Translator-oriented, corpus-driven technical glossaries: the case of cooking terms

Stella E.O. Tagnin¹ and Elisa Duarte Teixeira¹

Abstract

Although there are many English–Portuguese technical glossaries on the market, very few are designed to meet the specific needs of translators, whose main task is the production of an idiomatic text either in the mother tongue or in the foreign language. For that purpose, a mere list of monolexical terms (usually based on previous compilations or abstract conceptual schemata, and their equivalents) will not suffice. Most importantly, the producer of a text needs to know how a word is used in context, and this can be inferred from the observation of authentic examples. Corpus Linguistics has proved to be an invaluable tool in retrieving technical terms and phraseologies from corpora. In this paper, we employ a corpus-based methodology or, more precisely, the ‘corpus-driven approach’, to compile a bilingual, monodirectional English–Portuguese glossary of cooking terms (Teixeira and Tagnin, 2008), covering the various stages of the project, with an emphasis on the identification of candidate terms, and their subsequent validation through the generation of frequency and keyword lists using the lexical analysis software, WordSmith Tools.

1. Technical translation, Terminology and Corpus Linguistics

It is a known fact that translators need a variety of reference sources to produce a good quality translation. The more specialised the text to be translated, the greater the need for specific terminological sources. However, most existing reference sources for the English–Portuguese pair pose two problems: first, they do not meet the needs of translators as regards usage, (i.e., they do not provide the context in which the term occurs), and, secondly,
they are rarely up-to-date, because technology and knowledge now advance at such a pace that no published work can keep up with them. Thus, most bi- and multilingual technical glossaries tend to focus on the comprehension of technical texts by apprentices of specialised areas, rather than on the production of texts by translators. Indeed, they merely list monolexical terms and their ‘equivalents’, without examples from which the translator might glean some information as to their usage, as can be seen under Figure 1.

Many misconceptions regarding the practice of technical translation have been pointed out by various authors (Byrne, 2006: 2–7; and Korning Zethsen, 1999; among others), and all of them can be subsumed to one: the idea that technical texts can be reduced to the terminological content they convey. In technical translation, the correct use of terminology does not guarantee, alone, that the resulting text will be a good translation (Azenha Jr, 1999; Byrne, 2006; Korning Zethsen, 1999; and Wright, 1993)—that is to say, one that re-textualises the original message in an accurate, fluent and natural-sounding way. Wright (1993: 70, original emphasis) notes:

Documents must speak ‘the language’ of the target audience and should resemble other texts produced within that particular language community and subject domain […] These considerations frequently
require that translators move beyond merely correct strategies in terms of lexical and grammatical content in order to account for stylistically appropriate solutions.

One must bear in mind that a translator is a text producer and, therefore, needs to know how a word is used – how it associates with other words (Firth, 1957); and this information can be inferred from the observation of authentic examples.

In this sense, a further handicap of the majority of technical reference materials is that entries generally comprise only nouns and adjectives while little or no attention is devoted to words with other grammatical functions (Finatto, 2007). In addition, these reference materials seem to ignore the fact that technical language does not consist solely of monolexical terms; multiword expressions, as Estopá (1999) and Bevilacqua (2001) claim, may also enjoy terminological status and should, therefore, be registered as entries in their own right. That would be the case, for example, of including not only ‘pepper’ or ‘black pepper’ but also ‘freshly ground black pepper’ in a glossary of cooking terms. Moreover, the selection of which terms will be included in the glossary is usually the compiler’s choice, and the compiler sometimes relies on previous works, abstract schemata or the discretion of one or a few specialists, irrespective of the term’s relevance and frequency in actual texts from the field in question (Teixeira, 2008).

In short, most reference materials that are available are not translator-oriented: they do not provide context, collocations, phraseologies, translation suggestions in cases when there are no ‘direct equivalents’, or even information about the usage of a term in different genres and textual types (Fromm, 2008; Fuentes Morán and García Palacios, 2002; Gómez and Vargas, 2004; Salgado, 2006; Teixeira, 2004, 2008; and Varantola, 1998).

