

HOW TO USE LINGUISTIC CORPORA TO IMPROVE YOUR TRANSLATIONS

Naomi James Sutcliffe de Moraes
Just Right Communications Ltda.

Abstract: The use of corpora is one of the best ways for translators to improve their translations on their own. A corpus is a collection of texts either with or without their corresponding translations. I will describe the types of data extracted from corpora, how you can create your own or use ready-made corpora, or take advantage of the corpora built automatically by translation memory tools like DeJaVu, Trados and SDLX while you work. Corpora are perfect for the translator studying a new source language, working in a new sub-area (law, medicine, etc.) or out of their mother tongue, but we all can learn from them.

1. WHAT IS A CORPUS?

A corpus is a collection of texts (not necessarily complete texts, but neither random words) in a machine-readable format that, together, make up a representative sample of a language or sublanguage. There are many sources for ready-made corpora. To see some corpora available in many languages, see David Lee's site (Ref. 4). The basic idea is that the corpus is "experimental" data and corpus linguists are scientists analyzing the data collected. One of the main purposes of corpora is to find collocations for a given word. A collocation is "the habitual juxtaposition of a particular word with another word or words with a frequency greater than chance" (Concise Oxford Dictionary, 19th Ed.).

Another way to think of corpora and corpus tools is to think of all the pages on the internet as a corpus (a rather untidy one) and of Google as a corpus tool. By searching the Internet with a keyword or phrase you are effectively searching this massive corpus for the query text. Unfortunately, especially in specialized areas, a Google search gives us far more information than we can use.

2. TYPES OF CORPORA

Corpora can be categorized in many ways, but here I will only present the types I feel are most useful for translators. **General corpora** contain texts that do not belong to a single text type, subject field or register. **Specific corpora** can be as specific as the creator desires, containing all the texts of a given author, all the texts published in a given magazine or newspaper, or a collection of contracts from many sources. These are also referred to as sublanguage corpora.

In addition to the specification above, corpora can contain texts in only one language (**Monolingual**) or in two or more languages (**Bilingual** or **Multilingual**). Obviously, a bilingual or multilingual corpus would be of interest to the translator! But wait, there is an important subcategory: when a corpus contains originals and translations, it is called a **Parallel Corpus**, and when it contains similar original (non-translated) texts in two different languages, it is called a **Comparable Corpus**.

3. HOW TO USE CORPORA TO IMPROVE YOUR TRANSLATIONS

In the following sections I will explain how translators can use the various types of corpora described above in their work. As you read through the examples, try to think of similar things you have come across in your work. The most important uses of corpora are to discover collocations in the target language, whether the target language is your native language or not, and to find terminology in the target language either based on collocations or on a stored translation in a parallel corpus. They can also be used to understand unfamiliar expressions and collocations in the source language/dialect.

3.1 General Monolingual Corpora

General monolingual corpora are available on the Internet for many languages or can be purchased on CD from some sources (usually academic). Examples are the BNC (British National Corpus, Ref. 1) and the Cobuild Corpora (Ref. 2). For access to general corpora for different languages, see David Lee's Corpora site (Ref. 4) or search on the Internet for the language and the word corpus. Two websites allow corpus-type searches of documents on the Internet: WebCONC (Ref. 10) for European languages and WebCorp (Ref. 11) for English. Most countries have at least one monolingual corpus. The Hellenic National Corpus (HNC, Ref. 8), accessible online, allows you to search on two or three terms and indicate the number of words separating each term.

