Carmen Dayrell

# Investigating the Preference of Translators for Recurrent Lexical Patterns: A Corpus-based Study

## 1 Introduction

Several translation scholars have claimed that translated language is different from non-translated language, in other words, translated texts are said to show some distinctive features which make them different from texts originally produced in the language in question (among others, Even-Zohar 1978, 1990, Toury 1995: 103-105, Baker 1993, 1995, 1996, 2004). These features, Baker explains, are "patterns which are either restricted to translated texts or which occur with a significantly higher or lower frequency in translated texts than they do in originals [non-translated texts]" (Baker 1995: 235). They are usually referred to as *universals of translation* in the sense that they tend to occur in translations in general, irrespective of the source and target languages involved.

A number of studies have therefore been carried out with a view to exploring the nature of translated texts and relevant differences have been found between the lexical and syntactical make up of translated and non-translated texts across various languages. For instance, Mauranen (2000) identifies important differences between translated and non-translated Finnish with respect to their preferences for some unusual combinations of words. Baroni and Bernardini (2003) examine translated and non-translated Italian and look at the co-occurrence of items which are strongly bonded. No fundamental quantitative differences are found between the two collections. However, a qualitative analysis of the most frequent collocations in the two subcorpora indicates that translations show a stronger tendency towards topic-dependent sequences whereas non-translated texts opt for collocations which are not related to the topic of the texts. Baker (2004) examines recurring lexical patterns (such as *in other words*, *at the same time*, etc.) and phrases related to temporal and spatial orientation (*in the middle of*, *for the first time*, etc.) in translated and non-translated English. The results reveal that these types of lexical phrases tend to occur more frequently in translated than in non-translated texts. In terms of distribution across texts, the phrases seem less evenly distributed in the translated in comparison with the non-translated subcorpus. Jantunen (2004) and Nilsson (2004) take a slightly different approach and look at the collocational patterning of specific items. Jantunen (2004) focuses on three synonymous Finnish modifiers (*hyvin*, *kovin* and *oikein*), all roughly translated into English as "very". Nilsson (2004) examines the Swedish grammatical word "*av*" ("of" or "by" in English) in translated and non-translated Swedish fiction. Significant differences are found between the collocational patterns of translated and non-translated texts. However, some differences may be due to the influence of the source language in the translation process.

The present study focuses on Brazilian Portuguese. A small-scale study is carried out with a view to investigating whether translated texts show a stronger preference for recurring lexical patterns in comparison with non-translated texts of the same language. By recurring lexical patterns I refer to repeated continuous chunks of language which do not allow any significant lexical variation. I shall return to this point later and explain how these patterns are retrieved. What is important to explain here is that this hypothesis assumes that recurring lexical patterns are more frequent than flexible sequences which allow various items within it. This is in line with Sinclair's (1991: 6, 108, 2003: 3) suggestion that words do not occur randomly in a text but are instead closely associated with their surrounding context. Thus, a positive answer to the research question above could be interpreted as a tendency of translators to draw more heavily on standard forms of the language. In other words, it may be an indication that translations tend to "conform to patterns and practices which are typical of the target language" (Baker 1996: 176).

This paper is divided into five sections. The next section presents a brief overview of the corpus from which data has been collected. Section 3 details the methodology adopted here to retrieve recurring lexical patterns. Section 4 explains how the hypothesis is tested and it is followed by a discussion of the results and some concluding remarks. Last but not least, this paper includes two illustrative appendices.

## 2 The Brazilian Portuguese Comparable Corpus (BPCC)

The data analysed in this paper is drawn from a monolingual comparable corpus of Brazilian Portuguese which consists of two separate subcorpora designed according to the same criteria and specifications, one made up of translated Brazilian Portuguese and the other consisting of non-translated Brazilian Portuguese.

The BPCC (cf. Dayrell 2007 for a more detailed description) was designed to include only books categorised as fiction which have been rated best-sellers in Brazil during the period under analysis (1990 onwards). The main rationale behind this decision is that fiction was one of the most popular genres in Brazil during the period examined and hence more likely to include a reasonable number of translated and non-translated texts. The corpus contains only texts targeted at an adult audience and classified as "romance" in the Brazilian Cataloguing-in-Publication (CIP) categorisation system, which corresponds roughly to the category "novel" in the English system. All books have been cleared for permission by the copyright holders. The texts are included in full, rather than in the form of extracts, and an attempt has been made to diversify the selection of texts as much as possible in terms of authors, translators and publishers.

The translational corpus contains only direct translations from English, that is, translations from texts originally written in English. It includes only texts produced by professional translators whose mother-tongue is Brazilian Portuguese and priority is given to translations whose source text was also published from 1990 onwards.

Table 1 shows the present overall size of the BPCC. A detailed list of all books included in the corpus is presented in Appendix I.

| | Number of words (tokens) | Number of books | Number of authors/translators |
|---|---|---|---|
| Translated | 545,395 | 5 | 5 |
| Non-translated | 565,920 | 8 | 8 |

Table 1: Present overall size of the BPCC

An important point to stress here is that, in line with other corpus-based translation studies which are also based on comparable corpora (see, in particular, Olohan 2003; Baker 2004), the translated and the non-translated subcorpora are designed to be of similar size in terms of number of words (tokens) rather than with respect to number of texts and/or number of authors/translators. This means that there is an imbalance in terms of the number of texts in each subcorpus. The translated collection consists of five texts whereas the non-translated collection includes eight texts. I shall refer to this and other limitations of the corpus later and discuss their impact on the results of the analysis. For the time being, what is important to bear in mind is that the primary purpose of this study is to develop a corpus-based methodology for investigating the preference of translators for recurring lexical patterns. Thus, the focus is on the methodological procedures rather than on providing highly reliable findings on the collocational behaviour of translated and non-translated Brazilian Portuguese. Flaws in the corpus design are therefore an issue to be addressed in future studies.

## 3 Methodology

The methodological procedures for retrieving recurring lexical patterns involve three major steps: (1) selection of the words to be taken as *nodes*, that is, the words to be studied; (2) retrieval of their *collocates*, that is to say, "any word that occurs in the specified environment of the node" (Sinclair 1991: 115); and (3) identification and retrieval of the recurring lexical patterns. All procedures described below are carried out by means of the software package *WordSmith Tools*, version 3.0 (Scott 1999).

It is worth mentioning that, for the purposes of this study, no lemmatisation is applied and all nodes and collocates are selected taking into account individual word forms. Although lemmatisation is usually viewed as a useful procedure in collocational analysis (see, for instance, Stubbs 1995a, Berber-Sardinha 1999, 2000), some scholars (Sinclair 1991: 8, Mason 1997) are critical of the approach and argue that different word forms may manifest different collocational behaviour. I opt therefore to consider lemmatization in future research.

