Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 95 (2013) 637 – 643

Procedia
Social and Behavioral Sciences

5th International Conference on Corpus Linguistics (CILC2013)

Corpora and Interpretation: a Study of Coffee Field Terminology

Luciana Latarini Ginezi*

University of São Paulo, São Paulo, Brazil

Abstract

This paper aims at presenting the results of a research based on the work of professional consecutive interpreters in the coffee field. When working with specialized areas, interpreters are frequently faced with variation in terminology, especially due to the spoken language in use. By using Corpus Linguistics methodology, we have produced a vocabulary (Portuguese – English) of coffee terms and their variants, showing the importance of including oral data as input to terminology work for interpreters.

Keywords: Corpus Linguistics; oral corpus; interpreting studies, coffee terminology, spoken language.

1. Introduction

Brazil is the number one Arabica coffee producer and exporter in the world (ABIC, 2013). This leading position attracts international business to the country and, consequently, justifies the need for translators and interpreters in the area, which also demands specialized reference materials (dictionaries, glossaries, vocabularies, corpora etc.). However, few terminological resources are available for the coffee domain in Brazil. Terminological standardization is still very controversial in relation to language use, that is, it does not reflect language as it is actually used (Pearson, 1998). Published works in Brazil in the coffee domain do not make reference to the methodology used, and as they are usually written by coffee specialists, there is an underlying assumption that the terms were recognized intuitively. While I was working as a consecutive interpreter for a company that sells coffee machinery worldwide, based in Brazil, the need for building my own reference material was evident, due to the lack and poor quality of bilingual materials in this field.

* Corresponding author. Tel.: +55-11-3091-2929; fax: +55-11-3091-2929.
E-mail address: luginezi@usp.br; luginezi@uol.com.br

doi:10.1016/j.sbspro.2013.10.692

In the beginning, the corpus was built using digital informative texts from the company and other texts collected from specialized websites. The importance of building terminological reference materials using Corpus Linguistics is unquestionable, since we need authentic language to be the input for a work based on evidence. As Kennedy (1998, p. 8) says, “Any scientific enterprise must be empirical in the sense that it has to be supported or falsified on evidence and, in the final analysis, statements made about language have to stand up to the evidence of language in use.” The evidence, according to the author, may be based on the intuition of native speakers or on a corpus. The difference resides in the richness of a reliable and valid work.

But then another issue came to scene. Although written materials could provide the general and substantial terminological guide, there were terms used only orally by rural workers, farmers or technicians of the company’s plant in Brazil. Besides, foreigners coming from all over the world to learn about the Brazilian coffee production and processing also had “their” own English terms, which made me rethink my reference material. I would need oral data to improve the corpus, coming from conversations among different professionals, lectures from specialists as well as real interpreting material.

Therefore, I started to build the oral corpora in the coffee domain. The Brazilian Portuguese (BP) data was collected in the South of Minas Gerais state and in the Mogiana region of São Paulo State. Both places are referential in the Brazilian coffee market. In English (EN), oral data was collected from foreigners visiting the Brazilian companies for business, like Vietnamese, American, Australian, Indian etc. Therefore, English spoken as lingua franca was considered for this research, as interpreting is performed with non-native English speakers all the time. Halfway through data collection, I realized the BP corpus was far larger than the EN one. So I went to Ethiopia, the birth place of coffee and also one of the biggest producers in the world, in the year of 2007. There I attended the 4th African Fine Coffee Conference and Exhibition and could collect hours and hours of data.

1.1. Terminology, Interpreting and Corpus Linguistics

The interaction between terminology and interpreting is clear, given the specialized areas interpreters have to work with. The languages for specific purposes, or specialized languages, are mainly constituted by terms used in professional communication. Therefore interpreters should know the terms in both languages, besides other linguistic and extra-linguistic information. According to Gile (1995), “(…) the Knowledge Base of interpreters and translators of non-literary texts seldom contains all the information necessary for them to perform their work, so that Knowledge Acquisition is a regular and important part of their Translation operations”. To reach this knowledge, glossaries, vocabularies and specialized dictionaries may help the task. Sometimes interpreters have to go to field and talk to specialists, or even trust their intuition about terms. However, using corpora to provide all specialized language information interpreters need is the most efficient way, since they consult only one resource instead of multiple resources (Bowker & Pearson, 2002). Also, a corpus provides contexts and authentic language in use, and even if you have to count on a specialist help, this context will help you know the area beforehand, and formulate the right questions, saving time for other tasks during preparation step for interpreting jobs. Gile (1995) comments the correct terminological use in interpreting (and also translation) may be considered the essential element to the final users. However he points out that the glossaries developed by interpreters are not as reliable as the ones developed by translators, because the latter have more time available to prepare them. We believe interpreting may require more than written information resources, as we will see below.

