scale which described four different types of translational correspondence. These types of translational correspondence are related to each other and organised in a hierarchy. All the string pairs of the parallel corpus were classified according to these four types of translational correspondence which are presented subsequently. Some examples shown in the introduction to illustrate the correspondence types will be revisited and discussed in more detail.

Type 1 correspondences

Thunes defined a scale of complexity ranging from 1 to 4. Type 1 correspondences represent the lowest degree of translational complexity. In order to be classified as type 1, source and target strings necessarily need to share similarities on three distinct levels: Syntax, semantics, and pragmatics. These similarities result in implicational relations between the relations of equivalence: If there is equivalence in the syntactic level, this implies in semantic equivalence, which in turn implies in pragmatic equivalence (Thunes, 2011).

According to Thunes, translation tasks that conform to the characteristics of type 1 correspondences are considered computable and they can be characterized as linguistically predictable tasks. With the purpose of solving them, the following information sources are necessary:

> (…) firstly, sufficient information about the source language to identify all lexemes in the

> source string and to derive its constituent structure; secondly, sufficient information about the interrelations between source and target language to find out that each source string lexeme has a syntactically and semantically matching TL correspondent, and that the source string structure likewise has a match in the target language; thirdly, information about the word order of the source string in order to generate a target string where the sequence of words is identical to that of the source string, and sufficient information about morphological restrictions in cases where the lexical relations between SL and TL are not enough to identify the correct word forms in the target string. (Thunes, 2011, p. 145)

The processing effort required to produce such string pair and the information considered necessary during the translation task can be represented in a finite way. The most important part of the translation task, in these cases, is the syntactic analysis (or syntactic parsing) of a natural language expression. Thunes also assumed that the effort required by the type identification and generation of the target text is equivalent to the size of the translation task.

In summary, "translational correspondences of type 1 are cases of word-by-word correspondences" (Thunes, 1998, p. 25). Consider string pair (9) as an example of Portuguese English type 1 correspondence, because the translation contains the same number of words of the original. Therefore, in the translation there is one correspondent word to each word present in the original text:

(9a) The artist can express everything.
(9b) O artista pode exprimir tudo.
(PDG, Oscar Wilde)[22]

Furthermore, each lexeme from the source string lexeme has a syntactically and semantically matching TL correspondent. The definite article *the*, corresponds to a definite article *o* in Portuguese. The same happens to the other lexemes in terms of syntactic function and semantic

[22] http://www.scribd.com/doc/7155292/Oscar-Wilde-o-Retrato-de-Dorian-Grey.

content. In relation to linearity, the sequence of words of the target text is identical to that of the source string.

Type 2 correspondences

In the scale of complexity, type 2 correspondences are considered the second lowest degree of translational complexity. The relations of equivalence between source and target string are obligatory on the levels of syntax, semantics and pragmatics. In other words, in correspondences of type 2, it is almost possible to translate word by word but differences between source and target string in relation to word order and/or the use of grammatical function words impede this process. Nevertheless such divergences

> (…) cannot violate the requirements that source and target string must be equivalent with respect to the assignment of syntactic functions to constituents, and that all lexical words in the source string must have a target correspondent of the same category and with the same syntactic function. (Thunes, 2011, p. 153)

String pair (10) is an example of a Portuguese-English type 2 correspondence identified in the corpus:

(10a) Mad in pursuit and in possession so,
(10b) Insana ao perseguir, e assim na possessão

Here there are differences in relation to phrase order, but all lexical words in the source string have a target correspondent of the same category and with the same syntactic function in the target string.

Type 2 correspondences, just like type 1, also fall within the domain of linguistically predictable translation tasks, and as a consequence they are considered computable. With the purpose of solving these types of correspondence, the following information sources are necessary:

> (…) firstly, sufficient information about the source language to parse the source string; secondly, sufficient information about the interrelations between source and target language to find out that each lexical word in the source

> string has a syntactically and semantically matching TL correspondent, (…); thirdly, information about the constituent structure if the target string must be different, and, finally, sufficient information about morphological restrictions in cases where the lexical interrelations between SL and TL are not enough to identify the correct word forms in the target string. (Thunes, 2011, p. 154)

In relation to processing effort, the necessary information during the translation task can be represented in a finite way, but it involves an extra effort of evaluating the differences with respect to the linear order of constituents. Still, in general terms and similar to type 1, the effort required by the type identification and generation of the target text can be considered of modest complexity (Thunes, 2011).

Type 3 correspondences

On the scale which ranges from type 1 to type 4, type 3 translational correspondences "represent the second highest degree of translational complexity" (Thunes, 2011, p. 163). In those cases, source and target string maintain a relation of equivalence based on semantics and pragmatics. Another important aspect is that "implicational relations between such equivalence relations exist to a lesser degree than in the cases of types 1 and 2" (ibid) because although there is not syntactic equivalence between the entire strings, there is still equivalence on the semantic level, and consequently it is assumed that there is pragmatic equivalence in both texts.

In type 3 correspondences structural discrepancies between source and target string are greater if compared to those of type 2. According to Thunes, the main characteristic of type 3 is that, "while the two strings can be assigned equivalent semantic representations, there is at least one lexical word in one of the strings for which the other string lacks an equivalent word of the same lexical category and with the same syntactic function" (1998, p. 28). This is the case of Example (11), in this specific context there is no exact correspondence for the word *quiet*:

(11a) For thee and for myself no quiet find.
(11b) Não podem repousar, graças a ti e a mim.

However, the expression "*no quiet find*" can be considered correspondent to "*não podem repousar*" in terms of semantic content. Therefore, the strings are considered "semantically equivalent in the sense that the same informational content is linguistically encoded in both of them" (Thunes, 2011, p. 164).

Concerning computability, type 3 correspondences are considered computable because just like types 1 and 2, they fall within the domain of linguistically predictable tasks.

In order to solve these types of correspondence, the following information sources are indispensable:

> (…) firstly, sufficient information about the source language to identify all lexemes in the source string, to derive its constituent structure, and to derive a semantic representation containing all relevant components of meaning expressed by the source string; secondly, sufficient information about the interrelations between source and target language to find out that the target string is structurally different (…); thirdly, sufficient lexical, morphological, syntactic, and semantic information about the target language in order to generate a target string on the basis of the semantic representation of the source string. (Thunes, 2011, p. 164).

Similarly to types 1 and 2, in type 3 correspondences "all kinds of information required to solve the translation task can be represented in a finite way" (Thunes, 2011, p. 164) and the type identification is also solvable in linear time.

The subtask of analysis, as it was mentioned before, is the same for all types of correspondences, the only difference is the need for a semantic analysis of the source string in type 3 correspondences (Thunes, 2011).

With respect to the last subtask (generation), Thunes assumes that translation tasks of type 3 differ from types 1 and 2 "in the sense that whereas a modest processing effort is required by target string generation" (2011, p. 164) in lower types, "generation from semantic representations in type 3 is very resource-intensive" (ibid). From a computational point of view, these correspondences would probably be treated as intractable problems.

Type 4 correspondences

Type 4 correspondences differ from the lower types 1-3 in relation to the aspect of computability. Since they are not inserted in the domain of linguistically predictable translation tasks, they are considered noncomputable, in other words, they depend on the work of a human translator. These correspondences represent the highest degree of translational complexity on the scale proposed by Thunes (2011). In general terms,

> (…) there is not semantic equivalence between the entire source and target strings; pragmatic equivalence may exist, but not necessarily. Hence, there do not exist, as in string pairs of the lower types, any implicational relations between equivalence relations on different linguistic levels. (Thunes, 2011, p. 170)

Translational correspondences of type 4 "are cases where there are discrepancies between original and translation not only on the structural, but also on the semantic level. Type 4 is assigned to translational correspondences where we cannot derive equivalent semantic representations for source and target string" (Thunes, 1998, p. 28). Example (12) is an instance of a type 4 correspondence because it contains a mismatch on the semantic level:

(12a) Yet, do thy worst, old Time: despite thy wrong,
(12b) Faze o pior, porém: malgrado o teu rigor,

The expression *old Time* is an example of semantic component which has no correspondence in the target string. The semantic representation of the expression *malgrado o teu rigor* does not correspond to *despite thy wrong* in terms of meaning.

Therefore, in order to solve a translation task of type 4, it is necessary to have access "to the information linguistically encoded in the source text, as well as to general information about SL and TL, and their interrelations" (Thunes, 2011, p. 171). In addition to that, it is necessary to access other information sources, since an understanding of the source string depends on syntax and semantic knowledge. "Since it is necessary (…) to access information sources falling outside the finite,

prestructured domain, there is in principle no limit on the processing effort required to search for the needed in formation" (ibid).

The extracted string pairs taken from the parallel corpus were initially classified according to these four types of translational correspondence described by Thunes. The four types are summarised in Table 3, which contains an extra column which describes the type of information required from the translator during the process of translation.

Table 3. Definition of the correspondence types

TCT[23]	Definition	Requires …
Type 1	Cases of word-by-word correspondences	Information about the syntax of the source string, i.e., its constituent structure.
Type 2	Cases where it is nearly, but not quite, possible to translate word by word.	Information about the syntax of the source string. Information about which syntactic constructions in the source string must be changed when producing the target string.
Type 3	Cases where there are greater structural discrepancies between source and target string.	Information about the syntax of the source string (in order to find out that the syntax of a translation must be different and in order to derive the semantic representation). Information about the semantic representation of the source string. Information about the syntactic rules of the target language and about how a translation is generated from these rules together with the semantic representation of the source text.
Type 4	Cases where there are discrepancies between original and translation not only on the structural, but also on the semantic level.	Information about the syntax as well as the semantics of the source string, in order to see that the translation will differ from the original both syntactically and semantically. Types of information which must be derived from the linguistic expression alone which may include: •discourse information which must be derived from a wider linguistic context. •information about the utterance situation of the source string. •extra-linguistic background information, including domain specific technical information.

[23] Translational correspondence types.

2.7 Towards a Methodology for Investigating the Style of a Literary Translator

Baker (2000, p. 242) initiates her discussion towards the style of a literary translator affirming that "a number of translation scholars have attempted to apply various interpretations of the notion of style to the study of translation, mostly with a view to elaborating criteria for quality assessment".

As an example, she mentions House who "sets out to develop a model for describing the linguistic and situational peculiarities of the source text, comparing source and translation texts, and making informed statements about the relative match of the two" (Baker, 2000, p. 242). According to Baker, these statements have an evaluative purpose; their objective is to decide if the translation can be considered good, bad or indifferent. The proposed evaluation is based on the analysis of two distinct sets of 'situational dimensions': "the dimensions of language user and the dimensions of language use" (ibid).

The first dimension "covers geographical origin, social class, and time" (Baker, 2000, p. 242) while the second dimension "covers medium, participation, social role relationship, social attitude, and province" (ibid).

According to Baker,

> House in fact combines two of the most common interpretations of the notion of style: as variation in the level of formality, hence the borrowing of the categories from Jools, and as patterned choices across all linguistic levels. She does not attempt a systematic treatment of the notion of style as such, since ultimately what she aims to describe is not so much the style of the original text or author, and certainly not the style of the translation or translator, but where the two texts diverge along the two dimensions of language user and language use, and only along those two dimensions. Hers then is essentially a checklist of features designed to allow the scholar to formulate a statement of the relative match of source and target texts and the relative success of the latter in reproducing the 'style' of the original. (Baker, 2000, p. 242)

Baker (2000) signalises that, apart from House's study, several attempts to combine insights from linguistic and literary studies of style have been used to explain choices made by some translators or to determine "guidelines for the selection of specific translation strategies on the basis of broad stylistic categories formalized as text types or registers" (ibid, p. 243). Baker believes that,

> (…) this reflects the fact that the notion of style in both linguistic and literary studies has traditionally been associated with one of three things: the style of an individual writer or speaker (e.g. the style of James Joyce, or Winston Churchill), linguistic features associated with texts produced by specific groups of language users and in a specific institutional setting (e.g. the style of newspaper editorials, patents, religious sermons), or stylistic features specific to texts produced in a particular historical period (e.g. Medieval English, Renaissance French). (Baker, 2000, 243)

With respect to Translation Studies, it has clearly inherited distinct types of knowledge from literary studies and from linguistics. From the first one, it inherited "its preoccupation with the style of individual creative writers, but only insofar as describing the style of a writer can inform the process if translating his or her work" (Baker, 2000, p. 243). From the second, on the other hand, it "inherited the interest in studying the style of social groups of language users (more commonly known as register analysis)" (ibid). According to Baker, the classifications of style can be based on distinct criteria:

> (…) the context in which language is used (e.g. journal articles, radio broadcasts), subject matter (medical discourse, legal language), a combination of both (medical journal articles, law textbooks), or the nature of the message and addressor/addressee relationship (argumentative discourse, the language of instructions). (Baker, 2000, 243)

Nevertheless, independently from the chosen classification, the objective is usually to provide "a starting point for identifying the distinctive features of the source text in order to reproduce in the

translation either those same features or the typical features associated with the same text type in the target language" (ibid).

In relation to style of translation, the heritage received by Translations Studies from literary studies and linguistics is associated with the idea of 'original' writing. That means that the interest in studying the style of a translator, or a group of translators, or a corpus of translated material belonging to a specific period of time has been of little or no interest within the research community. As a consequence, the idea that a translator does not have the right to develop his or her own style becomes implicit and commonly accepted, reducing the translation task to the reproduction of the style of the original as much as possible (Baker, 2000). Baker questions this assumption claiming that "it is impossible to handle an object without leaving one's fingerprints on it" (ibid, p. 244).

In the present investigation, I share with Baker the idea of describing the translator style and the attempt "to develop a model for describing the linguistic and situational peculiarities of the source text, comparing source and translation texts, and making informed statements about the relative match of the two" (2000, p. 242). The basic difference is that, instead of adopting the 'typical dimensions' (of language user or of language use) traditionally used, I shall trace the differences of style between the two translators (Jorge Wanderley and Péricles Eugênio da Silva) based on the level of translational complexity identified on their respective translations.

2.8 Towards more objective evaluation of poetic translation

Britto (2001) discusses the issue of evaluating poetic translation, which is considered "a complex and delicate task".

> Poetic texts deal with language on all its levels – semantic, syntactic, phonetic, and rhythmic, among others. Ideally a poem should articulate all these levels, or at least several of them, in order to achieve a certain set of poetic effects. The translator of poetry must then re-create, using the resources of the target language, the effects of content and form in the original – or, again, at least a good number of them. (Britto, 2001, p. 1)

Based on all these elements, Britto sketches a methodology for the evaluation of poetic translations which includes "a systematic examination of the different levels of language involved in the poem" (ibid). In order to do so, the first step would be to define precisely the idea of correspondence, in other words, what is meant when one says that "a given element of a translated poem corresponds to a given element of an original poem" (Britto, 2001, p.1).

The author points out the necessity of understanding the concept of correspondence on various levels of exactness. In order to do that, first it is necessary to determine what are the formal and semantic features of the source text. In order to compare each of these features, he suggests the use of the antithetical concepts of "correspondence" and "loss", which means that "the greater the correspondence, or match, between a feature of the original and its counterpart in the translation, the smaller the loss" (Britto, 2001, p. 7) of information. This concept of loss is directly related to the subdivision of type 3 (as suggested by Thunes) into types 3.1 and 3.2 (presented in Chapter 4).

The concepts defined by Britto should be defined on the basis of the notion of levels of correspondence, that is, "the higher the one-to-one match between the components of a given feature of the original and the components of its counterpart in the translation, the smaller the loss" (Britto, 2001, p. 7). Besides the evaluation of this degree of loss suggested by Britto (which are also related to types 1 and 2 presented by Thunes, 1998), the following questions might also be considered:

> (1) How relevant is the feature in the original?
> (2) Is the maximum degree of correspondence feasible? When the target language lacks exact counterparts for the items in question, a close match cannot be reasonably expected.
> (3) How desirable is an exact match? There may be cases when it seems better to rely on functional rather than formal correspondence. (2001, p. 7)

This preliminary sketch of a method suggested by Britto (2001) represents a way to arrive at less subjective way of evaluating poetic translation. This is similar to the objective of the present study, since the adaptation of Thunes' model also relies on more objective aspects of data in order to quantify value-judgments expressed through concepts of correspondence between source and target strings.

2.9 Shakespeare Sonnets

Shakespeare's lyrical poetry includes 154 sonnets, two narrative poems (*A Lover's Complaint*, *The Rape of Lucrece*), two long poems (*Venus and Adonis*, *Phoenix and the Turtle*) and the poem *Funeral Elegy by W.S*. The sonnets were probably "written about 1593-1600, first printed by Thomas Thorpe in 1609. There are many interrelated problems connected with these 154 sonnets, the main one being the authenticity of the order of the sonnets" (Hodeck[24], 1971, p. 21).

According to Jones (2006, p. 1), "the public story of *Shakespeare's Sonnets* began late in 1598", when Francis Meres mentioned Shakespeare's name in the book *Palladis Tamia*. One year later, in 1599, *The Passionate Pilgrime* was published by William Jaggard. This small octavo volume contained twenty sonnets, but only four sonnets and one lyric are surely attributed to Shakespeare (although many readers assumed that all the sonnets published in that volume were Shakespeare's). The problem is that "*The Passionate Pilgrime* was surely disappointing, in both quality and quantity" (ibid, p. 2), and the readers concluded that Meres had "overpraised" those specific Shakespeare's Sonnets which had a "doubtful" quality. In 1609, Shakespeare finally "had assumed control of his own text of his *Sonnets*, by selling the collection to Thorpe and giving it the title *Shakespeare's Sonnets*" (ibid, p. 3).

As previously mentioned, there is no agreement in relation to the authenticity of the order of the sonnets. The period comprehended between 1608 and 1919 was marked by the plague outbreaks, which culminated in the closure of the public theatres. According to Jones (2006, p. 10), "during this phase of theatre closure it seems probable that Shakespeare turned once more to his sonnets, revising poems he had already written, and expanding and redesigning his sequence in a matter which suited the new culture heralded by the new reign". Just like "better-documented sonneteers", Shakespeare probably rewrote and reordered his Sonnets during the period comprehended between Meres' publication and the first authorised publication. However, some scholars just believe that "Shakespeare did begin to write at 1 and simply carried straight through to 154" (ibid, p. 16).

[24] Hodeck wrote the introduction of the *The Complete Works of William Shakespeare* published by Spring Books.

The present investigation relies on the analysis of a corpus composed of 45 sonnets by Shakespeare and two respective translations into Brazilian Portuguese. The chosen sonnets were translated by Péricles Eugênio da Silva Ramos and by Jorge Wanderley. The former published 45 sonnets in a bilingual edition titled *Sonetos* (2008), the latter also published a bilingual edition also titled *Sonetos* (1991) with all the 154 sonnets translated into Brazilian Portuguese.

Péricles Eugênio da Silva Ramos (1919-1992) was Born in Lorena-SP. His first poems were published in the newspaper called *Diário de Notícias* from 1936. The book "Lamentação Floral" (1946) marked his debut on the national literary scene. In 1945 he joined a group of writers and poets and founded journal called *Revista Brasileira de Poesia*, responsible for the dissemination of the 45 Generation's aesthetics. As a translator, he was responsible for the translations of several major authors (besides Shakespeare) such as Stéphane Mallarmé, François Villon, Luís de Góngora, Byron, among others. He also produced several anthologies of Brazilian poetry[25].

Jorge Wanderley was born in Recife, Pernambuco, in 1938. Physician, poet and translator, he started to write when he was 16, and he published his first book *Gesta e outros poemas* in 1960. Known primarily as a translator and literary critic, his own work is also considered important. He published numerous chronic and literary essays in magazines and newspapers, and a dozen translations, which include classics of literature, such as Dante, Shakespeare and Bukowski. In 1998, he completed the first part of his project which included a full translation d 'The Divine Comedy - with annotated translation of "Inferno", Dante. He also wrote essays about their prologues and translations, for which he received the prize called *Prêmio Jabuti de Tradução Literária* in 2004. He died in Rio de Janeiro, on December 11th, 1999[26].

In terms of metre, Péricles Eugênio da Silva uses iambic hexametre verses, that is, the "twelve syllable verse stressed in theory in the pair syllables, that will coincide or not, with the French alexandrine"[27] (2008, p. 12). The author justifies his choice claiming that, if he had maintained the original metre (iambic pentametre), he

[25] More information available at:
http://www.jornalolince.com.br/2008/nov/capa/pericleseugenio.php
[26] Source: http://www.dicionariodetradutores.ufsc.br/pt/JorgeWanderley.htm, my translation.
[27] My translation: "pelo verso de doze sílabas acentuado em tese nas sílabas pares, e que coincidirá, ou não, com o alexandrino de tipo francês" (p. 12).

would have to sacrifice many elements of the text in the translation. He also emphasises that, since it is very difficult to keep the same stress, in some cases the foot pattern will not succeed in the same order in the original and in the translation. According to him, sometimes the appearance of the same rhythm can be preserved, but not the rhythm itself. The maintenance of the same rhythm would imply the succession of the same foots in the same order, with the same caesuras, which definitely cannot be the same in English and in Portuguese.

Jorge Wanderley decided to keep the same metre of the original. He claims that a Shakespearean sonnet is basically a musical being. Since it was originally written in iambic pentametre, it should necessarily sound in such way: "the sound of a decasyllable (in this case the iambic pentametre) is not the sound of a dodecasyllable or any other verse. Just like the form that Beethoven creates for a quartet differs from a trio"[28] (1991, p.19). According to him, this is one of the main principles in poetry translation.

Although both translators explicitly justify their choices in terms of metre in the introduction of their respective editions, sometimes it is difficult to identify the iambic hexametre in the translations by Péricles Eugênio da Silva and the iambic pentameter in the translations by Jorge Wanderley. Consequently, the role of metre in the discussion of translational correspondences is likely to require sophistication in future research.