### 3.1 Selecting Nodes

The present study examines the lexical patterns of 10 nodes which have been selected on the basis of three criteria. The first criterion is a minimum frequency of 200 occurrences in each subcorpus, translated and non-translated. This criterion is adopted for purely methodological convenience, based on the fact that the analysis of repeated patterns, by its very nature, requires a sufficient body of data to yield useful insights.

The second criterion for the selection of nodes is that the frequencies of the item in the translated and non-translated subcorpora should be as similar as possible. Like the first criterion, it is also adopted for purely methodological convenience. It relies on the assumption that the frequency of the node may have an influence on the number of lexical patterns associated with it. The exact range of difference between the frequencies of the item in each subcorpus is defined by the data itself, taking into account the other criteria being adopted in the selection of nodes. I will return to this point shortly, once the third criterion is explained.

The third criterion establishes that nodes should be predominantly nouns. The assignment of nodes to grammatical categories is based on the classification provided by the *Houaiss* (Houaiss 2001) and *Aurélio* (Ferreira 1999) Dictionaries and on a superficial analysis of the collocates and concordance lines of potential nodes. In broad terms, homographs which belong to more than one grammatical class are still considered as long as the item is predominantly a noun. The aim therefore is to select 10 word types which are predominantly nouns, even though I may be including instances in which the node belongs to other grammatical categories.

Going back to the criterion of similarity of frequency in the translated and the non-translated subcorpora, the idea is to manually select, within the range of words with a minimum frequency of 200 occurrences in the translated and the non-translated subcorpora, 10 nouns whose frequencies in the two subcorpora are as similar as possible. Table 2 lists the resulting selection of nodes, ordered by the difference between the node frequencies in the two subcorpora.

| | NODES | Frequency in each subcorpus | | Difference between the node frequencies |
|---|---|---|---|---|
| | | Translated | Non-translated | |
| 1. | manhã ‘morning’ | 222 | 223 | 1 |
| 2. | rosto ‘face’ | 385 | 388 | 3 |
| 3. | trabalho ‘work’ | 209 | 212 | 3 |
| 4. | tarde ‘late’/‘afternoon’ | 284 | 300 | 16 |
| 5. | mão ‘hand’ | 517 | 540 | 23 |
| 6. | água ‘water’ | 221 | 247 | 26 |
| 7. | hora ‘hour/time’ | 245 | 271 | 26 |
| 8. | verdade ‘truth’ | 323 | 289 | 34 |
| 9. | quarto ‘room’ | 320 | 361 | 41 |
| 10. | noite ‘night’ | 593 | 545 | 48 |

Table 2: Selected nodes

## 3.2 Retrieving Collocates

Three criteria have been established for the selection of collocates. Firstly, preference is given to lexical items. Secondly, potential collocates have to co-occur with the node at least four times in a span of four words to the right and four words to the left (4:4), irrespective of structural boundaries. This is in line with other corpus-based studies which also focus on collocational patterns (see, in particular, Sinclair 1991: 106, 117, Stubbs 1995a,b). The third criterion considers the strength of their association with the node, which is estimated here by means of the mutual information index (hereafter MI) proposed by Church and Hanks (1990) and Church et al. (1991).[1] A minimum MI of five is used as a cut-off point. This threshold is established in order to avoid selecting as collocates very high frequency words such as the verbs *ser/estar* 'to be' and adverbs such as *já* 'already' and *ainda* 'yet', which may yield interesting linguistic patterns but are also likely to co-occur with almost any word in the corpus.

This study focuses on the co-occurrences of 10 selected nodes with their highest frequency collocates in the translated and non-translated subcorpora. For some nodes, the highest frequency collocate is the same in the translated and non-translated subcorpora. For instance, *fazer* 'to do'/'to make' is the most frequent collocate of the node *trabalho* 'work' in both the translated and the non-translated subcorpora. However, for many nodes, the highest frequency collocates in the two subcorpora are different. For example, *noite* 'night' has *meia* 'half' as the most frequent collocate in the translated and *dia* 'day' as the most frequent collocate in the non-translated subcorpus. In these cases, both collocations are examined provided that there are at least four instances of the collocation in each subcorpus. If evidence is scarce in one subcorpus (less than four instances), the collocate is discarded. I have also discarded all collocates with more than 50% of instances in one text only. For instance, the most frequent collocate of the node *quarto* [room] in the translated subcorpus is *hóspedes* 'guests', they co-occur 19 times. However, 18 out of these 19 instances (95%) come from text fntr05. This fourth criterion has been established as an attempt to minimize the influence of one single text on the data retrieved from the corpus, which is only to be expected due to the limited size of the corpus. Thus, when a collocate is discarded, we move downwards in the list of collocates and take the next item as a potential collocate to be further analysed.

[1] The MI calculation formalises Sinclair's (1987, 1991: 69-70) argument that the comparison between the actual frequency of co-occurrence (observed frequency) and the expected frequency if the items were to co-occur by chance (expected frequency) can indicate how likely the two items are to co-occur. In other words, it can provide a rough measure of the strength of attraction between relevant items. The higher the MI the less likely the co-occurrence between node and collocate is due to chance.

Table 3 presents the 14 collocational patterns selected for investigation.

| | Node | Highest Frequency Collocate | Frequency of co-occurrence in each subcorpus | |
|---|---|---|---|---|
| | | | Translated | Non-translated |
| 1. | manhã ‘morning’ | seguinte ‘following’ | 31 | 12 |
| 2. | manhã ‘morning’ | café ‘coffee’ | 29 | 12 |
| 3. | rosto ‘face’ | expressão ‘expression’ | 21 | 4 |
| 4. | rosto ‘face’ | mãos ‘hand’ | 13 | 16 |
| 5. | trabalho ‘work’ | fazer ‘to do’/‘make’ | 10 | 15 |
| 6. | tarde ‘late’/‘afternoon’ | noite ‘night’ | 14 | 10 |
| 7. | tarde ‘late’/‘afternoon’ | fim ‘end’ | 5 | 14 |
| 8. | mão ‘hand’ | esquerda ‘left’ | 37 | 21 |
| 9. | água ‘water’ | copo ‘glass’ | 13 | 14 |
| 10. | hora ‘hour/time’ | meia ‘half’ | 26 | 28 |
| 11. | verdade ‘truth’ | é ‘is’ | 77 | 84 |
| 12. | quarto ‘room’ | porta ‘door’ | 11 | 19 |
| 13. | noite ‘night’ | meia ‘half’ | 37 | 13 |
| 14. | noite ‘night’ | dia ‘day’ | 22 | 25 |

Table 3: Collocations selected for analysis

### 3.3 Identifying Recurring Lexical Patterns

Once the collocations have been selected, the next step is to retrieve all instances in which node and collocate co-occur in both the translated and non-translated sub-corpora. Recurring lexical patterns are identified by sorting the concordance lines by the different positions in which the collocate occurs and examining the items in the vicinity of the collocation, i.e. the items between the node and the collocate as well as the items on the left and on the right of the pattern.