Figured to example

A colleague translating out of English contacted me recently about the expression "figured to". It was in an article about a sports team, and said, basically, that the team was figured to win the game. He had never seen the expression and it was not in his English dictionaries. I knew what it meant, perhaps because I am American, but immediately thought of ways he could have discovered the meaning on his own. A search on Google was not very helpful:

Google search on "figured to"

1. 'Full-figured to fit' from Muscle & Fitness...
2. I went from full figured to fat, from fat to obese, ...
3. Adult Full Figured to 48 in the Everything Else

4. A little kid, about 4 feet tall, glided up to me on his snowboard -- I figured, to check if I was OK.
5. No individual shall in any case be entitled to more than four-fifths (4/5) of his benefit rate for that week, figured to the highest dollar.
6. Social Security and Medicaid Taxes: This is figured to be 15.3% of monthly labor expenses. Because the worker will be self-employed for informal and family ...
7. etc.

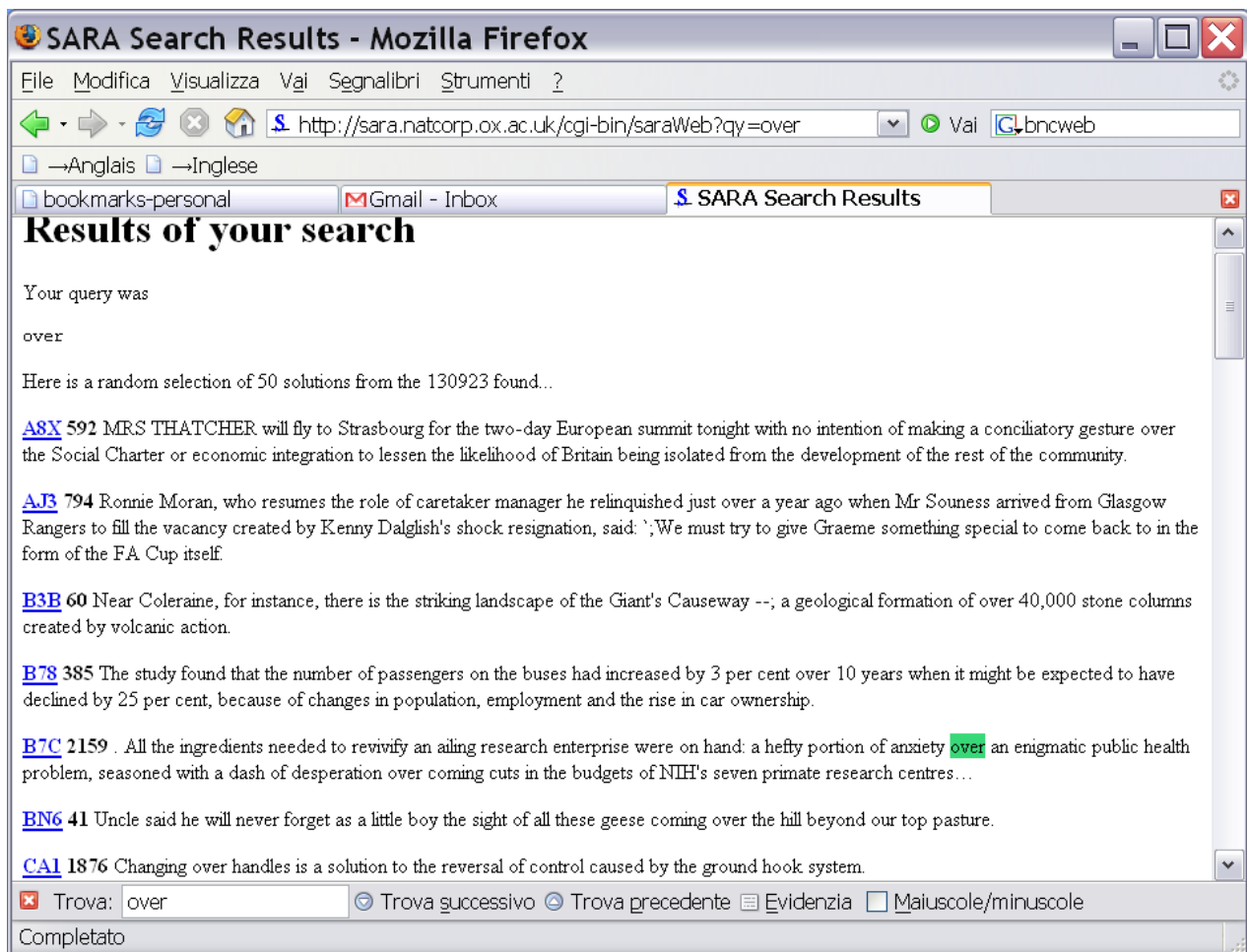
None of these hits is slightly useful. I then turned to WebCONC, which somehow seems to select web pages with the linguist in mind, and immediately found examples in context that made the meaning (were figured to = were expected to) very clear:

1. "We believe that if we stay healthy, and win a game we're not figured to win, we could possibly go 5-5."
2. Pennsylvania' s unbeaten football team figured to beat Harvard on Saturday, and both offenses figured to score often.
3. "Thanks. That there's just what we figured to do."

The fact that the first two examples are also sports-related also helps to confirm the definition of the term.

Over example

Another colleague recently asked me for advice on how to use "over" in the sense "with respect to". I searched for over in the BNC (British National Corpus), accessible on the Internet. It provided only 50 random context sentences, but this is probably more than a translator needs or has time to look at. I expected to find few uses of over in this sense, but was surprised to see that the first and fifth hit were good examples. The results do not highlight the search term in each sentence, but you can use the browser's search tool to find each occurrence.



Collocation information is particularly useful when translating out of your native language, as a way to study natural collocations in the target language. In Brazil, there is far more demand for translations into English than American and British translators, so the vast majority of translations into English are done by Brazilians. This is probably the case in many countries.

Of course, after living abroad for many years a corpus based on recent texts can help translators living outside their native country keep in touch with recent terminology and changes in grammar. This type of corpus is also invaluable for language learners, and many articles can be found on the web regarding how to use them in the classroom or for preparing foreign-language teaching material.

3.2 Specific Monolingual Corpora

Specific monolingual corpora are also available on the Internet, or you can create them yourself. If you translate many newspaper articles into language X, you can probably access a specific corpus with newspaper articles for that language or download articles yourself and create your own personal corpus. Wordsmith (Ref. 14) can be used for this. Wordsmith 4 even has a built-in tool to help you download texts off the Internet, or you

can use a website copying tool like WinHTTrack (Ref. 13). Texts must be converted into simple text format to be read by Wordsmith. The type of information extracted from specific corpora is the same as for general corpora, but the results are more specific to the area covered by the corpus. For a specific corpus containing newspaper articles, words like corruption, politics, and party would be more common than in a general corpus.

Why should you go through all this work to create your own corpora? I translate medical texts from Portuguese into English. Mostly journal articles. Copying a few journal articles on the same topic (often found in the references of the article being translated) into text files and searching them can facilitate translation. When I first started out in translation, I looked for articles like these and read them, underlining terminology that might appear in my source text in the other language. Now, instead of reading them, I can search for certain words and read only the sentences in which they appear. Corpora are also great for the "guess and check" method. If you are sure of one word in a 2-3 word expression, you can search on the one sure word to see with which words it collocates in the specific sublanguage of your text/corpus. For example, when the original article mentions some medical test performed, and I cannot find a translation through other means, searching for the words "test" and "tests" in a specific corpus containing only the articles in the references and looking at the context can at least give me a place to start. I can look at all the tests commonly used in this sub-area, look them up on the Internet, and see if one matches the description of the test in the original text. I could do this on the Internet, and spend all day looking. I could also read all the reference articles, but that would take a long time too. A specific corpus is also useful for the spelling of words that could end in -ic or in -ical (like electric and electrical). Note that we write electric train, electric switch, but electrical wiring and electrical installations. Medical terminology is similarly confusing.

