ABSTRACT: This paper compares “The Da Vinci Code” and its translation in Portuguese against the language of canonical and non-canonical literary works. Based on the notion of literariness, as forwarded by the Russian Formalists, the study looks at the lexical density of these works. Two corpora are compiled and formulaicity is checked with the help of computers. Classification of lexical bundles is then carried out and seen in the light of socio-cultural differences. The results show that formulaic expressions are used differently in these works and that there are slight changes in the types of lexical bundles in both corpora. This paper contributes to the characterization of the language of literary texts by using empirical observations rather than hermeneutic argumentation. In addition, it brings out the advantages of corpus analysis for the study of literature.

KEYWORDS: Corpus Linguistics; lexical bundles; literary theory; canon; translation

1. Introduction

Defining a literary text has proven to be one of the most complex issues in literary theory as it involves values and perceptions. In its simplest terms, a literary text is a collection of linguistic signs that function within a certain cultural context and operates on several dimensions: the existential, the reflexive and the functional (ZYGIER, 1994, p. 83). It would be out of the scope of this paper to hold a discussion of what determines a literary text. Instead, two questions generally asked of work of literature are dealt with: “Is the language used literary?” and “To what extent is a text canonical?” These two questions tend to be seen as interchangeable, but they refer to quite different issues, and the distinction is worth revisiting.

The process of selecting a body of texts to be taught in schools and used as models is inevitable. But the property of belonging to the canon, or canonicity, does not depend on inherent textual features. It is rather a product of the operation of a complex cultural system with social as well as political implications. Canonized texts are “accepted as legitimate by the dominant circles within a culture and whose conspicuous products are preserved by the community to become a part of its historical heritage.” (EVEN-ZOHAR, 1990, p. 15). Canonicity is not necessarily linked to language use.

The notion that literary texts had properties which could be the object of scientific investigation was first proposed by the Russian Formalists in the early years of the 20th century. The first scholar to systematize the concept of literariness was Jakobson as early as 1921 (see reprint in JAKOBSON, 1971). In his view, literature was a special use of language, which contained specific organization and structural devices. Since then, literariness has been interpreted in many different ways. Its existence as an intrinsic feature of texts is controversial (see EAGLETON, 1996), and it has come to be understood as a product of the relations between different uses of language, with the literary use standing out against a background of more ordinary uses. Today, literariness is not considered an exclusive feature of literature; in
fact, it can be found in texts which are not necessarily literary. However, while not adequate for drawing a sharp line between the literary and the non-literary, the term can be used to refer to the more or less distinct presence of certain linguistic features which can account for the quality of language use. In Carter and Nash’s (1990, p. 29) words, one can actually draw “a ‘cline’ or gradation of stylistic and discursive qualities” which will respond for the literariness of a text.

It is assumed in this study that literary uses of language are characterized by the innovative manipulation of linguistic patterns, thereby breaking with readers’ expectations. Literariness is thus closely related to unpredictability; on the other hand, it contrasts sharply with formulaicity, a characteristic of uses which are highly conventional and tend to agree with, rather than defy, the addressee’s expectations (see Wray, 2002). Whereas the more formulaic language of daily interactions tends to rely on predictability for economy of usage, the language used in literary works in prose is characterized by creativity, innovation, and unexpectedness.

Grounded on the notion that certain textual features may reveal elements of literariness, this study uses Corpus Linguistics to look at the relations between literariness and canonicity, and at how the quality of the language in an original text may be altered in the process of translation. The questions which guide the present research are the following:
(a) Are there elements of literariness in non-canonical works of literature?
(b) Does the process of translation have any influence over the literariness of The Da Vinci Code, a non-canonical work?

Based on the notions of literary, and using tools from Corpus Linguistics, it is focused on the novelty / formulaicity of language use in literary works in prose. This investigation is thus grounded on objectively collected data.

2. Theoretical background

This section, which is divided into two parts, covers the main theories that guide the present study. First, a review of literary theory is presented. Three main historical phases are described and their differences are discussed. The second subsection covers Corpus Linguistics so as to make clear the differences between this branch of Linguistics and Generativism. The importance of computers is highlighted and some concepts in this area of study are explained. The nature of language choices is also considered: it is discussed whether these choices are free or restricted in some way. Finally, a definition of the concept of lexical bundle is forwarded.

