Diversity in collocational patterns of translated texts: a quantitative approach

Carmen Dayrell

Centre for Translation and Intercultural Studies /The University of Manchester (UK)[1]
Project financed by CAPES (Brazil)
carmen_dayrell@hotmail.com

Abstract: *The primary objective of this paper is to propose a corpus-based methodology for comparing quantitative aspects of lexical patterning in translated and non-translated texts of the same language. This study focuses on diversity of collocational patterns and examines the overall number of collocates for a given node in translated and non-translated texts as well as their tendency to converge around specific collocates. The analysis uses a comparable corpus of Brazilian Portuguese which consists of two separate subcorpora: one made up of translated Brazilian Portuguese and the other consisting of texts which have been originally written in Brazilian Portuguese.*

Resumo: *O principal objetivo deste artigo é propor uma metodologia que utilize corpora para a comparação de aspectos quantitativos da padronização lexical em textos traduzidos e não-traduzidos, ambos no mesmo idioma. O estudo aborda a diversidade de colocações, e visa a investigar o número total de colocados para cada um dos nódulos em textos traduzidos e não-traduzido e a tendência de cada um dos nódulos a convergir para determinados colocados. Os dados são extraídos de um corpus comparável do português brasileiro, composto por dois subcorpora: um subcorpus de textos traduzidos para o português brasileiro e outro de textos originalmente escritos em português brasileiro.*

1. Introduction

This paper reports on part of a larger PhD research project whose primary overall objective was to propose a corpus-based research methodology for comparing lexical patterning in translated and non-translated texts of the same language. The study aimed to investigate whether collocational patterns tend to be less diverse (i.e. reduced in range) in the translated texts in comparison with non-translated texts of the same language.

The term ***collocation*** is understood here to mean 'the occurrence of two or more words within a short space of each other in a text' (Sinclair 1991:170). As Kenny (2001:87) explains, collocation refers to the 'syntagmatic relationship between at least two lexical items, though these lexical items are not usually thought of as having equal status'. This study also follows Sinclair's terminology for the description of collocations. Sinclair (1991:115) suggests the term ***node*** 'for the word that is being studied' and ***collocate*** for 'any word that occurs in the specified environment of the node', that is, any word which collocates with the node.

This paper focuses on quantitative aspects of collocational patterns in translated and non-translated texts of the same language. The following hypotheses are tested with respect to the collocates of each node:

(1) translated texts may exhibit a lower number of collocates for each node in comparison with non-translated texts of the same genre;

(2) translated texts may show a stronger tendency to draw heavily on a small number of collocates in comparison with non-translated texts of the same genre.

The comparison is made between collocational patterns of translated texts vis-à-vis non-translated texts of the same language as opposed to other studies which compare collocational patterns of source and target texts (see, for instance, Berber-Sardinha 1999, 2000 and Kenny 2001). Data is extracted from a comparable corpus of Brazilian Portuguese which has been designed and compiled as part of the research project. In line with Baker's (1995:234) definition of comparable corpora, the Brazilian Portuguese comparable corpus (hereafter BPC) 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.

2. The Brazilian Portuguese Comparable Corpus (BPC)

BPC has been designed on the basis of the extensive list of parameters suggested in the literature for selecting individual texts to be included in a corpus (Atkins *et al.* 1992, Biber 1993, 1994, EAGLES 1996a, 1996b, Laviosa-Braithwaite 1996, Laviosa 1997). The general features of BPC are summarised as follows:

FEATURES	BPC
Type of corpus	comparable
Language	Brazilian Portuguese
Time span	synchronic – 1980 onwards
Medium	written, best-selling published books
Size of texts	full-texts
Genres	fiction and self-help

Table 1: General Features of the Brazilian Portuguese comparable corpus (BPC)

BPC focuses on Brazilian Portuguese specifically and includes only books published in Brazil from 1980 onwards, with priority being given to texts published from 1990 onwards. All books have been rated best-sellers in Brazil during the period under analysis. The lists of best-selling books used here are retrieved from a major Brazilian weekly magazine *Veja*. In addition to these criteria, the BPC includes only texts targeted at an adult audience. All 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 to avoid over-representing any single factor.