Even if you always work into your own language/dialect, you may not always have the luxury of translating into an area you studied at school. Technical dictionaries often provide more than one translation for a given word. When translating a text on concrete, the subject was "fissuração" in Portuguese. My technical dictionary provided both fissuring and cracking as definitions in English. I created a specific corpus of civil engineering texts in just one file, specifically on concrete and structures. It only took a couple minutes to search the Internet and cut and paste some texts into a text file. It helped determine that cracking was the correct translation in my context and even brought shrinkage to my attention, which I used later.

Concordance					
N	Set	Tag	Word No.	File	%
1	83180	158E	CEB Design Manual	Cracking and Deformations (198	12.955 _en~1.bt 65
2		m	Sections * Post-serviceability	Cracking Stage and Limit State	14.848 _en~1.bt 73
3		rced	beam	subjected to shear. A crack-based analysis is propose	13.939 _en~1.bt 69
4		flexion	Behavior	of Beams * Pre-cracking Stage: Region I * Alter	14.817 _en~1.bt 73
5		calibration,	surface flaws, fatigue	crack growth and fatigue life of s	9.263 _en~1.bt 47
6		ctability.	Interprets pre- and post-cracking	elastic behavior. Organi	1.260 _en~1.bt 7
7		cking	Moment Evaluation * Post-cracking	Service Load Stage: Re	14.830 _en~1.bt 73
8		esigner	to control shrinkage and	cracking in a concrete slab or w	16.710 _en~1.bt 82
9		I * Effective	Moment of Inertia of	Cracked Beam Sections * Post-	14.842 _en~1.bt 73
10		ntal	techniques used in fracture,	crack tip stress fields, strain ene	9.248 _en~1.bt 47
11		spaced too far	apart or a zigzag	crack in the concrete may devel	19.630 _en~1.bt 93
12		Region I * Alternative	Methods of	Cracking Moment Evaluation * Po	14.825 _en~1.bt 73
13		in	elasticity; stress singularity at	cracks and corners; stresses and	10.198 _en~1.bt 52
14		Concrete is	weak in tension and	cracks easily when it shrinks or	15.678 _en~1.bt 77
15		tropic	hardening. After the initial	cracking of the concrete the she	13.772 _en~1.bt 69
16		ther.	It provides the resistance to	cracking, shrinkage, temperature	15.891 _en~1.bt 78
17		uate lap	length can cause severe	cracking in the concrete around	19.205 _en~1.bt 91
18		lancing,	partial prestressing and	cracking moment; design for she	8.350 _en~1.bt 42
19		orted	One-Way Slab * Tolerable	Crack Widths Topic 8 (Chapter	14.929 _en~1.bt 74
20		retaining	wall functions more for	crack and shrinkage control. Its	16.012 _en~1.bt 79

Michael Wilkinson has written a wonderful article entitled "Using a Specialized Corpus to Improve Translation Quality" (Ref. 12). His Finnish translation students often translate tourism texts into English, and he is teaching them to use specific corpora containing English-language tourist brochures to create more natural-sounding texts.

3.3 Parallel Corpora

Parallel corpora are available online for many languages, but they often include only fiction (usually books and translations already in the public domain) or newspaper/magazine articles due to copyright restrictions. Sometimes you must download the files to your computer and align them yourself. The COMPARA corpus (Ref. 3) provides English/Portuguese parallel corpora which can be searched online free. A search on the word *prazo* returned 32 hits. The first three are:

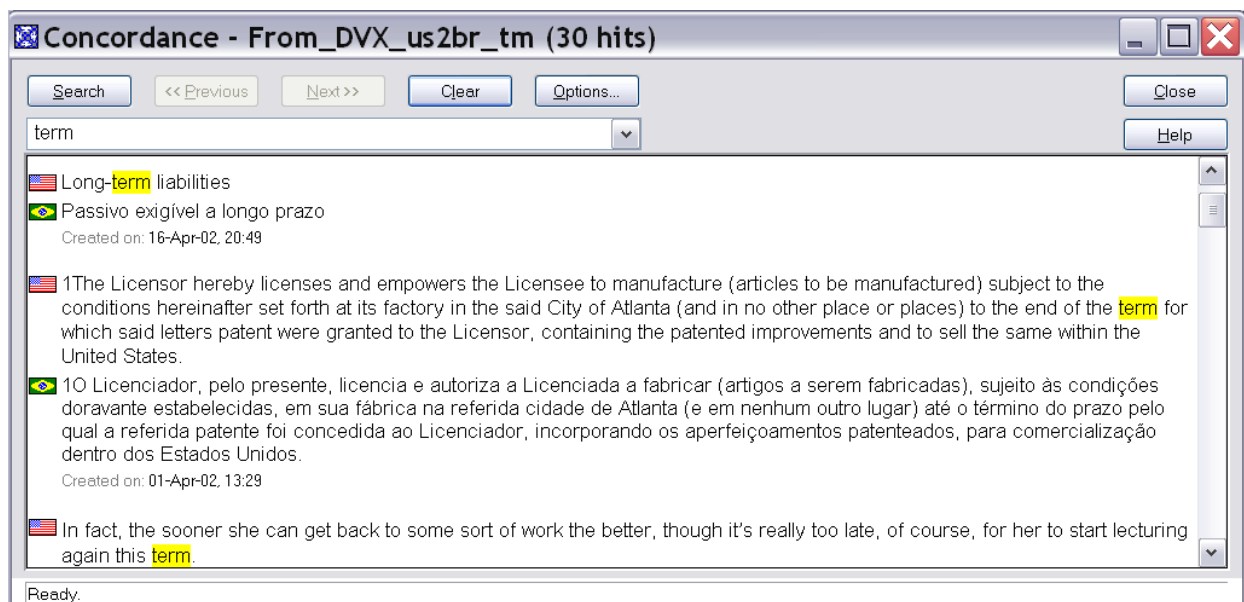
Esta coisa da acupunctura tem definitivamente que se lhe diga, embora eu não saiba se faz bem a longo prazo .	There's definitely something to this acupuncture business, though whether it does you any lasting good, I don't know.
Vendo as coisas a longo prazo , até a grande crise económica dos anos 30 foi temporária.	Even the Slump of the Thirties was «temporary», in the long run.

— Porque o produto será barato, de confiança e disponível a curto **prazo** — responde Vic.