In Corpus Linguistics Studies (CLS), due to the time consuming task required to compile an oral corpus, investigations focus mainly on written language. The same can be said about interpreting when compared to translation. But since interpreting is an oral and communicative task, glossaries, vocabularies or dictionaries should consider its spoken nature and give variants the right space. Additionally, spoken language is about interaction; speakers improvise, they rely on the other’s speech to make their own. This also happens in specialized languages, when the participants of a conversation use their turns to be understood, according to the information flow. Consecutive interpreters have to reproduce the meaning of this conversations or dialogues, using the specialized vocabulary the speakers are referring to in both working languages. What usually happens, though, is that workers from different professional categories or from different regions use variants of the so-called standardized terms. The situation is particularly acute when the specialized area covers rural and urban areas, as the coffee domain suggests in this study.

The objective of this paper, thus, is to present the construction of a vocabulary for interpreters, working from Portuguese into English, in the coffee domain, based on the needs of interpreters for consulting bilingual material which contains variants. I also present a contrast between written and oral corpora. The interaction among Corpus Linguistics, Terminology and Interpreting Studies is clear and necessary to develop this multidisciplinary study.

2. Methodology

2.1. Data collection and transcription

Before starting collecting oral data, the researcher should define some criteria. The first one is related to the topics in the domain or sub domain. The coffee area covers a broad range of topics, as plantation, harvesting, processing, machinery engineering, plant diseases, market, trading, coffee quality, social-economic working conditions etc. For this study, we have selected the topics harvesting and after-harvesting processing. In these two sub domains, there is a wide range of workers involved, as well as different opportunities for interpreters. The next decision is about the participants for the corpus compilation: who they are, how many we need, where the recordings will take place, how the recordings will be performed, and what genres are necessary for the corpus. Finally, we have to check if all variables are possible in both languages.

The participants for this research are rural workers, farmers, agronomists, machinery technicians, traders and coffee consultants. The recordings were performed as conversations (three or more participants), interviews (one participant and the researcher) or interpreting activity (minimum three: two speakers and the interpreter). The places, as mentioned above, were Brazil and Ethiopia, during visits to farms, meetings or conferences. The text length was not previously determined. In order for speakers to speak naturally, the conversation was not interrupted. For this reason, there were more words in BP than in EN, even with almost the same number of speakers: 29 in BP and 33 in EN. To reach this number, we based our criteria on the number of words we had for the corpus: about 40,000 words in BP and 30,000 in EN.

Needless to say that the recordings that took place outdoors were much more difficult to transcribe than the ones indoors. Also, recordings in English were harder to understand than in Portuguese, mainly because of the different accents. If the data was not transcribed right after the recording, we realized some information could be missing. Sometimes the recorder would stop working, due to battery fall, as in Sidamo fields, countryside region of Ethiopia, where there were no energy sources available everywhere; other times, the recorder with speech control would stop at a pause and restart only after the beginning of the talk. There were also cases of voice folding, when we could not understand the speech, and many other situations that limited the transcription task. Unfortunately, the transcription task consumed most of the time in the research. Some research has been developed in the area of voice recognition to facilitate transcription. However, the error margin is still too high and, for this study, the software available in English did not fit. In Brazilian Portuguese there was no software available at all.

Each 1-hour recording required an average 3-hour transcription. We still have 37 hours of material to be transcribed, which was not used in this study, due to time constraints.

2.2. Corpora design and exploration

This study is composed of comparable corpora, except for the Interpreting Corpus, which is multilingual. Comparable corpora are, according to McEnery and Wilson (1996), “collections of individual monolingual corpora with the same or similar sampling procedures and categories for each language but contain completely different texts in several languages”.

The Interpreting Corpus was tagged for BP and EN, with about 10,000 words for each language, as shown below.

Table 1. Coffee Corpora.