The analysis starts from the position in which the collocate occurs the highest number of times and the cycle moves from one position to another until all instances have been examined. Any recurring continuous sequence occurring at least three times is taken as a recurring lexical pattern. This means that in order to be regarded as a recurrent pattern, the chunk should occur at least three times in the corpus. Once a given pattern has been identified, we examine the remaining concordance lines and search for instances which may be regarded as slight variants of it. The procedure is carried out in the two subcorpora altogether and repeated as many times as necessary until all instances have been examined.

The collocational patterns of *trabalho* ‘work’ with *fazer* ‘to do’/‘to make’ are used here to illustrate how recurrent lexical patterns are retrieved. The following notations are used to describe patterns:

- Optional items are indicated between brackets;
- Lemmas are represented in capital letters (TER 'HAVE');
- The position of the collocate in relation to the node is identified by using L or R (left or right respectively), followed by a number which indicates the distance from the node. For example, L1 stands for the first position on the left of the node.

Thus, by sorting the concordance lines by position L2, we find six instances of the sequence *fazer o trabalho* 'to do the work'.

| | Concordance Line | Text[2] |
|---|---|---|
| 01 | ia. Ainda era cedo para **fazer o trabalho**. Chicão ligou o | fnnt01 |
| 02 | o pedira a Abraham para **fazer o trabalho** sujo que precis | fntr02 |
| 03 | enhum de confiança para **fazer o trabalho**. O chefe não quer | fnnt01 |
| 04 | de Mattos, para a velha **fazer o trabalho**. Quando chegou | fnnt01 |
| 05 | va de mim e eu não pude **fazer o trabalho** direito. Eu sabia | fnnt01 |
| 06 | formulada: "Ele poderia **fazer o trabalho**?" Para Rossini, | fntr03 |

We also find two lines which indicate that, in addition to the definite article (*o* 'the'), other items may be inserted between the collocate and the node: possessive pronouns (*seu* 'your' and *meu* 'my') and the adverb *bem* 'well'. The item *para* appears on the left of the collocate in 63% of instances (5 out of 8). These eight instances are summarised in the formula *(para) fazer (bem) o (seu/meu) trabalho* '(in order) to do the (your/my) work (well)'.

| | | |
|---|---|---|
| 07 | xarei você sozinho para **fazer o** seu **trabalho**. Ela deu mei | fntr04 |
| 08 | a que me preocupa é **fazer** bem **o** meu **trabalho**." O estômago | fnnt01 |

Three lines show the indefinite article between the node and the collocate (lines 09-11 below). Line 11 is regarded as a variation of the pattern since, in addition to *um* 'a'/'an', it also shows the adjective *bom* 'good'. These three lines yield the pattern *fazer um (bom) trabalho* 'to do some (good) work'.

| | | |
|---|---|---|
| 09 | de uma cafetina de luxo **fazer um trabalho** de abutre, com | fnnt01 |
| 10 | que o crioulo tinha ido **fazer um trabalho** no apartamento | fnnt01 |
| 11 | remendo demais para **fazer um** bom **trabalho** – e sua filha | fntr04 |

By sorting the concordance lines by position R2, we find five instances with *a* in position R1 and the lemma TER 'HAVE' on the left of the node. These five lines are summarised in the pattern *TER (um) trabalho a fazer* 'HAVE (some) work to be done'.

[2] Texts are identified according to the following structure: fn stands for fiction, tr for translated and nt for non-translated texts. The texts are then numbered so that they can be identified within each subcorpus (see Appendix I).

| | | |
|---|---|---|
| 12 | sso paciente. Você tem **trabalho a fazer**. Eu não o incomoda | fntr03 |
| 13 | Sé está vazia e temos **trabalho a fazer**. Claudio Stagni ti | fntr03 |
| 14 | definidamente. E tenho **trabalho a fazer** pelas Mães da Praça | fntr03 |
| 15 | ue sobrávamos tínhamos **trabalho a fazer** – uma hora para com | fntr04 |
| 16 | s meninos que tinha um **trabalho a fazer** e, assim como um | fntr05 |

The remaining lines are all discarded because they do not yield any recurring lexical pattern. Line 17 is the only instance in which the pattern appears in its uninterrupted form; however, it occurs only once. Lines 18-25 show various items between the node and the collocate.

| | | |
|---|---|---|
| 17 | cem contos, para **fazer trabalho** de responsabilidade com | fnnt01 |
| 18 | s não tinham se dado o **trabalho** de **fazer** Wharton vesti-lo | fntr04 |
| 19 | não esperava ter tanto **trabalho** para **fazer** uma coisa tão s | fnnt01 |
| 20 | , que se enriquece sem **trabalho**, para **fazer** pouco da gente | fnnt04 |
| 21 | o, que com a cabeça no **trabalho** que ia **fazer** prestara pouc | fnnt01 |
| 22 | as mãos livres para o **trabalho** que ia **fazer**. Retirou da m | fnnt01 |
| 23 | Imaginei ser colega de **trabalho** daquelas pessoas, **fazer** pa | fnnt05 |
| 24 | deixando o fio livre, **trabalho** de quem sabe **fazer**. – Poi | fnnt03 |
| 25 | **fazer** uma avaliação do **trabalho** que denominava "a miss | fnnt01 |

Once patterns have been identified, the next step is to count the number of instances that patterns appear in each subcorpus, translated and non-translated. Table 4 summarises the patterns yielded by the collocations of *trabalho* 'work' with *fazer* 'to do'/'to make', ordered by number of instances in the translated subcorpus.

| | Recurring Lexical Patterns | Number of instances in each subcorpus | |
|---|---|---|---|
| | | Translated | Non-Translated |
| i | TER (um) trabalho a fazer<br>'HAVE (some) work to be done' | 5 | 0 |
| ii | (para) fazer (bem) o (seu/meu) trabalho<br>'(in order) to do the (your/my) work (well)' | 3 | 5 |
| iii | fazer um (bom) trabalho<br>'to do a/some (good) work' | 1 | 2 |
| | Total | **9** | **7** |

Table 4: Number of recurring lexical patterns realised by *trabalho* 'work' and *fazer* 'to do'/'to make' in the translated and the non-translated subcorpora

A relevant point to stress here is that lexical patterns may vary in a wide range of ways. There may be cases in which it is by no means easy to decide whether to treat a given instance as a separate pattern or as a variation of a given pattern. Thus, some criteria have been established in order to introduce an element of consistency in the categorisation of patterns. First, different lexical items are not grouped by grammatical class or semantic category unless there are at least three different items of the category in a particular position within the pattern. For instance, if we look at pattern ii in table 4, we notice that the possessive pronouns *seu/meu* 'your'/'my' are represented as individual lexical items. By contrast, in the co-occurrences of *rosto* 'face' and *expressão* 'expression' (concordance lines below), we find various adjectives between the node and the collocate which are grouped together as a grammatical class (ADJ). These instances are summarised in the formula *expressão* (*bem*) (ADJ) *em o/meu/seu rosto* 'expression (very) (ADJ) on the/my/his face'.