As regards text genres, BPC includes fiction and self-help. These two genres have been chosen because they are the most popular genres in Brazil during the period analysed and hence more likely to include a reasonable number of translated and non-translated texts. This paper focuses on the fiction subcorpus and this is why no detail of the criteria adopted for selecting self-help books is provided. The selection of texts in

the fiction subcorpus follows the Cataloguing-in-Publication (CIP) categorisation and includes only books classified as 'romance' in the Brazilian system.

An additional set of parameters is applied in the selection of translated texts (Table 2) given that, as Laviosa-Braithwaite (1996:57) explains, translated texts have some specific characteristics which are not shared by non-translated texts. In addition to the criteria detailed above, the translational corpus is designed to contain only direct translations from English, that is, translations from texts originally written in English. Moreover, BPC 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 within the specified time span.

FEATURES	Translated Brazilian Portuguese Corpus
Translational Corpus Type	Direct translations
Source language	English
Translators	Professional translators
Time span	1980 onwards

Table 2: Additional Features of the Translated Brazilian Portuguese Corpus

The present overall size of the fiction subcorpus, in terms of number of words, number of books and number of authors/translators, is presented in Table 3:

	Number of words	Number of books	Number of authors/translators
Translated Fiction	545,395	5	5
Non-translated Fiction	565,920	8	8

Table 3: Present overall size of the fiction subcorpus

3. Identifying lexical patterns

This section explains the methodological procedures for the retrieval of the collocational patterns to be analysed. The software package *Wordsmith* tools, version 3.0 (Scott 1999) is used here to manipulate the data. As will be seen, the process involves two major steps: (1) selection of the words to be taken as nodes, and (2) retrieval of their collocates.

An important point to be made here is that linguists have not yet reached a consensus on criteria for the automatic retrieval of collocational patterns. Not surprisingly, different approaches have been put forward and various criteria and values have been suggested. Decisions are therefore in many cases arbitrary and choices vary according to the scope and aims of each programme of research.

It is also important to emphasise that this study intends to focus on lexical patterning of lexical items as opposed to grammatical or functional items. By focusing on lexical items, it is not my intention to suggest that functional items cannot reveal 'interesting' lexical patterns or preferences. Some studies have already suggested that functional words 'actually have a clear lexical presence, which amounts to treating them in the same way as 'vocabulary' words are treated' (Sinclair 2003:105). However, lexical and functional items behave differently on the collocational level and hence reveal different aspects of the way phrases build up. Co-occurrences in which one or both items are functional words tend to hold a stronger grammatical influence (Jones

and Sinclair 1974). Thus, items which belong to the following grammatical classes are not chosen as nodes nor as collocates: articles, prepositions, conjunctions, interjections and pronouns[2].

The first step in the identification of collocations is to select the word(s) to be taken as **node**(s). Three basic criteria are applied in order to select nodes which best suit the scope and aims of the present study:

(1) minimum frequency of the item in each subcorpus (translated and non-translated):
What is considered here is the raw frequency of the item irrespective of the size of each subcorpus. A minimum frequency of 200 occurrences is used as a cut-off point. 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.

(2) similarity in the frequency of the item in the translated and the non-translated subcorpora:
Like the first criterion, this also refers to the raw frequency of the item within each subcorpus. It is also adopted for methodological convenience, based on the assumption that node frequency may have an influence on the diversity of collocational patterns associated with the node.

(3) grammatical class of the word:
This study focuses on the collocational patterns of items which are predominantly nouns. The idea is to select, within the range of words with a minimum frequency of 200 occurrences in the translated and the non-translated subcorpora of the same genre, 10 nouns whose frequencies in the two subcorpora are as similar as possible.

Table 4 shows the resulting selection of nodes for the fiction subcorpus. A relevant methodological point to be mentioned here is that, for the purposes of this study, nodes are selected taking into account individual word forms, not lemmas[3].