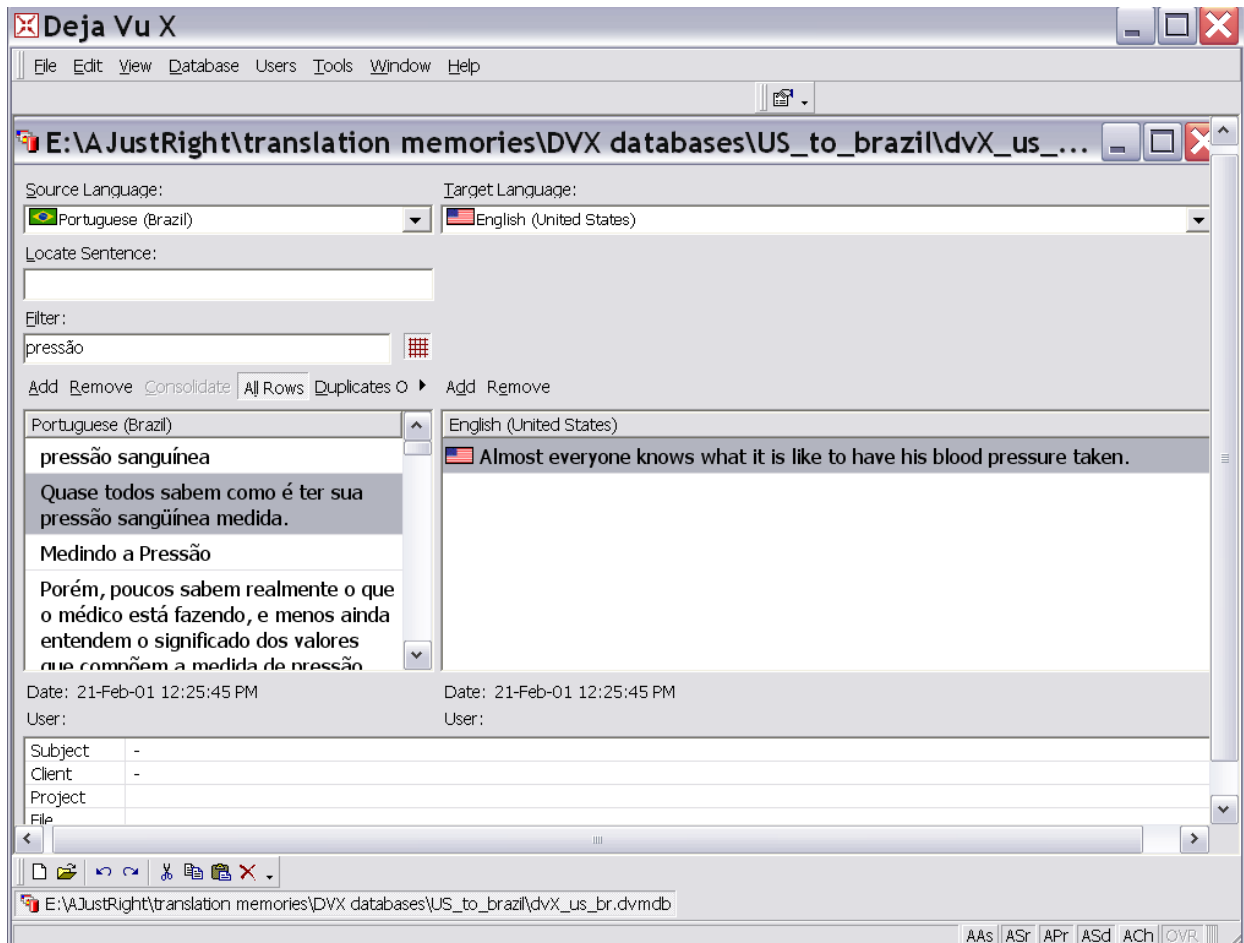
‘Because the product will be cheap, reliable and available at short notice,’ says Vic.

Note the interesting translation of *prazo* as **notice** in the third example, as **run** in the second and its total elimination in the first example. This corpus contains only literary translations.

Translation memories are a type of parallel corpora created as you work, with the only selection criteria being what you yourself have translated. Wouldn't it be great to have access to the translation memories of experienced translators in addition to your own translation memories? Many good translations are available on the internet and can be downloaded and added to your personal corpus. By aligning them, you are creating a valuable translation memory (a.k.a. parallel corpus) based on these translations. Needless to say, you must be careful what you add to your corpus and not rely on it blindly. You can also feed these translations into your translation memory using the environment's alignment tool (both DeJaVuX and SDL Trados come with an alignment tool, and other tools probably do too). These tools do basic pre-alignment based on paragraphs and periods which you can then correct, if necessary. Of course, you might want to create a separate TM to hold these texts translated by others. All translation memory tools have a menu option or button that allows you to look at previous occurrences of a source word in the parallel texts translated earlier—**Scan** in DeJaVu and **Concordance** in SDL Trados. However, not all users of these programs know that you can also search your translation memories when you are not using the tool to translate a text (for example, when the original is not in electronic format). If you use SDL Trados, open your TM in Translator's Workbench. Choose Concordance from the Tools menu. In the Concordance window you can search on any word



In DejaVu, open a TM and chose the source language and target language. Type the word or words in the filter and click on the brown "fence" symbol.



DejaVu's interface is more powerful (you can use SQL expressions if you know how), but not as easy to glance through to find what you are looking for in the target language—only the translation for the cell selected on the left is shown in the very large window on the right.