2.1. Literary theory

One of the most relevant notions that helped define literariness was ostranenie (i.e. “making strange”, defamiliarization), introduced by Shklovsky (1965 [1917]). Shklovsky, one of the leading Russian formalists, postulated that the purpose of art was to make objects unfamiliar in order to prolong viewers’ perceptions and thus allow them to experience artfulness. This concept led to the notion of foregrounding, that is, how certain elements stand out when readers experience them as strange or different (see review in Van Peer, 1986). In other words, formalists proposed a literary language which deviated from the norm (which norm they referred to was not clear) and they concluded that literariness could be found in the text itself.

1 It must be pointed out, however, that everyday language can also be highly unpredictable, as in advertisements or newspaper headlines.
In the second half of the 20th century, literary studies turned from the Formalists’ focus on texts to the primacy of readers. Reception theories, particularly in America and Germany, held that literariness depended on the act of reading, emphasizing readers’ roles in creating meaning. Reader-response theories recognized readers as active agents who imparted real existence to the work and completed its meaning through interpretation. As a result, literature should be viewed as a kind of performing art in which readers created their own text-related performances.

One of the leading voices in reception theory, Jauss (1978) held that literature was a dialectic process of production and reception; readers had a certain mental set, a horizon of expectations, which was activated during reading and took into account both social and cultural aspects. On another note, Iser (1974) stated that every literary work required an “implied reader”, i.e., a hypothetical construction which embodied all predispositions necessary for it to exercise its effect. Although readers and texts assumed similar conventions from reality, texts left great portions unexplained to readers, whether as gaps in the narrative or as structural limits of the representations of the world made by texts. This basic indeterminacy invited readers to participate in synthesizing meaning throughout the process of reading.

Although there was an effort in increasing the relevance of the act of reading in meaning-making, it was noticeable that the theories developed by the reader response scholars defined readers in terms of the text, confining their activities within the limits set by the literary work. Until that moment, there was still a need for the development of theories that focused not only on texts or readers, but on the social system in which Literature resides.

In 1973, Schmidt (1982) and other literary scholars in the University of Bielefeld in Germany created the NIKOL group (Nicht Konservativ Literaturwissenschaft, or Non-Conservative Science of Literature) as an attempt to develop a science of literature based on the view that it was a system in which four social roles can be identified: the producer, receiver, mediator and post-processor. Under this paradigm, literariness was not restrained to the formal characteristics of texts or psychological aspects of reading, but to a set of conventions defined by authors, readers, publishers and literary critics within a historical context. This basic assumption underlied the principles of the Empirical Study of Literature (henceforth ESL), a theory of systems that involved all the actions which affect human communication.

According to the ESL, the written text was considered a linguistic basis for the development of a KOMMUNIKAT\(^2\), i.e., a mental construct created by readers. Thus, to be considered literary, a text should be produced and received as such by the actors involved in the interpretative community, stressing the need for the development of empirical studies on the roles taken in the literary system.

Therefore, the notion of lexical density on which this study is grounded find resonance in the way the Russian Formalists defined literariness, as it focuses on language creativity and innovation in both canonical and non-canonical texts.

\subsection*{2.2. Corpus Linguistics}

Before the 80’s, the field of Linguistics differed greatly from that of today. As Sinclair (1991, p. 1) pointed out, “the tradition of linguistics has [had] been limited to what a single individual could experience and remember”. At that time, the Chomskyan theory had already influenced a great number of university centers, especially in the United States. Linguistic studies were not based on data, but on introspection. The emphasis was on competence rather

\footnote{The convention is to use capital letters here.}
than on performance, believed to be too fuzzy to be analyzed. This led to a focus on the possible general structures of languages, obtained by means of intuition.

As computers started to become accessible to an increasing number of people and larger amounts of data were made available, the arguments against the use of real data in linguistic studies were weakened. It became possible to probe data by means of computer tools, which made the analysis more reliable and less dependent on human beings. It was at this point that Corpus Linguistics made its (re)appearance in the academic setting.

Corpus Linguistics constitutes a new way of looking at language. It avoids intuitions and is grounded on samples of language in use. Within this field, “large quantities of ‘raw’ text are processed directly in order to present the researcher with objective evidence” (SINCLAIR, 1991, p. 1 – our emphasis).

As its name suggests, Corpus Linguistics concerns the study of a corpus which is considered to be representative of a given language or one of its features. This is achieved by making a principled selection of texts with a clear aim in mind. The objective of a corpus needs to be defined before the selection is carried out. As Corpus Linguistics entails the study of language with the help of computers, it is also necessary that the corpus be an electronic one.

