>>> revista tradumàtica
Traducció i Tecnologies de la Informació i la Comunicació
Número 07: L'aplicació dels corpus lingüístics a la traducció : desembre 2009 :
http://www.fti.uab.cat/tradumatica/revista : ISSN: 1578-7559

Google and Beyond: Web-As-Corpus Methodologies for Translators

Adriano Ferraresi, University of Naples "Federico II"

1. Introduction

In empirical linguistics, corpora, i.e. collections of texts assembled so as to represent a given language or language variety, have become the tools of the trade. Building a corpus, however, requires considerable time and effort. Texts may not be easily available for the language (variety) under scrutiny, or, if a corpus exists, it may prove inadequate because, e.g., it is too small, or too outdated. In this context, the exponential growth of the web was welcomed by (corpus) linguists, who variously defined it as "a fabulous linguists' playground" (Kilgarriff and Grefenstette 2003: 333), and "a potentially unlimited source of linguistic data" (Baroni and Ueyama 2006: 1), presenting "a serendipitous opportunity [of accessing] free, plentiful, up-dated and up-to-date" texts (Renouf et al. 2007: 48). So many seemed the potential advantages of turning to web data that a new strand of research developed, aimed at exploring the potential of using the web as a benchmark for linguistic research and to build new corpus resources rapidly and efficiently. This approach has come to be known as the "web-as-corpus" (WaC) approach.

Translators are confronted with needs somewhat similar to those of linguists when gathering and consulting reference materials for a translation task. They need resources affording a favourable trade-off between costs (in terms of time and effort required to build and use them), and advantages they offer (in terms of quality/reliability and effectiveness). In this context, too, the web has become a widely used resource. According to a survey carried out in 2007 among translation students and professionals (within the EU-funded MeLLANGE project), nearly 95% of the interviewees make use of the web for translation-related tasks, e.g. for terminological work.[1] Only less than half of them (mainly students probably) actually compile corpora.

In this paper an overview of current "web-as-corpus" approaches will be presented and their relevance for translation purposes will be highlighted. The aim is to introduce translators to some of the tools and resources out there, in the hope of contributing to raising their awareness of the advantages (as well as inevitable pitfalls) these offer for their work.

2. Google(-based) tools…

The first way in which the web can be used for linguistic purposes is by considering the whole of it as if it were a gigantic corpus. When translators use Google to find term definitions, or to search for phrases in the target language, they are using the web in this way: they access web texts as if they were a (virtual) corpus, and use search engines as their "concordancers", i.e. the software tools one usually employs to search corpora and retrieve examples of words and phrases in context.

According to the MeLLANGE survey, this is one of the most common uses that translators make of the web. This is unsurprising: using search engines relieves translators from the task of learning to use new software programs, such as concordancers, and, depending on the task, may provide enough information and thus make the download of reference materials superfluous. By consulting the whole of the web through search engines, they have the largest possible data set at their disposal, and query strategies relying on search engines' functionalities (e.g. the "search phrase" function, or the restriction of results to a certain site through the site: operator) may be applied on a one-by-one basis, according to the specific translation problem to be tackled. For an example of this kind of approach, see Gatto (2009: 45 ff).

Commercial search engines, however, are designed to find contents, not linguistic forms. Hence, the number of pitfalls that web-as-corpus research has highlighted with respect to their use for linguistic purposes (e.g. Kilgarriff 2007, Thewall 2005). First, the ranking and sorting of results are performed according to criteria like "popularity" of the websites, or geographic relevance, and a translator may have to navigate through a potentially vast set of irrelevant hits before finding relevant examples. Secondly, counts for the pages returned by a query sometimes prove to be inexact and unstable: it has been shown that the same search can yield radically different numbers of hits, depending on such unpredictable and uncontrollable factors as the time of the day, or the location from which the query is made (Kilgarriff 2007: 148). Finally, search engines present very limited options for searching and (re-)sorting results compared to even the most basic concordancer. The query language is linguistically unsophisticated, making it impossible, e.g., to perform substring matching (as Lüdeling et al. 2007 point out: how would one search for words ending with the suffix "*-itis*" using Google?), or to specify the span occurring between two search terms (the kleene star "*", used in conjunction with the search phrase function, may variably match one to an unpredictable number of words).

