Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 95 (2013) 336 – 343

Procedia
Social and Behavioral Sciences

5th International Conference on Corpus Linguistics (CILC2013)

Corpus-driven Methodology for Exploring Cultural References in the Areas of Cooking, Football and Hotel Industry

Rozane Rodrigues Rebechi*

Universidade de São Paulo, Av. Prof. Luciano Gualberto, 403, São Paulo, 05508-900, Brasil

Abstract

This paper investigates how corpora can reveal cultural aspects, so as to contribute to the construction of comprehensive bilingual Portuguese-English specialized glossaries of cooking, football and hotels. Despite posing a great challenge for translators, these areas lack bilingual terminological reference works. As a consequence, appropriateness and idiomaticity in the target texts may be compromised. The methodology used here shows that specialized corpora can expand the scope of terminological research by revealing cultural aspects of a special subject field through its linguistic patterns and also enable the search for equivalents in the target language.

Selection and peer-review under responsibility of CILC2013.

Keywords: Corpus linguistics; terminology; cultural references; translation;

1. Introduction

Culinary, football and hotel industry are, among others, areas which carry significant cultural features of the country they refer to. Like language, cooking is said to be "[…] a truly universal form of human activity" (Lévi-Strauss, 2007: 28), as it reveals traditions which are characteristic of a given community. Football in turn is the most popular sport in several countries, among them Brazil and England. Nevertheless, this sport reveals distinctive styles when played in different places. In relation to the hotel industry, establishments also show peculiarities regarding their place of origin. These cultural features certainly pose a great challenge for translators, since these professionals must not only transmit words but, above all, concepts from one culture to another.

* Corresponding author. Tel.: +55 11 49903056
E-mail address: rozanereb@gmail.com

1877-0428 © 2013 The Authors. Published by Elsevier Ltd.
Selection and peer-review under responsibility of CILC2013.
doi:10.1016/j.sbspro.2013.10.655

In Brazil, these three areas will be particularly related due to major sports events which will take place in the country in the next few years: the Confederations Cup in 2013, the FIFA World Cup in 2014 and the Olympic Games in 2016. More than ever, the country will attract foreign tourists who will certainly use mostly English to communicate. Menus, restaurant reviews, hotel websites, newspaper reports on match results etc. will be translated into English but studies have revealed a shortage of meaningful terminological works in those areas. Navarro (2011) discusses the inexistence of a bilingual reference work which could provide translators with examples, collocations and phraseologies in the area of hotel industry, thus enabling natural equivalence in the target language. With regard to the translation of football texts, Matuda (2011) argues that both the few online and the printed bilingual dictionaries available are not comprehensive enough to provide phraseological equivalents in the pair of languages English-Portuguese. Rather, they simply present lists of decontextualized equivalent terms in the target language. In relation to the English translation of Brazilian cooking texts, Rebechi (2010) identified problems such as mistranslations and wrong definitions of terms, substitution of ingredients and lack of translation standardization, just to mention but a few problems found.

As a contribution toward filling this gap, glossaries in the three areas are being built. These works focus on culture-specific terminological units extracted from authentic texts.

2. Corpus linguistics, terminology and translation

Zanettin (2012), citing Robert de Beaugrande (1994), states that, as a social phenomenon, language should be analyzed based on real data. The analysis of large amounts of data is enhanced by a methodology based on Corpus Linguistics (CL). CL has been of great relevance to terminological studies because its computational tools help with terminology recognition and definitions extracted from textual corpora (Cabré, 2005). However, in general it is not enough for translators to be aware of equivalents and meanings of terms. They also need to know what other words these terms occur with in order to produce fluent texts in the target language. Besides allowing for the identification of the terms actually used in a certain specialized domain, corpus linguistics electronic tools also enable the retrieval of useful contextual chunks (Pearson, 1998).

Thus, here we understand language as a probabilistic system which functions by means of pre-fabricated conventionalized units. Successful bilingual glossaries should help translators produce fluent translations in Portuguese by providing professionals not only with appropriate equivalents but also with contextualized examples, linguistic and textual patterns, collocations and other information.