| | | | |
|---|---|---|---|
| 01 | m uma **expressão** maligna no **rosto**, "na minha op | fntr02 | 'evil' |
| 02 | m uma **expressão** atônita no **rosto**, e eu estava | fntr02 | 'astonished' |
| 03 | uma **expressão** assustada no **rosto**. Mais um pedi | fntr02 | 'shocked' |
| 04 | ma **expressão** preocupada no **rosto**. "Melhor o se | fntr02 | 'worried' |
| 05 | os e uma **expressão** dura no **rosto**. O cliente an | fnnt05 | 'stern' |
| 06 | com uma **expressão** séria no **rosto**: – O que est | fnnt05 | 'serious' |
| 07 | **expressão** sardônica constante no **rosto**. A prim | fntr02 | 'sardonic' |
| 08 | ma **expressão** bem alegre no **rosto**; aliás, enqua | fntr02 | 'happy' |
| 09 | surdo, é?" A **expressão** no **rosto** de Lambajan | fntr02 | |
| 10 | lembrei-me da **expressão** no **rosto** de Vasco no di | fntr02 | |
| 11 | nou foi a **expressão** no seu **rosto**. Havia tranqü | fntr04 | |
| 12 | ao ver a **expressão** em meu **rosto**, quanto fui vi | fntr02 | |
| 13 | a **expressão** esperta em seu **rosto**. Ou de noite, | fntr01 | 'smart' |

There are other cases in which different lexical items are grouped semantically. For instance, within the pattern PASSAR *da meia-noite* 'it was after mid-night', we may find words which refer to how much time has gone by (*muito* 'a lot', *bastante* 'very much', *alguns minutos* 'some minutes'). These items are represented under the semantic category "time" – PASSAR "time" *da meia-noite* 'it was "time" after mid-night'.

| | | |
|---|---|---|
| 01 | nde delicadeza. **Passava de meia-noite**. Os convidados não | fntr01 |
| 02 | de ontem. Dois **Passava da meia-noite** quando Luca Rossin | fntr03 |
| 03 | os acompanhar." **Passava da meia-noite** quando chegaram ao | fnnt01 |
| 04 | ociedade. **Passava** muito **da meia-noite** quando ele acordou | fntr01 |
| 05 | ndo. Já **passava** bastante **da meia-noite**. Ela estava deitad | fntr01 |
| 06 | **Passavam** alguns minutos **da meia-noite** quando Chico pediu | fnnt01 |

For the purposes of this paper, I have discarded all patterns which show regularity in terms of grammatical or semantic categories. Here, the focus is on repeated continuous chunks of language which do not allow significant lexical variation.

It is also worth mentioning that we may find more than one recurring item in the vicinity of the collocation. For instance, in the co-occurrences of *hora* 'hour' with *meia* 'half', the pattern *meia hora* 'half an hour' can be followed by *depois* 'after', *antes* 'before' or *mais tarde* 'later'. *Manhã seguinte* 'following morning' may be preceded by *na* 'in the', *da* 'of the' or *até a* 'until the'. In these cases, each variation is treated as a separate pattern irrespective of whether the item is lexical or grammatical. The only condition is that it should appear at least three times in the corpus. Appendix II lists the recurring lexical patterns realised by all 14 collocations analysed in this paper.

## 4 Testing the Hypothesis

Since the number of times node and collocate co-occur in each subcorpus may be different, the hypothesis is tested by taking into consideration the overall percentage of recurring lexical patterns in each subcorpus, rather than the raw number of patterns. In the example above, *fazer* collocates with *trabalho* 10 times in the translated subcorpus and 15 times in the non-translated subcorpus (table 5).

| | Overall number of instances | Overall number of recurring lexical patterns | % of recurring lexical patterns |
|---|---|---|---|
| Translated | 10 | 9 | 90% |
| Non-Translated | 15 | 7 | 47% |

Table 5: Overall number and percentage of recurring lexical patterns realised by the collocations of *trabalho* 'work' and *fazer* 'to do'/'to make' in the translated and the non-translated subcorpora

Nine out of the 10 instances in the translated subcorpus (90%) are recurring lexical patterns whereas, in the non-translated subcorpus, only seven out of the 15 instances (47%) are recurring lexical patterns. The patterns of *trabalho* and *fazer* therefore confirm the hypothesis that translated texts show stronger a preference overall for recurring lexical patterns in comparison with non-translated texts.

Table 6 summarises the findings for all 14 collocations analysed in this study. For each pair of words, it shows the overall frequency of collocation, the number of recurring lexical patterns and the percentage of recurring patterns in relation to the overall number of times node and collocate co-occur in each subcorpus. Difference refers to the difference between the percentages of recurring lexical patterns in the two subcorpora, expressed in percentage points (pp).

| | Collocation | Translated | | | Non-translated | | | Differ-ence |
|---|---|---|---|---|---|---|---|---|
| | | Frequency of collocation | Number and % of recurring lexical patterns | | Frequency of collocation | Number and % of recurring lexical patterns | | |
| 1. | manhã & seguinte<br>‘morning’ & ‘following’ | 31 | 31 | 100% | 12 | 9 | 75% | 25pp |
| 2. | manhã & café<br>‘morning’ & ‘coffee’ | 29 | 27 | 93% | 12 | 10 | 83% | 10pp |
| 3. | rosto & expressão<br>‘face’ & ‘expression’ | 21 | 0 | 0% | 4 | 0 | 0% | 0pp |
| 4. | rosto & mãos<br>‘face’ ‘hand’ | 13 | 5 | 38% | 16 | 4 | 25% | 13pp |
| 5. | trabalho & fazer<br>‘work’ & ‘to do’/‘to make’ | 10 | 9 | 90% | 15 | 7 | 47% | 43pp |
| 6. | tarde & noite<br>‘late’/‘afternoon’ & ‘night’ | 14 | 11 | 79% | 10 | 5 | 50% | 29pp |
| 7. | tarde & fim<br>‘late/ afternoon’ & ‘end’ | 5 | 4 | 80% | 14 | 14 | 100% | 20pp |
| 8. | mão & esquerda<br>‘hand’ & ‘left’ | 37 | 35 | 95% | 21 | 18 | 86% | 9pp |
| 9. | água & copo<br>‘water’ & ‘glass’ | 13 | 12 | 92% | 14 | 11 | 76% | 16pp |
| 10. | hora & meia<br>‘hour/time’ & ‘half’ | 26 | 22 | 85% | 28 | 27 | 96% | 11pp |
| 11. | verdade & é<br>‘truth’ & ‘is’ | 77 | 52 | 68% | 84 | 69 | 82% | 14pp |
| 12. | quarto & porta<br>‘room’ & ‘door’ | 11 | 6 | 55% | 19 | 13 | 68% | 13pp |
| 13. | noite & meia<br>‘night’ & ‘half’ | 37 | 29 | 78% | 13 | 11 | 85% | 7pp |
| 14. | noite & dia<br>‘night’ & ‘day’ | 22 | 13 | 59% | 25 | 13 | 52% | 7pp |
| Totals | | **346** | **256** | **74%** | **287** | **211** | **74%** | **0pp** |