[28] My translation: "O som de um decassílabo (no caso o pentâmero iâmbico) não é o som de um dodecassílabo ou outro qualquer. Da mesma forma que o som que Beeethoven dá a um quarteto não é o que dá a um trio" (p. 19).

Chapter 3

Method

3.1 Overview

This chapter is divided into six main sections. The second section (3.2) briefly presents the text material. Section 3.3 explains how the alignment process was done. In Section 3.4, the criteria that define each of the four translational types are scrutinized to clarify how the annotation process was carried out. Section 3.5 discusses some methodological principles. Finally, Section 3.6 presents an annotated Sonnet which illustrates how the annotation process was carried out throughout the parallel corpus.

3.2 Text material

The centrepiece of the present investigation consists of an empirical investigation of selected parallel texts in English and Brazilian Portuguese. The data consists of a manually and annotated corpus of approximately 80,000 words. The corpus contains 45 Shakespeare's Sonnets and the respective translations into Portuguese made by the Brazilian writers Péricles Eugênio da Silva and Jorge Wanderley.

As mentioned before, the original sonnets were probably written between 1593 and 1600, and their first printed version was published in 1609. Jorge Wanderley published a bilingual edition titled *Sonetos* with all the 154 sonnets translated into Brazilian Portuguese in 1991, while Péricles Eugênio da Silva published a

reduced bilingual edition also titled *Sonetos* in 2008, containing 45[29] sonnets. Initially, I intended to analyse all the sonnets, but while Jorge Wanderley had translated all the sonnets, Péricles Eugênio da Silva Ramos only translated 45 of them. Therefore, in order to have at least two different translations of the same sonnets into Brazilian Portuguese in the parallel corpus, it was necessary to restrict the analysis to those 45 sonnets translated by Péricles Eugênio da Silva Ramos.

3.3 The alignment process

The method used for the purposes of this investigation relies on the analysis and annotation of all the string units identified in the parallel corpus, that is, the verses and their respective translations. Just as in Archer's investigation (2006), the approach used in this investigation can be classified as top-down, because the categories used to annotate the corpus are predefined: The four types of translational correspondence (as defined by Thunes) are used to annotate the corpus. The process of annotation in the present investigation is manual, not automatic. This investigation aims, therefore, at manual annotation of a parallel corpus composed of 45 Shakespeare's Sonnets and their respective translations into Brazilian Portuguese.

After choosing the translations that would be used in the investigation, the first step of the work was the digitalization of the texts. A twentieth-century edition of Shakespeare's Sonnets can be easily found in electronic format, but the two translations of the sonnets into Brazilian Portuguese are not available in such format in the internet. Thus, it was necessary to type the translations of the selected sonnets in order to save and store them as .txt documents.

Once all the Sonnets were digitalised, the next step was to align them with the translations. All the sonnets were aligned manually, but this was only possible because the corpus was relatively small. Some

[29] Sonnets V, XV, XVIII, XIX, XXII, XXIII, XXV, XXVI, XXVII, XXIX, XXX, XXXIII, LII, LIV, LV, LVII, LXI, LXII, LXVI, LXXI, LXXIII, LXXVI, LXXVI, LXXX, LXXXVII, XC, XCVII, XCVIII, XCIX, CIV, CV, CVI, CVII, CIX, CXVI, CXIX, CXXI, CXXIII, CXXVII, CXXIX, CXXX, CXLII, CXLIV, CXLV, CXLVI and CXLVII.

programs, such as the software WordSmith Tools[30], contain an aligner tool that automatically aligns source texts and translations based on punctuation parameters. Nevertheless, I detected that this tool would not work properly in this specific corpus because there are great differences between the punctuation of the original and the punctuation of the translated sonnets. Methodologically speaking, it was easier to align all the sonnets manually.

The process of alignment consists of the organisation of source and target strings, in such way that each original verse and its two respective translations appear in three-line sets as in Example (13). The order of alignment was always the same: original verse followed respectively by the translations by PS and by JW.

(13) Those hours that with gentle work did frame (original)
Aquelas horas que formaram meigamente (PS)
As horas que formaram gentilmente (JW)

Initially, the verses of the corpus were aligned just as in Example (13) and the verses were typed in different colours to facilitate the visualisation of the two distinct translations during the analysis. The translations by PS were always marked in blue, while the translations by JW were always marked in red in the parallel corpus. Differently from Thunes' model, where the unit of translation was the finite clause, the string pair in the parallel corpus was the verse. This was considered a coherent decision, since the translated sonnets maintained the same structural organisation of the original, 14 verses, except Sonnet XCIX, which contained 15 verses. Thus, each verse aligned with the respective translations was considered a distinct string pair, which was supposed to be classified individually according to one of the four translation types during the analysis. The organisation of the string pairs can be observed in Example (14):

[30] a software developed by Mike Scott and marketed by Oxford University Press which offers many tools to the analysis of electronic corpora.

(14) Sonnet XXII

My glass shall not persuade me I am old,
De minha idade o espelho não me pode argüir, (string pair 1)
Não me convence o espelho de estar velho (string pair 2)
So long as youth and thou are of one date;
Visto que a juventude e tu andais a par; (string pair 3)
Enquanto a juventude te acompanha. (string pair 4)
But when in thee time's furrows I behold,
Quando os sulcos do tempo em ti eu descobrir, (string pair 5)
Mas se mostra rugas teu espelho (string pair 6)
Then look I death my days should expiate.
Logo a morte virá meus dias consumar. (string pair 7)
Então a morte os dias me arrebanha. (string pair 8)
For all that beauty that doth cover thee
Pois toda essa beleza que te cobre assim (string pair 9)
Pois a graça que a ti se concedeu (string pair 10)
Is but the seemly raiment of my heart,
É a linda veste, apenas, de meu coração, (string pair 11)
Traz ao meu coração as vestimentas (string pair 12)
Which in thy breast doth live, as thine in me:
Que vive no teu peito, como o teu em mim: (string pair 13)
Se ele mora em teu peito e o teu no meu, (string pair 14)
How can I then be elder than thou art?
Como hei de ser mais velho do que tu, então? (string pair 15)
Posso ter a idade que aparentas? (string pair 16)
O, therefore, love, be of thyself so wary
Contigo, meu amor, tu deves ter cuidado, (string pair 17)
Por isso, amor, cuida de ti, atento, (string pair 18)
As I, not for myself, but for thee will;
Como contigo, e não comigo, eu hei de ter: (string pair 19)
Como eu cuido, no meu , do teu destino: (string pair 20)
Bearing thy heart, which I will keep so chary
Trago teu coração, e guardo-o desvelado, (string pair 21)
Levo o teu coração e lhe acrescento (string pair 22)
As tender nurse her babe from faring ill.
Como doce ama a criancinha a protejer. (string pair 23)
Cuidados de enfermeira a seu menino. (string pair 24)
Presume not on thy heart when mine is slain;
Quando meu coração morrer, do teu desiste: (string pair 25)
Não contes, morto o meu, com o coração (string pair 26)
Thou gavest me thine, not to give back again.
Foi para todo o sempre que mo transferiste. (string pair 27)
Que me deste – e não tem devolução! (string pair 28)

Most of the sonnets could be easily aligned with the two respective translations. That means that each verse was immediately followed by two translated and correspondent verses during the alignment process. Another example that illustrates this process of alignment can be observed in the Sonnet XXII. The order of the original verses was respected by both translators, an important aspect that facilitated the alignment process.

(15) Sonnet CV

Let not my love be call'd idolatry,
Oh! ninguém chame idolatria o meu amor,
Não chamem meu amor de idolatria
Nor my beloved as an idol show,
Nem dê por ídolo quem alvo é desse preito,
E que o meu bem um ídolo não lembre
Since all alike my songs and praises be
Porque todo o meu canto e todo o meu louvor
Por ser só dele a minha poesia
To one, of one, still such, and ever so.
São para alguém, de alguém, e sempre, e de um só jeito.
E eu louve um só e louve o mesmo e sempre.
Kind is my love today, tomorrow kind,
Meu amor hoje é afável, amanhã afável,
Suave é hoje e sempre revelado
Still constant in a wondrous excellence;
Sempre constante numa esplêndida excelência:
Na mesma maravilha em que cintila
Therefore my verse to constancy confin'd,
Logo meu verso, limitado ao invariável,
E o meu verso, à constância confinado
One thing expressing, leaves out difference.
Exprime uma só coisa, e exclui a impermanência.
Expressa o mesmo – e o diferente exila
Fair, kind and true is all my argument,
"Bom, belo e verdadeiro" – é um só meu argumento,
"O belo, o bem, a verdade", eis o tema;
Fair, kind, and true varying to other words,
"Bom, belo e verdadeiro" – em vária locução:
"O belo, o bem, a verdade" – a variação.
And in this change is my invention spent,
Nessa mudança absorvo tudo quanto invento,
E neste espaço inventa o meu poema
Three themes in one, which wondrous scope affords.

Três temas postos num, de amplíssima extensão.
Num só, três temas – que infinitos são.
Fair, kind, and true, have often lived alone,
"Bom, belo e verdadeiro" alheios têm vivido:
O belo, o bem, a verdade, antes sós,
Which three till now never kept seat in one.
Num ser ainda não se haviam reunido.
Agora assentam numa mesma voz.

However, as pointed out in Section 3.2, producing a link between two texts at sentence level in a corpus might not be a trivial task (Piao, 2000, 2002, qtd. in McEnery & Xiao, 2008), and once the alignment process advanced, two specific problems surfaced.

During the alignment phase, some sonnets presented a different pattern of organisation that made the process of alignment problematic. This happened when one of the translators decided to invert the order of the verses in the translations. One example of problematic alignment can be observed in the Sonnet XIX. The second and third verses were inverted in the translation by PS. Although JW kept the same order, it was not possible to align verse by verse due to this variation. Therefore, in order to visualise the original and the two translations, I decided to join the two original verses and to create six-line sets to facilitate the visualisation during the annotation of the parallel corpus.

(16) Sonnet XIX

Devouring Time blunt thou the lion's paws,
Cega, ó Tempo voraz, as garras do leão,
Tempo voraz, que ao leão lima as garras,
And make the earth devour her own sweet brood,
Pluck the keen teeth from the fierce tiger's jaws
E dos tigres arranca os dentes à maxila;
Faze que a terra coma a própria geração,
Que à terra faz comer filhos da terra
E ao tigre arranca as presas da bocarra,
And bum the long-liv'd Phoenix in her blood,
E a fênix, no seu sangue em flamas, aniquila!
E queima a fênix no sangue que encerra
Make glad and sorry seasons as thou fleet'st,
Fugindo, as estações alegra ou entristece;
E passa e deixa a estação bela ou triste,
And to whate'er thou wilt, swift footed Time,
Dispõe, Tempo dos pés velozes, do universo,

Faz como queiras, com teu pé veloz,
To the wide world and all her fading sweets:
E de quanta doçura, eu sei, nele esmaece;
Ao que declina, ao que no mundo existe;
But I forbid thee one most heinous crime,
Porém eu te proíbo um crime mais perverso:
- Mas te proíbo o crime mais feroz:
O carve not with thy hours my love's fair brow,
Nor draw no lines there with thine antique pen,
Não queiras entalhar de meu amor a fronte
Com tuas horas, nem riscá-la com tua pena
Com as horas talhar a fronte amada,
Vincá-la com teu cálamo maduro;
Him in thy course untained do allow,
Antiga; mas que puro, ó Tempo, ele defronte
Permite que em teu curso a meu bem nada
For beauty's pattern to succeeding men.
Os pósteros – padrão de formosura plena
Perturbe, que é padrão para os futuros;
Yet, do thy worst, old Time, despite thy wrong,
Faze o pior, porém: malgrado o teu rigor,
- Ou causa, tempo, os teus maiores danos;
My love shall in my verse ever live young.
Sempre jovem será em meus versos meu amor.
Meu verso traz meu bem à flor dos anos.

In that case, the six-line sets were used only to facilitate the visualisation of the original and the respective translations, but this should not interfere in the classification process. Nevertheless, each verse should be classified individually, independently from the verse order. The string pairs were those which were correspondent in terms of meaning, independently from the verse order:

(17)

And make the earth devour her own sweet brood,	(string pair 3)
Faze que a terra coma a própria geração,	(type 4)
Pluck the keen teeth from the fierce tiger's jaws	(string pair 4)
E dos tigres arranca os dentes à maxila;	(type 3)

In Sonnet CXVI, the order of the final verses was completely changed. This was the only Sonnet of the whole corpus which had verses that needed to be aligned in twelve-line sets, as observed below. Fortunately, this would not necessarily bring implications to the

classification process; as mentioned before, each verse should be later classified individually, independently from the verse order.

(18) Sonnet CXVI

Let me not to the marriage of true minds
Impedimentos não admito para a união
Ao casamento de almas verdadeiras
Admit impediments, love is not love
De corações fiéis; amor não é amor
Não haja oposição. Não é amor
Which alters when it alteration finds,
Quando se altera se percebe alteração
O que muda à mudança mais ligeira
Or bends with the remover to remove.
Ou cede em ir-se, quando é infiel o outro amador.
Ou, desertando, cede ao desertor.
O no, it is an ever-fixed mark
Oh! não, ele é um farol imóvel tempo em fora,
Oh, não, que amor é marca muito firme
That looks on tempest and is never shaken;
Que as tempestades olha e nem sequer trepida;
E nem a tempestade o desbarata;
It is the star to every wand'ring bark,
É a estrela para as naus, cujo poder se ignora,
É estrela para a nau, que o rumo afirme,
Whose worth's unknown, although his height be taken,
Malgrado seja a sua altura conhecida.
Valor ignoto – mas na altura, exata.
Love's not Time's fool, though rosy lips and cheeks
Within his bending sickle's compass come,
Love alters not with his brief hours and weeks,
But bears it out even to the edge of doom:
O amor não é joguete em mãos do tempo, embora
Face e lábios de rosa a curva foice abata;
Não muda em dias, não termina em uma hora,
Porém até o final das eras se dilate.
Não é do Tempo mera extravagância,
Amor, embora a foice roube o riso
'A face e ao lábio rosa; na constância,
Resiste até o Dia do Juízo.
If this be error and upon me proved,
Se isso for erro e o meu engano for provado,
Se há erro nisto e assim me for provado,
I never writ, nor no men ever loved.

Jamais terei escrito e alguém terá amado.
Nunca escrevi, ninguém terá amado.

This first problem was considered structural because it would not complicate the annotation process. Thus, the same methodological decision was applied in all the sonnets which presented different verse order.

The second problematic aspect was observed during the alignment of verses nine and ten from Sonnet XIX. The order of the verses was not changed, but one part of the original verse (*thy hours*) that originally appeared in the ninth verse was transferred to the tenth in the translation by PS (*com tuas horas*). At this point of the analysis, it was possible to foresee that this could complicate the classification of the string pairs:

(19) O carve not **with thy hours** my love's fair brow,
Nor draw no lines there with thine antique pen,
Não queiras entalhar de meu amor a fronte
Com tuas horas, nem riscá-la com tua pena

Methodologically speaking, I decided that this would not be taken into consideration at this point of the research. This issue is discussed later on Chapter 4. Interestingly, Thunes reports the same problem during the identification phase:

> In the majority of cases it is a straightforward task to see what part of the target text is the translational correspondent of a given source string. In other cases some piece of meaning expressed in a certain string may not have any match in the parallel text, as it happens that meaning can be added or deleted during translation. There may even be cases where two particular strings, although they do not correspond with respect to what they express, constitute a string pair simply because other possible correspondence relations are excluded, and since neighbouring strings clearly belong to other string pairs. (Thunes, 2011, p. 194)

Once the process of alignment was concluded and the unit of translation was determined, the next step was the annotation of the

corpus according to the translational correspondence types, as defined by Thunes (2011).

3.4 Translational correspondence types

This empirical investigation is a classification of translationally corresponding strings according to four different types as defined by Thunes (2011). According to the definitions briefly introduced in the review of literature,

> (…) type 1 correspondences are cases of a full linguistic match, structurally as well as semantically, between source and target string; type 2 correspondences allow minor mismatches on the structural level, but none on the semantic; in type 3 correspondences there can be major structural divergences while there is still a semantic match, and in type 4 correspondences there are semantic as well as structural mismatches between source and target string. (Thunes, 2011, p. 125)

The identification of each correspondence type is based on syntactic and semantic criteria, being related to each other in a hierarchy related to the amount of information considered necessary to produce the translation. The translational complexity is measured according to this amount of information in a scale that ranges from type 1 to type 4 (Thunes, 2011).

Before presenting each correspondence type, four important topics related to Thunes' approach to translational complexity will be explained: The notion of 'translation task'; important criteria for distinguishing and describing correspondence types; the notion of 'necessary information' and finally the need for general information sources. These topics are relevant to our discussion not only because they help to understand the correspondence types themselves, they also justify some methodological decisions. Furthermore, it is necessary to clarify that some of these topics will be treated differently from Thunes' approach during the present study, given the specificity of the parallel corpus used in this investigation.

Thunes considers the notion of 'translation task' important to an approach involving translational correspondences and here I adopt the same concept used in her study: Translation task covers "the task of translating anything from a single lexical item, or a sentence, to an entire document, such as a handbook or a novel" (Thunes, 2011, p. 126).

In Thunes' study, the issue of disambiguation was not explored, placed apart from the translation task. Her analysis of complexity exclusively involves translation tasks; therefore the problem of source text disambiguation is out of the scope of her study. The relevant interpretation of *aL1*[31] relates to the interpretation that lies behind the chosen translation *bL2*. The ways traced by the translator to identify the relevant interpretation are not included in her analysis (Thunes, 2011).

Another important aspect related to the approach is that Thunes recognises that a source expression corresponds with a set of possible target expressions and the translator needs to make a choice in order to translate the source text:

> (…) what we have aimed at in the analysis of translationally corresponding string pairs is to measure the complexity in a collection of concrete translation tasks (i.e. string pairs) where the chosen target expression is only one of a set of possible translations in *L2*. Thus, the complexity measurement applies to specific translation tasks *aL1* → *bL2*, and the analysis of each string pair is an attempt to describe the complexity of the selected task *solution* in relation to the source expression *aL1*, given its relevant interpretation. We do not consider the complexity of the translation task that is not solved yet; that would amount to analyzing the complexity of the general translation task (*aL1*→ *TL2*), which has a set of possible solutions. (Thunes, 2011, p. 127)

Therefore, it is clear that Thunes does not consider external evidence that might be part of the translational process. She focuses on the product of translation itself, leaving aside the set of possible solutions. Thunes focuses on the "identification of the complexity type

[31] Sometimes Thunes (2011) uses aL1 to SL and bL2 to TL.

(1, 2, 3, or 4) of the solution that has been chosen by a translator" (2011, p. 128), independently of the set of possible target expressions.

The focus of the present analysis is to measure the complexity in a collection of concrete translation tasks, but differently from Thunes, in some cases speculation on how the translator has identified the relevant interpretation will be within the scope of the analysis. This methodological decision is directly related to the genre of the texts. Poetry translation is different and can even be considered "special" when compared to other text types. Issues related to this specificity will be explained in detail in the discussion of the results.

With respect to the criteria used to distinguish and describe correspondence types, Thunes lists three items that are used to distinguish between the four types of translational correspondence:

> The first criterion pertains to the linguistic characteristics of the relation between source and target string, characteristics which show the degree to which there exist implications between relations of equivalence between source and target string. The second criterion concerns the amounts and types of information needed to produce the translation, and may be conceived of as the structure of the search task involved in translation. The third criterion deals with the processing effort required by the translation task, which may be seen as the *weight* of the search task. (Thunes, 2011, p. 129)

First, the author assumed that there are some linguistic properties linked to the relation established between source and target strings and these properties are used to identify each correspondence type. If we identify some structural similarities in a certain language pair (like English-Portuguese, for example), "there will be a certain set of linguistic structures in the source language sharing properties with translationally corresponding structures in the target language" (Thunes, 2011, p. 129). Therefore, when there is a high degree of similarities between source and target expressions, the translation task might be considered easy to solve. On the other hand, when original and translation are completely unrelated in terms of structure, the translation task might be considered harder to solve (Thunes, 2011).

> The presentation of the correspondence type hierarchy will show that in cases where similar structures of respectively SL and TL are translationally matched, there will exist relations of equivalence between source and target string, and also, implications between such equivalence relations. There relations of equivalence concern different linguistic levels: syntax, semantics, and pragmatics. The discussion of the correspondence types will illustrate that in cases where source-target equivalence with respect to syntax implies equivalence also with respect to semantics and pragmatics the degree of translational complexity is low, and that a translational complexity increases, such implications exist to a lesser degree. (Thunes, 2011, p. 130)

The second item listed by Thunes (2011) concerns the fact that the correspondence types should "be characterized with respect to the amounts and kinds of task-specific information required to translate source language strings" (ibid). In other words, one part of the analysis of the structure includes the identification of the information needed to interpret the source task and subsequently the search for information needed to produce the target expression.

Finally, "each correspondence type is characterized with respect to (…) the weight of the translation task, i.e. the amount of required processing effort" (Thunes, 2011, p. 131). According to Thunes, "the decomposition of the translation task into three subtasks is relevant also for these topics as the amount of required effort varies not only among the types of translational correspondences, but, (…), also among the subtasks" (ibid).

In relation to the notion of 'necessary information', Thunes first emphasises that the analysed string pairs were originally produced by a human translator and they are a representation of translation tasks solvable by competent language users that are bilingual. Some of these string pairs can be considered computable tasks because the use of some 'pre-structured linguistic information' seems to be enough to solve them. One of the aims in the classification of the string pairs is to identify the minimal necessary information needed to compute or to produce the target strings manually. In some cases, only the analysis on the level of syntax is enough to generate the target text. This is precisely what happens with types 1 and 2. But when it comes to type 3, further

analysis which includes the semantic level is required to generate the target text (Thunes, 2011).