	NODES	Frequency in the TRANSLATED fiction subcorpus	Frequency in the NON-TRANSLATED fiction subcorpus
1.	manhã [morning]	222	223
2.	rosto [face]	385	388
3.	trabalho [work]	209	212
4.	tarde [late/afternoon]	284	300
5.	mão [hand]	517	540
6.	água [water]	221	247
7.	hora [hour/time]	245	271
8.	verdade [truth]	323	289
9.	quarto [room]	320	361
10.	noite [night]	593	545

Table 4: Selected nodes in the fiction subcorpus

Once the nodes have been selected, the next step is to retrieve their **collocates**. Three main criteria are used:

(1) preference is given to lexical rather than grammatical items;

(2) the frequency of co-occurrence with the node:

In order to be selected as a collocate, items must co-occur at least 4 times with the node in a span of 4 words to the right and 4 words to the left of the node (4:4), disregarding structural boundaries. The choice of this specific window size follows Sinclair (1991).

(3) the strength of the association of node and collocate:
This is estimated by means of the mutual information index (hereafter MI) proposed by Church and Hanks (1990) and Church et al. (1991). MI formalises suggestions made in the literature (Sinclair 1987, 1991, Stubbs 1995) that the comparison between the actual observed frequency of co-occurrence and the expected frequency if the items were to co-occur by chance can provide a rough measure of the strength of attraction between relevant items. The threshold of 4 is used to select collocates.

It is important to emphasise that, for the purposes of this study, I have opted for keeping span constant and window size is not taken explicitly into consideration here for computing MI, following Stubbs (1995). However, it is worth mentioning that the issue will be addressed in future research, given that window size can clearly affect the retrieval of collocates.

4. Comparing collocational patterns in translated and non-translated texts

➢ Hypothesis (1): Translated texts may exhibit a lower number of collocates for each node in comparison with non- translated texts of the same genre.

Table 5 shows the overall number of collocates retrieved for each node selected from the fiction subcorpus. For 80% of the nodes (8 out of 10), the number of collocates is lower in the translated subcorpus. The exceptions are the node **rosto** which shows the same number of collocates in both subcorpora and the node **verdade** which shows a lower number of collocates in the non-translated subcorpus.

	NODES	TRANSLATED FICTION		NON-TRANSLATED FICTION	
		Node frequency	Number of collocates	Node frequency	Number of collocates
1.	manhã [morning]	222	**22**	223	27
2.	rosto [face]	385	38	388	38
3.	trabalho [work]	209	**9**	212	13
4.	tarde [late/afternoon]	284	**28**	300	32
5.	mão [hand]	517	**48**	540	62
6.	água [water]	221	**20**	247	24
7.	hora [hour/time]	245	**22**	271	24
8.	verdade [truth]	323	27	289	**25**
9.	quarto [room]	320	**40**	361	42
10.	noite [night]	593	**51**	545	55

Table 5: Total number of collocates for each node

One could argue, however, that the number of collocates may be influenced by the node frequency. For instance, **água** occurs 221 times in the translated subcorpus with 20 collocates and 247 times with 24 collocates in the non-translated subcorpus. This is a typical case in which the lower node frequency in the translated subcorpus may account for the lower number of collocates. To obtain a clearer picture of the influence of the node frequency, the next step is to calculate the difference in the number of collocates in relation to the frequency of the node in each subcorpus. This is done by

dividing the node frequency by the number of collocates in each subcorpus. Thus, in the case of **água**, 221 is divided by 20 and 247 is divided by 24, which renders the ratios 11.0 and 10.3 respectively. A higher ratio implies that each collocate would co-occur with the node a higher number of times; therefore, it reflects a lower number of collocates (Table 6). The abbreviation TR is used to indicate the translated subcorpus and NTR for the non-translated subcorpus.

	NODES	TRANSLATED ratio	NON-TRANSLATED ratio	Subcorpus showing a LOWER number of collocates
1.	manhã [morning]	**10.0**	8.2	**TR**
2.	rosto [face]	10.1	**10.2**	NTR
3.	trabalho [work]	**23.2**	16.3	**TR**
4.	tarde [late/afternoon]	**10.1**	9.4	**TR**
5.	mão [hand]	**10.8**	8.7	**TR**
6.	água [water]	**11.0**	10.3	**TR**
7.	hora [hour/time]	11.1	**11.3**	NTR
8.	verdade [truth]	**11.9**	11.6	**TR**
9.	quarto [room]	8.0	**8.6**	NTR
10.	noite [night]	**11.6**	9.9	**TR**

Table 6: Number of collocates in relation to the node frequency

For the nodes **hora**, **verdade** and **quarto,** when the number of collocates is assessed in relation to the node frequency, the lower proportion of collocates shifts to the other subcorpus (see Tables 5 and 6). This means that, for these three nodes, the lower number of collocates may be simply reflecting a lower node frequency. If we leave these three nodes aside and focus only on the remaining 7 nodes, we find that 6 nodes show a lower number of collocates in the translated subcorpus. In other words, 86% of the nodes confirm the hypothesis that translated texts tend to exhibit a lower number of collocates overall in comparison with non-translated texts.