Table 6: Overall number and percentage of recurring lexical patterns realised by all collocations in the translated and non-translated subcorpora

As can be seen in table 6, no difference is found between the overall percentages of recurring lexical patterns in the two subcorpora – 74% in both. However, the preference of translators for recurring lexical patterns becomes evident when we examine individual collocations. For 57% of the collocations (8 out of 14), translated texts show a stronger preference overall for recurring lexical patterns in comparison with non-translated texts. For 36% of the collocations (5 out of 14), the preference for recurring lexical patterns is stronger in the non-translated subcorpus. One collocation (#3) reveals a similar proportion of recurring lexical patterns in the two subcorpora. I have regarded as "similar" all those cases in which the difference between the percentages of recurring lexical patterns in the two subcorpora is no higher than five percentage points. Table 7 summarises these findings.

| Preference for Recurring Lexical Patterns | Number of Collocational Patterns |
|---|---|
| TRANSLATED texts show a stronger preference for recurring lexical patterns | 8 (57%) |
| NON-TRANSLATED texts show a stronger preference for recurring lexical patterns | 5 (36%) |
| Similar proportion of recurring lexical patterns in BOTH subcorpora | 1 (7%) |
| Total | **14 (100%)** |

Table 7: Preference for recurring lexical patterns

If we leave the collocations which show a similar proportion of recurring lexical patterns in the two subcorpora out of this calculation, the percentage of collocations where translated texts show a stronger preference for recurring lexical patterns rises to 62% (8 out of 13).

## 5 Discussion

The results of the analysis seem to indicate that translated Brazilian Portuguese does exhibit a more marked preference for recurring lexical patterns than non-translated Brazilian Portuguese. This tendency is even clearer when we look at the difference between the percentages of recurring lexical patterns in the two subcorpora (table 6). We find three cases in which the percentage of recurring lexical patterns is at least 25pp higher in the translated collection (#1, 5, 6). When the non-translated collection shows a stronger preference for recurring lexical patterns, the difference is no higher than 20pp (#7, 10-13).

One possible reason to explain this phenomenon is the influence of one single text on the overall number of recurring lexical patterns yielded by a given collocation. However, by taking into consideration the three collocations with a considerably higher percentage of recurring lexical patterns in the translated subcorpus (#1, 5, 6) and examining the distribution of instances across texts, we find that this is not the case. For collocation #1, the highest percentage of recurring lexical patterns in one single text is 29% in text fntr05. For collocation #5, 44% of instances come from text fntr03; however, in the non-translated subcorpus, 100% of the instances (all 7 instances) occur in one single text (fnnt01). For collocation #6, 36% of instances appear in text fntr02.

Another point worth commenting on is that 16 recurring lexical patterns occur in one subcorpus only. Ten patterns (63%) appear in the translated and show no evidence in the non-translated subcorpus while six patterns (37%) occur in the non-translated but not in the translated subcorpus. These figures reinforce the suggestion that translated texts tend to draw more heavily on recurring lexical patterns than non-translated texts.

Although the findings seem to confirm the hypothesis I put forward earlier, there are a number of points which deserve further discussion and clarification. For instance, one could argue that there may be cases in which the higher percentage of recurring lexical patterns is simply reflecting a higher frequency of collocation. The collocations of *tarde* 'late'/'afternoon' and *noite* 'night' can serve as an example to illustrate this point (table 6). In the translated subcorpus, the two items co-occur 14 times and 79% of these instances (11 occurrences) are recurring lexical patterns. In the non-translated subcorpus, they collocate 10 times and 50% of these instances (5 occurrences) are recurring lexical patterns. If our assumption that recurring lexical patterns are more frequent than flexible sequences holds true, a higher frequency of collocation may be expected to enhance the chance of yielding recurrent lexical patterns. I therefore further examine all collocations whose frequency in one subcorpus is at least 20% higher than its frequency in the other subcorpus (table 8). For example, *trabalho* 'work' collocates with *fazer* 'to do'/'to make' 10 times in the translated subcorpus. The number of instances is 50% higher in the non-translated subcorpus (15 instances). The threshold of 20% has been chosen arbitrarily and it is used to assure a reasonable difference between the frequencies of collocations in the two subcorpora.

We notice that, for 60% of these collocations, the subcorpus with a higher frequency of collocation also displays a higher percentage of recurring lexical patterns (#1, 2, 6, 7, 8, 12 – table 6). This could be interpreted as an indication that it is not entirely incorrect to say that the higher percentage of recurring may be simply reflecting a higher frequency of collocation. It is interesting to notice that four out of these six collocations refer to the translated subcorpus. However, if we examine the remaining four collocations, we find that two collocations are more frequent in the non-translated but show a higher percentage of recurring lexical patterns in the translated subcorpus (#4, 5). These two cases reveal a clear tendency of translated texts to display a more pronounced preference for recurring lexical patterns. Two collocations are more frequent in the translated subcorpus; however, only one exhibits a higher percentage of recurring lexical patterns in the non-translated subcorpus (#10). The other collocation does not yield any recurring lexical pattern (#3) and the number of instances in the translated subcorpus is more than four times the number of instances in the non-translated subcorpus. This suggests that a higher frequency of collocation does not necessarily mean a higher proportion of recurring lexical patterns.

| | Collocations | Frequency in each subcorpus | | Difference | Subcorpus with higher % of recurring lexical patterns |
|---|---|---|---|---|---|
| | | Translated | Non-translated | | |
| 1. | manhã & seguinte<br>'morning' & 'following' | 31 | 12 | 158% | TR |
| 2. | manhã & café<br>'morning' & 'coffee' | 29 | 12 | 141% | TR |
| 3. | rosto & expressão<br>'face' & 'expression' | 21 | 4 | 425% | SAME PROPORTION |
| 4. | rosto & mãos<br>'face' & 'hand' | 13 | 16 | 23% | TR |
| 5. | trabalho & fazer<br>'work' & 'to do'/'to make' | 10 | 15 | 50% | TR |
| 6. | tarde & noite<br>'late'/'afternoon' & 'night' | 14 | 10 | 40% | TR |
| 7. | tarde & fim<br>'late'/'afternoon' & 'end' | 5 | 14 | 180% | NON-TR |
| 8. | mão & esquerda<br>'hand' & 'left' | 37 | 21 | 76% | TR |
| 9. | quarto & porta<br>'room' & 'door' | 11 | 19 | 72% | NON-TR |
| 10. | noite & meia<br>'night' & 'half' | 37 | 13 | 184% | NON-TR |

Table 8: Collocations whose frequency in one subcorpus is at least 20% higher than its frequency in the other subcorpus

Another point to bear in mind when interpreting the results is that the analysis is based on a very restricted number of collocational patterns. The investigation of a higher number of collocations may however yield different results. It is also important to stress that this study has focused on the collocational patterns of individual word forms, both node and collocate. It would be interesting to examine whether the analysis of all variants of the same lemma would yield the same results. One may also wish to take flexible patterns into account, such as the cases discussed earlier which allow lexical variants of a given grammatical or semantic category. Translated and non-translated text may display different preferences with respect to their tendency to use flexible sequences. Further achievements could also result from examining the consequences of assigning different values to frequency of co-occurrence and window size in the selection of collocates. Other statistical calculations, such as the log-likelihood ratio (Dunning 1993) or t-score test (Church et al. 1991), could also be adopted to estimate the strength of association between node and collocate.