When the author describes this idea of 'necessary information', it should not be interpreted as an attempt to conceptualise the translation process itself. During the translation process, a human translator takes into consideration not only the syntactic structure, but also the semantic content linked to the contextual information, and he or she will only choose a literal translation if that is considered appropriate. An efficient translator needs to dedicate special attention to the meaning and context of the source text. When this translator consciously chooses a word-by-word translation, that is probably done because he or she considers it appropriate after having carefully considered the meaning and context. It is never an arbitrary decision (Thunes, 2011).

However, if we put the human translator aside, and think about an automatic translation system, we realise how complex and necessary is the task of processing all the possible types of information linked to a certain text, even a very short one (Thunes, 2011). The analysis of complexity in the present investigation is not related to the creation of a model used for automatic translation of poetry, because this would not be even desired. Similarly to Thunes' investigation, the identification of

> *the necessary information sources* for translation in relation to each correspondence type is a way of describing how the complexity of chosen translation task solutions is determined by how much and what kinds of information that must *at least* be available in order to produce them. (Thunes, 2011, p. 132)

The next issue to be discussed is the need for general information sources, since this is an important subject related to Thunes' model. As pointed out before, the translator needs to have certain information available before producing the target text: "information about source and target language and their interrelations, and various kinds of extra-linguistic background information" (Thunes, 2011, p. 132). Although these different types of information exist independently from the translation activity, they are intrinsically related to it, if we assume that they are used by the translator during the translation process. This is justified by Thunes' when she mentions that correspondence types 1, 2, 3, and 4 vary depending on the amount of the given information sources required during the translation task. If we accept that types 1 to 3

represent translation tasks which are solvable inside the prestructured domain of linguistic information, type 4 correspondences represent the only case where extra information sources are required in order to produce the target text. Furthermore, it is clear that syntactic and morphological information is considered sufficient for the lower types (1 and 2), while in relation to type 3, semantic information needs to be considered during the translation task (Thunes, 2011).

According to Thunes, the information sources can be classified in two different ways: Linguistic and extralinguistic information. The first one "represents a limited domain" (2011, p. 133) while the second one is related to "an open-ended domain" (ibid); it comprises any information about the world. Theoretically, information about source and target languages and their respective interrelations could be represented in a finite way in information modules for automatic translation systems. However, it would be impossible to compute and decide which pieces of world information should be included in the same system. But the human translator has no difficulty to deal with extralinguistic information:

> In cases where translation requires the processing of given, general *world* information, we assume that, in general, this is not a problem that the computer can solve: the information is not available in the pre-structured domain of linguistic information, and hence not accessible. It is only within artificially delimited domains that world information can be made accessible in finite ways. For the human translator, on the other hand, it is hardly an effort to make use of general, extra-linguistic background knowledge. (Thunes, 2011, p. 133)

Thunes describes three subtasks which can be considered as subtasks of the translation process: analysis, complexity measurement (or type identification) and generation. The first subtask is the analysis, which involves the syntactic parsing of the source string. In that case, "the parsing problem is solved by using information contained in the representations of the source language lexicon and grammar" (Thunes, 2011, p. 134). Once the information is accessed, it should be respectively correlated with the length and linguistic complexity of the source string. The initial part of the analysis includes the recognition of word forms, and the analyst should assume that "the information

structure representing the SL lexicon is organized by base forms, so that for each inflected word morphological analysis is necessary to identify the lexeme it belongs to" (ibid).

The second subtask is the measurement of the complexity itself, which "is done by combining the task-specific linguistic information given in the interpretation of the source string with general information about the interrelations between source and target language systems" (Thunes, 2011, p. 135). Here, the description of corresponding elements of source and target texts is based on rules of their respective grammars. In other words,

> (…) when a translation task is computed, the subtask of analysis provides the bilingual information needed to diagnose the complexity of the translation task. The underlying principle is that information about how SL and TL are interrelated entails information about translational correspondences between specific linguistic elements in the two languages, so that identifying a particular lexeme or a particular syntactic structure in a source text will provide direct access to information about translationally corresponding elements in the given target language and information about linguistic properties shared by source and target elements. (Thunes, 2011, p. 136)

The third subtask called generation "requires information retrieved from the representations of the target language lexicon and grammar" (Thunes, 2011, p. 136). Therefore, in Thunes' model the information and subtasks which had just been described constitute the criteria used to measure the complexity in a translation task. From now on, the correspondence type hierarchy is presented in detail in order to illustrate the exact criteria used to annotate the parallel corpus.

Type 1 correspondences

This is the least complex class of translational correspondence described by Thunes. They are described as "word-by-word translations". They are definitely possible in the language pair English-Portuguese, but we do not expect them to be very frequent in a parallel corpus of translated poetry. String pair (20), presented in Section 2.6

and repeated here, is an example of a Portuguese-English type 1 correspondence:

(20a) The artist can express everything.
(20b) O artista pode exprimir tudo.
(PDG, Oscar Wilde)[32]

The translation contains the same number of words of the original, that is, in the translation there is one correspondent word to each word present in the original text.

In relation to linguistic characteristics of type 1, the translationally matched structures "are so similar that there is equivalence between source and target string with respect to the sequence of translationally corresponding surface forms" (Thunes, 2011, p. 137). In order to be classified as type 1, Thunes describes three prerequisites that need to be present in the string pair:

> Firstly, the strings must be syntactically equivalent, i.e. equivalent with respect to the assignment of syntactic functions (subject, object, etc.) to constituents. Secondly, the syntactic structures have to be compositionally equivalent in the sense of having corresponding properties with respect to compositional semantics: predicates and arguments must be contributed by corresponding constituents. Such compositional equivalence will in the normal case be a consequence of syntactic functional equivalence. Finally, the strings have to be pragmatically equivalent in the sense of being used to perform corresponding pragmatic functions, or speech acts, in the given texts. (Thunes, 2011, p. 137)

A string pair will only be classified as type 1 if the requirements described above are fulfilled. These requirements are important because they specify the correspondent linguistic properties that must be shared by source and target string in order to be classified as type 1. Another important issue is that "word-by-word correspondences do not qualify as

[32] http://www.scribd.com/doc/7155292/Oscar-Wilde-o-Retrato-de-Dorian-Grey.

type 1 unless they also correspond syntactically, semantically, and pragmatically" (Thunes, 2011, p. 137). In those cases, the translation task could be solved "by translating word by word" (ibid, p. 138) and a deep analysis of the source string would not be necessary to complete the task. Consequently, these correspondences can be considered linguistically predictable:

> Given the extent to which linguistic properties are shared between original and translation in type 1 correspondences, in particular the sharing of semantic properties, it follows that type 1 correspondences are included among the linguistically predictable translational correspondences, (…). That is, a target string corresponding to the source string according to type 1 requirements is a member of the LPT set of the source string. (Thunes, 2011, p. 138)

If we treat type 1 correspondences as linguistically predictable translational correspondences, that means that in order to solve the translation task the translator needed to access the prestructure domain of linguistic information. If one thinks from a computational point of view, the task is very simple: It would only be necessary to replace the lexical items in the source string with the correspondent word forms of the target string (Thunes, 2011). That is exactly what the translation tool of Google Translator would do with the sentence "The book is on the table" if one asks to translate it into Brazilian Portuguese. In a parallel corpus, this string pair shown in Figure 2 would be classified as type 1:

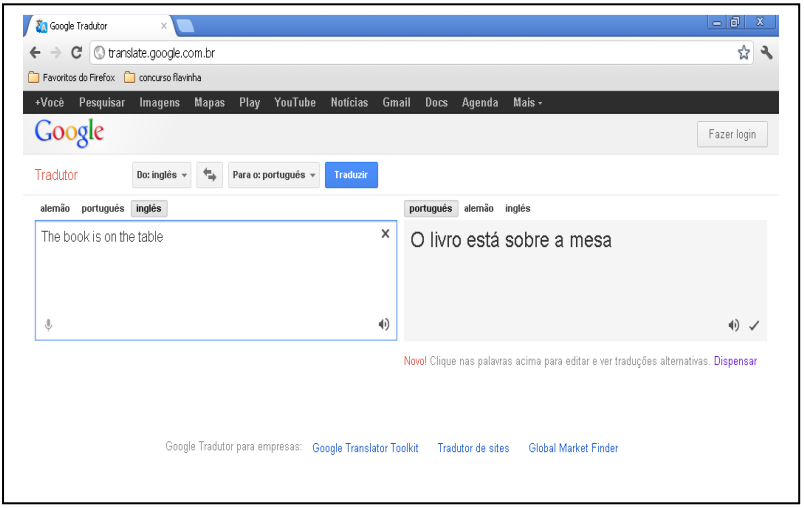

Figure 1. Example of type 1 correspondence produced by Google Translator, from a computational point of view.

Even in those cases, the interpretation of the source task "is an initial, indispensable subtask" (Thunes, 2011, p. 138) to determine if a given translation task really is a type 1 correspondence. What would, then, be necessary to identify a translation task as a type 1 case? In those cases, "it is necessary to compute a syntactic analysis of the source string" (ibid). This is consistent with the concept of parsing mentioned before, which "is solved by processing the information encoded in the source string together with given, general information about the source language system" (ibid).

Again, from a computational point of view, it would not be necessary to analyse the semantic structure of the source expression because morphological and syntactic information are sufficient in order to identify all lexemes (including function words) from the source string (Thunes, 2011).

In the subsequent subtask (type identification) it is necessary to check if the two following requirements are respected:

> Firstly, every lexical item in the source string must have a target language correspondent with syntactic and semantic properties matching those of the source item. (Thunes, 2011, p. 139)

If we take the string presented in Section 2.6 again, it is clear that every lexical item has a correspondent with the same syntactic and semantic constituents.

(20c)

The	**artist**	**can**	**express**	**everything.**
↓	↓	↓	↓	↓
O	**artista**	**pode**	**exprimir**	**tudo.**

> Secondly, in the target language there must be a structure which is equivalent to that of the source string with respect to the linear order of constituents and the assignment of syntactic functions to constituents. (Thunes, 2011, p. 139)

In relation to the second requirement, the order of constituents is linear and the syntactic functions of the constituents are the same. First there is an article (*the/o*), followed by a noun (*artist/artista*), a verbal utterance (*can express/pode exprimir*) and finally an indefinite pronoun (*everything/tudo*). The same order in relation to subject verb and complement is consequently respected.

(20d)

The	**artist**	**can**	**express**	**everything.**
O	**artista**	**pode**	**exprimir**	**tudo.**
subject	+	verb	+	complement

The final step described by Thunes is the generation of the target string. In a type 1 correspondence, it is necessary to look for the specific target language word forms which are necessary to replace those of the source string. At this point, the translator already knows that the order of constituents is going to remain the same and he/she has already accessed "information about lexical correspondence relations between SL and TL" (Thunes, 2011, p. 140). Another important aspect pointed out by Thunes is related to the inflexion of lexemes. When dealing with lexemes without inflection, the accessed information is considered enough to identify the correspondent word forms. The only problem is that when there is more than one form involved, extra information needs to be accessed in order to identify the correct word forms. Thunes

(2011) assumes that this subtask can be solved in linear time[33], since the required amount of information is proportional to the length and order of input, for example.

In type 1 correspondence cases, certain morphological discrepancies, like gender differences, are tolerated between corresponding lexical items:

> (…) it may be allowed within type 1 that corresponding word forms exhibit morphological differences which do not affect denotational properties, i.e. which do not influence the semantic properties of the expressions involved. (Thunes, 2011, p. 140-141)

This can be illustrated by gender marking in the language pair English-Portuguese. Portuguese has obligatory gender marking in relation to nouns and adjectives, while in English those categories are, in general, neuter. The word *teacher*, for example is neuter in English, while in Portuguese there are two forms, *professor* (male) and *professora* (female). Such cases would not interfere in the classification of type 1 correspondences.

In relation to the difference between type 1 and type 2 correspondences, what clearly distinguishes them in relation to the generation task is that type 2 requires the retrieval of the corresponding target language syntactic rules, whereas for type 1 the determination of the existence of these rules is enough to solve the task. This can be justified by the fact that as soon as the translation task is treated as type 1, it is immediately known that the translation can be generated based exclusively on the structure of the source string (Thunes, 2011).

With the purpose of summarising the different linguistic information sources required during the generation in translation tasks of type 1, Thunes proposes the following:

> In general, these sources include correspondence relations between the lexemes of SL and TL, morphological information derived from the word forms of the source string, information about the syntactic structure (which is derived from the

[33] By linear, we mean that "the sequence of word forms in the target string is already given by the word order of the same string" (Thunes, 2011, p. 140), in other words, the translator does not need to worry about the word order of the target string.

> source string and, in type 1 correspondences, shared with the target string), and information about morphological restrictions in the target language. (Thunes, 2011, p. 143)

Therefore, the three subtasks mentioned before (source text analysis, type identification, and target text generation) are used as criteria to characterise the weight of a translation task in terms of required processing effort. Table 4 summarises how Thunes (2011) defines the information required during each subtask:

Table 4. Required information for each subtask in type 1 correspondences

Type 1 correspondences	
Subtask	Required information
Source string analysis	The analysis requires sufficient lexical, morphological, and syntactic information about the source language to identify all lexemes in the source string, and to derive its constituent structure.
Type identification	It is necessary to check, first, if every lexical item in the source string has a target language correspondent with syntactic and semantic properties matching those of the source string and, secondly, if the structure is equivalent to that of the source string with respect to the linear order of constituents.
Target text generation	Identification of the correct word forms to replace each word form in the target string.

Type 2 correspondences

When compared to type 1, type 2 correspondences are considered more complex. Although it is not possible to translate word by word, the degree of complexity "is low enough to allow translation constituent by constituent" (Thunes, 2011, p. 146). Examples (21) and (22) are classified as type 2 correspondences:

(21a) Kissing with golden face the meadows green,
(21b) Beijar com face de ouro o prado verdejante,

(22a) That every word doth almost tell my name,
(22b) Que cada termo meu quase o meu nome fala,

Thunes reported that "string pairs of type 2 are not frequent with respect to the pair of languages English and Norwegian" (2011 p. 146), and, personally, I expected that this would be true for the English-Portuguese pair as well. Contradicting expectations, string pairs of this type are relatively frequent in the parallel corpus used in this study[34]. As well as in type 1, here there is still "a high degree of structural relatedness between original and translation" (ibid).

But what are the linguistic characteristics of type 2 correspondences? Since the four correspondence types are organised in a hierarchy which reflects the increase of complexity degree from type 1 to 4, type 2 correspondences consequently "are subject to the same restrictions as those applying to type 1" (Thunes, 2011, p. 147), except for two basic differences or deviations that are not tolerated in that case:

> (…) in string pairs of type 2 there may be differences between source and target string with respect to the sequence of constituents, and/or with respect to the occurrence of function words.

The first deviation can be observed in Example (21) presented previously: In (21a) the adjective *golden* precedes the noun *face*, while in the target text these lexemes are inverted. The second deviation can

[34] And this type of correspondence is very likely to be even more frequent in other types of text involving the English-Portuguese pair.

be observed in Example (22), where there is no form matching the definite article *o* (22b) in the source string (22a).

(22c)

That	**every**	**word**	**doth**	**almost**	**tell ? my name,**
Que	**cada**	**termo**	**meu**	**quase**	**fala o meu nome,**

In these types of correspondences, source and target strings are still similar, but the equivalence is not present in the entire string pair, as it could be observed in Example (22). Another aspect to be observed is that the sequence of translationally corresponding words is different. Nevertheless the requirements related to type 1 still have to be fulfilled: "source and target string have to be equivalent with respect to the assignment of syntactic functions to constituents" (Thunes, 2011, p. 147).

In order to make the distinction between types 1 and 2 clearer, Thunes adds that:

> (…) in type 2 correspondences every source string lexeme *with semantic content* must have a translational correspondent in the target string which is equivalent to the source lexeme with respect to both lexical category and syntactic function. In this connection the relevant distinction is between lexical words and function words. (Thunes, 2011, p. 147)

Function words are considered semantically light lexemes, they have little semantic content of its own and they basically indicate a grammatical relationship. Some examples are prepositions, conjunctions or articles. Lexical words like nouns, verbs, adjectives and most adverbs[35], on the other hand, are considered semantically heavy in terms of content (Thunes 2011).

> The use of function words is predictable from information about the grammatical structure of a language, and the requirements of type 2 correspondences are not violated by source-target deviations with respect to the occurrence of function words. (Thunes, 2011, p. 148)

[35] Some adverbs like *then* and *why* are considered function words.

Nevertheless the other requirements shared by type 1 and type 2 still need to be fulfilled: “the syntactic structures of respectively source and target string have to be equivalent with regard to compositionally derived semantic properties, and the two strings need to be pragmatically equivalent” (ibid). Similarly to type 1, type 2 correspondences are considered linguistically predictable translational correspondences.

Thunes reinforces that in type 2 correspondences, implications between equivalence relations should be observed according to the following information: “there is syntactic near-equivalence between source and target string which implies also semantic equivalence, which in turn implies pragmatic equivalence, between the two strings” (2011, p. 148). In those cases, similar to type 1 correspondences, it is not necessary to go through a deep linguistic analysis of the source string in order to solve the translation task.

Again, if one thinks from a computational point of view, the task can still be considered simple: It is necessary to replace the lexical items in the source string with the correspondent word forms of the target string, but some extra information concerning syntactic information of the target language needs to be computed in order to solve the task. In cases of type 2 correspondences, Google Translator would need to know that in Portuguese adjectives usually appear after the nouns, different from English, when the adjectives always precede the nouns. If one uses Google Translate to translate the sentence “The black cat is chasing the rat”, a type 2 correspondence would be obtained:

Figure 2. Example of type 2 correspondence produced by Google Translator, from a computational point of view.

In relation to the structure of translation task, type 2 correspondences are considered linguistically predictable, which means that linguistic information taken from a prestructured domain is sufficient to solve them.

> The structure of the translation task is similar to that of type 1 correspondences, but somewhat more complex since it involves computing certain minor structural differences between source and target string. (Thunes, 2011, p. 148)

In type 2 correspondences, the subtask of analysis "involves the same kind of parsing task as the analysis step in type 1 correspondences does" (Thunes, 2011, p. 148) and it also requires the same types of information that were previously described. Thus, at this level, it is not possible to distinguish the translational complexity of these correspondence types.

The subtask of type identification is directly related to the result of the previous task (source string analysis). Just like in type 1 correspondences, the type identification is based on "the amount of

bilingual information present in the constituent structure derived for the source string" (Thunes, 2011, p. 149).

Finally, the generation of the target string "requires a constituent structure in order to compare the linear sequence of surface word forms – this holds for all four types of translational correspondences" (Thunes, 2011, p. 150). But at this level, there is a significant difference between both types:

> (…) with respect to the subtask of generating the target string, types 1 and 2 differ in the sense that while generation in type 1 cases can be based directly on the constituent structure of the source string, generation in cases of type 2 requires also some processing of syntactic information specific to the target language. But to the extent that syntactic structure is shared between the source string and the corresponding rules of the TL grammar it is unnecessary to derive again syntactic structure already identified by the analysis of the source string. (Thunes, 2011, p. 150)

During the generation subtask, in addition to identifying the correct forms which replace each word form in the target string, "it is necessary to retrieve the information given by the relevant syntactic rule(s) of the target language grammar" (Thunes, 2011, p. 150). According to Thunes, these TL grammar rules are important because they also clarify divergences related to function words:

> (…) either the generation of the target string requires introducing a function word not found in the target string, or a certain function word occurring in the source string is *not* matched by a function word in the target string, and these facts will follow from syntactic information about the target language. (Thunes, 2011, p. 150)

In conclusion, in type 2 correspondences, "the task of identifying the correct target word forms requires the same kinds of information (…) needed in type 1 cases" (Thunes, 2011, p. 150). Considering the restrictions that were previously discussed, there are two possibilities in relation to the words identified in the target string: They "will either be TL-specific function words or words which correspond translationally to

the lexemes identified in the source string according to the same restrictions as those applying to lexical correspondences in type 1 cases" (ibid).

In relation to the generation task, the difference between type 1 and type 2 is that in the latter there is a point "where solving the translation task demands a larger amount of information" (Thunes, 2011, p. 151). While the need for information in the previous subtask is on the same level as in type 1 correspondences, in terms of generation "information about how source and target must be structurally different and about how the correct target structure is derived" (ibid) is also required.

Therefore, just as it was done in the description of type 1 correspondences, the subtasks related to type 2 (source text analysis, type identification and target text generation) will also be used as criteria to characterise the weight of translation task in terms of required processing effort. Table 5 summarises how Thunes (2011) defines the information required during each subtask. Observe that

> with respect to the amount of effort needed in order to access and process the necessary information sources, the requirements of type 2 are mostly the same as those of type 1, but differ on one point, reflecting how the two types vary with respect to the structure of the translation task". (Thunes, 2011, p. 151)

Table 5. Required information for each subtask in type 2 correspondences

Type 2 correspondences	
Subtask	Required information
Source string analysis	The analysis requires sufficient lexical, morphological, and syntactic information about the source language to identify all lexemes in the source string, and to derive its constituent structure.
Type identification	It is necessary to check if every lexical item in the source string has a target language correspondent with syntactic and semantic properties matching those of the source string and also evaluate the differences with respect to the linear order of constituents.
Target text generation	Identification of the correct word forms to replace each word form in the target string and identification of how source and target strings are structurally different to derive the correct target structure.

Type 3 correspondences

According to Thunes (2011), type 3 correspondences are the second most complex type of translational correspondences. In those cases, the target text still conveys the same meaning of the source task, but the structure violates the restrictions established for type 1 and type 2 correspondences (Thunes, 2011).

In type 3 correspondences there are "greater structural discrepancies between source and target string than in those of type 2" (Thunes, 1998, p. 27). If we observe Example (23) extracted from Sonnet LXXVI, it is possible to see that there are some lexical words in

the source string that do not have a corresponding word of the same lexical category and with the same syntactic function in the target string:

(23a) Why is my verse so barren of new pride?
(23b) Por que meus versos são tão nus de ornato novo?