Investigating a corpus by means of the computer tool allows researchers to have several features of the corpus at one mouse click. Virtually all computer programs which analyze corpora will provide their tokens, types and type/token ratio. ‘Token’ is the term used to refer to the running words in a corpus. ‘Type’, on the other hand, refers to the different words which can be found in a corpus. If the number of different words is divided by the number of running words, the result will be a measure of lexical variety at word level in a corpus. Such a measure is called ‘type/token ratio’ and is expressed in percentage terms. The more varied the lexical items in a corpus are, the higher its ratio is. However, as one may argue, this ratio depends on corpus size. There is a probability that the larger the corpus, the lower its type/token ratio. This is why the standardized type/token ratio is more commonly used. Different from the first one, this ratio is not calculated for a whole corpus, but for groups of \( n \) tokens. The final result, that is, the standardized type/token ratio, is an average of the individual ratios.

One important concept in Corpus Linguistics is the notion that words do not generally occur on their own, but in the presence of absence of others. It is also believed that while some language choices are made on the level of isolated words, most choices are made on sequences of words. Sinclair (1991) explains that there are two principles in action: the idiom and the open-choice principles. The latter accounts for the choices made on the level of the word, that is, words are chosen individually – one by one – as the situation presents itself. The only constraint in this case is the grammar of that language. However, Sinclair (1991) argues that most choices are not based on words, but on longer sequences that include more than one word. This is explained by the idiom principle:

The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments […] However it arises, it has been

---

3 There were investigations based on real data before Generativism, but they were not carried out in the same way it is done nowadays for researchers did not have the necessary equipment at that time. Even under the influence of Generativism, data were still used in phonological studies.

4 Thus the result of the previous division needs to be multiplied by 100.

5 In this study, the value of \( n \) corresponds to 1,000 tokens.

6 This general idea would help explain the concepts of collocation and colligation, for instance. Both of them refers to the use of one word together with another one (collocation), or with a grammatical class or structural pattern (colligation).
relegated to an inferior position in most current linguistics, because it does not fit the open-choice model (SINCLAIR, 1991, p. 110).

More recently, Biber et al. (2004) brought to the forefront the sequences of words that are used by speakers. In their study, they question the assumption that language is strictly compositional, arguing that users resort to multi-word units more often than it is believed. With this approach to language, they decide to investigate the lexical bundles in their research corpus. Lexical bundles, in Biber et al.’s (2004, p. 376)\(^7\) words, correspond to “the most frequent recurring lexical sequences in a register”\(^8\). In other words, they are sequences of words that occur a number of times in a research corpus. It is important to stress here that not every single sequence of words listed by the computer is considered a lexical bundle. In Biber et al.’s (2004) study, frequency and dispersion criteria were used, so a sequence had to occur at least 40 times per million words and in five different texts.

3. Lexical bundles: structural and functional taxonomies

In addition to working with the concept of lexical bundle, Biber et al. (2004) proposed two taxonomies for classifying bundles: structural, which takes into account the types of grammatical patterns found in bundles, and functional, which covers the meanings conveyed by bundles.

The structural taxonomy contains three major types. Figure 1 below illustrates them with examples from the research corpora for this study. It should be noted that the examples in Portuguese do not correspond to the ones in English. They are intended to be different.

<table>
<thead>
<tr>
<th>Type 1</th>
<th>O Código Da Vinci (Portuguese)</th>
<th>The Da Vinci Code (English)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 2</td>
<td>tenho certeza de que que o Santo Graal</td>
<td>that ought be on he was about to</td>
</tr>
<tr>
<td>Type 3</td>
<td>do outro lado da o Priorado de Sião</td>
<td>the Priory of Sion of the Holy Grail</td>
</tr>
</tbody>
</table>

Figure 1: Structural taxonomies – some examples

Type 1 bundles are those which contain a simple verb phrase. Given the scope of lexical bundles (only four words in this study), most bundles will not contain a complete phrase, but just a fragment. This can be seen in the bundle ‘the holy grail is’, which starts with a noun phrase and is followed by a verb. The verb itself is the only element of the verb phrase that can be found in the bundle. Bundles from the second type consist of those sequences that incorporate (fragments of) dependent clauses in addition to simple verb phrases. These bundles can start with the dependent clause itself (as is the case of ‘that ought be on’) or with the main clause followed by a particle which signals the dependent clause (as in the bundle ‘he was about to’. Finally, Type 3 lexical bundles contain noun phrases (‘the Priority of Sion’) or prepositional ones (‘of the Holy Grail’) either complete or incomplete.