Within web-as-corpus research, tools have been developed that overcome some of these limitations. Instruments like KWiCFinder,[2] or WebCorp[3] rely on search engines to retrieve results, and then post-process output so as to present it in a "linguist-friendly" format, i.e. usually the KWiC, "Key-Words-in-Context", format, which will be familiar to all those who have worked with corpora and concordancers (see Figure 1 and 2 for an example). They also provide convenient functionalities such as alphabetical re-sorting of concordance lines (e.g. ordering the output according to the words preceding or following the search item, so as to make patterns stand out more clearly) and calculation of word co-occurrence statistics, e.g. for identifying key phrases associated with a given search-word (for further uses of WebCorp for documentation purposes, see Sánchez-Gijón 2003).

While surely representing a potentially interesting attempt at making the web more similar to a "proper", traditional corpus, this is probably not the most promising way of exploiting the potential of the web. Since they rely solely on search engines, the tools mentioned suffer from the same problems in terms of data selection as the search engines themselves, i.e. they allow for very little control on the data that they display. These data will be those that Google, Yahoo!, etc. pass on to the post-processing tools, with limited possibility of enlarging and/or modifying the pool of webpages that the results are extracted from. Thus, for example, the most "popular" – or best advertised – pages, such as those from the websites of big commercial companies, are likely to saturate the output, effectively hiding pages that may be less relevant from a web-user perspective, yet more informative for a translator. In the next section a more promising approach to the use of the web for linguistic and translation purposes is described.

3. … and beyond.

A second approach to the exploitation of the web as a corpus consists in using it not as a corpus *per se*, but as a source of textual data that are downloaded and post-processed according to one's needs. Accordingly, the focus is not on finding efficient ways of drawing directly upon search engines, but rather on developing tools that (automatically) select potentially relevant webpages and subsequently include them in a "standalone" corpus.

This is perhaps a less well-known approach among translation students and professionals. Recall that according to the MeLLANGE 2007 survey, only 40% of the interviewees declared that they use corpora. Moreover, most of them (65%) claimed they do not use concordancers to consult them, but rather search facilities within standard text processors (like Microsoft Word), witnessing to a general underuse of corpus methodologies in translation practice (see also Bernardini 2006). In this sense, web-as-corpus

methodologies would seem to have a lot to offer translators, by providing either (free) ready-made corpora, or tools that allow them to efficiently build corpora suited for their needs.

Some of the advantages of using web-derived corpora are the same as those associated with the use of search engines and search engines' post-processors. By turning to web data, translators have at their disposal an enormous amount of textual material, which is constantly updated and up-to-date, and which is available for virtually every language and specialised topic or domain. On the other hand, turning web data into a corpus makes it possible to overcome the limitations that search engines impose, and to fully exploit the methods of linguistic enquiry associated with the use of a traditional corpus (cf. section 4). Compared to search engines, corpora and concordancers allow fuller control over the data that are consulted: webpages to be included in a corpus can be pre-selected (or expunged if they turn out to be undesirable), thus producing more reliable results, and releasing corpus users from having to browse through irrelevant pages at every search. Moreover, they can be browsed offline, and consulted through the concordancer one is most familiar with (several user-friendly tools are freely available on the web, see section 4).

A wide range of corpora and corpus-building tools have been created within this general paradigm, from multi-lingual comparable corpora to algorithms for automatically building parallel corpora (for further examples see Hundt et al. 2007). Due to space limitations, here only two such resources will be discussed, selected among those which seem to be of greatest relevance to translators. In the next section an example will be provided in which these resources are applied to solve a real translation problem. The example should highlight the kinds of insights one can gain from using different web-derived corpora for specialised translation.

The first resource is a (web-derived) reference corpus, i.e. a collection of texts gathered so as to include the greatest possible variety of registers and domains, meant to represent general language. Sharoff (2006) created and made freely available large (~ 100 million words) reference corpora, called the Leeds Internet Corpora.[4] These are available for a number of languages (English, Spanish, Italian, to name but a few), and can be consulted through an on-line interface.[5] As we will illustrate in the next section, such corpora can be profitably applied to tackling "general language" translation problems, such as recognising the connotations and typical contexts of use of (non-specialised) phrases. The second resource is WebBootCaT, a web interface to a set of Perl scripts (Baroni and Bernardini 2004) that, taking as input a few key terms, draw upon web data to automatically build a specialised corpus for the domain of interest. This corpus will be applied to the identification of plausible Italian equivalents for a technical term in an English source text.