Why is considered a function word, and it corresponds to two lexemes in Portuguese (*por que*), which characterises the first deviation if compared to type 1 and type 2 correspondences. The word order is also different, given the grammatical structure of wh questions in English. While in English the order of constituents is adverb + verb + possessive pronoun + noun + adverb + adjective + preposition + adjective + noun, in Portuguese the order is adverb + possessive pronoun + noun + verb + adverb + adjective + preposition + noun + adjective.

The linguistic characteristics of type 3 correspondences can be defined as follows:

> (…) in a string pair of type 3 it is the case that for at least one lexical word in one of the strings there is no correspondent in the same string of the same category and/or with the same syntactic function as that lexical word. Source-target divergences of this kind will cause greater differences in constituent structure between source and target string than the differences allowed within type 2 correspondences, but they must not violate the requirement of semantic equivalence between original and translation. (Thunes, 2011, p. 155)

In type 3 correspondences, there need to be equivalence between source and target strings in relation to the sets of expressed predicates and arguments and the way that they relate to each other. But differences in relation to syntactic functional equivalence are tolerated, what is not accepted in type 2 correspondences. The characteristic shared by types 1, 2, and 3 is semantic correspondence, that means that "the same information content is linguistically encoded in the source string, as well as in the target string" (Thunes, 2011, p. 56). Thunes considers this a central principle of her analytical framework and, thus, she summarises the types of correspondence as follows:

> (…) in translational correspondences of type 3 we do not find, as in types 1 and 2, syntactic

> functional equivalence between source and target string. But in order to fall within type 3, source and target string must be equivalent with respect to compositionally derived semantic properties, and in the given texts they must be pragmatically equivalence in the sense of being used to perform corresponding pragmatic functions, or speech acts. The structural divergences between source and target text in type 3 correspondences show that the degree to which there exist implicational relations between equivalence relations on different linguistic levels is smaller in translational correspondences of type 3 than in those of lower types. (Thunes, 2011, p. 156-157)

Given this definition, in type 3 correspondences source and target strings have differences on the structural level but semantically speaking they are still correspondent. Consequently, this implicates in pragmatic equivalence between the texts, placing type 3 correspondences among the linguistically predictable translations, just like types 1 and 2 (Thunes, 2011).

If these correspondences are linguistically predictable, they are also "solvable within the domain of linguistic information, as it is not necessary to process extra linguistic information or information from the textual context of the given translation task in order to generate a semantically and pragmatically equivalent target task" (Thunes, 2011, p. 157). The translation task is considered more complex if compared to types 1 and 2 because there are more structural differences that need to be solved. When compared to the lower types, the process of type identification is very similar to what was described so far, but the need for information in the subtasks of analysis and generation is more significant (Thunes, 2011).

The first subtask of analysing the source string is equal for all correspondence types. Consequently, type 3 correspondences require exactly the same types of information of the lower types: "sufficient morphological and syntactic information to identify all lexemes in the source string and to derive its constituent structure" (Thunes, 2011, p. 157).

The subtask of type identification depends on the result obtained during the previous stage, that is, the analysis task. At this point, the required information to solve type identification is the amount of bilingual information linked with the constituent structure which in turn

is derived for the source string (Thunes, 2011). Once a type 3 is identified during the translation task, the following step is to derive information related to the semantic structure of the source text: This is necessary to compute the target string.

> (…) a semantic representation of the source string must be produced, and this is derived compositionally from the syntactic representation together with semantic information associated with the lexemes identified in the source string. (Thunes, 2011, p. 158)

In order to derive this semantic representation, three specific types of information are considered necessary. First, "information about the constituent structure of the source string" (Thunes, 2011, p. 158); second, information about "the assignment of syntactic functions to constituents" (ibid); and finally, information about "any components of meaning encoded linguistically in the source text (e.g. predicate-argument relations, spatial and temporal relations)" (ibid).

According to Thunes, generation task of type 3 correspondences differ from the lower types in the sense that:

> (…) the generation of the target string in cases of type 3 must be based on information about the semantic structure of the source string because type 3 correspondences involve structural source-target divergences of a kind that is qualitatively different from those found in type 2. (Thunes, 2011, p. 159)

Since the task of generation is related to a semantic representation of the target string, it involves the necessity of making choices within the scope of the entire lexicon and grammar of a given target language. During this process, "lexical units and grammar structures are not selected independently of each other in natural language generation, as there are always close interconnections between meaning and structure in linguistic expressions" (Thunes, 2011, p. 160). Thus, the objective of this selection is "to extract elements of the TL lexicon ad grammar in order to cover all of, but no more than, the components of meaning contained in the semantic representation of the source text" (ibid). Therefore, Thunes assumed that the necessary information to generate the target text is totally provided by the semantic representation.

In relation to the weight of translation task in terms of required processing effort, just as before, Table 6 summarises how Thunes (2011) defines the information required during each subtask.

Table 6. Required information for each subtask in type 3 correspondences

Type 3 correspondences	
Subtask	Required information
Source string analysis	The analysis requires not only syntactic parsing, but also a semantic analysis of the source string. It is necessary to process a larger amount of the source language information available prior to translation.
Type identification	It is necessary to check if every lexical item in the source string has a target language correspondent with syntactic and semantic properties matching those of the source string and also evaluate the differences with respect to the linear order of constituents.
Target text generation	It is more demanding to access the required information because generating the target string from the semantic representation of the source string involves a number of choices for which the space is the entire TL language description.

Type 4 correspondences

These types of correspondences constitute the most complex class of translational correspondences in Thunes' hierarchy of correspondence types. Type 4 correspondences represent translation tasks where there are linguistic differences between the string pair which violate the rules established for the other correspondence types (1, 2, and 3). The major

difference is that the divergences exist not only on the structural, but also on the semantic level (Thunes, 2011).

Thunes affirms that "with respect to the language pair English-Norwegian, type 4 cases represent the most frequent class of translational correspondence" (2011, p. 165) and I hypothesize that this will be true for the language pair English-Portuguese due to specific characteristics intrinsic to the process of poetry translation[36]. If translational correspondences of type 4 "are cases where there are discrepancies between original and translation not only on the structural, but also on the semantic level" (Thunes, 1998, p. 28), they represent cases "where we cannot derive equivalent semantic representations for source and target string" (ibid). Example (24) is an instance of a type 4 correspondence because it contains a mismatch on the semantic level:

(24a) Yet, do thy worst, old Time: despite thy wrong,
(24b) Faze o pior, porém: malgrado o teu rigor,

Besides the mismatch on the semantic level, "divergences between source and target [strings] violate the restrictions on types 1, 2, and 3" (2011, p. 165). There is a difference in the word order of the constituents; at least one lexical item in the source string does not have a correspondence in the target string: The lexemes of the expression *old Time* belong to the class of content words and do not have any correspondent in the target string. *Thy* has a correspondent word that belongs to the same grammatical category (*teu*), but the words *malgrado* and *rigor* do not correspond semantically to *despite* and *wrong*, respectively. Observe in Example (24) the differences in relation to word order and the lack of correspondence. The symbol ɵ represents the lack of correspondent lexemes in the strings:

(24c)

Yet	**do**	**thy**	**worst**	**old**	**Time:**	**despite**	**thy**	**wrong,**
Porém	**faze**	**o**	**pior**	**ɵ**		**malgrado**	**o teu**	**rigor,**

Consequently, just by observing this example it is possible to conclude that linguistic characteristics of type 4 are different from the lower types:

[36] This will be discussed in detail during the discussion of the results.

> (…) source and target string are not equivalent with respect to constituent structure as in type 1 cases, or they are not equivalent with respect to the assignment of syntactic functions to constituents as in type 2, or they are not equivalent with respect to compositional semantic properties. (Thunes, 2011, p. 165-166)

Once assumed that type 4 translational correspondences differ from the other types, as a consequence they "are not solvable within the pre-structured domain of linguistic information, and the need of information required to translate is larger in type 4 correspondences than in any of the other types" (Thunes, 2011, p. 167). The need for this increased amount of information can be observed especially during the subtasks of analysis and generation.

If the prestructure domain of linguistic information is not considered enough to solve these correspondence types, they are not computable as well. The noncomputability happens because

> (…) there is no principle for delimiting a representation of the information sources lying outside the pre-structured domain, granted that our scope is the translation of general language, and not translation within a restricted semantic area. Thus, there is no principled limit on the amount and types of information that could be needed to solve a task of type 4. (Thunes, 2011, p. 167)

Because of these characteristics, Thunes treats type 4 as cases that depend on human translation, since the translator has the capacity of collecting all the information required by the task. The translator has the possibility of enlarging the textual context, by searching extra background information of several types, or he/she can even ask other translators for help to eventually produce a target text. Considering these aspects, translations tasks of type 4 are considered translatable, but definitely they are not computable (ibid).

So far it was assumed that translations types 1, 2, and 3 were computable. Type 4, differently from the other lower types is noncomputable, it needs to be solvable exclusively by humans. Consequently, the translation task must be described in a different way. But just like in Thunes' study, the focus is not to study the human

translation process. In order to cope with the different nature of type 4 correspondence, she proposes that the descriptive approach should not be altered in relation to the structure of type 4 tasks, but the focus must be on the aspects that place those cases outside the computable domain (Thunes, 2011).

With respect to the subtask of analysis, correspondences of type 4 require the same kinds of linguistic information as those described for type 3 with the purpose of creating a constituent structure and semantic representation of the source string. But differently from the lower types, "an understanding of the source string which goes beyond a syntactic and semantic analysis" (Thunes, 2011, p. 168) is necessary. This task also demands "sources of information included neither in the pre-structured domain of linguistics information nor in the information that is explicitly encoded in the linguistic form of the source string" (ibid).

In cases of type 4, if one observes the source string alone, it is not possible to identify precisely the extra information necessary to solve the task. Thunes presents some examples of possible additional information sources that might be required during the process:

> (…) general information about the world, domain-specific technical information, task-specific linguistic information about reference relations, as well as task-specific extra-linguistic information about the utterance situation of the source text, and about the described situation of the source text. (2011, p. 168)

In relation to the subtask of type identification, just like in the other types, Thunes assumed that its solution is closely related to the result of the analysis task. During the analysis of the lower types, the analysis itself provides "information about the translational properties, with respect to the target language, of the linguistic items identified in the source string" (Thunes, 2011, p. 169). Another important aspect related to type 4 is that in those cases, there are two possibilities related to the translator's choices:

> (…) either the translator has chosen a target string deviating semantically from the source string although a literal translation could have been produced, or the analysis will reveal that for at least some subpart of the source string there is no

> linguistically predictable correspondence in the specific target string. (Thunes, 2011, p. 169)

During the generation task additional sources of information are required. First, information about the semantic structure of the source string, combined with information about the semantic differences between source and target strings, and second, information about the lexicon and grammar of the target language. Therefore, besides the subtask of analysis, at least one of the pieces of information presented above is considered necessary. Similarly to type 3 cases, the generation is a matter of selecting the most appropriate lexemes and structures in order to create the target string. (Thunes, 2011) An important distinction to be established here is that

> in type 3 this is done by choosing elements of the TL lexicon and grammar in order to cover all of, but no more than, the components of meaning contained in the semantic representation of the source string. In type 4 additional information must contribute to deciding which of those semantic components of the source string that are expressed in the target string, and which are not – as well as which components, if any, that are expressed instead. (Thunes, 2011, p. 169)

In order to summarise the weight of the translation task in type 4 in terms of processing effort, Table 7 shows how Thunes (2011) defines the information required during each subtask. With respect to the subtasks of generation and analysis, Thunes emphasises that because some of the information considered necessary to translate a type 4 correspondence is not part of the finite domain of prestructured linguistic information, the size of the search job required to compile the necessary information has no limits at all. She also stresses that within the approach to translation complexity adopted by her there is no framework available that can cope with the description of the quantity of computational resources necessary to access and process this additional information.

Table 7. Required information for each subtask in type 4 correspondences

Type 4 correspondences	
Subtask	Required information
Source string analysis	The analysis requires not only syntactic parsing, but also a semantic analysis of the source string. It is necessary to process a larger amount of information prior to translation. It also requires additional information that is not available in the finite domain of linguistic information sources.
Type identification	It demands no more effort than in the lower correspondence types.
Target text generation	It is more demanding to access the required information because generating the target string from the semantic representation of the source string involves a number of choices for which the space is the entire TL language description. It is also necessary to access additional information that is not available in the finite domain of linguistic information sources.

Type 4 tasks are definitely treated as cases where human translation is indispensable, but the effort required by a human to solve a specific translation task will be equivalent to his/her individual competence as a translator (Thunes, 2011).

All the string pairs were analysed according to the approach that was detailed in this section. Initially, all the string pairs of the parallel corpus were classified according to the four types of translational

correspondences suggested by Thunes (2008, 2011). In total, 1,290 translation pairs[37] were analysed and annotated manually.

3.5 Methodological principles

In the method applied in the present study, the string pairs were extracted from parallel texts and classified according to the measure of translational complexity as defined by Thunes (2011). The process of linguistic analysis and subsequent annotation of all the string pairs was done manually by a bilingually competent human annotator.

The concept of 'translational correspondence' just like in Thune's study "covers a pair of translationally related linguistically units of two different languages" (2011, p. 193).

The main difference between the methodological principles adopted here is related to the definition of the unit of translation. Thunes adopted very complex syntactic criteria to identify translational units in the parallel texts. The choice of a limited set of syntactic units was necessary because one of the purposes of her study was to obtain results that could be relevant to the field of machine translation. As pointed out before, the present investigation does not aim at obtaining such results. Because of that, a detailed criterion to establish the unit of translation was not necessary. As the alignment process showed, most of the verses could be easily matched. Consequently, the choice of the verse as the translation unit seemed to be the best option, and this was confirmed during the annotation process.

3.6 An example of the annotation process

The aligned Sonnet (25) represents how the parallel corpus was annotated. Each string pair was analysed and classified according to one of the four translational types described in the previous section (3.3).

[37] Considering that each aligned sonnet contained 28 translation pairs, since all the sonnets had 14 verses and two respective translations. Only one of the sonnets (XCIX) contained 15 verses. Thus, 28 translation pairs x 45 sonnets + 30 = 1,290.

Once the type was identified, its identification was typed next to the respective target string (verse), as observed in the annotated Sonnet V:

(25)
Those hours, that with gentle work did frame
Aquelas horas que formaram meigamente type 3
As horas que formaram gentilmente type 3
The lovely gaze where every eye doth dwell,
Teu aspecto gentil, que todo olhar procura, type 3
Tua feição, que todo olhar procura, type 3
Will play the tyrants to the very same
Hão de tiranizar-te ainda amargamente, type 4
Serão tiranas, quando, indiferentes, type 4
And that unfair which fairly doth excel:
Desformoseando o que é sem par em formosura: type 4
Ao formoso roubarem formosura: type 4
For never-resting time leads summer on
Ah! pois o tempo sem descanso leva estio type 3
Que o tempo é sem repouso, e do verão type 4
To hideous winter and confounds him there;
Ao coração do inverno odioso, onde o oblitera; type 4
Leva ao inverno – e ali verão se esgota, type 4
Sap cheque'd with frost and lusty leaves quite gone,
As tenras folhas vão-se, a seiva, entanca-a o frio; type 3
Gelada a seiva; e na devastação, type 4
Beauty o'ersnow'd and bareness every where:
Jaz nevada a beleza, e a desnudez impera: type 3
- O belo sob a neve e as folhas mortas ... type 4
Then, were not summer's distillation left,
A liquid prisoner pent in walls of glass,
Se entre muros de vidro o estio destilado type 4
Então não perdurasse, olente prisioneiro
Se a essência do verão não se guardasse type 4
E líquida prisão, muros de vidro,
Beauty's effect with beauty were bereft,
Nor it nor no remembrance what it was:
Ter-se-ia da beleza extinto o resultado, type 4
Sem memória deixar de seu fulgor primeiro type 4
Havia de findar, sem que o lembrassem, type 4
O efeito da beleza, desvalido. type 4
But flowers distill'd though they with winter meet,
Mas em vão, destilada a flor, o inverno a ameaça: type 3
Mas a essência da flor, chegado o inverno, type 4
Leese but their show; their substance still lives sweet.
Perdida a forma, em sua essência ela não passa. type 4

E extinta a forma, resta um bem eterno. type 4

All the sonnets of the parallel corpus were annotated according to the procedure described in this section. The results and discussion of the process of annotation are presented in Chapter 4.

Chapter 4

Discussion and Results

This chapter is divided into four sections. In the first section, translational complexity across the data is presented based on the distribution of the four correspondence types along the corpus. In the second section, I explain how the correspondence types can be used to analyse differences between translators in relation to individual style.
In Section 4.3, examples of correspondence types identified across the data are discussed and the creation of subtypes is explained. Finally, in Section 4.4, the translators' styles are discussed based on the correspondence types identified in the parallel corpus.

4.1 Translational complexity across data

Table 8 shows the amount of translational correspondence types identified in the whole corpus:

Table 8. Correspondence types across data

Correspondence types	Number of correspondences identified in the parallel corpus	%
Type 1	1	0.07%
Type 2	78	6.04 %
Type 3	537	41.62%
Type 4	674	52.24%

As expected, since literal translations tend to be rare in poetry translation, types 1 and 2 are less frequent in the corpus. There was just one occurrence of type 1 correspondence in the whole corpus, while type 2 correspondences represent only 6.04% of the cases. Type 3 correspondences represent 41.62 % of the cases while type 4 is the most frequent type, representing 52.24% of the cases.

During the analysis, type 3 and type 4 were respectively divided into two subcategories. These categories will be discussed in Section 4.3. Table 9 shows the amount of correspondence types identified in each translation:

Table 9. Correspondence types in each translation

Correspondence types	Translation by PS	Translation by JW	Total
Type 1	1	0	1
Type 2	60	18	78
Type 3.1	139	189	328
Type 3.2	174	35	209
Type 4.1	55	49	104
Type 4.2	216	354	570
Total of string pairs	645	645	1,290

As observed in this table, there is a gradual increase in the amount of correspondence types of greater complexity across the data. Interestingly, the amount of type 2 and type 3 correspondences in PS's translation is significantly higher when compared to JW's translation. This fact already signalizes that there are differences between the two translators in terms of writing style. These results are discussed in detail in Section 4.4.

4.2 Relating correspondence types with the translator style

The amount of correspondence types in a corpus can be considered both an indicative of how the language pairs are related in

terms of grammar and syntactic similarities and the translators' writing styles. In relation to the first issue, it is expected that the lower types of correspondence be very frequent in a parallel corpus of languages that pertain to the same family (like Spanish and Portuguese, for example) and share structural similarities. On the other hand, if there are many structural differences, the lower types might be rare and types 3 and 4 must be predominant. This is not the main focus of this specific study, but the results might point, to some extent, to the degree of similarity between English and Brazilian Portuguese on the structural level. In relation to writing styles, some of the translators' choices can be explained based on the correspondence types identified in their translations. Subsequently, I will propose how this relation between correspondence types and styles can be traced.

As mentioned before, "type 1 correspondences are cases of a full linguistic match, structurally as well as semantically, between source and target string" (Thunes, 2011, p. 125). Consequently, the massive occurrence of type 1 correspondences in a parallel corpus might indicate first, that there are structural similarities between the languages systems that allow such constructions, and second, that the translator tried to build a target text as similar as possible to the source text in terms of structure and content. This is not expected to happen in poetry translation, since there are many other aspects involved in this task. For instance, if metre is considered, it is not expected that type 1 correspondences will be produced, because the number of syllables needs to be accounted during the process, and this might result in the addition or reduction of words in the target string. Sometimes there is a word in the target language that corresponds exactly to the meaning of the source language; it might even be a cognate, but it cannot be used because its number of syllables does not fit into the metre chosen. If we observe Example (26), it is clear that among the sets of possible translations for the word *woe* (*aflição, angústia, preocupação, pena, pesar, dor, mágoa*)[38] the translator chose the shortest one probably to conform to the metre of the verse:

(26) Sonnet XC

And other strains of woe, which now seem woe,
E outras formas de dor, que ora parecem dor, → type 1

[38]http://michaelis.uol.com.br/moderno/ingles/index.php?lingua=inglesportugues&palavra=woe

Curiously but not unexpectedly, this was the only occurrence of type 1 correspondence identified in the corpus.

Type 2 correspondences are similar to type 1 in relation to the aspects discussed above, but they "allow minor mismatches on the structural level, but none on the semantic" (Thunes, 2011, p. 125). The occurrence of type 2 correspondences in a parallel corpus might indicate that the translator attempted to build a target text as similar as possible to the source text, but he/she had to care about other elements (like word order, for example), in order to produce the adequate translation. Word order is an important aspect in poetry because it influences the rhythm and musicality of the verses. Word order also might influence metre, since in iambic hexameter verses, for example, the stress in on the pair syllables. Example (27) would be an instance of type 1 correspondence if the same word order were maintained*, but PS decided to change the order of lexemes:

(27) Sonnet CV

Kind is my love today, tomorrow kind,
**Afável é meu amor hoje, amanhã afável,*
Meu amor hoje é afável, amanhã afável, → type 2

With respect to type 3 correspondences, "there can be major structural divergences while there is still a semantic match" (Thunes, 2011, p. 125). In those cases, the translator's objective was probably to convey a message similar to the one conveyed by the source text, but not necessarily keeping the same structure of the original text. If we observe Example (28), it is visible that a literal translation* would be possible, but the translator chose a different construction:

(28) Sonnet XV

When I perceive that men as plants increase,
**Quando eu percebo que os homens como plantas crescem*
Se os homens sei que como as plantas arborescem → type 3

Here, the translator felt free to change the meaning of some words during the translation process, something that does not happen when he/she produces type 1 or type 2 correspondences. It is clear that the literal translation would not be suitable here because of the number of

syllables. The literal translation *Quando eu percebo que os homens como plantas crescem* contains more than twelve syllables. This would not fit the metre chosen (iambic hexameter). It is clear that the choice of words here is not only a matter of freedom[39], it is also directly related to the metre chosen.