Translators working with Canadian French can create an interesting legal corpus by downloading laws published in both French and English from the Canadian Department of Justice's web site (<http://laws.justice.gc.ca>). They make it easy... search for what you want in English, then click on the menu option for French and the exact same legislation is shown in French. They even provide definitions and translations of terms in the margins! The French government also has an official English translation of its Civil Code (Ref. 7). The European Union also has many publications, including laws (Ref. 6) and patents, translated into many languages. All trade communities (like NAFTA) have some documents published in all applicable languages. Some universities have already separated out this information and compiled it. All you would need to do is download the text files and align them. The

Europarl Parallel Corpus (Ref. 5) is an example: download text files in most European languages, labeled and ready for alignment.

As an example of the use of a specific parallel corpus, I have searched on the word "prazo" in Portuguese documents translated into English (from my personal legal corpus):

Portuguese original	English translation
...durante o [prazo] ajustado para esta locação...	...during the agreed upon <u>Term</u> of the Agreement...
...dentro do [prazo] contratado...	...by the agreed upon <u>deadline</u> ...
durante o [prazo] contratual	during the contractual <u>period</u>
Início e [prazo] da locação	Commencement and <u>term</u> of lease...
pelo [prazo] de 6 (seis) meses	for a 6 (six) month <u>period</u> ...
no [prazo] de até 10 (dez) dias contados do recebimento...	within ten (10) working days from the receipt...
para que cumpra no [prazo] de um mês	...so the latter can comply...within one month...
outorgado por [prazo] indeterminado.	granted for an indeterminate <u>period of time</u> .

Prazo is a very complicated word to translate, since the sentence structure must often be changed and the meaning can vary slightly. Here we see prazo translated as "term", "deadline", "period", "period of time", and effectively eliminated by changing the structure of the sentence to use "within". In the second example the word "date" could also have been used. Note the difference between these results from a legal corpus and those from the general corpus above.

3.4 Comparable Corpora

The objective of a comparable corpus is to compare texts without the distortions created by translation. If you are translating into language X, you collect documents similar to the document you are translating but originally written in language X. For example, if you need to translate a text on soil mechanics into English, you could search for various sites on the subject, copy the texts of a few good web pages to text files, and immediately use them as a corpus for your job (instead of searching for something in Google, search for it in your new specific corpus). In the "prazo" example above, a search in an English corpus for "term", "deadline", "period", "period of time", "within" and "date" might indicate which structures are more natural sounding in English and are commonly used by native speakers.

Another use of comparable/monolingual corpora is to discover connotations of words and their translations. This is important when translating into a foreign language or into a dialect other than your own. For example, the word "smell" in English usually carries a negative connotation unless some

positive adjective modifies the word (good smell, nice smell, pleasant smell) or unless the speaker's tone of voice or expression indicates a positive feeling ("what's that smell" as speaker enters the kitchen, smiling). Note that "scent" is almost always positive unless related to animals. In contrast, in Portuguese, the word "cheiro" is positive by default, and requires a negative modifier to become negative ("cheiro ruim"). This is exactly the kind of information that can be extracted from a comparable corpus, and will be the basis for future bilingual dictionaries. It helps to choose between synonyms.

4. HOW TO BUILD YOUR OWN CORPORA

If you use Wordsmith to create a monolingual (or comparable) corpus, the files will need to be converted into text format. Most programs have a Save As... text option. For a parallel corpus, just align the texts within your translation tool.

4.1 File Formats and Organization for the Corpus Program Wordsmith

The corpus program Wordsmith needs corpus files in simple text (*.txt) format. If you have the text in word, just save it as a simple text file. If the file is in a searchable pdf file, click on the I icon, select the text and copy it into a word file. Normally, when a pdf file is pasted into Word this way a hard carriage return is inserted at the end of each line (a ¶). The good thing is that these are easy to remove. In the Word edit menu, chose replace. Type ^p in the "find what" field and a space in the "replace with" field. Normally, I go through the file replacing the hard returns one by one so that the paragraphs do not all run into each other. There are probably programs that convert pdf files into Word, but I have never needed one—I do not need to do this that often. If the text is in HTML, it is easier to copy and paste the text to Word or Notepad than to save the file and then have to remove all the embedded HTML codes.