Concerning the search for target-language equivalents and clusters, translation-driven corpora, defined by Zanettin (2012: 8) as "[…] those which are created and/or used for some translation-related purpose", are useful for the analysis of typical linguistic behavior as well as for individual translation choices. Tognini-Bonelli (2001) emphasizes that, by using this approach, the linguist is committed to evidence. In this study we use a corpus-driven methodology in the sense that we extract from the corpus data which are worth analyzing, according to the criteria established for the compilation of each glossary.

3. The study corpora

The study corpora of the three areas mentioned in this paper were compiled according to criteria defined for each research. We will briefly describe the composition of each corpus.

3.1. The football corpus

The football corpus, compiled by Matuda (2011), is a comparable Portuguese-English corpus with texts available on the internet. Its distribution is shown in Table 1:

Table 1. The football corpus.

	Newspaper reports on match results		Laws of the game		Live minute by minute commentaries		Commentaries by sports journalists and by football fans via Twitter and Facebook		**Total**	
	Tokens	Texts	Tokens	Texts	Tokens	Texts	Tokens	Texts	**Tokens**	**Texts**
Portuguese	544,002	1,335	24,593	1	311,147	284	37,331	21	**917,073**	**1,641**
English	584,931	947	22,583	1	322,895	138	72,488	17	**1,002,897**	**1,103**

It is interesting to observe that for balancing the corpus in relation to size – approximately one million tokens in each language –, the Portuguese subcorpus demanded more texts than the English subcorpus. Matuda (2011) explains that this is a consequence of the difference between high-context culture and low-context culture, as pointed out by Hall (1976), for whom the amount of linguistic and contextual information necessary to transmit meaning varies from culture to culture. In high context cultures, like Brazil, many things are left unsaid, very little is in the explicit part of the message; while in lower context cultures, like England, the communicator needs to be much more explicit, that is, the mass of information is vested in the explicit code.

3.2. The culinary corpus

Regarding the Brazilian cooking research, a comparable corpus and a parallel corpus were built. The comparable corpus comprises four cookbooks originally written in Portuguese and eleven in English, whereas the parallel corpus consists of four cookbooks originally written in Portuguese and translated into English. Because this is printed material, it had to be digitalized so as to be analysable by a computational tool. It is also important to mention that, in general, these cookbooks, mainly the ones targeted to a foreign audience, also provide additional information about ingredients, dishes, customs etc. Since this content is useful for the extraction of term definitions and comments, it is also part of the corpus. This content was saved separately, under the name 'informational'. Table 2 shows the distribution of the culinary corpus:

Table 2. The culinary corpus.

	Comparable corpus		Parallel corpus	
	Recipes	Informational texts	Recipes	Informational texts
	Tokens	Tokens	Tokens	Tokens
Portuguese	176,112	76,763	65,367	39,155
English	280,151	148,139	68,155	42,463

All the recipes in Portuguese, totalizing 241,479 tokens, were used for the retrieval of terms which are characteristic of Brazilian cooking, whereas equivalents, definitions, clusters, examples and comments were retrieved from both the recipes and the informative texts in English, which explains why it is important for the English subcorpora to be larger.

3.3. The hotel industry corpus

For the hotel field, Navarro (2011) built a comparable corpus comprising texts extracted from Brazilian and American hotel websites. The corpus is divided into five hotel categories. Table 3 shows the distribution of the corpus:

Table 3. The hotel industry corpus.

English			Portuguese		
	Tokens	Texts		Tokens	Texts
Hotels (1-5 stars)	115,608	50	Hotels (1-5 stars)	109,873	135
Resorts	110,615	20	Resorts	110,444	61
Bed and Breakfast	100,125	74	*Pousadas* [inns]	107,219	183
Suites	100,723	96	*Hotéis-fazenda* [farm hotels]	101,368	150
Condos	119,035	81	Flats	85,545	181
Total	**546,106**	**321**	**Total**	**514,449**	**710**

As can be seen in Table 3 above, hotel categories that make up the corpus are not 100% equivalent across both languages, this is so because the hotel industry in Brazil and in the US have distinct characteristics, consequently the corpus reflects two different realities. Moreover, the corpus had to be balanced according to the number of words rather than the number of texts. In order to reach a similar number of tokens in both subcorpora, the Portuguese subcorpus required over twice the number of texts of the English subcorpus. This fact demonstrates another cultural aspect: while Brazilian hotel websites favor a more concise type of communication, American establishments tend to give more detailed descriptions of their premises. We can therefore say that the compilation of the corpus revealed cultural aspects typical of both countries it represents, in addition to cases of non-equivalence at the idiomatic level.