In type 4 correspondences there are semantic as well as structural mismatches between source and target string (Thunes, 2011, p. 125). These mismatches may exist for two different reasons that are going to be discussed in detail in the next section (see 4.3). The first possibility is actually the one described by Thunes: Cases "where we cannot derive equivalent semantic representations for source and target string" (Thunes, 1998, p. 28). Example (29) represents those cases (henceforth classified as type 4.1), where the exact meaning of the source text cannot be conveyed:

(29) Sonnet V

And that unfair which fairly doth excel:
Desformoseando o que é sem par em formosura: → type 4.1
Ao formoso roubarem formosura: → type 4.1

Differently from Example (29), where a literal translation is not possible and it seems difficult to convey the same meaning, in (30) it would be possible to build a similar construction[40] able to keep the meaning, but both translators chose to create target texts that differ from source text on structural and semantic levels[41].

(30) Sonnet XVIII

Thou art more lovely and more temperate
Vencendo-o em equilíbrio, és sempre mais amável: → type 4.2
Tens mais doçura e mais amenidade: → type 4.2

Therefore, occurrences classified as type 4 in the parallel corpus, might indicate that the translator created a verse different from the original for

[39] Freedom of choosing vocabulary.
[40] Por exemplo, uma possível tradução seria "*Tu és mais amável e mais comedida*". This translation would even conform to the iambic hexameter.
[41] Cases henceforth classified as type 4.2.

two distinct reasons: The first possibility is that it was not possible to convey a message similar to the one conveyed by the source text due to structural differences between the two languages. The second possibility is that he/she felt free to build a construction that seemed to fit the poem better, despite the semantic differences in relation to the source text.

Based on the assumptions described above, I traced a parallel between Thunes' model of complexity and how its analysis might explain differences between distinct translators' writing styles. Table 10 summarises what probably happens during the translation process when a certain correspondence type is produced:

Table 10. Relation between Thunes' model and the translators' styles

A correspondence types *x* is produced	**when …**
Type 1	The translator tries to build a target text similar (or even identical) to the source text in terms of structure and content.
Type 2	The translator attempts to build a target text as similar as possible to the source text in terms of structure and content.
Type 3	The translator does not attempt to build a target text similar to the source text in terms of structure and he/she allows himself/herself to express meaning in a more flexible way.
Type 4	The translator creates a target text different from the original either because it is not possible to convey a message similar to the one conveyed by the source text or because he consciously chooses to build a verse that fits the poem better, despite the semantic differences in relation to the source text.

Despite the fact that translators do not bear in mind the type of correspondence that they are producing, the classification proposed by Thunes defines clear criteria that can be used to approach writing styles of different translators. It focuses on the structural and semantic level and it can be a good option for researchers who do not want to go through an interpretative or literary analysis, which is usually more subjective.

4.3 Discussion of correspondence types identified in the parallel corpus

This section presents some examples extracted from the parallel corpus that illustrate the major difficulties related to the annotation process. It also suggests an adaptation of Thunes' model, so that it can be used to analyse any parallel corpus of poetry written in Brazilian Portuguese and English.

The examples of each correspondence type will be presented in subsections. The symbol ɵ used in the analysis indicates a spot where there is no correspondence between the strings. In other words, it represents a lexical item from one of the strings that lacks correspondence in the other string.

4.3.1 Type 1 correspondences

As pointed out in Chapter 2, "translational correspondences of type 1 are cases of word-by-word correspondences" (Thunes, 1998, p. 25). I did not expect to find any examples of this correspondence type in the parallel corpus, even knowing that these correspondences are possible for the language pair English-Portuguese. As mentioned before, one correspondence of this type was identified in Sonnet XC (Example 31) translated by PS. The translation contains the same number of words of the original, and there is one correspondent word to each word of the source text. Each lexeme from source string has a syntactically and semantically matching TL correspondent. Although *now* was not literally translated as *agora*, the translation *ora* plays the role of adverb and the meaning of the whole string is maintained:

(31) Sonnet XC

And other strains of **woe**, which now seem **woe**,
E outras formas de **dor**, que ora parecem **dor**, → type 1

Correspondences of this type, thus, represent only 0.07% of the correspondences identified in the parallel corpus.

4.3.2 Type 2 correspondences

Correspondences of type 2 represent cases where almost all the structural elements can be translated word by word, except for differences between source and target string related to word order and/or the presence of grammatical function words (Thunes, 1998). Table 11 contains examples of grammatical function words. Different from content words, that have their own meaning, function words have little meaning on their own, because they are used to create grammatical or structural relationships into which the content words may fit.

Table 11. Function words

Function Words	**examples**
Prepositions	*of, at, in, without, between*
Pronouns	*he, they, anybody, it, one*
Determiners	*the, a, that, my, more, much, either, neither*
Conjunctions	*And, that, when, while, although, or*
Modal verbs	*can, must, will, should, ought, need, used*
Auxiliary verbs	*be (is, am, are), have, got, do, does (doth)*
Particles	*no, not, nor, as*

(Available at: http://psychol.ucl.ac.uk/transcription/intro.html)

In Example (32), the target string contains one correspondent word to each lexeme of the source string, except from the preposition *de*. Since differences regarding word order and the use function words

are tolerated in type 2 correspondences, string pair (32) was classified as a type 2 correspondence.

(32) Sonnet XXIII

The perfect ceremony of love's rite,
A perfeita cerimônia em (de amor) ritual
A cerimônia exata em ritual de amor, → type 2

In many other target strings identified in the corpus, word order was the only structural difference between source and target strings. In (33), the adjective *prophetic* is placed before the noun *soul*, conforming to the rule of adjective position in English. In Portuguese, it is possible to place the adjective before the noun, but this is not very usual. PS placed the adjective *predizente* after the noun *alma*, and this difference led to a type 2 correspondence.

(33) CVII

Not mine own fears, nor the **prophetic** soul,
Nem meu próprio temor, nem a alma **predizente** → type 2

Similarly to Example (33), in (34) the correspondent of *most heinous* is placed after the correspondent of *crime* in the target string:

(34) Sonnet XIX

But I forbid thee one **most heinous** crime,
Porém eu te proíbo um crime **mais perverso**: → type 2

In Example (35) the difference of word order is not related to the position of adjectives, but the subject of the sentence is placed in initial position in the translation by PS, while in the original verse it appears after the verb *is*:

(35) Sonnet CV

Kind is **my love** today, tomorrow kind,
Meu amor hoje é afável, amanhã afável, → type 2

String pair (36) is another example of type 2 correspondence. Here the determiner *os* has no correspondent in the source string. Considering that *os* is a function word, this would not prevent the string from being classified as type 2. Syntactically speaking, *both* and its respective translation *os dois* work as objects of the transitive verbs *defy* and *desafiar*:

(36) Sonnet CXXIII

Thy registers and thee I **both** defy,
Teus registros e tu, **os dois** eu desafio; → type 2

The classification of string pair (37) raised two issues: One related to grammatical category and another related to semantic content.

(37) Sonnet LXXIII

As **the** death-bed, whereon **it must ɵ expire**,
Como **em** leito final onde ɵ **haja de expirar**, → type 2

In relation to grammatical categories, the first thing observed was that the determiner *the* was translated into the preposition *em*. In principle it could be considered a deviation that would change the classification of the string pair to type 3. But since both determiners and prepositions are function words and this difference is tolerated according to the description of type 2 correspondences the string pair was still classified as such. Another difference is that the pronoun *it* has no correspondent in the target string while the preposition *de* has no correspondent in the source string. Considering that pronouns and prepositions are function words and in Portuguese the former can be omitted in many cases, this would not implicate in a deviation of type 2 correspondence requirements. In relation to semantic content, the literal translation of death-bed would be *leito de morte*, but PS chose the expression *leito final*. Since there is equivalence between these two expressions, I concluded that his choice did not alter the semantic content of the translated verse when compared to the original.

Another aspect to be considered when a type 2 correspondence was identified was in relation to the infinitive form of verbs. In Portuguese, the infinite is represented by only one lexeme, while in English the infinitive form is indicated by the function word *to* that

signalises that the following verb is an infinitive[42]. Thus, cases like Example (38) which contained verbs in the infinitive form were classified as type 2, despite of the fact that *to* had no correspondent in the target string. This decision is coherent with the criteria proposed by Thunes (1998) who states that differences between source and target string with respect to the use of function words in type 2 correspondences can be tolerated.

(38) Sonnet XV

To change your day of youth to sullied night,
Mudar teu jovem dia em noite desluzente. → type 2

Although type 2 correspondences are not difficult to be identified, in some cases, like Example (39), it was not easy to decide whether it should be classified as type 2 or type 3.

(39a) Sonnet XVIII

Shall I compare thee to a summer's day?
A um dia de verão como hei de comparar-te? → type 2?

If we dismember the target string, it is easy to visualise that it contains one correspondent to each content word of the source string, except for the prepositions *de* which are function words.

(39b) Sonnet XVIII

Shall	I	ɵ	compare	thee	to	a	summer's day?
Como	hei	**de**	comparar-	te	a	um	dia **de** verão?

In terms of syntactic structure there are similarities even in relation to the order of constituents. But considering that type 2 correspondences "must be equivalent with respect to the assignment of syntactic functions to constituents, and that all lexical words in the source string must have a target correspondent of the same category and with the same syntactic function" (Thunes, 2011, p. 153), the translations of the lexemes *Shall* and *I* violate this rule. The modal verb *shall* was translated as *como*, which in this specific case is an adverb. The pronoun

[42] http://www.merriam-webster.com/dictionary/to

I was replaced by *hei*, which is the first person conjugation (singular) of the verb *haver*. Given that these lexical words (shall/I) do not have correspondents of the same category and with the same syntactic function, the string pair in Example (39) cannot be considered an instance of type 2 correspondence. Since they can be considered correspondent in terms of meaning, this would be an example of type 3 correspondence.

Some methodological decisions were also necessary to be taken in relation to the translation of adverbs. In Example (40), there is one correspondent to each lexeme in the target string, but the adverb *sometime* corresponds to the adverbial locution of time *às vezes* in Portuguese. Similarly to the infinitive form of verbs, differences regarding the translation of adverbs were tolerated for type 2 correspondences. Consequently, string pair (40) was considered an example of type 2 correspondence:

(40) Sonnet XVIII

Sometime too hot the eye of heaven shines,
Às vezes, muito quente, o olho do céu fulgura, → type 2

Cases where the only difference between source and target string was the omission of a pronoun were still labeled as type 2 correspondences, since pronouns are considered function words. In Example (41), the only deviation is that the archaic pronoun *thou* does not have a correspondent in the target string. The pronoun *tu* could be added before the verb *cresces*, but PS omitted it in his translation to avoid redundancy[43]:

(41) Sonnet XVIII

When in eternal lines to time **thou** grow'st:
Quando em verso imortal, no tempo ɵ cresces. → type 2

Thus, all cases identified in the corpus where the only deviation between the strings was the omission of the pronoun were classified as type 2.

In (42), there is one difference in relation to word order. While the adjective *linda* (*seemly*) was placed before the noun *veste* (*raiment*) in the translation (obeying the same order of the source string), the

[43] In Portuguese, the presence of the pronoun *tu* would indicate redundancy.

adverb *apenas* was placed right after the lexeme *raiment*, differently from the original where the adverb *but* appears after the verb *is*:

(42) Sonnet XXII

Is **but** the seemly raiment of my heart,
É a linda veste, **apenas**, de meu coração,→ type 2

The literal translation of the first verse of Sonnet XXII (43) would be *Senhor do meu amor, a quem em vassalagem*. Structurally speaking, the only difference between the strings is that there is no correspondent to the preposition *in* in the target string. In terms of semantics, *apreço* singly does not seem to be the most adequate translation for *love*, but considering the context, the expressions *lord of my love* and *senhor do meu apreço* share the same meaning. Since *in* is a function word and both strings share the same semantic content, the target string was also classified as type 2.

(43) Sonnet XXVI

Lord of my love, to whom in vassalage
Senhor de meu apreço, a cuja ɵ vassalagem → type 2

Finally, Examples (44) and (45) contain mismatches in relation to word order and the use of grammatical function words. In (44) there is a difference in relation to the order of constituents and the lexeme *te* has no correspondent in the source string. In (45) there is also a difference in relation to the order of constituents and the pronoun *I* was omitted in the target string:

(44) Sonnet LXXI

Nay if you read this line, **remember not**
Não, **não te lembres**, se tu leres estas rimas, → type 2

(45) Sonnet LXXI

The hand that **writ it**, for I **love yo**u so,
A mão que **o escreveu**: pois ɵ **te amo** tanto, → type 2

The requirements of type 2 correspondences seemed to work well in the classifications of English-Portuguese string pairs. Yet some aspects need to be observed by the researcher who wants to use this framework on this specific language pair: To observe attentively function words, decide whether the omission of personal pronouns (which are common in Portuguese) will be considered a restriction for this type of correspondence, decide whether the infinitive form in Portuguese is going to represent a restriction, and finally decide to what extent the translation of one adverb into more than one lexeme in Portuguese might be accepted. Another aspect to be remembered is that even when the production of a type 2 correspondence is possible, the translator will only choose this type of construction if the number of syllables fits the metre chosen.

4.3.3 Type 3 correspondences

In type 3 correspondences there are "greater structural discrepancies between source and target string than in those of type 2" (Thunes, 1998, p. 27). "While the two strings can be assigned equivalent semantic representations, there is at least one lexical word in one of the strings for which the other string lacks a correspondent word of the same lexical category and with the same syntactic function" (ibid, p. 28).

As soon as type 3 correspondences started to be observed in the corpus, some differences in relation to the string pairs classified as such were identified. Observe Examples (46) and (47):

(46) Sonnet LV

Nor dare I chide the world without end hour,
Não ouso censurar ɵ a hora interminável → type 3

(47) Sonnet XV

Cheered and check'd even by **the selfsame sky**:
E **o céu** que lhes dá aplauso é **o céu** que os vem vaiar → type 3
(the sky) (the sky)

The string pair (46) was classified as type 3 because the noun *world* has no correspondent in the target string: It was omitted by PS in his translation. While in (46) a content word was omitted (*world*) in the target string, resulting in a string pair with the same semantic content but fewer content words, in (47), instead of translating *the selfsame sky* as *o mesmo céu*, the translator rendered *selfsame sky* simply as *céu*. Therefore, there is an important structural rearrangement, the use of relative clauses in the target string instead of non-finite participial clauses, which would already justify the classification of the string pair as type 3.

Given the differences mentioned above, it is important to make a distinction between cases where the translator clearly omitted or added extra information in the target string. Based on these differences, type 3 correspondences were divided into two subtypes described in Table 12:

Table 12. Subtypes of type 3 correspondences

Type 3.1	Cases where there is equivalence on the semantic level, but at least one content word of the source text was omitted in the target string.
Type 3.2	Cases where there is equivalence on the semantic level, but one content word that was not part of the source string was added or repeated in the target string.

The important aspect is that both subtypes conform to one of the main characteristics of type 3, that is, "there is at least one lexical word in one of the strings for which the other string lacks an equivalent word of the same lexical category and with the same syntactic function" (Thunes, 1998, p. 28).

According to this new categorisation, string pairs (48) and (49) would be classified in the following way:

(48) Sonnet LV

Nor dare I chide **the world** without end hour,
Não ouso censurar **ɵ** a hora interminável → type 3.1

(49) Sonnet XV

Cheered and check'd even by the selfsame **sky**:
E o **céu** que lhes dá aplauso **é** o **céu** que os vem vaiar → type 3.2

String pairs (50) and (51) also represent the difference between types 3.1 and type 3.2 correspondences:

(50) Sonnet LV

When wasteful war shall statues overturn,
Quando a ɵ guerra as estátuas devastar → type 3.1

(51) Sonnet LV

Not marble, **ɵ** nor the gilded monuments
De mármore **não sei**, nem de áureos monumentos → type 3.2

In (50), the adjective *wasteful* was also omitted in the translation by JW, resulting in a target string with less information than the source string. In (51), on the other hand, the expression *não sei* has no correspondent in the source string, but it was added in the translation by PS, resulting in a 3.2 correspondence.

One aspect that needs to be considered during the analysis is that the difference between types 3.1, 3.2, and 2 is computed in terms of the presence of content words only, not the total amount of lexemes which includes function words as well. Observe string pair (52):

(52) Sonnet XCVIII

From you **have** I **been absent** in the **spring**,
Ausentei-me de ti na **primavera** → type 3.1

Here, the source strings contain nine lexemes, while the target string contains only six. In terms of content words, there are three in the source string (*have been, absent, spring*) while in the target string there are only two (*ausentei* and *primavera*). This probably happened because the information conveyed by the verb *have been* and the adjective *absent* was contracted into the verb utterance *ausentei-me* in Portuguese.

Interestingly, the idea of dividing type 3 correspondences into these two subtypes can be related to the concepts of lexical simplification and explicitation described in Section 2.6. In type 3.1 correspondences the translator creates a target string which corresponds to the source string on the semantic level omitting one or more content words. This is similar to the concept of lexical simplification described by Laviosa as "the process and/or result of making do with *less* words" (1998, p. 288). In type 3.2 correspondences, on the other hand, the translator creates a target string which corresponds to the source string on the semantic level, but also repeats content words or adds extra information, making the original content from the source string more explicit in the target string. According to Laviosa, the feature of explicitation was initially observed by Blum-Kulka (1986), who noticed that "shifts occur in the types of cohesion markers used in the target texts and records instances where the translator expands the target text by inserting additional words" (1998, p. 289).

The objective of this comparison is not to affirm that correspondence types 3.1 and 3.2 represent respectively the same phenomena of simplification and explicitation described by Laviosa (1998). My aim was just to point to these similarities shared by the two approaches. Besides that, I do not intend to use these terms for the purpose of this analysis, especially because, as it was pointed out in Section 2.6, although the search for universals of translation has been object of study of many scholars, this idea is considered by many authors a controversial issue.

The important issue to be stated here is that the subdivision of type 3 correspondences into these two subcategories can be an alternative for researchers who are interested in the phenomena of simplification and explicitation, but do not want to explore them from the viewpoint of features of translation. This adaptation of Thunes' model is an alternative which defines clear criteria for this kind of analysis.

String pair (53) shows how this distinction between types 3.1 and 3.2 might point to differences in relation to how the message was

conveyed by the translators in two distinct ways. In the translation by PS, the noun *carne* is added to the target string reinforcing the idea of death, when the body is compounded with clay. The word *carne* (which means *flesh*) has no correspondent in the source string. JW, on the other hand, chose to summarize the message using less content words:

(53) Sonnet LXXI

When I perhaps compounded am with clay,
Quando eu tiver com a argila a carne confundida,→ type 3.2
Quando ao barro eu for parte reunida, → type 3.1

The problem involving this subcategorisation was that in some examples of type 3 correspondences the amount of content words was exactly the same in both source and target strings. In (54) there are five content words in the source string: the verb *holds,* the nouns *perfection* and *moment*, the adverb *but*, and the adjective *little*. In the target string there are also five content words: the adverb *só*, the verb *is*, the adjectives *perfeito* and *breve*, and finally the noun *instante*.

(54) Sonnet XV

Holds in **perfection but** a **little moment**.
Só é perfeito por um **breve instante** → type 3

The string has the same number of lexemes, that is, each word of the target string has a correspondent in the source string. It cannot be classified as type 2 because some lexical words have a target correspondent that belongs to a different grammatical category and/or plays a different syntactic function. The verb *holds* was translated as the adverb *só*, the preposition *in* was replaced by the verb *is*, and the adverb *but* was replaced by the preposition *por*.

At this point it was necessary to decide if a third subcategory of type 3 correspondence should be created in order to classify such cases. During the annotation process, few examples of type 3 correspondences where the amount of content words was exactly the same in both source and target strings, were identified in the corpus. I could not see how a third subtype would bring significant differences for the analysis. Thus, these cases were also classified as type 3.2, and its definition was modified to cope with cases similar to the string pair (53). The new

definition for type 3.2, as well as the definition for type 3.1, are shown in Table 13.

Table 13. Definitions of subtypes 3.1 and 3.2 revisited

Type 3.1	Cases where there is equivalence on the semantic level, but at least one content word of the source text was omitted in the target string.
Type 3.2	Cases where there is equivalence on the semantic level, but at least one content word that was not part of the source text was added or repeated in the target string. This subtype also includes cases where there is equivalence on the semantic level and the number of content words is exactly the same, but one of the correspondent content words belongs to a different grammatical category and/or plays a different syntactic function in the sentence.

The next subsections present and discuss examples of type 3.1 and type 3.2 correspondences in order to make the distinction between them clearer.

4.3.3.1 Type 3.1 correspondences

In string pair (55), the message of the source string is conveyed by five content words (*gain, ills, thrice, more, have spent*). Remembering that *have spent* is treated as one content word because

have is an auxiliary, not the main verb. The target string produced by PS contains only four content words (*ganho, mal, triplo, expendido*). This happened because the expression *thrice more than* was replaced by the adjective *triplo* in the target string.

(55) Sonnet CXIX

And gain by ills thrice more than I have spent.
E ganho pelo mal o triplo expendido. → type 3.1

Target string pair (56) was characterised as type 3.1 because the adjective *little* has no correspondent in the target string; it was omitted by PS in the target string. The preposition *in* was translated as the determiner *a,* not only because it fits better with the verb *alcança*, but also for the syntactic reason that *alcançar* is a direct transitive verb.