More importantly, one cannot afford to ignore that the corpus from which data is retrieved has a number of limitations. Firstly, the BPCC is very limited in size and in number of publications. In addition, the translated and the non-translated subcorpora are not balanced in terms of the number of texts in each component. The translated subcorpus consists of five texts whereas the non-translated subcorpus includes eight texts. This inevitably implies that the non-translated component is more diverse with respect to topics and authors than the translated subcorpus. Such mismatch cannot be ruled out as a potential reason to justify the differences in the collocational patterns of the two subcorpora. Another limitation of the corpus is that it comprises one text genre only (fiction) and hence it does not enable the researcher to determine whether the differences identified are genre-dependent. Moreover, the translational subcorpus includes translations from English only, which raises the issue of whether the choices made by translators had been influenced by source-language patterns. Some differences identified here may be related to specific features of the languages involved, in this case, English and Portuguese. In short, in order to be able to identify features which are specific to translated texts irrespective of the source language influence or preferences of individual translators, we would need access to a robust comparable corpus, consisting of a wide range of authors and translators as well as diverse source languages and genres.

## 6 Concluding Remarks

This paper has indicated that translated texts exhibit a more marked preference overall for recurring lexical patterns in comparison with non-translated texts of the same language. This finding may be related to the tendency of translators to produce "uniform" texts and resort to patterns which are frequently used in the target language. It provides evidence to support the suggestion that translations in general, irrespective of the source and target languages involved, tend to conform to typical and standard forms of language.

However, as in any corpus-based research, the data retrieved from the corpus is influenced by the selection of texts. Thus, it is crucial to interpret the results according to the composition and balance of the corpus. Here, a note of caution was added regarding the various limitations of the BPCC, which do not allow the researcher to reach firmer conclusions on the lexical patterning of translated and non-translated Brazilian Portuguese. It is also important to stress that the analysis is based on a very restricted number of collocational patterns. Even more importantly, the present study focused on Brazilian Portuguese specifically and only includes translations from English source texts. In fact, some differences identified here may not be due to the universal features of translated texts but instead they may be specific to the English-Portuguese language pair. This means that valid conclusions on the impact of the translation process on the language produced by translators can only be drawn if similar studies are carried out across different languages. This paper is therefore an initial step and this is why I have attempted to describe the methodology in as much detail as possible so that other researchers are able to support or refute the tendencies displayed here.

### *Acknowledgements*

I am very grateful to my PhD supervisor, Professor Mona Baker (The University of Manchester, UK), who has closely followed this study. All publishers, authors and translators who have granted permission for the copyright material to be included in corpus are gratefully acknowledged. Thanks are also due to CAPES (Brazil) for providing the financial support needed in this research.

## References

Baker, Mona (1993): "Corpus Linguistics and Translation Studies: Implications and Applications." Mona Baker, Gill Francis, Elena Tognini-Bonelli (eds): *Text and Technology: In Honour of John Sinclair*. Amsterdam/Philadelphia: Benjamins, 233-250

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

Baker, Mona (1996): "Corpus-based Translation Studies: The Challenges That Lie Ahead." Harold Somers (ed.): *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Amsterdam/Philadelphia: Benjamins, 175-187

Baker, Mona (2004): "A Corpus-based View of Similarity and Difference in Translation." *International Journal of Corpus Linguistics* 9 [2]: 167-193

Baroni, Marco; Silvia Bernardini (2003): "A Preliminary Analysis of Collocational Differences in Monolingual Comparable Corpora." Dawn Archer, Paul Rayson, Andrew Wilson, Tony McEnery (eds): *Proceedings of Corpus Linguistics 2003*. (UCREL Technical Report 16 Special Issue.) Lancaster: Lancaster University, 82-91

Berber-Sardinha, Tony (1999): "Estudo Baseado em Corpus da Padronização Lexical no Português Brasileiro: Colocações e Perfis Semânticos." *PROPOR'99. IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*. Évora: Universidade de Evora, 269-287

Berber-Sardinha, Tony (2000): "Semantic Prosodies in English and Portuguese: a Contrastive Study." *Cuadernos de Filologia Inglesa* 9 [1]: 93-110

Church, Kenneth W.; Patrick Hanks (1990): "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics* 16 [1]: 22-29

Church, Kenneth; William Gale, Patrick Hanks, Donald Hindle (1991): "Using Statistics in Lexical Analysis." Uri Zernik (ed.): *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Hillsdale: Erlbaum, 115-164

Dayrell, Carmen (2007): "A Quantitative Approach to Compare Collocational Patterns in Translated and Non-translated Texts." *International Journal of Corpus Linguistics* 12 [3]: 375-414

**trans-kom**

**trans-kom** ist eine wissenschaftliche Zeitschrift für Translation und Fachkommunikation.

**trans-kom** veröffentlicht Forschungsergebnisse und wissenschaftliche Diskussionsbeiträge zu Themen des Übersetzens und Dolmetschens, der Fachkommunikation, der Technikkommunikation, der Fachsprachen, der Terminologie und verwandter Gebiete.

Beiträge können in deutscher, englischer, französischer oder spanischer Sprache eingereicht werden. Sie müssen nach den Publikationsrichtlinien der Zeitschrift gestaltet sein. Diese Richtlinien können von der **trans-kom**-Website heruntergeladen werden. Alle Beiträge werden vor der Veröffentlichung anonym begutachtet.