---

\(^7\) It should be made clear that Biber first proposes this concept in Biber et al. (1999).

\(^8\) As Biber et al. (2004, p. 372) put it, these sequences of words have been studied by a number of researchers and have been labeled differently: ‘fixed expressions’, ‘formulas’, ‘lexical bundles’, ‘lexical phrases’, ‘prefabricated patterns’, ‘prefabs’ and ‘routines’, for instance.
As far as the functional taxonomy runs, there are also four main categories: stance, discourse organizer, referential and conversational. These divisions are exemplified in Figure 2 below.

<table>
<thead>
<tr>
<th></th>
<th>O Código Da Vinci (Portuguese)</th>
<th>The Da Vinci Code (English)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stance</strong></td>
<td>tenho certeza de que</td>
<td>that ought be on</td>
</tr>
<tr>
<td></td>
<td>não podia deixar de</td>
<td>he was about to</td>
</tr>
<tr>
<td><strong>Discourse</strong></td>
<td>à medida que o</td>
<td></td>
</tr>
<tr>
<td><strong>Referential</strong></td>
<td>de Custódia de Zurique</td>
<td>the back of the</td>
</tr>
<tr>
<td></td>
<td>com todo o cuidado</td>
<td>the Salle des États</td>
</tr>
<tr>
<td><strong>Conversational</strong></td>
<td>está me dizendo que</td>
<td>I beg your pardon</td>
</tr>
<tr>
<td></td>
<td></td>
<td>it speaks of Rosy</td>
</tr>
</tbody>
</table>

Figure 2: Functional taxonomy – some examples

Stance bundles are defined by Biber et al. (2004, p. 384) as those which “express attitudes or assessments of certainty that frame some other proposition”. They can signal, among other things, how certain/uncertain somebody/something is, what somebody wants to be done, what is important/unimportant or necessary/unnecessary to be done, for instance. Discourse organizing bundles are those which help readers/listeners understand the relationship between what has been said/written and the information which follows it. There are no instances of discursive bundles in *The Da Vinci Code*. In its translation in Portuguese, there is one type: the bundle ‘à medida que o’, which corresponds in English to ‘to the extent that’. The referential category “makes direct reference to physical or abstract entities, or to the textual context itself, either to identify the entity or to single out some particular attribute of the entity as important” (BIBER et al. 2004, p. 384). The last type of functional bundle is that which comprises sequences of words that are generally used in conversations such as ‘I beg your pardon’ to signal politeness and ‘it speaks of Rosy’, which concerns reported speech.

Both the structural and functional taxonomies, which have been outlined in this section, were used to classify the lexical bundles in *The Da Vinci Code* and *O Código Da Vinci*. The results of this analysis are detailed in Section 5.

4. Methodology

In order to investigate the language used in *The Da Vinci Code* both in its original version in English and in its translation into Portuguese, the texts had to be in electronic form, more specifically in TXT files. This procedure was divided into three phases. First, the books were digitized manually by one of the researchers with the help of a scanner and an OCR program. Second, the resulting files were read by another researcher with a view to minimizing the number of mistakes in the files. It was also during this phase that the texts

---

9 In Biber et al.’s (2004) functional taxonomy, each category is divided in a number of subcategories. The latter will not be discussed here as they will not be used in the present study.
10 The examples in Portuguese are not translations of the English ones.
11 We are thankful to Suzana de Deus who helped us in the very first part of the formatting of the translated version of *The Da Vinci Code* in Portuguese.
12 For instance, one common problem that had to be dealt with in this phase was to change the number ‘1’ to the letter ‘l’ since the OCR program automatically transformed the letter into the former.
were formatted accordingly. Finally, a third researcher read the texts once more to ensure no problems were left in the files. The final results of this three-phase procedure were the digital copies of *The Da Vinci Code* in English and in Portuguese in two separate corpora.

Once the texts had been transformed into electronic files, they could be probed with a computer program, namely, WordSmith Tools (SCOTT, 1999). This program has three main tools: WordList, Concord and KeyWords. In this research, the most used tool was WordList, which compiles lists of words or sequences of words in frequency or alphabetical order. This tool also provides analysts with a number of other features about the research corpus such as the number of types and tokens, the type/token ratio and the standardized type/token ratio, which have been discussed in Section 2.2. To a lesser extent, the Concord tool was also used when it was necessary to investigate the context – that is, the linguistic environment – of a specific lexical bundle before it was classified.