Hypothesis (1): Number of Collocates	Number of nodes in the FICTION subcorpus
Translated texts exhibited a LOWER number of collocates for each node	6 (86%)
Translated texts did NOT exhibit a lower number of collocates for each node	1 (14%)
Total number of nodes	7 (100%)

Table 7: Findings of hypothesis (1) in the fiction subcorpus

- Hypothesis (2): Translated texts may show a stronger tendency to draw heavily on a small number of collocates in comparison with non-translated texts of the same genre.

The second hypothesis is tested by examining the overall distribution of collocations in relation to a given node in the translated and non-translated subcorpora. This starts by applying a test of statistical significance – **chi-square** – in order to determine whether collocates of a given node are evenly distributed in each subcorpus. Chi-square is adopted here to indicate whether the difference between the actual observed distribution of collocations of a given node and a hypothetical mean distribution of collocations in the same subcorpus is statistically significant. The mean distribution assumes that all collocates of the node occur in equal proportion in that particular subcorpus and is

calculated by dividing the overall number of occurrences by the total number of collocates of the node in that subcorpus.

The node **manhã** will serve as an example. In the translated subcorpus, **manhã** has 22 collocates with an overall number of occurrences of 174 (see Appendix I). In an entirely homogeneous distribution of collocations, each collocate would co-occur with the node 7.9090 times (174 divided by 22). The chi-square compares the actual observed frequencies of collocations (f(n,c)) with a hypothetical homogeneous distribution in which all 22 collocates would co-occur with the node 7.9090 times. This gives a p-value of 0.0000 in the translated subcorpus. In the non-translated subcorpus, **manhã** has 27 collocates with an overall number of occurrences of 178. If the distribution of collocations were entirely homogeneous, each collocate would co-occur with the node 6.5925 times (178 divided by 27). The p-value is 0.5044 in the non-translated subcorpus. Table 8 shows the resulting p-value for the distribution of collocations of the 10 nodes extracted from the fiction subcorpus.

	NODES	p-value in the TRANSLATED subcorpus	p-value in the NON-TRANSLATED subcorpus
1.	manhã [morning]	0.0000	0.5044
2.	rosto [face]	0.0000	0.0394
3.	trabalho [work]	0.4046	0.1481
4.	tarde [late/afternoon]	0.0000	0.0035
5.	mão [hand]	0.0000	0.0000
6.	água [water]	0.4634	0.6986
7.	hora [hour/time]	0.0000	0.0000
8.	verdade [truth]	0.0000	0.0000
9.	quarto [room]	0.0048	0.0000
10.	noite [night]	0.0000	0.0000

Table 8: Resulting p-value (chi-square test) for the distribution of collocations in the fiction subcorpus

The resulting p-values are interpreted in relation to a pre-established ***level of significance***[4]. It is standard procedure in many research fields, and social sciences in particular, to adopt the threshold 0.05 for the level of significance. In our case, a resulting p-value $\geq$ 0.05 shows that the difference between the actual and mean distributions of collocations is not statistically significant. Collocates are therefore homogeneously distributed in the relevant subcorpus, in other words, there is no tendency to draw heavily a small number of collocates in that particular subcorpus. Conversely, a p-value $\leq$ 0.05 reveals that the difference between the observed and the mean distribution of collocations in the same subcorpus is statistically significant. This means that there is a tendency to converge around a smaller number of collocates. However, it is important to point out that a p-value $\leq$ 0.01 indicates that the difference between the actual and the mean distributions of collocations is highly significant, that is, there is a strong tendency to draw heavily on a small number of collocates. On the other hand, a p-value between 0.01 and 0.05 shows that there is a statistically significant difference between the actual and mean distributions of collocations; however, the tendency to converge around a small number of collocates may not be as highly pronounced.