**trans-kom** wird ausschließlich im Internet publiziert: http://www.trans-kom.eu

Dunning, Ted (1993): "Accurate Methods for the Statistics of Surprise and Coincidence." *Computational Linguistics* 19 [1]: 61-74

Even-Zohar, Itamar (1978): "The Position of Translated Literature within the Literary Polysystem." James Stratton Holmes, José Lambert, Raymond van den Broeck (eds): Literature and Translation: New Perspectives in Literary Studies. Leuven: Acco, 117–127 – revised version in Even-Zohar (1990)

Even-Zohar, Itamar (1990): "Polysystem Studies." *Poetics Today* (Special Issue) 11 [1]: 9-26

Ferreira, Aurélio B. H. (1999): *Dicionário Aurélio Eletrônico – Século XXI*. CDROM Versão 3.0. Rio de Janeiro: Editora Nova Fronteira/Lexikon Informática

Houaiss, Antônio (2001): *Dicionário Eletrônico Houaiss da Língua Portuguesa*. CDROM Version 1.0. Rio de Janeiro: Editora Objetiva

Jantunen, Jarmo H. (2004): "Untypical Patterns in Translations: Issues on Corpus Methodology and Synonymity." Anna Mauranen, Pekka Kujamäki (eds): *Translation Universals: Do They Exist?* Amsterdam/Philadelphia: Benjamins, 101-126

Mason, Oliver (1997): "The Weight of Words: An Investigation of Lexical Gravity." Barbara Lewandowska-Tomaszczyk, Patrick James Melia (eds): *PALC '97: Practical Applications in Language Corpora. Proceedings*. Łódź: Łódź University Press, 361-375

Mauranen, Anna (2000): "Strange Strings in Translated Language: A Study on Corpora." Maeve Olohan (ed.): *Intercultural Faultlines. Research Methods in Translation Studies I: Textual and Cognitive Aspects*. Manchester: St Jerome, 105-118

Nilsson, Per-Ola (2004): "Translation-specific Lexicogrammar? Characteristic Lexical and Collocational Patterning in Swedish Texts Translated from English." Anna Mauranen, Pekka Kujamäki (eds): *Translation Universals Do They Exist?* Amsterdam/Philadelphia: Benjamins, 129-141

Olohan, Maeve (2003): "How Frequent Are the Contractions? A Study of Contracted Forms in the Translational English Corpus." *Target* 15 [1]: 59-89

Scott, Mike (1999): *WordSmith Tools Version 3.0*. Oxford: Oxford University Press

Sinclair, John (1987): "The Nature of Evidence." John Sinclair (ed.): *Looking up. Account of the Cobuild Project in Lexical Computing*. London: Harper-Collins, 150-159 – revised version in Sinclair (1990: 67-79)

Sinclair, John (1991): *Corpus Concordance and Collocation*. Oxford: Oxford University Press

Sinclair, John (2003): *Reading Concordances*. London: Pearson Education, Longman

Stubbs, Michael (1995a): "Collocations and Semantic Profiles: On the Cause of Trouble with Quantitative Studies." *Functions of Language* 2 [2]: 23-55

Stubbs, Michael (1995b): "Corpus Evidence for Norms of Lexical Collocation." Guy Cook, Barbara Seidlhofer (eds): *Principle and Practice in Applied Linguistics*. London: Oxford University Press, 245-256

Toury, Gideon (1995): *Descriptive Translation Studies and Beyond*. Amsterdam/Philadelphia: Benjamins

## Appendix I

This appendix presents the details of all books included in the Brazilian Portuguese Comparable Corpus (BPCC).

### *Translated Fiction*

| File name | Translation Title | English Title | Author | Translator | Date | Ed. | Publisher | Number of words |
|---|---|---|---|---|---|---|---|---|
| fntr01 | O Paciente Inglês | The English patient | Michael Ondaatje | Rubens Figueredo | 1994 | 10th | Editora 34 | 86,571 |
| fntr02 | O Último Suspiro do Mouro | The moor's last sigh | Salman Rushdie | Paulo Henriques Britto | 1996 | 1st | Cia das Letras | 161,435 |
| fntr03 | A Eminência | Eminence | Morris West | Maria dos Anjos Rouch | 1999 | 1st | Record | 97,949 |
| fntr04 | A Espera de um Milagre | The green mile | Stephen King | Marcos H. C. Côrtes | 2000 | 1st | Objetiva | 140,836 |
| fntr05 | Klone e Eu | The klone and I | Danielle Steel | Heitor Pitombo | 2000 | 1st | Record | 58,604 |
| Total | | | | | | | | **545,395** |

### *Non-translated Fiction*

| File name | Title | Author | Date | Ed. | Publisher | Number of words |
|---|---|---|---|---|---|---|
| fnnt01 | Agosto | Rubem Fonseca | 2002 | 2nd | Cia das Letras | 92,264 |
| fnnt02 | O Xangô de Baker Street | Jô Soares | 1995 | 1st | Cia das Letras | 66,242 |
| fnnt03 | Saraminda | José Sarney | 2000 | 1st | Siciliano | 60,097 |
| fnnt04 | A Muralha | Dinah Silveira Queiroz | 2000 | 1st | Record | 113,681 |
| fnnt05 | Bala na Agulha | Marcelo Rubens Paiva | 2001 | 9th | Siciliano | 41,374 |
| fnnt06 | Inferno | Patrícia Melo | 2001 | 1st | Cia das Letras | 103,325 |
| fnnt07 | Rapina | Ivan Sant'Anna | 1996 | 1st | Record | 51,713 |
| fnnt08 | Benjamim | Chico Buarque | 1995 | 1st | Cia das Letras | 37,224 |
| Total | | | | | | **565,920** |

## Appendix II

This appendix lists all recurring lexical patterns realised by the collocations analysed in this paper. TR stands for translated subcorpus and NON-TR for non-translated subcorpus.

1. Collocations of *manhã* 'morning' and *seguinte* 'following'

| | Recurring Lexical Patterns | Number of instances in each subcorpus | |
|---|---|---|---|
| | | TR | NON-TR |
| i | `na manhã seguinte`<br>`'in the following morning'` | 23 | 9 |
| ii | `da manhã seguinte`<br>`'of the following morning'` | 5 | 0 |
| iii | `até a manhã seguinte`<br>`'until the following morning'` | 3 | 0 |
| | Total | **31** | **9** |

2. Collocations of *manhã* 'morning' and *café* 'coffee'

| | Recurring Lexical Patterns | Number of instances in each subcorpus | |
|---|---|---|---|
| | | TR | NON-TR |
| i | `no café da manhã 'breakfast'` | 8 | 3 |
| ii | `de (o) café da manhã 'breakfast'` | 7 | 0 |
| iii | `(o) café da manhã 'breakfast'` | 5 | 4 |
| iv | `TOMAR (o) café da manhã 'HAVE breakfast'` | 4 | 2 |
| v | `para (o) café da manhã 'breakfast'` | 3 | 1 |
| | Total | **27** | **10** |

3. Collocations of *rosto* 'face' and *expressão* 'expression'

No recurring lexical patterns

4. Collocations of *rosto* 'face' and *mãos* 'hand'

| | Recurring Lexical Patterns | Number of instances in each subcorpus | |
|---|---|---|---|
| | | TR | NON-TR |
| i | `COBRIR o rosto com as mãos`<br>`'COVER the face with the hands'` | 3 | 1 |
| ii | `o rosto em as/suas mãos`<br>`'the face in the/your hands'` | 2 | 3 |
| | Total | **5** | **4** |