WordSmith Tools allows the corpora used here to be described. *The Da Vinci Code* totalled 142,655 tokens and 11,123 types, with a standardized type/token ratio of 46.17%. *O Código Da Vinci* (the version in Portuguese) amounted to 148,214 tokens and 14,774 types. Its standardized type/token ratio was 49.21%.

As far as lexical bundles were concerned, the researchers decided to work with four-word sequences following Biber et al.’s (2004) study so as to avoid repeated instances of overlapping, which would occur if lexical bundles were to contain three words only. A cut-off frequency of six times was adopted. In other words, for a sequence of words listed by the computer to be considered a lexical bundle and thus included in the analysis, it had to occur at least six times in the research corpora. Finally, another criterion was adopted to identify lexical bundles: they had to start with a phrase – either nominal, prepositional or verbal. Consequently, sequences of words like ‘depositary bank of Zurich’ and ‘far end of the’, for instance, were excluded from the analysis. However, similar sequences, containing the definite article ‘the’, were considered lexical bundles: ‘the depositary bank of Zurich’ and ‘the far end of’.

5. Results

If literariness in prose is defined by means of creativity, innovation and unexpectedness as pointed out in the introduction to this paper, it is expected that the linguistic choices made by authors should indicate variety. Within Corpus Linguistics, it is the

---

13 Due to certain constraints, the formatting procedure is not going to be detailed in this article, but some of the rules which had been followed can be mentioned here. Splitting syllables was not allowed even if they had been split in the original text otherwise split words would be considered different from their non-split forms. Periods were inserted after the heading of a chapter, the prologue or the epilogue. This was done so that the computer program recognized it as the end of that specific sequence of words. All types of visual elements were discarded as they could not be handled in TXT files, but it was inserted a code (‘<sp’ without the single quotes) to mark the deletion of an element in the print version. The code in this case is ignored by the computer program. In the case of tables, their textual content was copied and periods inserted at the end of each cell.

14 As pointed out in Section 2.2, the (regular) type/token ratio is not going to be used in this study because it is subject to the size of the corpus. In order to exemplify the point which has been made previously, the (regular) type/token ratios for both *The Da Vinci Code* and *O Código Da Vinci* can be mentioned: 7.80% and 9.97%, respectively.

15 This absolute frequency corresponds to 4 times in 100,000 tokens.

16 A dispersion constraint was not adopted in this study in Biber et al.’s (2004) because each research corpus consists of only one text.

17 In Portuguese, two of the bundles which were discarded are ‘banco de custódia de’ and ‘outro lado da sala’ for not starting with a preposition. This means that they were included in the analysis as ‘do banco de custódia’ and ‘do outro lado da’.
standardized type/token ratio which signals how diverse the words in a specific corpus are.\textsuperscript{18} The richer the lexical choices, the higher the ratio. However, not only should the standardized type/token ratio be taken into account, it is important that the words are arranged in different ways, forming distinct combinations. It is at this stage that the number of lexical bundles matters to the present study. The fewer bundles, the richer the language. Therefore, literariness can be attributed to a high standardized type/token ratio and a low relative frequency\textsuperscript{19} of lexical bundles. Figure 3\textsuperscript{20} shows how both \textit{O Código Da Vinci} and \textit{The Da Vinci Code} vary along these two variables.

![Figure 3: The Da Vinci Code and O Código Da Vinci compared](image)

Figure 3 shows that \textit{O Código Da Vinci} has a standardized type/token ratio of 49.21\% and 4.86 lexical bundles per 10,000 tokens. \textit{The Da Vinci Code}, on the other hand, has a standardized type/token ratio of 46.17\% and 14.72 lexical bundles per 10,000 tokens. It seemed at first that the translation process could have had an influence in the original text, making it lexically richer than it is in English. There was, however, another possibility: \textit{O Código Da Vinci} could have scored higher than \textit{The Da Vinci Code} because of the Portuguese language. Texts in this language may have a tendency to use more varied lexical items, which are also ordered in different ways.