4. Analysis

Cultural markers are textual and discursive elements that point to a given culture (Zavaglia et al., 2011). The semiautomatic analysis of the corpora used in this study helped reveal features of the culture they come from. Using the software *WordSmith Tools 6.0* (Scott, 2012), we retrieved keywords from the Portuguese language corpora for cooking and football and from the English language corpus for hotels. Table 4 presents a list of the first ten keywords of the three corpora, in descending order of keyness:

Table 4. First ten keywords in the three corpora.

	Culinary	Football	Hotel
1	MANDIOCA	BOLA	SUITES
2	CHEIRO-VERDE	GOL	ROOM
3	DÁ	FALTA	HOTEL
4	COCO	ESCANTEIO	INTERNET
5	KG	BATIDO	ROOMS
6	AZEITE-DE-DENDÊ	MINUTOS	SUITE
7	CHARQUE	COBRANÇA	OUR
8	PIRÃO	JOGO	AMENITIES
9	CARNE-DE-SOL	PARTIDA	SPA
10	COLORAU	CHUTE	CENTER

As an exemplification of the methodology employed for building the specialized glossaries, we selected one keyword from each corpus: *mandioca*, from the cooking corpus, *gol*, from the football corpus, and *room*, from the hotel corpus. By looking at those words in their contexts (KWIC), as well as their collocates and clusters, we retrieved their main terminological patterns, i.e. combinations which form complete units of meaning. The next step was to look for their equivalents by analyzing lists of keywords, collocates, clusters and concordance lines in the

target language subcorpora. In the following sections, we analyze each keyword separately to show how they were used in the construction of entries for the three glossaries.

4.1. Mandioca

In spite of having the highest keyness in the Portuguese cooking subcorpus, the word *mandioca* does not always correspond to a single term. By retrieving the clusters with this keyword, we observed that from its 409 entries in the Portuguese subcorpus, 299 refer to the combination *farinha de mandioca*, a type of flour made from this edible starchy tuberous root which grows abundantly in Brazil.

All over Brazil the term *mandioca* refers to a kind of this root which is poisonous and, therefore, has to be detoxified before being transformed into flour and other byproducts. In different regions of the country *aipim* and *macaxeira* are names given to the non-poisonous root, which is simply cooked and pureed or fried. Nevertheless, in the state of São Paulo the term *mandioca* refers to both types.

In relation to the search for equivalents, the English subcorpus also provides three possibilities: 'manioc', 'cassava' and 'yuca', the latter usually misspelled as 'yucca'. However, these variants are not always synonyms and, therefore, interchangeable in every context. The analysis of the English subcorpus shows that 'cassava', 'yuca' and 'manioc' are used to refer to the non-poisonous root, but 'yuca' is never used to refer to the flour.

Using the software *TshwaneLex*, we built an entry and two subentries for *mandioca*, considering the needs of translators and writers, thus providing examples of use, clusters and useful comments extracted from the corpus, as shown in Fig. 1:

mandioca *(noun)* *(Manihot utilissima)* manioc, cassava, yuca A tropical plant with green leaves and starchy tuberous roots.// Basically two types are used: a non-poisonous type, which is known as *aipim* in Rio and the southern states, *macaxeira* in the North and Northeast and *mandioca* in São Paulo, and a poisonous type, known as *mandioca* throughout Brazil.