4. An example: applying web-derived corpora to a translation task

Imagine having to translate the following paragraph:

> *Unlike **a** true **global**, class attributes should not be accessed directly. Instead, their state should be inspected, and perhaps altered, only through the mediated access of class methods. These class attributes accessor methods are similar **in spirit** and function to accessors used to manipulate the state of instance attributes on an object.* (From: perltooc. My emphasis)[6]

Notice that this text was actually used during a technical translation course (English => Italian) at the School for Translators and Interpreters of the University of Bologna, Italy, in the academic year 2008/2009. The course focused on the translation of reference materials on the Perl programming language, a domain the students were utterly unfamiliar with, and that proved especially challenging for them. Here, we will concentrate on two examples of translation problems they were faced with. These problems differ in nature: one relates to the use of a phrase (i.e. "*in spirit*") belonging to general English, while the other ("*a […] global*") involves knowing the appropriate technical term. In particular, we will suggest possible ways of

drawing upon web-derived corpora to search for plausible equivalents in the target language, in our case Italian.

As for the translation of "*in spirit*", a problem arises in Italian concerning the preposition to use in conjunction with the noun "*spirito*" to render the sense of the English phrase. Several options are available, but the most obvious ones are "*in*" and "*nello*" (the latter being a combination of the preposition "*in*" and determiner "*lo*" ("*the*")); both "*in spirito*" and "*nello spirito*" are intuitively plausible target language equivalents of "*in spirit*". But how can a translator assess whether these are "good" target language equivalents, and that, e.g., they are not calques? As was mentioned above, reference corpora may be a valuable source of evidence to check intuitions about the meaning and typical contexts of use of such "general-language" phrases. The Leeds Internet Corpora, in particular, have the advantage of being freely accessible, unlike other reference corpora such as the well-known *British National Corpus;*[7] moreover, they come with a user-friendly web-based interface featuring basic and advanced search options, which is suited for expert and non-expert users alike.

A search for "*in spirito*" and "*nello spirito*" in the Italian corpus from the Leeds Internet collection reveals that the latter phrase is much more frequent (711 examples vs. 192, a proportion of nearly 3:1). Moving on to the analysis of concordance lines for "*in spirito*", some regularities emerge (Figure 1):

essere santa sia in corpo che **in spirito** . Ma colei che é sposat
nei confronti di Dio. I poveri **in spirito** attendono ogni aiuto da
uomo. Con il lavoro realizzato **in spirito** di amore, glorificò Dio
ina fra le altre, in libertà e **in spirito** di collaborazione, [l'
zione. Uniti nella preghiera e **in spirito** di collaborazione, racc
i da conseguire uniti insieme, **in spirito** di collaborazione e di
uindi lo aiuta, « in libertà e **in spirito** di collaborazione », a
e altri ad accostare la Bibbia **in spirito** di fede, di preghiera e

Figure 1. Selected concordance lines for "*in spirito*" in the Italian Leeds Internet corpus

Most occurrences of "in *spirito*" seem to be found in texts variously dealing with religious issues – signalled by words like "santa" ("*saint*"), "*Dio*" ("*God*"), and "*fede*" ("*faith*") occurring in the immediate co-text –, suggesting that the contexts of use of this phrase in Italian are not the same as those of the English phrase. The only other pattern that stands out is "*in spirito di (e.g. collaborazione)*" ("*in spirit of collaboration*"). No evidence therefore emerges that "*in spirito*" can be used in the same contexts and with the same meaning as "*in spirit*" is used in English. On the contrary, a search for "*nello spirito*", with results sorted first by the left co-text, and then by the right co-text, returns several examples that are coherent with the sense of the English phrase (see Figure 2):

CONTRADDICE, nella lettera e nello spirito , il passo citato nel
conoscendosi nella lettera e nello spirito dei suoi fini istituzi
ligiosi, ma nella sostanza e nello spirito in tutto uguali ad ess
te. La decisione s' inquadra nello spirito e nella lettera del co
idee e posizioni diverse ma nello spirito e nella prassi della n
ie di presidi e di servizi (nello spirito e nella sostanza di ci
e d' oggi " non corrisponde, nello spirito e nella realizzazione,

Figure 2. Selected concordance lines for "*nello spirito*" in the Italian Leeds Internet corpus

Interestingly, these results reveal extended phraseological sequences, such as "*nello spirito e nella prassi*" ("*in spirit and in practice*"), and "*nello spirito e nella sostanza*" ("*in spirit and in substance*"), that could be

considered as potential equivalents for the source text extended phrase "*in spirit and in function*". Finally notice that, given the popularity of religious sites on the web, all the first hits we get with Google for both "*in spirito*" and "*nello spirito*" refer to the Holy Spirit, and are thus irrelevant to our purposes – and, in the case of "*nello spirito*", are even potentially misleading.