With a view to answering the previous question, there was a need for to enlarge the number of literary works being analyzed. Another non-canonical literary work originally written in Portuguese – \textit{O Alquimista} by Paulo Coelho – and its translated version in English

\textsuperscript{18} This ratio has been explained in Section 2.2.
\textsuperscript{19} As both corpora have different sizes (148,214 tokens for \textit{O Código Da Vinci} and 142,655 tokens for \textit{The Da Vinci Code}), it is necessary to work with relative frequencies of lexical bundles to ensure that the comparison can be made. For the sake of this study, frequencies will be expressed per 10,000 tokens.
\textsuperscript{20} As has been stated, literariness is connected with a high standardized type/token ratio. This means that the most literary work should be on top of the graph if axis Y is considered. However, in relation to lexical bundles, literariness is defined in the other way around. The minimum this value assumes, the more literary a text is. Therefore, a richer literary work will be on the left side of axis X. As a summary, the ideal literary work and also the richest one in terms of lexical items and combinations is that which has 100\% of standardized type/token ratio and a relative frequency of lexical bundles equal to 0. Graphically, this literary work would be displayed on top and on the left of the graph.
– *The Alchemist* were added. Two other literary works considered to be canonical were also included – *Grande Sertão: Veredas* by Guimarães Rosa and *Memórias Póstumas de Brás Cubas* by Machado de Assis – as well as their English translations – *The Devil to Pay in the Backlands* and *Epitaph of a Small Winner*, respectively.  

Figure 4 shows how these eight literary works vary in terms of both standardized type/token ratio and relative frequency of lexical bundles.

![Figure 4: Four literary works and their translations compared](image)

Considering each pair of literary works (in Portuguese and in English), it can be noticed in Figure 4 that the Portuguese version scores higher than the English version irrespective of the language in which it was originally written. In other words, *O Código Da Vinci*, *Grande Sertão: Veredas*, *Memórias Póstumas de Brás Cubas* and *O Alquimista* are more diverse in terms of words and their sequences than their counterparts in English. This seems to have answered the previous question whether it is the translation process or the language itself which matters to the literariness of a given literary work. It may be inferred from the results shown in Figure 4 that the four texts in Portuguese are richer than the four texts in English. Even though *The Da Vinci Code* was originally written in English and then translated into Portuguese, the translation does not seem to have changed its nature. The Portuguese version of *The Da Vinci Code* is richer because that may be a feature of Portuguese.

Another look at Figure 4 reveals an interesting finding about the literary works analyzed in relation to their status in the canon. It appears to have two different groups in Figure 4: one containing those literary works which are regarded as canonical by literary critics and another one encompassing *O Alquimista* and *The Alchemist*, which differs greatly from the first group. If these two groups are recognized, there is an unexpected result: both *O*...
Código da Vinci and *The Da Vinci Code* seem to be closer to the canonical literary works than to the non-canonical ones, as would be expected at first. *O Código Da Vinci* has a standardized type/token ratio of 49.21%, which is lower than the one presented by *Grande Sertão: Veredas* (49.86%) and slightly higher than that of *Memórias Póstumas de Brás Cubas* (49.20%). However, as far as lexical bundles are considered, *O Código Da Vinci* has only 4.86 lexical bundles per 10,000 tokens whereas *Memórias Póstumas de Brás Cubas* makes use of 10.71 lexical bundles per 10,000 tokens. *Grande Sertão: Veredas*, however, scores even lower than *O Código Da Vinci*: it contains only 2.30 lexical bundles per 10,000 tokens. In English, *The Da Vinci Code* presents a standardized type/token ratio of 46.17%, which is higher than the ones which are presented by *The Devil to Pay in the Backlands* (43.74%) and *Epitaph of a Small Winner* (43.82%). In relation to lexical bundles, *The Da Vinci Code* also presents a better result, that is, has a relative frequency of 14.72, which is lower than those relative frequencies of bundles in *The Devil to Pay in the Backlands* (15.13 bundles per 10,000 tokens) and *Epitaph of Small Winner* (25.68 bundles per 10,000 tokens). The second group which contains *O Alquimista* and *The Alchemist* is quite different. The former has a standardized type/token ratio of 42.61% and makes use of 113.59 bundles per 10,000 tokens. Its translation into English, *The Alchemist*, has a standardized type/token ratio of 36.37% and totals 215.97 bundles per 10,000 tokens. The fact that *The Da Vinci Code* and its translation into Portuguese are placed in the first group seems to fire the notion of a canon and the way literary works are traditionally classified. One would expect that both *O Código da Vinci* and *The Da Vinci Code* would be similar to *O Alquimista* and *The Alchemist* as the four of them are considered to be popular, non-canonical literary works. Nevertheless, this corpus-based study seems to go in the opposite direction when lexical choices and their ordering are considered.