After establishing whether the distributions of collocations of a given node in both translated and non-translated subcorpora are homogeneous, the next step is to compare the results in two subcorpora. For the nodes under analysis here, we find the following differences between the resulting p-value in the translated and non-translated fiction subcorpora:

(1) one subcorpus renders a p-value ≥ 0.05 and the other renders a p-value ≤ 0.05:
This is the case of the node **manhã**. The chi-square test shows that the translated subcorpus displays a strong tendency to draw heavily on specific collocates (p-value = 0.0000), whereas the same does not hold true for the collocational patterns of the node in the non-translated subcorpus (p-value = 0.5044).

(2) one subcorpus shows a p-value between 0.01 and 0.05 and the other subcorpus renders a p-value ≤ 0.01:
The distribution of collocations of the node **rosto** gives a p-value of 0.0000 in the translated and 0.0394 in the non-translated subcorpus (Table 8). These figures show that both subcorpora reveal a tendency to converge around specific collocates. However, a p-value of 0.0000 in the translated subcorpus shows that there is a strong tendency to draw heavily on a small number of collocates. By contrast, a p-value of 0.0394 shows that, even though there is a statistically significant difference between the actual and mean distributions of collocations in the non-translated subcorpus, the tendency to converge around a small number of collocates is not as highly pronounced as in the translated subcorpus.

(3) both p-values are ≥ 0.05:
As can be seen in Table 8, this is the case of the nodes **trabalho** and **água**. In these two cases, both the translated and the non-translated subcorpora reveal an even distribution of collocations. In other words, from a statistical standpoint, neither subcorpus shows a tendency to draw heavily on a small number of collocates.

(4) both p-values are ≤ 0.01:
This is the case for the remaining six nodes extracted from the fiction subcorpus. For these six nodes, both subcorpora display a strong tendency to draw heavily on a small number of collocates.

In the two cases where there is a statistical difference between the p-values in the translated and non-translated subcorpora – (1) and (2) above –, the researcher can be confident to state which subcorpus demonstrates a stronger tendency to converge around specific collocates. Thus, the nodes **manhã** and **rosto** support the hypothesis that translated texts draw more heavily on a smaller number of collocates than non-translated texts of the same genre.

For the remaining 8 nodes, no statistically significant difference is found between the distributions of collocations in the translated and non-translated subcorpora. Notwithstanding, the tendency to converge around specific collocates can still be assessed by examining individual frequencies of collocations. I focus here on the top 3 collocates of the node in each subcorpus, based on the assumption that they are more likely to reflect major quantitative differences between the collocations of a given node in the two subcorpora. What is considered here is frequency percentages of the collocates in each subcorpus. The frequency percentage is calculated by dividing each collocation frequency by the overall number of occurrences and the resulting figure is converted into a percentage figure. Taking the node **manhã** as an example (Appendix

I), we find that it co-occurs with **seguinte** 31 times in the translated subcorpus (out of 174 occurrences), representing 18% of collocations.

Thus, for the nodes with no statistically significant difference between the distributions of collocations in the translated and non-translated subcorpora, the hypothesis is tested by comparing the sum of the frequency percentages of the top 3 collocates in each subcorpus (summarised in Table 9). The higher the resulting sum the stronger tendency to draw heavily on a smaller number of collocates. As can be seen in Table 9, for only one node (**quarto**), the sum of the frequency percentages of the top 3 collocates is higher in the non-translated than in the translated subcorpus.

	Nodes	**Translated**	**Non-translated**
1.	trabalho [work]	**51%**	36%
2.	mão [hand]	**23%**	13%
3.	tarde [late/ afternoon]	**29%**	21%
4.	água [water]	**28%**	22%
5.	verdade [truth]	**54%**	49%
6.	quarto [room]	18%	**21%**
7.	hora [time/ hour]	**33%**	31%
8.	noite [night]	**18%**	16%

Table 9: Sum of the frequency percentages of the top three collocates of a given node in the translated and non-translated subcorpora

Table 10 summarises the results of hypothesis (2) in the fiction subcorpus. 90% of the nodes (9 out of 10) support the hypothesis that translated texts tend to draw more heavily on a smaller number of collocates than non-translated texts of the same genre. As discussed above, **quarto** is the only node which does not confirm the hypothesis.