(56) Sonnet XV

Holds in perfection but a little moment.
Apenas um ɵ momento alcança a perfeição → type 3.1

As mentioned in Chapter 3, there were some cases where part of the translation that should pertain to one specific verse is located in the previous or in the subsequent verse. In those cases, the dislocated part of the verse was treated as if it was part of the original. Consider Example (57):

(57) Sonnet XIX

O carve not **with thy hours** my love's fair brow,	(verse 9)
Nor draw no lines there with thine antique pen,	(verse 10)
Não queiras entalhar de meu amor a fronte	(verse 9)
Com tuas horas, nem riscá-la com tua pena	(verse 10)

In this string pair, the expression *with thy hours* was originally part of verse 9, but in the translation by PS the expression was dislocated to verse 10. This would implicate in a different classification of the string pair because of the omission of this part of the verse. Then, as shown in Example (58), *Com as tuas horas* was considered during the analysis as a constituent part of the verse 9 in the target string.

(58) Sonnet XIX

O carve not with thy hours my love's fair brow,
Nor draw no lines there with thine antique pen,
Não queiras entalhar de meu amor a fronte [Com tuas horas,]→ type 3.1
nem riscá-la com tua pena → type 4.2

Considering that interjections are treated as content words, source string (58) contains seven content words, while the target string of verse 9 contains only six. Verse 10 was classified as 4.2 because the verb *draw* does not correspond semantically to the verb *riscar* (*riscá-la*).

In the translation by JW, the correspondent expression of *with thy hours* was maintained in verse 9, but the correspondent word of the verb *carve* (*vincá-la*) was dislocated to the subsequent verse:

(59a) Sonnet XIX

O carve not with thy hours my love's fair brow,
Nor draw no lines there with thine antique pen,
Com as horas talhar a fronte amada,
Vincá-la com teu cálamo maduro;

(59b) Sonnet XIX

O carve not with thy hours my love's fair brow,
Nor draw no lines there with thine antique pen,
Com as horas talhar a fronte amada, [Vincá-la] → type 3.1
com teu cálamo maduro; → type 4.2

In this case, *vincá-la* was treated as if it were part of verse 9 for the purpose of the analysis, resulting in a type 3.1 correspondence. Verse 10 was classified as type 4.2 because although one of the meanings attributed to the word *cálamo* in figurative and poetic usage is *pena*[44], the adjective *maduro* is semantically quite different from *antique*. Semantically speaking, the lexeme *pena* used by PS is closer to the original than JW's choice for *cálamo maduro*.

[44] According to the on-line version of Caldas Aulete dictionary available at aulete.uol.com.br/cálamo.

4.3.3.2 Type 3.2 correspondences

In (60), both target strings were classified as type 3.2 because they contain lexemes that are not present in the source string. PS added the expression *perde a ação*, which has no correspondent in the source string. JW added the adjective *aberta*, which does not have a correspondent either. It was difficult to decide if *on the stage* would be considered correspondent to *em cena*, but the semantic content of the strings is not different enough to be classified as type 4. Therefore, considering that both strings contain one extra content word, if compared to the original, they were classified as type 3.2.

(60) Sonnet XXIII

As an unperfect actor on the stage
Como o ator imperfeito em cena perde a ação, → type 3.2
Como o ator imperfeito em cena aberta → type 3.2

The next example represents cases where both strings contain the same number of content words. Source and target strings contain five content words, the verb *join*, the nouns *spite* and *fortune*, and the verbs *make* and *bow*. The target string contains three content words too, the nouns *sorte* and *cruel*, and the verbs *une*, *vem* and *humilhar*. Thus, the string pair conforms to the specification determined for type 3.2 correspondences.

(61) Sonnet XC

Join with the spite of fortune, make me bow,
Une-te à sorte cruel, vem humilhar-me

In Example (62), the string pair by PS represents cases where the target string contains more content words than the source string. While the source string contains six content words (*'Tis, better, be, vile, vile, esteemed*), the translation by PS contain seven content words (*melhor, ser, mesmo, vil, vil, ser, tido*). In contrast, JW reduced the amount of content words to six, conveying the message in a more compact verse, and thus producing a type 3.1 string pair.

(62) Sonnet CXXI

‘Tis better to be vile than vile esteemed,
Melhor ser mesmo vil do que por vil ser tido, → type 3.2
Melhor ser vil que ter a fama, → type 3.1

In (63), there is equivalence on the semantic level, but two content words that were not part of the original were added in the target string. The expression *bem sei* has no correspondent in the source string. Therefore, the target string corresponds semantically to the source string, but PS added extra information that was not present in the original. This extra information is represented by the content words *bem sei*, as observed below:

(63) Sonnet CXIX

That better is, by evil still made better.
Bem sei que o mal melhora ainda o que é melhor: type 3.2

Example (64) raised a question related to what extent source and target string maintain a relation of equivalence if the verbal tenses used in the translation are different from the original. In the original verse, the verb is in the simple present tense (*live*), while in the target string, PS conjugated the verb *to live* in the future (*viverás*). The methodological decision was to ignore issues related to tenses if the translated verb is correspondent to the original one. Thus, target string (64) was considered correspondent to the source string in terms of meaning:

(64) Sonnet LV

You **live** in **this**, and **dwell** in **lovers**’ **eyes**.
Em meu **verso** e no **olhar** dos que **amam viverás**. → type 3.2

Again, based on the amount of information measured through the quantity of content words in both strings, string pair (64) was classified as type 3.2.

In target string (65), PS used more lexemes to convey the message of the source string, but the amount of content words is the same. There are six content words in the source string (*give, not, windy, night, rainy,*

morrow) and six in the target string (*não, dês, manhã, chuva, noite, vento*).

(65) Sonnet XC

Give not a windy night a rainy morrow,
Não dês manhã de chuva à noite com seu vento, → type 3.2

Interestingly, what seems to make the target string longer in many cases is not the use of content words, but the need for function words to connect the content words in Portuguese. This can be observed in (66). In order to translate the expression *my mistress' eyes* it was necessary to add two function words (*o* and *da*).

(66) Sonnet CXXX

My mistress'eyes are nothing like the sun,
O olhar da amada sol não é, pois brilha menos; → type 3.2

Here the translator also added extra information to the target string that is not present in the source string. The information that the mistress' eyes is not like the sun because *it shines less* (*pois brilha menos*) is not present in the original verse. But the rest of the verse still corresponds to the source string in terms of meaning. Thus, this is another example of type 3.2 correspondence.

Similarly to Example (66), in Example (67) PS also added extra information to the target string that is not present in the source string. He translated *And whether that my angel be turn'd fiend,* into the string *Se o anjo se fez demônio*, a translation which can be considered correspondent in terms of meaning. But he also added *eis ponto alto encoberto*, which lacks a correspondent in the source string. Because of this extra information, the string pair was classified as type 3.2.

(67a) Sonnet CXLIV

And whether that my angel be turn'd fiend,
Se o anjo se fez demônio, eis ponto alto encoberto: → type 3.2

At this point of the analysis, I realised how important it was to adopt the verse as the unit of translation. As a starting point, I used the verb sentences as the unit of translation, but this would implicate in a

significant loss of information related to some decisions made by the translators, especially those probably related to metre. For example, if the unit of translation was not the verse, but based on other criteria like punctuation, for example, the second part of the verse would not be considered in the analysis:

(67b)
And whether that my angel be turn'd fiend, ɵ
Se o anjo se fez demônio, eis ponto alto encoberto:
string pair ɵ

This decision would interfere dramatically in the analysis and in the adaptation of the model itself. The parts of the verse with no correspondent would not be taken into consideration during the analysis, which seemed not to be a good decision, since in many cases the addition of information was probably done to adjust to the meter chosen for the translation. Furthermore, issues related to metre cannot be ignored when it comes to poetry translation, unless in cases where the translator chooses free verses. The implication of this decision is that, in order to use the adaptation of the model proposed here, the researcher necessarily needs to adopt the verse as the unit of translation.

At this point of the discussion, it was possible to foresee that the chances of type 3.1 correspondences would probably be less frequent in the translations that use twelve syllables.

4.3.4 Type 4 correspondences

Type 4 correspondences differ from the lower types 1 to 3 in relation to the aspect of computability. Since they are not inserted in the domain of linguistically predictable translation tasks, they are considered noncomputable, in other words, they depend on the work of a human translator. These correspondences represent the highest degree of translational complexity on the scale proposed by Thunes (2011). In general terms,

> (…) there is not semantic equivalence between the entire source and target strings; pragmatic equivalence may exist, but not necessarily. Hence,

> there do not exist, as in string pairs of the lower types, any implicational relations between equivalence relations on different linguistic levels. (Thunes, 2011, p. 170)

Translational correspondences of type 4 "are cases where there are discrepancies between original and translation not only on the structural, but also on the semantic level. Type 4 is assigned to translational correspondences where we cannot derive equivalent semantic representations for source and target string" (Thunes, 1998, p. 28). Example (68) is an instance of a type 4 correspondence because it contains a mismatch on the semantic level:

(68a) Sonnet XV

When I consider everything that grows
Se tudo quanto cresce (eu fico a meditar)

In the translation by PS, the expression *(eu fico a meditar)*, does not correspond to *When I consider* in the target string. Given the criteria described for each correspondence type, this semantic difference is sufficient to classify the string pair as a type 4 correspondence. Interestingly, a literal translation that could maintain the semantic content of the source string would be part of the sets of possible interpretation for each unit in this case. This construction is possible in Portuguese language:

(68b)	When	I	consider
	Quando	eu	considero

If the translation *Quando eu considero tudo que cresce* had been chosen in this case, we would have another example of type 1 correspondence. In spite of the fact that the suggested translation is part of the set of possible interpretations, for some reason this was not chosen by the translator. The issue of metre here would not be a justification because the verse *Quando eu considero tudo quanto cresce*[45] has twelve syllables, it would fit the iambic hexameter.

Thunes initially defended that this type of analysis should be based on the product of translation, and discussing the translation

[45] This verse would be part of the set of possible translations.

process or method would be not part of the scope of the analysis, but she recognises the importance of considering the sets of possible interpretation during the analysis:

> It should be noted that when measuring differences in the amount of linguistically expressed information in translationally corresponding units, we consider the sets of possible interpretation for each unit. Previously we have argued that when the translational complexity of given string pairs is analysed, we consider the target expression in relation to the relevant interpretation of the source expression, since we keep source text disambiguation apart from the translation task. However, in order to quantify differences in the amount of linguistically encoded information, it is necessary to take into account the sets of possible interpretations of both units. To consider only the relevant interpretation of the source string would mean an increase in uncertainty in every case where more than one interpretation is possible for the target string. (Thunes, 2011, p. 346)

From my point of view, taking into account the sets of possible interpretation is important, especially when it comes to type 4 correspondences. By doing so, in my analysis, it was clear that in many cases the characteristics of the string pairs fit into type 4 correspondence, but the existing discrepancies between original and translation both on the structural and on the semantic level existed because it was the translator's choice to create a target string different from the source string on these levels. The discrepancies existed not because of the impossibility of deriving equivalent semantic representations for source and target strings, as defined by Thunes.

Considering these two possibilities that could result in a type 4 string pair, it was necessary to take a methodological decision similar to the one adopted for type 3 correspondence: To make a distinction between these two possible cases of type 4 correspondences, because this could be a valuable source of information in relation to the translator's style. Thus, two subcategories (4.1 and 4.2) were created to differentiate type 4 correspondences in relation to cases where the language pair does not allow a similar construction and cases where it

was the translator's choice to change the structure and meaning. The subtypes are described in Table 14:

Table 14. Subtypes of type 4 correspondences

Type 4.1	Cases where it is not possible to produce a target text with similar syntactic structure and that conveys the same meaning.
Type 4.2	A target text with similar syntactic structure and that conveys the same meaning is possible for that language pair, but it is clear that the translator chooses to change the semantic content.

Based on the new subcategories, Example (69) should be classified as type 4.2, since the literal translation *Quando eu considero tudo quanto cresce* is possible for the language pair English-Portuguese.

(69) Sonnet XV

When I consider everything that grows
Se tudo quanto cresce (eu fico a meditar) → type 4.2

Given that a literal translation is viable in this case, it is possible to affirm that the creation of a target string with different meaning happened because the translator chose to do so, and not because the structure of the Portuguese language did not allow a similar construction.

4.3.4.1 Type 4.2 correspondences

String pair (70) represents, according to the new subtypes suggested in the previous section, another example of 4.2 correspondence. There are discrepancies between original and translation not only on the structural, but also on the semantic level. There is no correspondent for the expressions *aos desígnios* and *mais distantes*, which leads to a mismatch on the semantic level. But, by observing PS's translation, which was classified as subtype 3.2[46], it is possible to conclude that a translation that conveys the same meaning is possible, but it was the translator's choice to change the structure and the meaning of the verse:

(70) Sonnet XV

Whereon the stars in secret influence comment.
Se os astros vêm, com influência oculta, comentar → type 3.2
Aos desígnios dos astros mais distantes; → type 4.2

Consequently, if one of the string pairs was classified as type 3, as in Example (70), and the other had a different meaning, it would necessarily be classified as type 4.2, because the presence of a type 3 already shows that a similar construction is possible for that language pair. That means that in those cases the analyst does not necessarily need to think about the set of possible translations because the parallel corpus itself provides this information.

In (71), the same problem of dislocated information described in some cases of type 3 correspondences was identified.

(71) Sonnet XXXIII

Anon permit the basest **clouds** to ride (verse 5)
Mas logo permitir que seja percorrida
With ugly rack on his celestial face, (verse 6)

[46] Although the verb *vêm* is not present in the source string, the meaning of both strings is still the same.

Por negras **nuvens** sua face refulgente,

The lexeme *clouds* (*nuvens*), that originally belonged to the fifth verse was dislocated to the sixth verse in the translation by PS, but for the sake of analysis it was considered as part of verse 5, as observed below:

(72) Sonnet XXXIII

Anon permit the basest clouds to ride
Mas logo permitir que seja percorrida [Por negras nuvens] → type 4.2
With ugly rack on his celestial face,
sua face refulgente, → type 3.1

The target string does not correspond to the source string in terms of meaning because the expression *basest clouds* was translated as *nuvens negras*[47] instead of *nuvens mais vis*.

String pair (73) is definitely an example of 4.2 correspondence because there are discrepancies between original and translation on both structural and semantic level:

(73) Sonnet XVIII

Thou art more lovely and more temperate
Vencendo-o em equilíbrio, és sempre mais amável: type 4.2
Tens mais doçura e mais amenidade: type 4.2

The literal translation of this string pair would be *Tu és mais amável e mais comedida*. This possible translation indicates that a similar construction is viable for the language pair, but both translators chose to produce a string pair that differs on these two levels. The mere choice of translating the verb *to be* into the respective verbs *to win* and *to have* indicates a significant change in terms of meaning. Even so, JW's translation could be considered closer to the original meaning.

In (74), *Death* was translated as *Fim* by PS. Even considering that this could be a possible translation in terms of pragmatics, the rest of the information present in the target string does not correspond to the information of the source string. There is no target correspondent for the verb *to brag*, and the expression *não te verá* also lacks a correspondent in the source string. JW chose a shorter translation, but he created a

[47] Which means "dark clouds".

target string that is correspondent to the original by reducing its amount of information. The idea of walking off aimlessly in one's shade was replaced by the single verb *ensombrecer-te*, thus producing a type 3.1 correspondent, given that the number of content words was significantly reduced.

(74) Sonnet XVIII

Nor shall Death brag thou wander'st in his shade,
Vagando em sua sombra o Fim não te verá, type 4.2
Nem a morte rirá de ensombrecer-te, type 3.1

Example (75) was classified as type 4.2 because the translation of the second part of the string (*de tal modo se entristecem*) which literally means *so grieves* does not correspond to *'t is with so dull a cheer*. This verse could be literally translated as *Ou caso eles cantem, é com alegria tão maçante*, but the translator preferred to convey the idea of sadness instead of dull cheer.

(75) Sonnet XCVII

Or if they sing, 't is with so dull a cheer,
Ou, caso cantem, de tal modo se entristecem, type 4.2

4.3.4.2 Type 4.1 correspondences

In the following verses, two examples of 4.1 correspondences can be visualised. In (76), it is not possible to derive equivalent semantic representations of source string. What reinforces this idea is the fact that a literal translation is not possible. A literal translation of the verb *engraft* would not be appropriate in Portuguese, leading the translators to use the expression *te acresento* instead of the verbs *enxertar*, *imprimir* or *implantar*, which are examples of possible literal translations that would not sound poetic translations.

(76) Sonnet XV

And all in war with Time for love of you
E, por amor de ti, em Guerra o Tempo enfrento:
Então, por teu amor, o tempo enfrento
As he takes from you, I engraft you new.
Quanto ele em ti suprime, é quanto te acrescento. → type 4.1
E quanto ele te rouba, te acrescento. → type 4.1

Here, in both translations the verb *to take* cannot be literally translated, so the verbs *suprime* and *rouba* were chosen to replace it. The verb *engraft* was translated as *acrescento* and the adjective *new* lacks a correspondent in both target strings.

In (77), just like in the previous example, it is not possible to translate the verse literally. The verb *will play* has no semantic correspondent in both target strings, while PS transformed the expression *play the tyrants* into the transitive verb *tiranizar*, JW translated it as *serão tiranas*, that is, the verb *play* was replaced by the verb *ser*. The respective lexemes *amargamente* and *indiferentes* do not correspond to the expression *the very same* either. So, it seems that in these cases the structure of the target language does not allow an equivalent semantic representation of source string, resulting in type 4.1 correspondences.

(77) Sonnet V

Will play the tyrants to the very same
Hão de tiranizar-te ainda amargamente, → type 4.1
Serão tiranas, quando, indiferentes, → type 4.1

The subsequent verse of Sonnet V was also classified as type 4.1. The meaning of the original content words (*unfair* and *excel*) were not maintained in the target strings. The noun *formosura* used by both translators in the target string has no correspondent in the source string. The original lexemes have literal correspondents in Portuguese, but if the target string were translated literally it would be meaningless.

(78) Sonnet V

And that unfair which fairly doth[48] excel:
Desformoseando o que é sem par em formosura: → type 4.1
Ao formoso roubarem formosura: → type 4.1

In string pair (79) it is not possible to derive equivalent semantic representations of source string and the translator had to create a different syntactic structure to convey the message.

(79) Sonnet CXXX

And yet by heaven I think my love as rare,
No entanto, pelos céus! tão rara a considero → type 3.1 (verse 13)
As any she beli'd with false compare.
Como as belas que exalta um símile insincero. → type 4.1 (verse 14)

In (80), it would be difficult to translate the expression *past reason hunted*. PS did not translate the lexeme *hunted* in his target string. JW, on the other hand, did, but the meaning conveyed by his target string is not the same as the one conveyed by the source string. Here again, many content words of the target strings have no correspondent in the source string. Differently from type 3 correspondences, where this would be an indicative of extra information when the rest of the string corresponds to the original, the whole translated verse is considered distinct on the semantic and structural level.

(80) Sonnet CXXIX

Past reason hunted, and no sooner had
Buscada além do juízo, e, assim que desfrutada, → type 4.1
É caça além do siso, relutante, → type 4.1

Finally, in the translations of verse 5 from Sonnet CXXIX by PS and JW, the meaning of the original content words (*enjoyed, sooner, despised* and *straight*) were not maintained in the target strings. Here, just as Example (78), the original lexemes have literal correspondents in

[48] *Doth*, in Archaic English, is the third person singular present tense of *do*, which plays the role of auxiliary verb.

Portuguese, but if the target string were translated literally it would be meaningless.

(81) Sonnet CXXIX

Enjoy'd no sooner but despised straight,
Relegada ao desprezo logo que fruída; → type 4.1
Lenta em fruir-se, mas logo esquecida, → type 4.1

By considering the analysis of the proposed subtypes of type 4 correspondences, it seems that to differentiate types 4.1 and 4.2 it is necessary to resort to a literal translation of the source string when the target strings have a different structure and a different semantic content. This procedure conforms to Thunes' statement that "when measuring differences in the amount of linguistically expressed information in translationally corresponding units, we consider the sets of possible interpretation for each unit" (2011, p. 346). As a consequence, if a literal translation of the target string is possible and it has meaning, we are probably facing an example of type 4.2 correspondence. On the other hand, if a literal translation is not possible or it is meaningless, indicating that an equivalent semantic representation for source and target string is not possible, the string pair corresponds to type 4.1.

4.4 Relating correspondence types with the translator style

Once the analysis of the string pairs which represent the correspondence types and their respective subtypes was completed, it was necessary to establish how this information could be used to explain differences between the translators' styles.

In string pair (82), PS preferred to create a target string as close as possible to the source string, thus producing a type 1 correspondence. JW, on the other hand, decided to modify the original message creating a target string with different semantic content. That means that JW was less attached to the original meaning than PS; he allowed himself to work with more freedom when producing his target string. This can be considered a significant difference between the two translators which can be explained by the type of correspondence eventually produced.

(82) Sonnet XC

And other strains of woe, which now seem woe,
E outras formas de dor, que ora parecem dor, → type 1
E outras mágoas, que em dor se me oferecem, → type 4.2

Similarly to Example (82), in (83) PS also produced a target string whose meaning is closer to the source string and he did not add extra information that was not present in the source string. JW kept part of the original meaning, but he altered the original verb tense (present perfect) to a infinite form (*profanar*) and added the expression *os vias* which lacks a correspondent in the source string. Here again, JW seems to be less attached to the original meaning than PS.

(83) Sonnet CXLII

That have profan'd their scarlet ornaments,
Que profanaram seus purpúreos ornamentos → type 2
Que a profanar o ornato rubro os vias, → type 3.2

In (84), although JW translated literally the two first lexemes of the target string (*Whilst I = Enquanto eu*), the meaning of the rest of his target string does not correspond to the original message. PS again, produced a target string whose meaning corresponds to the source string in spite of the fact that there are structural differences like word order and use of content words.