Corpus	Nº of words	Type/Token ratio

BP oral	39,859	10
EN oral	29,390	14
Interpreting (BP/EN)	20,893	13
BP written	100,274	10
EN written	102,379	09

The written reference corpora were constituted of coffee non-related areas, as Social Sciences, Religion, Human Sciences and Health. We extracted the data from the corpora available at Lácio Web (2013) for BP and a journalistic corpus compiled from the same areas of knowledge for EN. The amount of words was nearly four million words for each language. The oral corpora were analyzed using the written reference corpus, because during the research we could prove that there was no difference in using an oral reference corpus in EN, the Micase (2006), with two million words, or the journalistic corpus, in the results of keywords. It was important to use the reference corpus in written form to avoid discrepancy with the BP corpora, since the only reference oral corpus available in BP was NURC – SP (2002), but unfortunately it was not available in readable format for this research at the time. Besides, it was considered small for a reference corpus, with 45,000 words.

The tool used for the analysis was the WordSmith Tools (WST), versions 4-6 (Scott, 2013).

2.3. Analysis

The first step was to analyse the oral corpora. We started with the BP corpus, since the vocabulary is Portuguese-English. After obtaining wordlists and their respective keywords, we had 209 term candidates. After that, we retrieved the data from the EN oral corpus, which amounted to 78 term candidates. The same methodology was applied to the written corpora: there were 396 term candidates in BP and 417 in EN.

With the keyword lists, we identified *prima facie* equivalents. But for some terms, however, there were doubts; this was when the concordance line analysis helped find the equivalents. Tagnin (2007) explains this methodology, which can be summarized as follows: first, analyse the concordance line of the term for which you want to find the equivalent. Then, identify the collocates for this term. Next, analyse the concordance lines of the collocates in the target language and finally identify the equivalent term.

One example is the term “terreiro”, the first place where the harvested coffee is dried after being washed. If we need to know the equivalent, or equivalents, we go to Concordance tool. There, select the collocate list. Look at the results for “terreiro”.

Table 2. Collocates for *terreiro*

N	Word	With
1,00	TERREIRO	terreiro
10,00	CAFÉ	terreiro
11,00	SECAR	terreiro
17,00	PRO	terreiro
20,00	TAMBÉM	terreiro
24,00	TEMPO	terreiro
25,00	LÁ	terreiro
26,00	PRA	terreiro
27,00	QUANDO	terreiro
28,00	TEM	terreiro
29,00	SECADOR	terreiro
30,00	TRADICIONAL	terreiro

The results show “café” (“coffee”) and “secar” (“dry”) as the first and second options (and not grammatical words). If we look for the word “coffee” in the EN corpora, we will find many collocates, since this is the most frequent word in the corpora. But the word “dry” may help. In the EN oral corpus, by analysing the concordance lines of “dry”, we finally found “patio” as the equivalent for “terreiro”. For the written corpus, the equivalents were “yard” and “terrace”. The collocation “chão do terreiro” was also quite common in the oral corpus, as well as “patio floor” in EN.

The search for variants followed the same methodology. The BP variants could be classified in two main groups: linguistic (phonological, morphological, syntactic and lexical) and discourse register. The EN variants were not classified for this study.

Table 3. Examples of variants

Phonological	colheta/colheita coiê café/colher café terreiro/terrero
Morphological	botão/botãozinho terreiro/terreirão
Syntactic	colher com a mão/colheita manual colheita mecanizada/colher com máquina
Lexical	café cereja/cereja grão verde/verde
Discourse register	barra/saia café catado no chão/café de varrição

3. Result analysis

The bilingual vocabulary design consists of the entry, in BP, its equivalent in EN, and the variants, when existent.

In a broader study, one would also include the collocations, definition and examples of usage. However, interpreters usually do not have the time to search for definitions while working, due to the immediacy of the interpreting task. Thus, the bilingual vocabulary we show below is a model we consider enough for consecutive interpreters, and even though it is not as complete as it could be, it is reliable.

Table 4. Excerpt from the bilingual vocabulary

BP	**EN**
Apanhar café (v.) *Var. colher café; coiê café; panhar café*	to pick coffee; to harvest coffee
Café bóia (s.m.) *Var. bóia; grão bóia*	floater; lights
Café de varrição (s.m.) *Var. café catado no chão; café de chão; café do chão; varrição*	gleanings

We had 178 entries, taken from oral and written corpora (BP). There were 92 variants, for the BP terms, and 51% of them were exclusively present in the oral corpora, as in the entries below.