5. Collocations of *trabalho* 'work' and *fazer* 'to do'

| | Recurring Lexical Patterns | Number of instances in each subcorpus | |
|---|---|---|---|
| | | TR | NON-TR |
| i | TER (um) trabalho a fazer<br>'HAVE (some) work to be done' | 5 | 0 |
| ii | (para) fazer (bem) o (seu/meu) trabalho<br>'(in order) to do the (your/ my) work (well)' | 3 | 5 |
| iii | fazer um (bom) trabalho<br>'to do a/some (good) work' | 1 | 2 |
| | Total | **9** | **7** |

6. Collocations of *tarde* 'late'/'afternoon' and *noite* 'night'

| | Recurring Lexical Patterns | Number of instances in each subcorpus | |
|---|---|---|---|
| | | TR | NON-TR |
| i | tarde da noite 'late at night' | 5 | 4 |
| ii | mais tarde naquela/nessa noite<br>'later that night' | 4 | 0 |
| iii | até tarde da noite<br>'until late at night' | 2 | 1 |
| | Total | **11** | **5** |

7. Collocations of *tarde* 'late'/'afternoon' and *fim* 'end'

| | Recurring Lexical Patterns | Number of instances in each subcorpus | |
|---|---|---|---|
| | | TR | NON-TR |
| i | no fim da/daquela tarde<br>'in the end of the/that afternoon' | 2 | 5 |
| ii | fim da tarde 'end of the afternoon' | 2 | 1 |
| iii | fim de tarde 'late afternoon' | 0 | 8 |
| | Total | **4** | **14** |

8. Collocations of *mão* 'hand' and *esquerda* 'left'

| | Recurring Lexical Patterns | Number of instances in each subcorpus | |
|---|---|---|---|
| | | TR | NON-TR |
| i | a mão esquerda<br>'the left hand' | 11 | 5 |
| ii | minha/sua mão esquerda<br>'my/your left hand' | 8 | 0 |
| iii | com a mão esquerda<br>'with the left hand' | 7 | 6 |
| iv | da mão esquerda<br>'of the left hand' | 6 | 3 |
| v | dedos da mão esquerda<br>'fingers of the left hand' | 3 | 1 |
| vi | na mão esquerda<br>'on the left hand' | 0 | 3 |
| | Total | **35** | **18** |

9. Collocations of *água* ‘water’ and *copo* ‘glass’

| | Recurring Lexical Patterns | Number of instances in each subcorpus | |
|---|---|---|---|
| | | TR | NON-TR |
| i | (um) copo (cheio) de água<br>‘(a) glass (full) of water’ | 7 | 7 |
| ii | um copo d’água ‘a glass of water’ | 5 | 4 |
| | Total | **12** | **11** |

10. Collocations of *hora* ‘hour’/‘time’ and *meia* ‘half’

| | Recurring Lexical Patterns | Number of instances in each subcorpus | |
|---|---|---|---|
| | | TR | NON-TR |
| i | meia hora ‘half an hour’ | 9 | 9 |
| ii | meia hora depois ‘half an hour after’ | 5 | 7 |
| iii | meia hora antes ‘half an hour before’ | 1 | 2 |
| iv | meia hora mais tarde<br>‘half an hour later’ | 3 | 1 |
| v | por (mais) meia hora<br>‘for (more) half an hour’ | 3 | 1 |
| vi | daqui/dali a (approximadamente) meia hora ‘in (approximately) half an hour’ | 0 | 5 |
| vii | (uma) hora e meia ‘(an) hour and half’ | 1 | 2 |
| | Total | **22** | **27** |

11. Collocations of *verdade* ‘truth’ and *é* ‘is’

| | Recurring Lexical Patterns | Number of instances in each subcorpus | |
|---|---|---|---|
| | | TR | NON-TR |
| i | é verdade ‘it is true’ | 12 | 30 |
| ii | a verdade é (que) ‘the truth is (that)’ | 10 | 14 |
| iii | é a (pura) verdade<br>‘it is the (plain) truth’ | 8 | 2 |
| iv | é bem/mesmo verdade (que)<br>‘is actually/indeed true (that)’ | 7 | 1 |
| v | não é verdade ‘it is not true’ | 5 | 10 |
| vi | isso (só) é verdade<br>‘this is (only) true’ | 4 | 0 |
| vii | é verdade que ‘it is true that’ | 3 | 8 |
| viii | é na verdade ‘it is in fact’ | 3 | 0 |
| ix | é (bem) verdade (ou não) o que<br>‘it is (actually) true (or not) what’ | 0 | 4 |
| | Total | **52** | **69** |

12. Collocations of *quarto* 'room' and *é porta* 'door'

| | Recurring Lexical Patterns | Number of instances in each subcorpus | |
|---|---|---|---|
| | | TR | NON-TR |
| i | a porta (fechada) do (seu) quarto<br>'the (closed) door of the (his) room' | 3 | 8 |
| ii | à porta do quarto<br>'at the door of the room' | 3 | 1 |
| iii | na porta do quarto<br>'on the door of the room' | 0 | 4 |
| | Total | **6** | **13** |

13. Collocations of *noite* 'night' and *é meia* 'half'

| | Recurring Lexical Patterns | Number of instances in each subcorpus | |
|---|---|---|---|
| | | TR | NON-TR |
| i | à meia-noite 'at midnight' | 9 | 0 |
| ii | meia-noite 'midnight' | 6 | 1 |
| iii | depois de(a) meia-noite<br>'after midnight' | 5 | 5 |
| iv | por volta da meia-noite<br>'before midnight' | 3 | 1 |
| v | a meia-noite 'midnight' | 3 | 0 |
| vi | da meia-noite 'midnight' | 2 | 2 |
| vii | antes da meia-noite 'before midnight' | 1 | 2 |
| | Total | **29** | **11** |

14. Collocations of noite 'night' and é dia 'day'

| | Recurring Lexical Patterns | Number of instances in each subcorpus | |
|---|---|---|---|
| | | TR | NON-TR |
| i | dia e noite 'day and night' | 6 | 4 |
| ii | da noite pro/para o dia 'overnight' | 5 | 1 |
| iii | (de) o dia e/ou (de) a noite<br>'(of) the day and/or (of) the night' | 1 | 3 |
| iv | noite e dia 'night and day' | 1 | 2 |
| v | um/o dia e uma/a noite<br>'a/the day and a/the night' | 0 | 3 |
| | Total | **13** | **13** |

## *Author*

Carmen Dayrell currently holds a post-doctoral research position at the University of São Paulo (Brazil). She has a PhD degree in Translation Studies from the University of Manchester (UK).
E-mail: dayrellc@gmail.com
Website: http://www.fflch.usp.br/dlm/comet/