(84) Sonnet XXIII

Whilst I, whom fortune of such triumph bars,
Mas, como desse triunfo a sorte me separa, → type 3.2
Enquanto eu, desses prêmios esquecido, → type 4.2

PS's target string (85) was classified as type 3.1 because although it can be considered correspondent to the source string semantically speaking, the translator used less content words to express the message. JW created a target string completely different from the source string:

(85) Sonnet XXVI

Till then not show my head where thou mayst prove me.
Por ora, escondo a fronte, ou poderás provar-me. → type 3.1
Não antes, que o proíbe o meu temor. → type 4.2

JW's choice could also be related to the metre, since sometimes is not possible to achieve the number of desired syllables with the literal translation, but this is just a speculation since we do not have access to what happened during the translation process. What we do know, is that PS opted for iambic hexametre verses which contain twelve syllables while JW chose the decasyllable verse. This might explain the addition or subtraction of information in type 3 correspondences.

By observing Example (86), it is clear that PS created a verse which is closer to the original in terms of meaning, if compared to JW's verse:

(86) Sonnet XXXIII

The region cloud hath mask'd him from me now.
Encobriram-no logo as nuvens da região. → type 3.2
Que o mascarou, se pôs de mim defronte. → type 4.2
Yet him for this my love no whit disdaineth;
Nem por isso o despreza o meu amor profundo: → type 3.2
Que o amor do sol nem mesmo assim decaia: → type 4.2
Suns of the world may stain when heaven's sun staineth.
Se o sol do céu se ofusca, assim os sóis do mundo. → type 3.1
Nublam-se os térreos se o do céu desmaia. → type 4.2

While the former produced three type 3 correspondences, the latter produced three type 4.2 correspondences. So, it becomes clear that the amount of type 4.2 correspondences might indicate how closer to the original in terms of meaning the translation is.

Example (87) shows again how JW is less attached to the original meaning when compared to PS. PS produced a type 3.2 because he kept the meaning of the original source string and added extra information (*não vês*), probably in an attempt to adjust the metre. JW simply replaced the expression *O know sweet love* by *eis a verdade*, producing a semantically different target string.

(87) Sonnet LXXVI

O know sweet love I always write of you,
Só escrevo sobre ti, meu doce amor, não vês? → type 3.2
Escrevo sobre ti – eis a verdade. → type 4.2

In (88), the expression *the fleeting year* was translated by PS as *ano que jamais perdura*, which is closer to the original than JW's translation *no tempo que se escoa*. Therefore, just from the level of vocabulary choice, JW shows his tendency of changing the original meaning in his target strings.

(88) Sonnet XCVII

From thee, the pleasure of the fleeting year!
De ti, delícias do ano que jamais perdura! → type 3.2
Que és o prazer, no tempo que se escoa! → type 4.2

In (89), both translators attempted to create a target string similar to the source string on the semantic level, but they produced respectively correspondence types 3.2 and 3.1. This example might confirm the speculations which relate type 3 correspondences to metre. PS needs more syllables, so he added extra information (*bem sei*) to the target string.

(89) Sonnet CXIX

That better is, by evil still made better.
Bem sei que o mal melhora ainda o que é melhor: → type 3.2
Que o melhor, pelo mal fica melhor → type 3.1

The following examples interestingly show that in some cases JW was the translator who created verses closer to the source string on the semantic level. In (90) he kept the content words *love* (*amor*) and *recompense* (*recompense*) in the target string while PS turned away from the original vocabulary, translating *plead* as *advoguem-me* and the expression *look for recompense* became *esperem premiá-la*. PS's choice of vocabulary significantly changed the meaning of the target string, which was then classified as a type 4.2 correspondence.

(90) Sonnet XXIII

Who plead for love and look for recompense
Advoguem-me a paixão e esperem premiá-la. → type 4.2
Que pede amor e pede recompensa → type 3.2

Similarly to what was previously described, in (91) the translation by JW was closer to the source string on the semantic level, while PS produced a correspondence type 4.2:

(91) Sonnet XXIII

More than that tongue that more hath more express'd.
Mais do que alguém que mais tem dito com mais fluência, → type 4.2
Mais do que a língua que mais pode o diz. → type 3.2

Subsequently, we might check whether the points discussed so far are consistent with the amount of correspondence types identified in each of the translations. With the purpose of doing so, consider again the data shown in Table 15, which contains the distribution of each correspondence type across the parallel corpus:

Table 15. Correspondence types across the translation pairs

Correspondence types	Translation by PS	Translation by JW	Total
Type 1	1	0	1
Type 2	60	18	78
Type 3.1	139	189	328
Type 3.2	174	35	209
Type 4.1	55	49	104
Type 4.2	216	354	570
Total of string pairs	645	645	1,290

First, in relation to the distribution of lower correspondence types, that is, types 1 and 2, PS produced 76.92% of all the type 2 correspondences identified in the parallel corpus. This is an indicative of his attempt to produce target strings as similar as possible to the source string. JW, on the other hand, produced only 18 target strings classified as type 2.

Second, in relation to the amount of type 3 correspondences, the amount of type 3.1 is more significant in JW's translation, since he had the tendency to reduce the amount of information of the source string, producing target strings with less content words. This might be related to the metre chosen: He probably had to reduce the amount of lexemes in order to obtain the desired decasyllable verses. In PS's translation, on the other hand, the amount of type 3.2 correspondences is more significant. This might also be related to the metre chosen by PS: Differently from JW he opted for iambic hexameter verses, which contain twelve syllables. The larger number of type 3.2 correspondences in PS's renderings is likely to be an attempt by PS to produce extra syllables and conform to the metre.

At this point of the discussion, one could easily question the creation of subcategories for type 3 correspondences by claiming that the metre chosen already implicates in the addition or reduction of words. In that case, it would not be necessary to create these subcategories (types 3.1 and 3.2) to conclude, for example, that the chances of type 3.1 correspondences appear in the translations that use twelve syllables are much reduced.

From my point of view, it is clear that metre has an important role during the translation process and it probably influences many decisions taken by the translators, but it is not the only aspect involved in the process. The suggested subcategories for type 3 indicate a very important characteristic in terms of writing: the amount of content words used in the translation, which interfere in the semantic content of the verses. From my point of view, the semantic content conveyed by the translated verse is important to determine differences in terms of translator's styles.

The total of type 4.2 correspondence in JW's translation confirms the assumption that he was less worried with maintaining the original meaning conveyed by the source string than PS, something that was observed during the discussion of the results.

Finally, the small quantity of type 4.1 correspondences along the corpus might indicate that there are small differences between English

and Portuguese, given that in most cases it would be possible to create equivalent semantic representations of the source string.

Chapter 5

Conclusion

5.1 The research questions

The present study proposed the investigation of topics of translation complexity, a possible adaptation of Thunes' model and finally to what extent the model of complexity could point to stylistic differences between PS and JW. Hence, the first objective was to verify if Thunes' model of complexity could be used to analyse two distinct translations of Shakespeare Sonnets and if so, what kind of adaptations would be necessary so that the framework could be used in any parallel corpus of poetry. Finally, I intended to verify to what extent the issue of complexity could explain the structural and semantic differences between PS's and JW's translations.

This final chapter will draw some conclusions on the basis of the present investigation, which will be centered on Thunes' framework, the method presented in Chapter 2, and the results obtained during the annotation process. Finally, I suggest further applications of the analytical approach developed during the investigation.

5.2 The framework (what needs to be adapted)

Similarly to Thunes' study, "the present work is a product-oriented approach to complexity in translation" (2011, p. 433) and I also have studied "intersubjectively available relations between source texts and existing translations" (ibid). Thunes states clearly that the scope of her investigation "does not include aspects related to translation methods, or to the cognitive processes behind translation", but differently from her I approached issues related to the translation process during the analysis of the results.

Thunes' model of complexity was originally developed with the objective of investigating to what extent it would be possible to automatise translation in a selection of English-Norwegian parallel texts. By this she meant "the computing of translations with no human intervention" (2011, p. 433) and "an approach to machine translation based on linguistic knowledge" (ibid).

The objective of using Thunes' model in the present investigation was distinct from what she had originally proposed. In my investigation I was not concerned with machine translation. As a consequence, many theoretical concepts used for the purpose of her analysis were not considered during my analysis. It is important to make this clear in order to avoid misunderstandings when comparing the two approaches. What I proposed was an adaptation of a consistent model that can be used for different linguistic analysis which exceeds the scope of machine translation properly.

My conclusion in relation to the application of this model in a parallel corpus of poetry is that it can be satisfactorily used to annotate any parallel corpus involving the language pair English-Portuguese, given the suggested adaptations. Thus, the answer for the research question "How can we adapt Thunes' model in order to be used to analyse any parallel corpus made up of Brazilian translated poetry?" will be discussed subsequently.

Type 1 correspondences

As pointed out in Chapter 2, "translational correspondences of type 1 are cases of word-by-word correspondences" (Thunes, 1998, p. 25). This kind of correspondence is possible for the language pair English-Portuguese, but only one example of this correspondence was identified in the whole corpus.

(92) Sonnet XC

And other strains of **woe**, which now seem **woe**,
E outras formas de **dor**, que ora parecem **dor**, → type 1

This result was already expected and I was surprised to find one example of this correspondence in the parallel corpus. This fact raised a methodological issue: To decide if this category should be maintained in the framework. Cases like this could be perfectly included in the group of type 2 correspondences. Given the small size of the corpus used in

the investigation and its specificity, it is very likely that in other types of corpora this type of correspondence might be identified more frequently. Consequently, my decision was to maintain type 1 correspondence as an individual type in the framework.

According to the analysis of results, the amount of type 1 correspondences might indicate to which extent the translational relation between English and Portuguese is complex. If these correspondences are very frequent, this could indicate that the languages present similarities in terms of grammar. Given the results of the present study, the translational relation between English and Portuguese can be considered of high complexity.

In relation to translator's style, type 1 correspondences represent the attempt of the translator to keep the same structure and the same meaning of the original text when the structure of both languages allows similar constructions and when it conforms to the metre chosen.

Type 2 correspondences

In the scale of complexity, "type 2 correspondences represent the second lowest degree of translational complexity" (Thunes, 2011, p. 153). The relations of equivalence between source and target string are obligatory on the levels of syntax, semantics and pragmatics. The procedures used to identify type 2 correspondences in the corpus worked well, the only issue was in relation to the use of grammatical function words. Therefore, it is necessary to pay special attention to prepositions, pronouns, determiners, conjunctions, modal verbs, auxiliary verbs and particles during the analysis to avoid misinterpretations. In my analysis, the function words presented in Table 11 (Chapter 4, Section 4.3.2) were considered. If a target string lacked a correspondent for these function words, but all the other content words had a correspondent, the string pair was classified as a type 2 correspondence.

One of the difficulties involving type 2 and the subsequent correspondence types was to decide whether there was equivalence on the level of semantics and pragmatics. In some cases, like the one presented in Example (93), the word order is the same and the target string only lacks a correspondent for the preposition *in*. So far this would not violate the restrictions established for type 2, but to what extent the word *apreço* can be considered correspondent to *love*? From my point of view, the source and target strings can be considered correspondent in terms of meaning, but this is a personal interpretation.

At this point I missed a theoretical basis to support my interpretation. What criteria could be used to determine the scope of correspondence between two lexemes of the same grammatical category?

(93) Sonnet XXVI

Lord of my **love**, to whom in vassalage
Senhor de meu **apreço**, a cuja ɵ vassalagem → type 2?

This difficulty pointed out to the necessity of delimiting the scope of semantic and pragmatics in a clearer way. Some of the criteria established for each correspondence type can be considered simple and easy to use: Checking word order, counting lexemes and checking their grammatical category and respective syntactic function did not represent a real challenge. But when it comes to semantics and pragmatics, the procedures to establish whether there is equivalence between the string pairs become less objective, especially when it comes to literary texts.

Type 3 correspondences

Type 3 correspondences are those cases in which there are "greater structural discrepancies between source and target string than in those of type 2" (Thunes, 1998, p. 27). "While the two strings can be assigned equivalent semantic representations, there is at least one lexical word in one of the strings for which the other string lacks an equivalent word of the same lexical category and with the same syntactic function" (ibid, p. 28). During the identification of type 3 correspondences, I observed that there were two major groups of target strings, and the omission or addition of content words pointed out to significant differences between the translators' styles. The identification of these two major groups was also intrinsically related to the metre chosen. Therefore, the first adaptation proposed was to subdivide type 3 into two subtypes, presented in Table 13 (Chapter 4, Section 4.3.3)

Similarly to type 2 correspondences, in some cases it was difficult to check whether there was equivalence on the level of semantics and pragmatics, which reinforces the necessity of delimiting the scope of semantic and pragmatics in a clearer way. One example is the target string (94): Initially, it was difficult to decide in which category it would fit. My first impulse was to classify it as type 1, because the translation contains the same number of words of the original and the same word order. However some lexemes do not belong

to the same grammatical categories. The verb *holds* was replaced by the adverb *só*, the preposition *in* was replaced by the verb *é*. But in terms of semantic equivalence, both strings share the same meaning, so according to the framework the string pair was classified as type 3.2.

(94) Sonnet XV

Holds in perfection but a little moment.
Só é perfeito por um breve instante → type 3.2

Type 4 correspondences

Within the group of type 4 correspondences, I also observed that there were two possible motivations for the production of these string pairs. The first group matches the definition proposed by Thunes: "cases where there are discrepancies between original and translation not only on the structural, but also on the semantic level" (1998, p. 28) because it is not possible to "derive equivalent semantic representations for source and target string" (ibid). The second groups includes cases where there are discrepancies between original and translation on both the structural and the semantic level, but it would be possible to derive an equivalent semantic representation if the translator wanted to. Thunes' model does not predict this second possibility, but this is consistent with what she proposed for her analysis, which was exclusively based on the product of translation. She clearly states that discussing the translation process or method would not be part of the scope of her analysis. As mentioned in the review of literature, in the present investigation I decided to go beyond the scope of the product of translation in some cases because this was necessary in order to explain some of the translators' choices.

In order to be used to analyse a parallel corpus of poetry, the second adaptation proposed for Thunes' model was the subdivision of type 4 into two subtypes, as shown in Table 14 (Chapter 4, Section 4.3.4).

String pairs (95) and (96) are examples of the respective subcategories of type 4 proposed for the purpose of analysis:

(95) Sonnet CXXIX

Enjoy'd no sooner but despised straight,
Lenta em fruir-se, mas logo esquecida, → type 4.1

(96) Sonnet XXIII

And dumb presagers of my speaking breast,
_ Intérpretes sem voz - de um coração que fala; type 4.2

In general terms, few adaptations are necessary so that Thunes' model can be used to analyse any corpus of translated poetry made up of the specific language pair English-Portuguese. The subdivision of types 3 and 4 into two subtypes and the delimitation of what lexemes will be treated as function words might be efficient to classify any parallel corpus of poetry.

5.3 The method

The method used in the present investigation consisted of the extraction of "translationally corresponding strings" (Thunes, 2011, p. 436) from a parallel corpus of 45 Shakespeare's Sonnets. During the annotation process, each string pair was classified according to one of the types and subtypes proposed after the adaptation of Thunes' model. Similarly to Thunes, the analysis was "applied to running text, omitting no parts of it" (ibid).

The primary unit of analysis was the verse, and this is a prerequisite for the use of the proposed adaptation of the model. If a different unit of analysis is chosen, some of the suggested subtypes might not be used because extra information in the case of 3.2 correspondences, for example, would not be considered during the analysis. Thus, this is a restriction in relation to the use of the model. If a different unit of analysis is chosen, a new adaptation of the model might be necessary.

In relation to the process of annotation, the same procedures adopted by Thunes were used:

> The identification of translational units, as well as the classification of each correspondence, have been done manually. The assignment of correspondence type to string pairs is an elimination procedure where we start by testing each correspondence for the lowest type and then

> move upwards in the hierarchy if the test fails. The analysis is an evaluation of the degree to which linguistic matching relations hold in each string pair. (Thunes, 2011, p. 437)

With regard to the annotation process itself, it was necessary to make decisions when information of the source string was dislocated to the previous or subsequent verse in the target text. My methodological decision was to consider the dislocated parts of the verse as if they pertained to the target string which corresponded to the source string, just as in Example (97). The expression *thy soul* was originally placed in verse 8, but in the target string translated by PS this information was dislocated to the previous verse (verse 7). Considering that the expression *que tua alma* belongs to verse 8, the target string which corresponds to verse 7 is classified as type 3.1 because it contains fewer content words than the target string:

(97) Sonnet XXVI

But that I hope some good conceit of thine (verse 7)
Mas espero que **tua alma**, em sua boa graça,
In **thy soul**'s thought, all naked, will bestow it; (verse 8)
Acolhimento a um pobre nu consiga dar.

Just as mentioned in Chapter 4, this might have happened as an attempt to conform to the metre chosen. It is possible to adopt a different procedure if the analyst desires to, but this might change the correspondence type during the analysis and, consequently, alter the results. Nevertheless, from my point of view, the verse was an appropriate translational unit for the purposes of this analysis.

As previously stated in Chapter 1, I expected that the adaptation of Thunes' model proposed here could be used to analyse any Portuguese-English parallel corpus of poetry. In relation to that, I share the same concerns that Thunes raised in relation to her method:

> The fact that only a small corpus of about 68 000 words has been analysed in the present study, raises the question whether the present approach could be applied to large parallel corpora. Since the method is time-consuming, and implemented manually, scaling up would require either

> automatisation or using a team of annotators. (Thunes, 2011, p. 439)

The corpus used in the present investigation is even smaller than Thunes' corpus. It would be difficult to annotate large corpora using this method because the analysis involves linguistic interpretation and it is difficult to guarantee consistency when a team of annotators is involved. Automatisation is not an alternative because it was not my objective to check if poetry translation could be done automatically, this is a task that needs to be done by a bilingual human translator. But the adaptation proposed here seems to work well for manual annotation of small parallel corpus of poetry in English-Portuguese. The small amount of data offered sufficient information to discuss stylistic differences between JW and PS and it might work well to discuss stylistic differences of other translators.

5.4 The results

The complexity measurements obtained from the annotation of the parallel corpus show that types 3 and 4 are predominant in the corpus, representing 93% of the analysed string pairs. This is an expected result, and one could easily argue that it is not necessary to carry out research to conclude that.

The aspect to be considered relevant here is that while there is an increase in the degree of complexity in the translational relation, there is not necessarily a decrease in the extent of implicational relations between types of translational correspondence related to different linguistic levels. By dividing types 3 and 4 and considering some of the decisions taken by the translators during the translation process, I automatically broke up with some principles originally established by Thunes. As a consequence, the hierarchy originally proposed by the author cannot be strictly used as a reference anymore.

The results highlight the fact that in most cases involving the language pair English-Portuguese, it is possible to produce a message similar to the one conveyed by the source text, but not necessarily keeping the same structure of the original text. The changes on the structural and semantic level are more related to translators' choices than to the impossibility of producing a target string due to structural

differences involving the two languages. The fact that only 8.06% of the translational pairs were classified as type 4.1 correspondences shows that in only few cases it was not possible to render the source text into the target language.

In relation to the first research question, it is possible to adapt Thunes' model so that it can be used to analyse any small parallel corpus made up of Brazilian translated poetry. The adaptation proposed is basically the subdivision of types 3 and 4 into two subtypes to clarify stylistic differences between the two translators.

Type 3 correspondences were subdivided into two subtypes, 3.1 and 3.2. Type 3.1 represents cases where there is equivalence on the semantic level, but at least one content word of the source text was omitted in the target string.

(98) Sonnet LXXI

When I perhaps compounded am with clay,
Quando ao barro eu for parte reunida, → type 3.1

Correspondences classified as type 3.2 represent cases where there is equivalence on the semantic level, but at least one content word that was not part of the original was added or repeated in the source string:

(99) Sonnet XV

Cheered and check'd even by the selfsame **sky**:
E o **céu** que lhes dá aplauso **é** o **céu** que os vem vaiar → type 3.2

This subtype also includes cases where there is equivalence on the semantic level and the number of content words is exactly the same, but one of the correspondent content words belongs to a different grammatical category and/or plays a different syntactic function in the sentence, as in Example (100), where the verb *holds* was replaced by the adverb *só*:

(100) Sonnet XV

Holds in **perfection but** a **little moment**.
Só é perfeito por um **breve instante** → type 3

It would be possible to create a third subcategory to classify cases like this one, but because there were not many cases like this, I realised that one more subdivision would not interfere significantly in the results. But yet, this is a possibility for the analyst who decides to use the model.

Type 4 correspondences were also subdivided into two subtypes, 4.1 and 4.2. Type 4.1 corresponds to the original description proposed by Thunes: "cases where there are discrepancies between original and translation not only on the structural, but also on the semantic level" (Thunes, 1998, p. 28) because it is not possible to "derive equivalent semantic representations for source and target string" (ibid), as in Example (101):

(101) Sonnet XV

As he takes from you, I engraft you new.
Quanto ele em ti suprime, é quanto te acrescento. → type 4.1

Consequently, in correspondences of type 4.1, it is not possible to produce a translation with similar syntactic structure and that conveys the same meaning.

Type 4.2 represents those cases where a translation with similar syntactic structure and that conveys the same meaning is possible for that language pair, but it is clear that the translator chose to change the meaning. String pair (102) is an example of type 4.2 correspondence:

(102) Sonnet XCVII

Or if they sing, 't is with so dull a cheer,
Ou, caso cantem, de tal modo se entristecem, → type 4.2

The creation of this subcategory points to an interesting aspect related to the translator style, because when he/she chooses a different structure on both semantic and syntactic level, even when the language structure allows a similar construction, this shows that he/she is not so attached to the original.

The challenge during the analysis was the lack of criteria to determine the semantic correspondence between source and target string pairs. The analysis was interpretive, but I believe that a theoretical basis on semantics could refine the model to establish more objective criteria

to guide the analyst during the annotation process. Maybe this would lead to the creation of more subtypes within type 4 correspondences.