All of the shortcomings of reference materials we have mentioned above are evident in the translation of most technical texts and become even more evident in the translation of cooking recipes (Colina, 1997; and Teixeira, 2004, 2008). Contrary to popular belief, cooking is a highly technical area, not only in its practice – an inadequate procedure may put the dish to waste – but especially when it comes to translating texts. First, recipes may convey culture-specific habits, ingredients or even procedures which may be strange to the target audience. Secondly, some ingredients or appliances may not exist in the target culture, which raises the question of how they should be translated. A literal translation may produce awkward and sometimes hilarious renderings such as 1/2 xíc. de chá de suco de lima for ‘1/2 cup lime juice’. In Portuguese, ‘lime’ is limão (the most common variety, Tahiti) whereas ‘lemon’ is limão siciliano; and lima is a type of orange with a slightly bitter taste. Another example of a strange rendering would

---

2 The fact that it can be rendered as ‘pimenta-do-reino moida na hora’ in Brazilian Portuguese and ‘pimenta preta moida na altura’ in European Portuguese is good evidence that this is a cohesive translation unit.
be chocolate semi-doce (‘semi-sweet chocolate’) instead of the conventional phraseology chocolate meio-amargo (‘half-bitter chocolate’); a cook would probably understand what chocolate semi-doce means, but would realise, nevertheless, that ‘that is just not the way we say it’.

Corpus Linguistics is an empiricist approach to language (McEnery and Wilson, 1996), and has been playing an important role in Translation Studies and, more specifically, in the practice and teaching of technical translation (Maia, 1997; Tagnin, 2002; and Varantola, 2002). In addition, it has also proved to be a valuable tool for Terminology in that it allows for the retrieval of technical terms and phraseologies, and their equivalents, from corpora (Bowker, 1996; Bowker and Pearson, 2002; Pearson, 1998; and Tognini-Bonelli, 2002) by means of special software that identifies recurrent patterns of language.

Corpus Linguistics views language as a probabilistic system, since it is concerned not only with what it is ‘possible’ to produce in a language, but mainly with what will ‘probably’ occur. It relies on the observation of linguistic phenomena in large quantities of text – called ‘corpora’ – in order to produce generalisations about that language. Thus, it seems natural to assume that a glossary based on a corpus may provide the answer to some of the needs of a translator, mentioned above, because: (a) a corpus can be built according to the translator’s needs; (b) it can be constantly updated; and (c) when carefully compiled, it offers authentic examples from natural-sounding and accurate texts – all of which help to reassure the translator that the term chosen is the most appropriate one.

In this paper, we will present the corpus-driven approach used to compile a bilingual monodirectional English–Portuguese glossary of cooking terms that we published in Brazil (Teixeira and Tagnin, 2008). We will discuss the various stages of the project, with an emphasis on the identification of candidate terms using frequency and keyword lists, and their subsequent validation through concordance lines, with the lexical analysis software, WordSmith Tools 3.0 (Scott, 1999).

2. Corpus-driven Terminology

Corpus Linguistics can be viewed from two perspectives: as a methodology and as an approach. In Brazil, it has mostly been used by terminologists as a methodology—that is, electronic corpora are compiled and then perused to retrieve definitions, examples and/or equivalents for selected terms (see, for example, Alves, 1998; and Krieger et al., 2006). Although some glossaries are a step forward in the sense that they use Corpus Linguistics tools to generate Keyword lists from which some of the entries for the glossary are selected (e.g., Perrotti-Garcia and Rebechi, 2007), the corpora they are

---

3 In this paper, the interrogation mark preceding a combination of words indicates that this is a doubtful rendering, (i.e., a non-natural sounding combination).
based on contain mainly didactic texts, in which defining contexts abound, rather than state-of-the-art materials that circulate among the specialists of a particular field of knowledge.

As stated before, traditional Terminology focusses mainly on nouns and adjectives; verbs have only recently received some attention (Bevilacqua, 2001; Estopá, 1999; and Finatto, 2007). Corpus-based terminology does the same: emphasis is usually given to the conceptual schema of the area under investigation, and these are defined by specialists who sometimes disagree about them. The meaning of terms is not, thus, deduced from the contexts in which they actually occur in authentic texts, but from the conceptual relations they have in the schema’s hierarchy.