Table 5. Excerpt from the bilingual vocabulary

BP	EN
pé de café (s.m.) *Var. árvore; árvore de café*	coffee tree
fazenda de café (s.f.) *Var. propriedade*	coffee farm
terreiro (s.m.) *Var. terrero; terreirão; terreiro tradicional*	patio

The EN equivalents were found in both oral and written corpora. There were variants, but as the vocabulary was Portuguese into English, we decided to present the terms and variants in EN at the same level. After concluding the matches between terms and their equivalents, a coffee specialist checked the reliability of the work. The complete vocabulary can be found in Ginezi (2007).

The amount of variants found only in the oral corpora determines the importance of compiling oral corpora to produce a glossary, vocabulary or dictionary for specialized language, especially when the user will work with spoken language. It is obvious that written corpora are necessary, at first because they are faster to compile and also because, in the coffee domain, statistics demonstrate they provide the larger amount of data.

As the oral and written corpora have different sizes, we cannot tell the lexical richness from the amount of term candidates we have found. But if we take a look at the standardized TTR, a statistical data provided by the WordList in the WST, we conclude how dense languages are in the corpora. The standardized TTR is the proportional pattern between types and tokens. This calculation is done on 1,000 basis, it means that for each 1,000 words, the rate of textual lexical richness is calculated. In other words, the standardized TTR calculates the distinct forms found at each 1,000 words. With this statistical tool, it is possible to compare different size corpora, as shown below.

Table 6. Standardized TTR.

Corpus	Standardized TTR
BP oral	33.27
EN oral	32.11
BP written	51.17
EN written	41.68

The lexical richness in the coffee domain, though, is higher in the written corpora, although the difference inside the written corpora is also representative. The EN written corpus is 29.8% lexically richer than the EN oral corpus, and the BP written corpus is 22.8% lexically richer than the EN written corpus. It is not a great difference (5%), and perhaps the BP written corpus should be evaluated in terms of balance. Further research can be developed, so as to identify the text topics that comprised all corpora. As we mentioned before, because of the recording hours still not transcribed, we believe all corpora can be increased.

4. Conclusion

Investigation in spoken language is time demanding in several aspects, from data collection to transcription, as well as the corpora balance in terms of participants or planning the participants' speech. Despite the problems concerning spoken language, the main objective of this study has been reached, with the results showing the importance of including oral data as input to build glossaries, vocabularies or specialized dictionaries for interpreters.

New research can be conducted using the coffee corpus, with interaction of other areas, as cultural translation, for example. And much more is yet to be done with the remaining material collected during this research.

Corpus Linguistics methodology proved to be an efficient theoretical and practical referential for this study, with its technological tools. As we observed, the major constraint to perform more investigation with oral corpora is the voice recognition or transcription, which should be technologically improved and available for investigation.

Acknowledgements

Special acknowledgements to the Department of Modern Languages (DLM), Linguistics and Literary Studies in English area, at the University of São Paulo, for providing me with financial aid to attend CILC 2013.

References

Abic. (2013). Associação Brasileira da Indústria de Café. http://www.abic.com.br/publique/cgi/cgilua.exe/sys/start.htm?sid=59&infoid=2492.
Bowker, L. & Pearson, J. (2002). Working with specialized language. A practical guide to using corpora. London & New York: Routledge.
Gile, D. (1995). Basic concepts and models for interpreter and translator training. Amsterdam & Philadelphia: John Benjamins Publishing Company.
Ginezi, L. L. (2007). Brazilian coffee – Portuguese and English variants in the spoken language. Master's Degree Dissertation. Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo: São Paulo. Available at http://www.teses.usp.br/teses/disponiveis/8/8147/tde-03042008-134508/pt-br.php
Kennedy, G. (1998). An introduction to corpus linguistics. London & New York: Longman.
Lácio-Web. (2013). Available at: http://www.nilc.icmc.usp.br/lacioweb/corpora.htm.
McEnery, T. & Wilson, A. (1996). Corpus Linguistics. Edinburgh: Edinburgh University Press.
Micase (2006). Michigan Corpus of Academic English. Cd.
Nurc-SP. (2002). Projeto Nurc-SP. CD-ROM.
Pearson, J. (1998). Terms in context. Amsterdam & Philadelphia: John Benjamins Publishing Company.
Scott, M. (2013). WordSmith Tools 6.0. Available at: www.lexically.net.
Tagnin, S. (2007) A identificação de equivalentes tradutórios em corpora comparáveis. Available at www.fflch.usp.br/dlm/comet.