mandioca

mandioca *(noun)* manioc, cassava, yuca Edible root which must be cooked before being fried or puréed.//Ex.: *Add the cooked manioc to the pan and shake a couple of times to coat the pieces with the butter and the onion. [TBT]*Ex.: *Peel cassava root, cut into slices and boil in water with 1 tablespoon salt until it is fork tender. [Braz.Cook.]*// The non-poisonous tubers are simply cooked and fried or mashed, as they are used in the preparation of stews, such as *bobó*.//Clust. manioc root, (peeled); sweet manioc; grated manioc; fried manioc **Also known as aipim, macaxeira, mandioca-doce, mandioca-mansa See also bobó**

mandioca *(noun)* manioc, cassava A tropical starchy tuberous root that is poisonous when raw.//Ex.: *Cooked, dried, then ground into flour or a coarse meal, it is one of the key ingredients, along with beans, fish, and rice, used in the cooking of Brazil. [TCOB]*// The hydrogen cyanide present in these tubers must be removed by extensive washing and cooking before they become edible. The juice extracted from the pulp is detoxified by boiling and becomes the basis for *tucupi*. Several kinds of flour originate from the dried pulp. A dish called *maniçoba* is prepared with its leaves, called *maniva*. **Also known as mandioca-amarga, mandioca-brava See also farinha de mandioca, maniçoba, maniva, polvilho, puba, carimã**

Fig. 1. Entry for *mandioca* in the Brazilian cooking glossary.

4.2. Gol

If lemmatized, *gol* and *gols* would account for the word with the highest keyness in the football corpus. Therefore, Matuda (2011) decided to analyze the phraseological units formed with this keyword. With *WordSmith Tools*, clusters with *gol(s)* were identified in the Portuguese subcorpus and phraseologies such as [*fazer/marcar*] [*o/um*] *gol* appeared among the most frequent ones. In order to find equivalents for these clusters, a similar procedure was followed in the English subcorpus with the word 'goal(s)'. This search revealed a much larger variation of phraseologies with the meaning 'score a goal': '[snatch/deliver/blast in/thunder/poke/fire/hammer/slot] a goal' are the most recurrent. However, it is important to emphasize that both in Portuguese and in English the action of 'scoring a goal' is not always transmitted by using the word 'goal'. *Empatar*, *finalizar*, 'equalise' and 'finish', for example, also convey this action if contextualized.

An example of the football glossary entry is shown in Fig. 2:

FAZER {o ★ | um} gol
*Eguren emendou para a rede e **fez o gol** do alívio uruguaio.*
*Ainda não foi desta vez que a Grécia **fez um gol** em Mundiais.*

= MARCAR {o+|um} gol
= MARCAR o tento
= FINALIZAR
= EMPATAR

SCORE ([{a | the} goal])
*The old warhorse, Pavel Nedved, equalized for Juve and **scored** again to give the Italian giants the lead they wanted.*

FIRE {home | in (a goal) | (the ball) past [goalkeeper] | into the net}
*Sheff Utd James Harper **fires in a goal** from close in low into the middle of the goal.*

Fig. 2. Glossary entry for *gol* adapted from Matuda (2011).

4.3. Room

The hotel industry glossary is a bilingual English-Portuguese reference work based on collocations. Using the software *WordSmith Tools*, the researcher identified 'room' as being the most frequent word in the corpus, besides having the highest keyness if lemmatized with the plural form 'rooms'. Therefore, phraseologies with this word were chosen to demonstrate the methodology used in building the glossary. The researcher identified 33 main collocations, among them 'accessible room', 'banquet room', 'book [a/your/this] room' and 'in-room safe', which would make up entries in the glossary. It was observed that there is not a single equivalent for the term 'room' in Portuguese. Rather, Portuguese real contexts showed that phraseologies with 'room' should many times be reformulated in order to be idiomatic in Portuguese.

Let us take 'in-room safe' as an example. In order to identify a suitable equivalent for the term, the researcher looked for the main clusters with the word *cofre* (the *prima facie* translation of 'safe') in the Portuguese subcorpus and observed that *cofre individual* is the most recurrent collocation, followed by collocations which specify different kinds of 'safe', such as *cofre digital* ['digital safe'] and *cofre eletrônico* ['electronic safe']. This analysis reveals that in Brazilian websites it is not necessary to mention that the object is located in the room. Having this in mind, Navarro (2011) built a glossary entry for the term, as shown in Fig. 3:

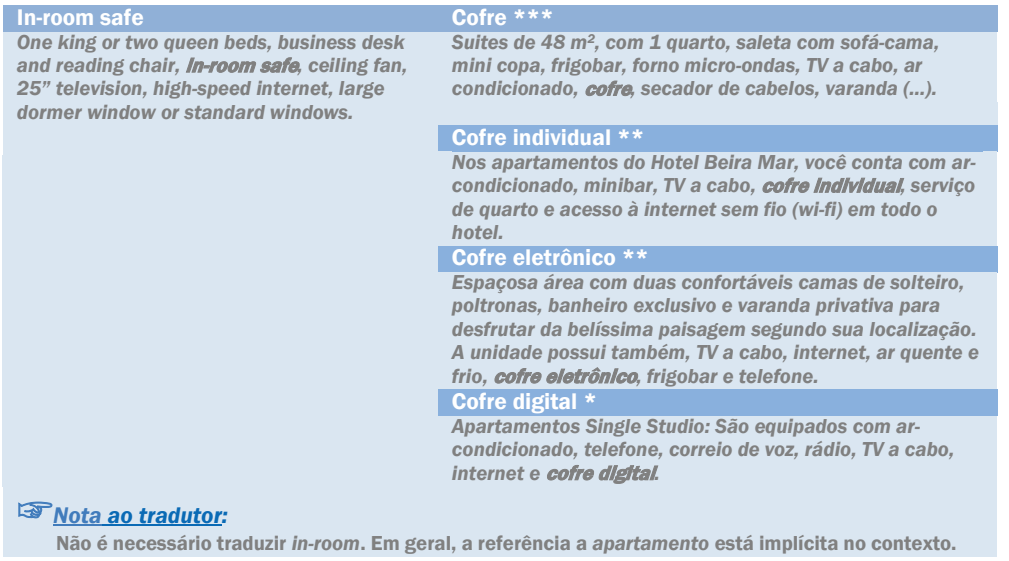

Fig. 3. Glossary entry for 'in-room safe'.

5. Concluding remarks

This study sought to demonstrate that specialized corpora can expand the scope of terminological research by revealing cultural aspects of a special subject field through its linguistic patterns. These cultural aspects should be systematically included in reference materials, especially bilingual publications, in order to increase translator's cultural awareness. It is not enough for professionals to know the equivalent terms in the target language. They also need to understand how these terms are really used in order to produce a text which is fluent and sounds natural in the target language. Therefore, a comprehensive glossary should also provide translators with examples, clusters and comments extracted from real contexts of use.

Acknowledgements

This research is supported by a grant from FAPESP (2011/19609-0).

I wish to thank Sabrina Matuda and Sandra Navarro for providing data derived from their studies on football and hotel industry, respectively, and to my supervisor, Stella Tagnin, for revision, advice and suggestions.

References

Cabré, M. T. (2005). *La Terminología: representación y comunicación*. Barcelona: Universitat Pompeu Fabra.

Hall, E. T. (1976). *Beyond culture*. New York: Anchor Press/Doubleday.

Lévi-Strauss, C. (2007). The culinary triangle. In C. Counihan (Ed.). *Food and culture: a reader* (pp. 28-35). New York: Routledge.

Matuda, S. (2011). *A fraseologia do futebol: um estudo bilíngue português-inglês direcionado pelo corpus*. Dissertation – FFLCH, USP, São Paulo.

Navarro, S. L. M. (2011). *Glossário bilíngue de colocações de hotelaria: um modelo à luz da Linguística de Corpus*. Dissertation – FFLCH, USP, São Paulo.

Pearson, J. (1998). *Terms in context*. Amsterdam/Philadelphia: John Benjamins.

Rebechi, R. R. (2010). *A imagem do brasileiro no discurso do norte-americano em livros de culinária brasileira*. Dissertation – FFLCH, USP, São Paulo.

Scott, M. (2012). *Wordsmith Tools 6.0*. Oxford: Oxford University Press.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Zanettin, F. (2012). *Translation-driven corpus: corpus resources for descriptive and applied translation studies*. Manchester/Kinderhook: St. Jerome.

Zavaglia, A., Azenha, J., and Reichmann, T. (2011). Cultural markers in LSP Translation. In K.-D. Baumann (Ed.), *Fach Translat Kultur. Interdisziplinäre Aspekte der vernetzten Vielfalt* (pp. 785-808). Berlim: Frank & Timme.