Last but not least, corpus-based glossaries, like traditional glossaries, will include phraseologies only if they contain a term (nouns or adjectives, but very rarely verbs). For example, *enquanto isso* (‘meanwhile’), which is a recurring Brazilian Portuguese phrase in cooking recipes, (and would be rendered as *entretanto* in European Portuguese recipes), would not be considered a valid entry because it does not include a content word. This is evidence that such theoretical constraints can do a disservice to the translator.

In previous works (Teixeira, 2008; and Teixeira and Tagnin, 2008), we proposed the use of Corpus Linguistics as an approach to bilingual Terminology, as opposed to the corpus-based methodology used by some terminologists. By corpus-based methodology we mean that, although these researchers use a corpus, they do so to speed up traditional methods of terminology compilation. In such instances, the corpus is merely used as a ‘fish pond’ (Hanks, 2008: 220, citing Sinclair, 1987) from which the terminologist yields examples and definitions to fill in a list of terms that has been compiled previously; the list is usually based on a domain area schema devised and/or approved by a specialist. For instance, Farias and Bezerra (2008) resort to magazines, newspapers and catalogues to extract equivalents for their trilingual glossary of fashion terms based on a previously compiled monolingual Portuguese Glossary of Fashion Terms (Farias, 2003). The authors explicitly state that they used Farias (2003) as the ‘corpus’ from which they extracted the list of source language terms to be used as the basis for their trilingual glossary. Besides, any doubts resulting from this kind of procedure are, again, submitted to one or more specialists who will, ultimately, decide what enters and what does not enter the dictionary. In contrast, by corpus-driven approach we mean that even the list of terms should be retrieved from the corpus; in this sense, the corpus will ‘tell’ us which are the terms most commonly used in the area being investigated, not the specialist.

If we were to draw an analogy, we could say that in the same way an ‘armchair linguist’ (Fillmore, 1992) relies on a native speaker to validate intuitions, a corpus-based terminologist relies on a specialist to validate the information to be entered into a technical dictionary. We propose, instead, that a specialised corpus (and not one or a few specialists) be
the main source of information used to build the dictionary entries and, especially, that the corpus be used to generate the list of headwords for the dictionary. The specialist can, then, have a more peripheral—and perhaps more effective—role in the compilation of translator-oriented technical dictionaries, such as helping the terminologist to choose the best sources of technical texts to build a sound, up-to-date and representative corpus.

Our Vocabulário para Culinária Inglês–Português (‘English–Portuguese Glossary of Cooking Terms’) was built according to this approach. We departed from the information contained in a carefully built comparable corpus of cooking recipes to select our entries, and we used a corpus-driven approach to extract information from the corpus using WordSmith Tools. In what follows, after briefly describing the corpus contents, we detail the methodology we adopted to build the list of entries in English and to find equivalents in the Portuguese comparable subcorpus. We then give some examples of the resulting ‘translator-oriented entries’, which are followed by our concluding remarks.

3. Compiling a corpus-driven glossary

3.1 Corpus building

In order to produce a bilingual corpus-driven glossary, we first built a comparable corpus—two corpora, one in each language, consisting of authentic texts covering the same domain, the same specific areas, and consisting of similar texts (in terms of genre, text typology, content, time span, extension, etc.). This would guarantee that all the resulting data were based on authentic language use. Parallel corpora (original texts and their translations) could also have been used as a source of prima facie equivalents for the entries identified in a corpus of authentic texts, but such corpora are rarely available on the Internet and, besides, require extra time to process, as they need to be aligned. So, our hypothesis was that a novice or even an experienced translator/terminologist working in a new subject area would be able to build a reliable bilingual glossary using well-built comparable corpora.