The second part of this analysis offers a qualitative appraisal of the nature of lexical bundles. The taxonomies for classifying bundles, proposed by Biber et al. (2004), have been described in Section 3.

As far as the structural taxonomy runs, bundles can be classified into three categories: Type 1 (verb phrases), Type 2 (dependent clauses) and Type 3 (noun/prepositional phrases). Figure 5 displays the results for *O Código da Vinci* and *The Da Vinci Code*.

![Figure 5: Structural analysis](image-url)
Both literary works have shown a very similar distribution of bundles. The most frequent type is that which contains fragments of noun/prepositional phrases, totaling 68.67% in Portuguese and 66.84% in English. Then there are bundles which incorporate simple verb phrases: 17.99% in Portuguese and 21.56% in English. Finally, Type 2 lexical bundles corresponds to 13.34% of the instances analyzed in Portuguese and to 11.60% of the instances analyzed in English. The results obtained for *O Código Da Vinci* and *The Da Vinci Code* seem to point to the use of similar structural patterns in both works.

As far as functions are concerned, Biber et al.’s (2004) describe four major categories: stance, discourse, reference and conversation. Figure 6 illustrates the results obtained for the pair of literary works under analysis here.

![Figure 6: Functional analysis](image)

Once again a similar distribution can be seen between *O Código Da Vinci* and *The Da Vinci Code*. First, there are referential bundles, which account for 89.95% in Portuguese and 83.73% in English. Second, there are stance bundles which appear 8.10% in Portuguese and 14.00% in English. Conversational bundles are more common in the original (2.27% of the bundles analyzed) than in the translated version (only 0.90%). Discourse bundles, on the other hand, only appear in the Portuguese version of *The Da Vinci Code*, totaling 1.05% of all instances.

Although there are similar distributions in both structural and functional classifications, there are some differences between the use of bundles in *O Código da Vinci* and *The Da Vinci Code*. The original version in English makes more use of stance and conversational bundles, which would explain the fact that there are more bundles incorporating verb phrases in English than there are in Portuguese. The Portuguese version, however, makes use of discourse organizing bundles, which do not appear in the original version. There are also more referential bundles in Portuguese than in English. This pairs up with the fact that there are more Type 3 bundles in Portuguese than in English as noun/prepositional phrases are generally used to describe the world. If it is considered that attitudinal and conversational bundles are expressions of individual ideas whereas the use of referential and discourse organizing bundles reflect a more informational nature, it may be concluded that the original version of *The Da Vinci Code* in English is more involving than its translation into Portuguese. This may be the result of the translation process: as the translator is not the author him/herself, s/he is not as involved with the literary works as the author is.
As a consequence, his/her linguistic choices are more informational than the ones made by the author. The translator also seems to be worried with how the literary work is going to be read and may insert some sequences to explicitly signal to readers what kind of relationship there is between prior and coming discourse.

6. Conclusion

This paper examined a measure of lexical density in non-canonical as well as canonical texts, and the impact of translation on the behavior of the indexes. It was found that, contrary to expectations, translations do not necessarily impoverish literary texts. Additionally, language may be biased in that lexical density tends to be higher in the texts in Portuguese, which obviously does not imply that the texts written in this language must necessarily be more literary. It must be made clear, however, that translations are texts written anew. They will always depend on the verbal artistry of the translator. Having said that, readers believe they are reading the literary work itself when they use a translation. For instance, when readers pick *Hamlet* in Portuguese, they will not say that are reading something else by Barbara Heliodora or Millôr Fernandes, who have translated the texts. They will say they are reading the play by Shakespeare. In this sense, comparing the language of translated texts with their originals is justified.

This study has also shown that using Corpus Linguistics to bring some light to the notion of literariness has proven fruitful, pointing the way towards the development of an automated index in further studies. As for the relations between literariness and canonicity, it was observed that non-canonical works may, in terms of literariness measured through lexical density indexes, behave quite similarly to the canonical works – as was the case with *The Da Vinci Code* and its translation into Portuguese. What these findings suggest is that not only may Corpus Linguistics show the way to more systematized analyses of literariness, but it may also lead to reconsider what makes a text canonical.

References