Hypothesis (2): Tendency towards specific collocates	**Number of Nodes in the FICTION Subcorpus**
Translated texts display a STRONGER tendency to draw heavily on a small number of collocates	9 (90%)
Translated texts did NOT display a stronger tendency to draw heavily on a small number of collocates	1 (10%)
Total number of collocational patterns	10 (100%)

Table 10: Findings of hypothesis (2) in the fiction subcorpus

5. Conclusion

This paper proposes a corpus-based research methodology for investigating diversity of collocational patterning in translated and non-translated texts of the same language. The study focuses on quantitative aspects of collocational patterns and examines the overall number of collocates in relation to a given node in the translated and non-translated subcorpora as well as the distributions of collocations in the two subcorpora. The results indicate that translated texts tend to show a lower number of collocates for each node in comparison with non-translated texts of the same genre. It is also suggested that translated texts tend to display a stronger tendency to draw heavily on specific collocates in comparison with non-translated texts of the same genre.

The originality of this study is to examine collocational patterns of translated texts vis-à-vis non-translated texts of the same language. This is therefore an alternative approach for assessing collocational patterns in translated texts and an attempt to shed some new light on the complex nature of translated language.

6. References

Atkins, Sue, Jeremy Clear, Nicholas Ostler (1992) 'Corpus Design Criteria' in *Literary and Linguistic Computing*, 7 (1): 1-16.

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

Berber-Sardinha, Tony (1999) 'Estudo Baseado em Corpus da Padronização Lexical no Português Brasileiro: Colocações e Perfis Semânticos' in *PROPOR'99. IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*, Évora, Portugal, pp. 269-287. Available at: http://lael.pucsp.br/~Tony (accessed in Sep/2004).

_____ (2000) 'Semantic Prosodies in English and Portuguese: a Contrastive Study' in *Cuadernos de Filologia Inglesa*, 9 (1): 93-110, Spain. Available at http://lael.pucsp.br/~Tony (accessed in Sep/2004).

Biber, Douglas (1993) 'Representativeness in Corpus Design' in *Literary and Linguistic Computing*, 8 (4): 243-257.

_____ (1994) 'An Analytical Framework for Register Studies' in *Sociolinguistic Perspectives on Register*, Douglas Biber and Edward Finegan (eds.), Oxford: Oxford University Press, pp. 31-56.

Church, K. and P. Hanks (1990) 'Word Association Norms, Mutual Information, and Lexicography' *in Computational Linguistics* 16 (1): 22-29.

Church, K., W. Gale, P. Hanks and D. Hindle (1991) 'Using Statistics in Lexical Analysis' in Uri Zernik (ed.) *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, New Jersey: Lawrence Erlbaum Associates Publishers, pp. 115-164.

EAGLES (1996a) 'Preliminary Recommendations on Corpus Typology' J. Sinclair (ed.) in *EAGLES document EAG-TCWG-CTYP/P*, May/1996. Available at http://www.ilc.cnr.it/EAGLES96/home.html (accessed in Dec/2004).

EAGLES (1996b) 'Preliminary Recommendations on Text Typology' J. McH Sinclair and J. Ball (eds.) in *EAGLES document EAG-TCWG-TTYP/P*, Jun/1996. Available at http://www.ilc.cnr.it/EAGLES96/home.html (accessed in Dec/2004).

Jones, Susan and John Sinclair (1974) 'English Lexical Collocations: A Study in Computational Linguistics', *Cahiers de Lexicologie* 24:15-61.

Kenny, Dorothy (2001) *Lexis and Creativity in Translation: a Corpus-based Study*, Manchester: St. Jerome Publishing.

Kurtz, Norman R. (1999) *Statistical Analysis for the Social Sciences*, Boston: Allyn and Bacon.

Laviosa, Sara (1997) 'How Comparable Can Comparable Corpora Be?' in *Target* 9(2):289-319.

Laviosa-Braithwaite, Sara (1995) *The English Comparable Corpus (ECC): a Resource and a Methodology for the Empirical Study of Translation*, unpublished PhD Thesis, Manchester: UMIST.