For the purpose of our cooking glossary, we gathered two subcorpora, one in English and one in Brazilian Portuguese, consisting of cooking recipes collected from reliable sources on the Internet, (i.e., websites containing recipes originally written, collected and/or edited by respected professionals in both languages). For this corpus in particular, we did not consult a specialist to help to locate representative and accurate texts on the Internet, since, being a corpus of home cooking recipes, the authors themselves—as home cooks and as home cookbook translators—agreed they should be able to evaluate which websites contained adequate texts. In the English language, some of the websites we used were The Great British
Translator-oriented, corpus-driven technical glossaries

Table 1: Composition of the comparable corpus

<table>
<thead>
<tr>
<th></th>
<th>Portuguese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of recipes</td>
<td>c. 8,300</td>
<td>c. 7,400</td>
</tr>
<tr>
<td>Tokens</td>
<td>1,520,864</td>
<td>1,578,125</td>
</tr>
<tr>
<td>Types</td>
<td>14,635</td>
<td>28,903</td>
</tr>
<tr>
<td>Type/token ratio</td>
<td>0.96</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Table 1: Composition of the comparable corpus

Kitchen by Helen Watson, and the Food section of the BBC website; for the Portuguese, the cooking section of the Terra portal, as well as the Basilico website, and the Gula magazine website were some of the main sources.

The recipes were copied manually and/or using offline browsers, such as HTTrack. The html files were then converted into txt files using the Text Converter utility of WordSmith Tools (Scott, 1999). This corpus is now part of the CorTec, the Technical Comparable Corpus of the CoMET project, and is available online. For detailed information about its compilation and its contents, see Teixeira (2008: 206–18). Information concerning its size can be seen under Table 1.

3.2 Corpus-driven approach

Wordsmith Tools was also the software used in our analysis. Since our aim was to produce a bilingual glossary in the English→Portuguese direction, we began by using the WordList function to generate a wordlist of the English corpus, (i.e., a list of all the words in all the texts of the corpus with their frequencies). Figure 2 shows the three screens produced by the Wordsmith’s WordList tool with the results: statistics (S), words in alphabetical order (A) and words in frequency order (F).

In order to have a more accurate picture of the vocabulary specific to the cooking area, we then generated a KeyWord list. This is done by comparing the WordList of the English subcorpus with a WordList of a larger and more general corpus. This procedure removes the vocabulary that is commonly used in the texts of both corpora and highlights the vocabulary peculiar to the area under investigation – in our case, cooking recipes.

---

4 See: http://www.httrack.com/. These browsers automatically extract all files from a chosen Internet address and store them in the user’s computer. This allows the user to surf the website even when the Internet connection is off, and means that they end up with a copy of the website in their computer (something that, unfortunately, most websites do not allow nowadays). The only problem is that one usually has to scan the files after the download is complete in order to delete unwanted ones (those that do not contain the type of text you are trying to collect). Even so, it is a much quicker process than copying each file manually.

5 See: http://www.fflch.usp.br/dlm/comet/consulta_cortec.html
Figure 2: Results generated by the WordList tool

used the written part of the American National Corpus (ANC)\(^6\) as a reference corpus. This produced over 3,670 keywords—words which occur in the English corpus of cooking recipes more often than we would expect them to occur by chance alone; the first thirty-two words can be seen under Figure 3.

Due to time constraints, we decided to investigate the first 300 of these keywords: one author focussed on nouns and adjectives, the other on verbs and adverbs. Our glossary would, therefore, consist of the 300 most probable words (and the multiword combinations they are part of) to occur in recipes—rather than an ad hoc collection of terms based on pre-conceived assumptions or abstract conceptual schemata. A more comprehensive glossary, though, should certainly have a much higher threshold (which has still to be defined in the literature), and should also accommodate the same process of making wordlists and keywords for multiword units, such as bigrams, trigrams, quadrigrams, etc., because there is no guarantee that any key combination of two or more words will contain at least one that is key in a keyword list of unigrams. Take, for instance, *enquanto isso*: neither *enquanto* nor *isso* are in the unigram keyword list in the Portuguese corpus of recipes when it is compared to Lácio-Ref,\(^7\) but the

---

\(^6\) See: http://americannationalcorpus.org/. The ANC project aims to create a massive corpus of American English, ‘including texts of all genres and transcripts of spoken data produced from 1990 onward.’ The version used for our glossary was the second release which contained about 18.5 million words of written American English.