Looking at text files in the notepad application is a pain, since the lines do not word wrap and each paragraph goes on and on off the page on a single line. I have found it worth my hard disk space to keep a copy of the original file in word or pdf together with the text format so I can see the context better, especially when the original file had complicated formatting (e.g. a financial statement with tables).

Wordsmith is easier to manage with short file names. Save yourself a headache and give text files a short code name. I have used the following system for both monolingual and multilingual corpus files:

Axxx_yy_z, where

- A is the type of document: C for contracts, F for financial statements, P for powers-of-attorney, etc.
- xxx is the number of the document of that type. So I could have a C001_br_o and a P001_br_o
- yy is the country code. I carefully note which country produced the original or the translation, since sometimes I will only want to look for collocations/vocabulary specific to one country. For example, Brazilian and Portuguese legal language is quite similar, but their engineering terms are not.
- z is "s" for a source text, "t" for a target text and "o" for when I only have that document (no translation available).

Using some kind of numbering system (and a spreadsheet to remember what each file contains) will help you, especially when creating corpora on the fly in Wordsmith, where you must select the corpus files every time you run the program. HTML headers are standard for corpus files in corpus linguistics, and the corpus tools can be configured to ignore the headers when processing text.

4.2 File Formats and Organization for Corpora Stored as Translation Memories

If you are creating a parallel corpus with a TM tool, you can leave the files in a compatible format (word, html, etc.). Read the tool manual to learn how to align the files and your parallel corpus will be ready to use.

5 HOW TO USE YOUR CORPORA DURING A JOB

So, in summary, there are various ways to use corpora when translating. I will present them again in order of time invested:

1. Access a monolingual general corpora in the target or source language available on the internet. You can find these and store them in your bookmarks for easy access.
2. Access a bilingual or multilingual corpus with your working languages available on the internet, if one exists. This may only work well if the subject area of the corpus matches the subject area of your job.
3. Create specific monolingual corpora on the fly as you work, with the texts as similar as possible to the job text in subject area and register.
4. Create specific bilingual corpora when you are not working on a job (or when a job is long-term or recurrent) with reliable translations provided by the client, a colleague or the Internet. Be very careful what you use, since a bad reference is worse than no reference at all.

6. REFERENCES

1. BNC Online: <http://sara.natcorp.ox.ac.uk/lookup.html>
2. Cobuild Corpora: <http://www.collins.co.uk/Corpus/CorpusSearch.aspx>

3. COMPARA Corpus Port/Eng:
<http://www.linguateca.pt/COMPARA/BuscaSimples.html>
4. David Lee's Corpora Site: <http://devoted.to/corpora> (most links are broken, but with the names of the corpora and their descriptions you can find them by searching the Internet).
5. Europarl Parallel Corpus:
<http://people.csail.mit.edu/koehn/publications/europarl/>
6. European Union laws: <http://europa.eu.int/eur-lex/lex/en/index.htm>
7. French Civil Code in English
http://www.legifrance.gouv.fr/html/codes_traduits/code_civil_somA.htm.
8. Hellenic National Corpus: <http://hnc.ilsp.gr/en/find.asp>
9. Sardinha, Tony Berber. *Lingüística de Corpus*. Barueri, SP: Manole, 2004.
10. WebCONC: <http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi>
11. WebCorp: <http://www.webcorp.org.uk/cgi-bin/webcorp2.nm>
12. Wilkinson, Michael. "Using a Specialized Corpus to Improve Translation Quality", In *Translation Journal*.
(<http://accurapid.com/journal/33corpus.htm>)
13. WinHTTrack: www.httrack.com
14. Wordsmith: <http://www.lexically.net/wordsmith/>