Scott, Mike (1999) *Wordsmith Tools version 3.0*, Oxford: Oxford University Press.

Sinclair, John (1987) 'The Nature of Evidence' in *Looking Up*, London: Harper Collins.

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

_____ (2003) *Reading Concordances*, London: Pearson Education Ltd., Longman.

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

Appendix I

This Appendix lists all collocates retrieved for the node **manhã** in the translated and non-translated fiction subcorpora. The following abbreviations are used: f(n) for the frequency of the node; f(c) for the frequency of the collocate; f(n,c) for the frequency of the collocation; MI for the mutual information value of the collocation; and % f(n,c) for the percentage frequency of the collocation.

	Translated: f(n) = 222						Non- Translated: f(n) = 223					
	Collocate	English glossary	f (c)	f (n,c)	MI	% f (n,c)	Collocate	English glossary	f (c)	f (n,c)	MI	% f (n,c)
1	seguinte	following	116	31	9.36	18%	café	coffee	136	12	7.81	7%
2	café	coffee	84	29	9.73	17%	seguinte	following	83	12	8.52	7%
3	horas	hours/o'clock	168	13	7.57	7%	depois	after	1,128	11	4.63	6%
4	hoje	today	135	11	7.65	6%	horas	hours/o'clock	239	11	6.87	6%
5	cinco	five	145	9	7.25	5%	já	already	1,439	10	4.14	6%
6	primeira	first	254	8	6.27	5%	cinco	five	191	8	6.73	4%
7	cedo	early	69	7	7.96	4%	dia	day	753	8	4.75	4%
8	casa	house	563	6	4.71	3%	havia	there was	825	8	4.62	4%
9	duas	two	354	6	5.38	3%	onze	eleven	37	7	8.91	4%
10	quatro	four	157	6	6.55	3%	oito	eight	99	6	7.26	3%
11	amanhã	tomorrow	66	4	7.22	2%	quatro	four	157	6	6.60	3%
12	Aurora	Aurora	576	4	4.09	2%	sol	sun	162	6	6.55	3%
13	dia	day	597	4	4.04	2%	tinha	had	906	6	4.07	3%
14	hora	hour/o'clock	245	4	5.33	2%	três	three	323	6	5.56	3%
15	ligar	to call	33	4	8.22	2%	acordou	woke up	36	5	8.46	3%
16	ligou	called	32	4	8.26	2%	Alzira	Alzira	393	5	5.01	3%
17	manhã	morning	222	4	5.47	2%	cedo	early	83	5	7.26	3%
18	nove	nine	40	4	7.94	2%	certa	certain/right	111	5	6.84	3%
19	Peter	Peter	399	4	4.62	2%	duas	two	356	5	5.16	3%
20	seis	six	106	4	6.53	2%	eram	they were	396	5	5.00	3%
21	sol	sun	101	4	6.60	2%	feira	(weekday)	115	5	6.79	3%
22	votação	voting	25	4	8.62	2%	hora	hour/o'clock	271	5	5.55	3%
23							sete	seven	88	5	7.17	3%
24							dez	ten	193	4	5.72	2%
25							hoje	today	284	4	5.16	2%
26							meia	half	104	4	6.61	2%
27							tarde	late/afternoon	300	4	5.08	2%
				174						178		

Notes:

[1] I am very grateful to my statistics consultant, Dr. Fatima Sanchez Cabo, whose help and interest were vital for developing the statistics analysis presented in this study. Thanks are also due to my supervisor Professor Mona Baker for her relevant comments on an earlier version of this paper.

[2] The list of items discarded in this study is based on lists of Portuguese functional words used by the following Brazilian researchers (personal communication): Tony Berber-Sardinha (PUC-SP) and Helena Caseli and Juliana Greghi (NILC - USP). I gratefully acknowledge their cooperation.

[3] The term ***lemma*** refers to 'a label under which all the inflected forms of a word can be gathered, where inflections are understood as minor and predictable changes in the shape of a word' (Kenny 2001:34). Kenny uses the example of the forms *write, writes, writing, written* and *wrote* which are all inflected forms of the lemma WRITE.

[4] The level of significance is 'the point at which the difference between what is observed and what is expected is too great to be due to chance or random variation' (Kurtz 1999:151).