\(^7\) See: www.nilc.icmc.usp.br/lacioweb. This corpus contains almost 10 million written words from various domains, such as Biology, Hard Sciences, Humanities, Social Sciences, Religion, etc. It was used as the reference corpus for our Portuguese keywords.
bigram *enquanto isso* occupies the 658th position in the bigram keyword list (it occurs 352 times in the corpus), in spite of the fact that it is also a rather common expression in the language overall (there are sixty occurrences in the Lácio-Ref corpus).

The next step was to extract concordance lines for each of these words and then analyse the co-text—words to the left and right of the target—with which they co-occurred in order to identify the multiword units they were part of. For this purpose, we used Wordsmith’s Concord function, which allows the lines containing the search word in the centre to be sorted according to the words to its left and/or right. This procedure makes it quite easy to visualise recurrent patterns. Figure 4 illustrates some of the collocations found for the search word ‘oil*’, such as *oiled [SURFACE]*, *basil oil*, *chilli oil*, *corn oil*, *excess oil*, *drizzle of oil*, *olive oil*, *sesame oil*, *vegetable oil* and *drizzle with oil*. All recurrent collocations and phraseologies that contained one of the 300 keywords analysed were included in the glossary.

Once all English entries were selected (words, collocations and phraseologies) it was then necessary to establish their translations in Portuguese. The first step we took was to look for equivalents in the Portuguese KeyWord list, which was obtained by comparing the Portuguese WordList with a WordList of Lácio-Ref, which was used as our Portuguese reference corpus. This showed several candidates for translation, such as *minutes* (‘minutos’), *cream* (‘creme’) and *garlic* (‘alho’). But it also showed variants like *heat* (‘fogo’, ‘quente’), *stir* (‘coloque’, ‘junte’, ‘acrescente’) and so on (see Figure 5). In the case of *oil*, for instance, we identified *azeite*
Figure 4: Selected, sorted concordance lines for *oil*

Figure 5: Translation candidates in the English–Portuguese comparable corpus of recipes
Translator-oriented, corpus-driven technical glossaries

<table>
<thead>
<tr>
<th>Collocate of finely</th>
<th>No. of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>chopped</td>
<td>2,528</td>
</tr>
<tr>
<td>sliced</td>
<td>443</td>
</tr>
<tr>
<td>diced</td>
<td>379</td>
</tr>
<tr>
<td>grated</td>
<td>280</td>
</tr>
<tr>
<td>chop</td>
<td>167</td>
</tr>
<tr>
<td>shredded</td>
<td>85</td>
</tr>
</tbody>
</table>

Table 2: Collocates of finely in the English corpus

('olive oil’) as the twenty-third keyword and, further down, óleo ('oil’) and the adjective untado/a, which can be an equivalent for ‘oiled’ in contexts such as ‘Place skin-side up on an oiled baking tray and grill for 4–5 minutes’.

We then generated concordance lines in Portuguese to make sure the candidates for translation occurred in similar contexts and with a similar frequency in the corpus. When the frequencies of equivalents were clearly different, we used the methodology devised by Tagnin (2007) – we looked for the collocates of the English term and generated concordance lists for them. For instance, finely occurred over 3,000 times in the English corpus while its Portuguese cognate, finamente, occurred only 236 times. Since the corpora were built to be comparable, it was expected that most corresponding terms would have similar frequencies. So, we looked for the most frequent collocates of ‘finely’ in the English corpus, which are shown under Table 2.

We then generated concordances for their candidate translations in Portuguese (for example, picado for ‘chopped’), to determine their collocates. This, in the case of picado, led us to bem picado (‘well chopped’) or (bem) picadinho (‘very well chopped’) as the best Portuguese equivalents for ‘finely chop(ped)’. Thus, although the adverbs ‘finely’ and bem would probably never be regarded as equivalents in other contexts, they do seem to play the same role as qualifiers of the corresponding participles chopped and picado in cooking recipes.

The second most frequent collocate of finely in the English corpus – the verb slice – could be translated as ‘fatiar’, but this possibility did not occur many times in the Portuguese corpus. So we used a wild card and searched for fatia*, which gave us ‘cort* em fatias finas’ (‘cut into fine slices’) as a better equivalent for finely sliced. A further search for cortar (‘cut’) produced ‘cortar em rodelas finas’ (‘cut into fine round slices’), usually used with onions, carrots and other round-shaped ingredients.