Since approximately 93% of the analysed string pairs are included in correspondences types 3 and 4, the subdivision of these categories into more subtypes could make the model even more efficient in pointing out differences between the translators' styles.

Finally, in relation to the extent that the analysis of complexity can point to stylistic differences between the translators, the adaptation of the model explains some of the translators' choices related to metre and to the extent that they were attached to the original text. Establishing clearer criteria to determine the semantic correspondence between the string pairs also could explain more precisely the differences between PS and JW on the semantic level.

5.5 Relevance of the study

Based on my review of the literature, I realised that some researchers resist the idea of using poetry for linguistic analysis. It seems that this type of literary text is not meant to be used as data if the analysis is not strictly interpretative. I hope that this investigation shows that poetry can also be subject of analysis in other levels of linguistic analysis.

The adaptation of Thune's model can be an alternative for those analysts who want to work with the translation of literary texts and approach the structural aspects of literary texts. Although the issue of subjectivity is still present during the analysis, the model also offers clear criteria to deal with this type of data and to cope with differences between the translators' writing styles, without taking the risk of falling into the issue of translation quality, which might be problematic.

Finally, with the purpose of clarifying the value of the present investigation and reinforcing value of corpus-based translation studies, I use Aijmer and Altenberg's words. First, "they give new insights into the languages compared – insights that are not likely to be gained via the study of monolingual corpora" (Aijmer & Altenberg, 1996, qtd. in McEnery & Xiao, 2008, p.18). Second, "they can be used for a range of comparative purposes and increase our knowledge of language-specific, typological and cultural differences, as well as universal features" (ibid). Third, "they illuminate differences between source texts and translations

and between native and non-native texts" (ibid). Finally, "they can be used for a number of practical applications, e.g. in lexicography, language teaching and translation" (ibid).

5.6 Further application

The adaptation of Thunes' model on translational complexity can be used to analyse other parallel corpora of the language pair English-Portuguese. It can be extended to other literary genres beyond poetry. It could be used as an instrument to identify stylistic differences between two or more translations of short stories, chronicles, theatre plays, the scope is unlimited. The only restriction is in relation to the corpus size, since the process of annotation needs to be manual.

It can also be an alternative for researchers interested in the analysis of features of explicitation and simplification. I am not claiming that the results obtained here somehow solve this issue that has been exhaustively discussed in Translation Studies and is still considered controversial. My point is that type 3 correspondences approach and explain the inclusion or omission of information during the translation process, without the need of mentioning the issue of universals of translation. As I said, it is an alternative, not a solution for a phenomenon that still needs to be investigated.

Finally, further application could involve an investigation that refines the issue of semantic correspondence, establishing clearer criteria or even a scale of equivalence to distinguish the target strings from the source string in a more subject way.

References

Primary sources

Shakespeare, W. (1971). *The Complete Works of William Shakespeare*. London; Spring Books.

Shakespeare, William. (2006). *Shakespeare's Sonnets* (K. Duncan-Jones, Ed.). London: The Arden Shakespeare.

Shakespeare, William. *Sonetos*
Trans. Jorge Wanderley. Rio de Janeiro: Civilização Brasileira S.A., 1991.

Shakespeare, William. *Sonetos*
Trans. Péricles Eugênio da Silva Ramos. São Paulo: Hedra, 2008.

Secondary sources

Aarts, J., & Meijs, L. (1984). *Corpus Linguistics*. Amsterdam: Rodopi.

Anderman, G., & Rogers, M. (2008). The Linguist and the Translator. In G. Anderman & M. Rogers (Eds.), *Incorporating Corpora The Linguist and the Translator* (pp. 5-17). Clevedon: Multilingual Matters Ltd.

Archer, D; Culpeper, J; Rayson, P. (2006). Love, - 'a familiar or a devil'? An Exploration of Key Domains in Shakespeare's Comedies and Tragedies. Retrieved June 10, 2009, from http://eprints.lancs.ac.uk/12671/

Azevedo, F. (2007). Desambiguação do Item Lexical Correto Através de Etiquetadores Semânticos. Dissertação de Mestrado (2007).

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). Amsnterdam/Philadelphia: John Benjamins Publishing Co.

Baker, M. (1995). Corpora in translation studies: an overview and some suggestions for future research. *Target – International Journal of Translation Studies, 7(2),* 223-243.

Baker, M. (1996). 'Corpus-based Translation Studies: The Challenges That Lie Ahead', in H. Somers (ed.) *Terminology, LSP and Translation*, Amsterdam and Philadelphia: John Benjamins.

Baker, M. (Ed.) (1998). Translation Studies. In *Routledge Encyclopedia of Translation Studies* (pp. 277-280). London and New York: Routledge.

Baker, M. (2000). Towards a Methodology for Investigating the Style of a Literary Translator. Retrieved June 25, 2011, from http://www.tracor.ufsc.br/seminario/uploads/texto/texto_target-2000-style_2011_05_29_20_26_08.pdf

Britto, P. H. (2000). A Humble Form. Retrieved November 30, 2012, from http://www.phbritto.org/2011/07/humble-form.html

Britto, P. H. (2001). Towards more objective evaluation of poetic translation. Retrieved November 30, 2012, from http://www.phbritto.org/2011/07/towards-more-objective-evaluation-of.html

Chandler, J. (2011). Otto Jespersen by Niels Haislund, 1943. Retrieved April 26, 2011, from http://interlanguages.net/haislund.html

Crystal, D. (1995). *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

Dyvik, H. (1990). *The PONS Project: Features of a Translation System. Skriftserie fra Institutt for fonetikk og lingvistikk* 39, B. University of Bergen.

Dyvik, H. (1999). On the complexity of translation. In: Hasselgard and Oksefjell (eds), 1999, 215-230.

Fromkim, V.; Rodman, R; Hyams, N. (2007). *An Introduction to Language*. Boston: Thomson Wadsworth.

House, J. (2008). Beyond Intervention: Universals in Translation? Retrieved June 10, 2009, from http://www.trans-kom.eu/bd01nr01/trans-kom_ _01_01_02_House

Johansson, S. (2007). Seeing through multilingual corpora. In R. Facchinetti (Ed.). *Corpus linguistics 25 years on* (pp. 1-21). Amsterdam: Rodopi.

Kilgarrif, A.; Grefesntette (2003). Introduction to the Special Issue on Web as Corpus. Retrieved May 15, 2012, from http://www.kilgarriff.co.uk/Publications/2003-KilgGrefenstette-WACIntro.pdf

Kilgarrif, A. (2010). A corpus factory for many languages. Retrieved May 15, 2012, from http://trac.sketchengine.co.uk/wiki/AK/Papers

Laviosa-Baithwaite, S. (1998). Universals of Translation. In *Routledge Encyclopedia of Translation Studies* (pp.288-291). London and New York: Routledge.

Leech, G. (1992). Corpora and theories of linguistic performance. In D. Svartvik (Ed.), *Directions in Corpus Linguistics Proceedings of Nobel Symposium 82 Stockholm*, 4-8 August 1991 (pp. 106-122). Morton de Gruyteer: Berlim.

Leech, G. (1997), 'Introducing corpus annotation', in Garside, Leech and McEnery Corpus Annotation: Linguistic Information from Computer Text Corpora, London: Longman.

Leech, G. (1998). *Corpus Annotation*. London & New York: Longman.

Magalhães, C., & Maria da Conceição Batista. (2002). Features in Translated Brazilian Portuguese Texts: a Corpus-Based Research. *Cadernos de Tradução* 10, 81-129.

McEnery, T., Xiao, R. (2008). Parallel and Comparable Corpora: What is Happening? In G. Anderman, & M. Rogers (Eds.), *Incorporating Corpora The Linguist and the Translator* (pp. 18-31). Clevedon: Multilingual Matters Ltd.

Rayson, P., Archer, D., Baron, A., Culpeper, J., & Smith, N. (2007). Tagging the Bard: Evaluating the Accuracy of a Modern POS Tagger on Early Modern English Corpora. In *Corpus Linguistics Conference (CL2007), July 2007*, (pp. 27-30). University of Birmingham, UK.

Sardinha, T. (2004). *Linguística de Corpus*. Barueri: Editora Manole Ltda.

Sarma. M. (2008). Translating Shakespeare: Intervention and Universals in Translation. Retrived June 10, 2009, from http://www.trans-kom.eu/bd01nr01/trans-kom_01_01_06_Sarma_Translating_Shakespeare

Scott, M. (1995). WordSmith Tools, Oxford University Press.

Thunes, M. (1998). Classifying translational correspondences. In S. Johansson and S. Oksefjell (Eds.). *Corpora and cross-linguistic research* (pp.25-50). Amsterdam: Rodopi.

Thunes, M. (2011). *Complexity in Transaltion, An-English-Norwegian Study of Two Text Types*. Unpublished doctoral dissertation, University of Bergen, Norway.

Appendix

Sonnet CV

Let not my love be call'd idolatry,
Oh! ninguém chame idolatria o meu amor,
Não chamem meu amor de idolatria
Nor my beloved as an idol show,
Nem dê por ídolo quem alvo é desse preito,
E que o meu bem um ídolo não lembre
Since all alike my songs and praises be
Porque todo o meu canto e todo o meu louvor
Por ser só dele a minha poesia
To one, of one, still such, and ever so.
São para alguém, de alguém, e sempre, e de um só jeito.
E eu louve um só e louve o mesmo e sempre.
Kind is my love today, tomorrow kind,
Meu amor hoje é afável, amanhã afável,
Suave é hoje e sempre revelado
Still constant in a wondrous excellence;
Sempre constante numa esplêndida excelência:
Na mesma maravilha em que cintila
Therefore my verse to constancy confin'd,
Logo meu verso, limitado ao invariável,
E o meu verso, à constância confinado
One thing expressing, leaves out difference.
Exprime uma só coisa, e exclui a impermanência.
Expressa o mesmo – e o diferente exila
Fair, kind and true is all my argument,
"Bom, belo e verdadeiro" – é um só meu argumento,
"O belo, o bem, a verdade", eis o tema;
Fair, kind, and true varying to other words,
"Bom, belo e verdadeiro" – em vária locução:
"O belo, o bem, a verdade" – a variação.
And in this change is my invention spent,
Nessa mudança absorvo tudo quanto invento,
E neste espaço inventa o meu poema
Three themes in one, which wondrous scope affords.
Três temas postos num, de amplíssima extensão.
Num só, três temas – que infinitos são.
Fair, kind, and true, have often lived alone,
"Bom, belo e verdadeiro" alheios têm vivido:
O belo, o bem, a verdade, antes sós,
Which three till now never kept seat in one.
Num ser ainda não se haviam reunido.
Agora assentam numa mesma voz.

Sonnet XXIX

When, in disgrace with fortune and men's eyes,
Desvalido em fortuna e aos olhos dos mortais,
Quando à margem da sorte e dos olhares
I all alone beweep my outcast state
Quando choro sozinho ao ver-me rejeitado
Dos homens todos choro o meu estado
And trouble deal heaven with my bootless cries
E os surdos céus perturbo em vão com altos ais
E o surdo céu perturbo erguendo aos ares
And look upon myself and curse my fate,
E amaldiçôo a sorte olhando meu estado,
Este inútil lamento ante o meu fado,
Wishing me like to one more rich in hope,
E almejo ser alguém, bem mais esperançado,
Sonho ser outro, com mais esperança,
Featured like him, like him with friends possess'd,
De que tivesse o aspecto e as ricas amizades,
Cheio de amigos e bem parecido,
Desiring this man's art and that man's scope,
E com o que fruo mais o menos contentado,
Querendo as artes de um, o que o outro alcança,
With what I most enjoy contented least;
Quero a arte deste, e doutro as oportunidades:
E sou, do que mais amo, desvalido.
Yet in these thoughts myself almost despising,
Quase que me desprezo, em coisas tais cuidando;
Mas mesmo assim, quase me desprezando,
Haply I think on thee, and then my state,
Mas penso em ti, e logo a minha condição,
Eu me lembro de ti e o meu destino
Like to the lark at break of day arising
Qual cotovia na alva a terra abandonando,
(Qual cotovia na manhã se alçando)
From sullen earth, sings hymns at heaven's gate;
Ergue às portas do céu hinos de gratidão;
Da terra exausta ao céu levanta um hino:
For thy sweet love remember'd such wealth brings
Pois traz-me tal riqueza o teu amor lembrado,
Que tendo o teu amor, recusarei
That then I scorn to change my state with kings.
Que desdenho trocar com os reis o meu estado.
Meu destino trocar pelo dos reis.

Sonnet XIX

Devouring Time blunt thou the lion's paws,
Cega, ó Tempo voraz, as garras do leão,
Tempo voraz, que ao leão lima as garras,
And make the earth devour her own sweet brood,
Pluck the keen teeth from the fierce tiger's jaws
E dos tigres arranca os dentes à maxila;
Faze que a terra coma a própria geração,
Que à terra faz comer filhos da terra
E ao tigre arranca as presas da bocarra,
And bum the long-liv'd Phoenix in her blood,
E a fênix, no seu sangue em flamas, aniquila!
E queima a fênix no sangue que encerra
Make glad and sorry seasons as thou fleet'st,
Fugindo, as estações alegra ou entristece;
E passa e deixa a estação bela ou triste,
And to whate'er thou wilt, swift footed Time,
Dispõe, Tempo dos pés velozes, do universo,
Faz como queiras, com teu pé veloz,
To the wide world and all her fading sweets:
E de quanta doçura, eu sei, nele esmaece;
Ao que declina, ao que no mundo existe;
But I forbid thee one most heinous crime,
Porém eu te proíbo um crime mais perverso:
- Mas te proíbo o crime mais feroz:
O carve not with thy hours my love's fair brow,
Nor draw no lines there with thine antique pen,
Não queiras entalhar de meu amor a fronte
Com tuas horas, nem riscá-la com tua pena
Com as horas talhar a fronte amada,
Vincá-la com teu cálamo maduro;
Him in thy course untained do allow,
Antiga; mas que puro, ó Tempo, ele defronte
Permite que em teu curso a meu bem nada
For beauty's pattern to succeeding men.
Os pósteros – padrão de formosura plena
Perturbe, que é padrão para os futuros;
Yet, do thy worst, old Time, despite thy wrong,
Faze o pior, porém: malgrado o teu rigor,
- Ou causa, tempo, os teus maiores danos;
My love shall in my verse ever live young.
Sempre jovem será em meus versos meu amor.
Meu verso traz meu bem à flor dos anos.

Sonnet XVIII

Shall I compare thee to a summer's day?
A um dia de verão como hei de comparer-te?
- Comparar-te com um dia de verão?
Thou art more lovely and more temperate
Vencendo-o em equilíbrio, és sempre mais amável:
Tens mais doçura e mais amenidade:
Rough winds do shake the darling buds of May,
Em maio o vendaval ternos botões disparte,
Flores de maio, ao vento rude vão
And Summer's lease hath all too short a date:
E o estio se consome em prazo não durável;
Como o estio se vai, com brevidade:
Sometime too hot the eye of heaven shines,
Às vezes, muito quente, o olho do céu fulgura,
O sol às vezes em calor se exalta
And often is his gold complexion dimm'd,
Outras vezes se ofusca a sua tez dourada;
Ou tem a essência de ouro sem firmeza
And every fair from fair sometime declines,
Decai da formosura, é certo, a formosura,
E o que é formoso, à formosura falta,
By chance, or nature's changing course untrimm'd:
Pelo tempo ou o acaso enfim desadornada:
Por sorte ou por mudar-se a natureza
But thy eternal summer shall not fade,
Mas teu verão é eterno, e não desmaiará,
Mas teu verão eterno brilha a ver-te
Nor lose possession of that fair thou ow'st,
Nem hás de a possessão perder de tuas galas;
Guardando o belo que em ti permanece.
Nor shall Death brag thou wander'st in his shade,
Vagando em sua sombra o Fim não te verá,
Nem a morte rirá de ensombrecer-te,
When in eternal lines to tome thou grow'st:
Pois neste verso eterno ao tempo tu te igualas:
Quando em verso imortal, no tempo cresces.
So long as men can breathe or eyes can see,
Enquanto o homem respire, e os olhos possam ver,
Enquanto o homem respire, o olhar aqueça,
So long lives this, and this gives life to thee.
Meu canto existirá, e nele hás de viver.
Viva o meu verso e vida te ofereça.

Sonnet LXI

Is it thy will, thy Image should keep open
My heavy eyelids to the weary night?
Mandas a imagem tua, e as pálpebras pesadas
Fazes que não as cerre, em noite de cansar?
Só por tua vontade estão abertas
As pálpebras que à noite espero tombem?
Dost thou desire my slumbers should be broken,
É tua decisão que o sono meu se evada
Desejas que eu não durma e da desperta
While shadows like to thee do mock my sight?
E sombras – cópias tuas – zombem deste olhar?
Minha visão as tuas sombras zombem?
Is it thy spirit that thou send'st from thee
So far from home into my deeds to pry,
Para me espiar longe de sua residência
É mesmo o teu espírito que tu me envias,
É teu espírito que a mim me mandas
De lá de longe a me espiar em casa,
To find out shames and idle hours in me,
A fim de achar vergonhas e horas de indolência
Ver se o que faço, a preguiçar desanda?
The scope and tenure of thy jealousy?
Que sejam alvo e teor de quanto desconfias?
Este o cuidado, a posse que te abrasa?
O no, thy love though much, is not so great,
Oh, não! teu grande amor não é tão grande assim:
Oh, não, tens muito amor, mas não tão grande:
It is my love that keeps mine eye awake,
Quem meus olhos descerra é apenas meu amor,
Meu é o amor que insone me mantém,
Mine own true love that doth my rest defeat,
É meu amor que em meu repouso põe um fim;
Meu verdadeiro amor que em mim comande
To play the watchman ever for thy sake.
E eu para te vigiar me fiz tresnoitador.
Seja eu vigia sempre, por teu bem.
For thee watch I, whilst thou dost wake elsewhere,
Além não dormes, e por ti fico desperto,
Por ti vigio, enquanto estás, desperto,
From me far off, with others all too near.
Que longe estás de mim, mas de outros muito perto.
Longe de mim, com outros muito perto.

Sonnet LV

Not marble, nor the gilded monuments
De mármore não sei, nem de áureos monumentos
Nem mármore nem áureos monumentos
Of princes shall outlive this powerful rhyme,
Que sobrevivam ao meu canto poderoso:
De príncipes, meus versos desmerecem;
But you shall shine more bright in these contents
Than unswept stone, besmear'd with sluttish time.
O tempo mancha a pedra, enquanto em meus acentos
Tu sempre ostentarás um brilho vigoroso.
Nos versos tens mais brilho e mais alento
Que em pedra rude, que o tempo enegrece.
When wasteful war shall statues overturn,
Quando estátuas a Guerra infrene derruir
Quando a guerra as estátuas devastar
And broils root out the work of masonry,
E as próprias construções das bases arrancar,
E virar pelo avesso alvenarias,
Nor Mars his sword, nor war's quick fire shall burn
Não poderão espada ou fogo derruir
Nem Marte, espada, ou fogo militar
The living record of your memory.
Este arquivo imortal que te há de relembrar.
A queimar-te a memória bastariam.
'Gainst death, and all-oblivious enmity
Indiferente a morte e a olvido hás de viver,
Contra a morte e o maligno esquecimento
Shall you pace forth, your praise shall still find room,
E encontrará guarida o teu louvor supremo
Avançarás, teu valor indiviso,
Even in the eyes of all posterity
No olhar das gerações que se hão de suceder
E aos olhos do futuro, o acatamento
That wear this world out to the ending doom.
Até que o mundo atinja o seu momento extremo.
Num mundo gasto é o final Juízo.
So till the judgment that yourself arise,
Assim, até o juízo em que despertarás,
Portanto, até que a te julgar te chamem
You live in this, and dwell in lovers' eyes.
Em meu verso e no olhar dos que amam viverás.
Vives no verso e aos olhos de quem ame.

Sonnet CIX

O never say that I was false of heart,
Falso, o meu coração? Não formes tal conceito,
Não digas falso o coração em mim,
Though absence seem'd my flame to qualify.
Se, quando longe, o ardor pareço moderar:
Mesmo se ausente à chama ele se faz:
As easy might I from myself depart,
As from my soul, which in thy breast doth lie:
Ir de minh'alma, que reside no teu peito,
É a mesma coisa que de mim eu me afastar.
A mim mesmo eu faltara, leve assim,
Ou a minha alma, que em teu peito jaz.
That is my home of love, if I have rang'd,
Se desse lar de amor eu me separo e vago,
Mora ali meu amor: e se eu vagueio,
Like him that travels I return again,
Como um viajante vou, e logo eis-me tornado:
Sou viajante já da volta escravo,
Just to the time, not with the time exchang'd,
So that myself bring water for my stain:
A água que lave minha mancha eu mesmo trago,
No prazo certo, e pelo tempo não mudado.
Que veio a tempo e sem mudança veio:
Trago-me a água com que as manchas lavo.
Never believe, though in my nature reign'd,
Não creias, posto minha índole apresente
Não creias, mesmo vendo a me cercar
All frailties that besiege all kinds of blood,
Toda a fraqueza que assedia a humanidade,
A fraqueza que o humano sangue atenta,
That it could so preposterously be stain'd,
Que eu pudesse manchar-me tão absurdamente,
Fosse o meu sangue – absurdo! – se manchar
To leave for nothing all thy sum of good:
Trocando por um nada a suma da bondade.
Dando por nada os bens que representas.
For nothing this wide universe I call,
Pelo nome de "nada" o do universo eu mudo,
Que o mundo é nada, eu digo – e está desnudo,
Save thou my rose, in it thou art my all.
Menos tu, minha rosa: nele, és o meu tudo.
Exceto, rosa, em ti – que és nele tudo.