Another collocate of finely in English is grate. Common collocates for this verb are ‘parmesan’, ‘cheese’ and ‘chocolate’. The corresponding verb in Portuguese, ralar, occurs with similar ingredients, but a noticeable difference is that while in English it is usual to say ‘finely grated parmesan cheese’ (fourteen occurrences in the corpus), in Portuguese parmesão (‘parmesan cheese’) does co-occur with ralado (‘grated’) but usually with no
This may point to a relevant cultural difference – parmesan cheese is ‘finely grated’ by default in Brazil, so there is no need, apparently, to specify that. However, when the cheese is to be grated differently, this fact is made clear: *ralado grosso* (‘coarsely grated’). In other words, for the Portuguese text to sound natural, *finely* should be omitted in the translation.
translator-oriented, corpus-driven technical glossaries

Figure 7: Published entry for CITRUS FRUIT in Teixeira and Tagnin (2008)

Shredded, another collocate of finely in English, co-occurs with cabbage while the Portuguese equivalent for this vegetable, repolho, co-occurs with picar in phraseologies such as ‘picado bem fino’, ‘picado bem fininho’ (‘finely chopped’). All this information appears in the final entry for finely in our published glossary (see Figure 6).

3.3 Translator-oriented entries

If we go back to the translator’s needs, as mentioned above, we will see that they are met in our glossary. With very few exceptions, all entries feature examples, all of which are extracted from the corpus, and their corresponding translations into Portuguese, as can be seen under Figure 5. Information on usage is also provided by listing common collocates and the type of recipe or the part of the recipe the words are more commonly associated with; for instance, we highlighted the fact that fold usually co-occurs with egg whites in the procedure part of recipes for cakes, soufflés, etc.
Culture-specific information is given in notes, as in the case of ‘buttermilk’, an ingredient that is not available in Brazil. In such instances, an explanation is given, as are suggestions on how to replace the ingredient in the recipes (specialists were consulted to provide a substitute for the ingredient). Several illustrations and charts throughout the glossary also add useful information, such as charts showing the differences between cuts of beef, pork and lamb meat, which were included in the corresponding entries. Conversion charts for the basic measuring units, and a colour-illustrated table of herbs and spices, appear at the end of the book.

Another feature of the glossary is the use of lemmas, or categories, as entries. For instance, after analysing the concordance lines for lemon, lime, orange and grapefruit, we realised that many collocates and multiword expressions were common to all or some of these words, which are all citrus fruits, such as ‘candied [CITRUS FRUIT] rind / peel’, ‘[CITRUS FRUIT] wedges’, ‘freshly squeezed [CITRUS FRUIT]’, etc. So, we created an entry, with capital letters to distinguish it from the other entries, for CITRUS FRUIT, listing all the phraseologies commonly associated with citrus fruits in the corpus (see Figure 6). This entry is cross-referenced in all entries of the actual citrus fruits (lemon, lime, orange, etc.).

Cross-references like the one mentioned above, as well as synonyms, language variants (British versus American) and possible pitfalls in translation—such as translating red onion as ‘cebola vermelha’ instead of ‘cebola roxa’ (‘purple onion’)—are all indicated by coloured icons which the translator can easily identify in the entry.

4. Conclusion

Our aim in this paper was to present the methodology used to compile a translator-oriented, bilingual glossary from a comparable corpus using a corpus-driven approach. We believe that such a glossary reflects the actual vocabulary of the area in a more reliable way, from the point of view of text production, than most glossaries that have been produced for the English–Brazilian Portuguese language pair using traditional methodologies in Terminology. As argued in the introduction, a reference source for this language professional should include context, collocates, as well as usage and culture-specific information so as to meet the needs of the translation task. The cooking glossary we have built complies with these requirements.

We hope other researchers may show an interest in testing this approach in other specialised areas, adjusting it to genre and text-specific features of other technical texts in other pairs of languages. This would help not only to improve the methodology but could also suggest ways to address other important issues in terminology compilation, such as writing definitions and creating conceptual representations using the corpus-driven approach.
References