A reference corpus of the general language typically does not help when one is faced with the second problem mentioned above, i.e. finding an Italian equivalent for the noun phrase "*a [...] global*". To this end we need a specialised corpus of the target language.[8] The corpus was built with WebBootCaT (using the service provided within the SketchEngine),[9] a tool which, based on user input, automatically selects and downloads potentially relevant webpages, thus making it possible to gather *ad hoc* corpora literally within minutes. Here is how the BootCaT procedure works:

1. users select "seed terms", i.e. keywords for the domain of interest; in our case these were "*perl*", "*programmazione*" ("*programming*"), "*linguaggio*" ("*language*") and "*codice*" ("*code*");
2. seeds are used to form "tuples", i.e. they are randomly combined into different multi-word sets by the programme (e.g. "*perl lingugaggio codice*");
3. tuples are sent as queries to the *Yahoo!* search engine, which returns a list of URLs matching the search terms; in our case we decided to take a maximum of 15 pages per tuple. In this phase, users can also decide to inspect the documents which will be included in the corpus, by simply clicking on their link in the results' page. Pages that are deemed to be irrelevant can be discarded at this stage (see Figure 3);
4. finally the webpages are downloaded and put together in a single file (the corpus), that users can save onto their computers and access through an offline concordancer, or browse online through the SketchEngine.

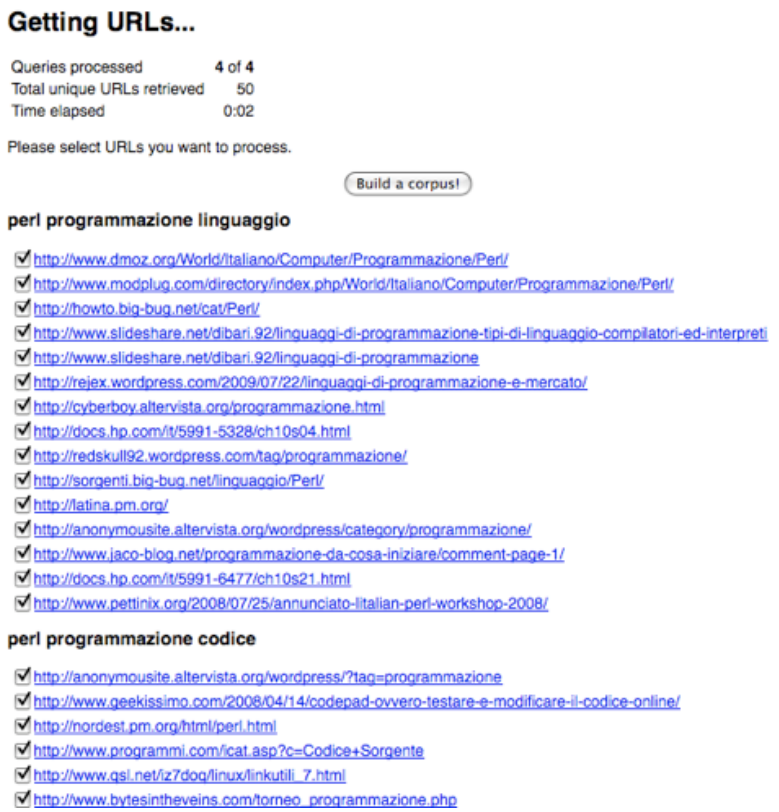

Figure 3. WebBootCaT's URL list

The corpus that resulted from this procedure was built in less than ten minutes processing time (about 2 minutes with user input), and contains 121 documents, for a total of nearly 375K words. The corpus was downloaded and accessed through the AntConc concordancer, a user-friendly concordancer which is downloadable free of charge.[10]

A search for the word "*globale*" in this specialised, *ad hoc* corpus, with results sorted according to the preceding co-text, would allow even a novice translator tackling programming texts for the first time to realise at a glance that the word "*globale*" cannot be used as a noun (Figure 4). No occurrence is found, e.g., of "*un globale*" or "*il globale*" ("*a/the global*"), which would attest to nominal uses of what is – in general Italian, and even in this technical register – primarily an adjective.

Figure 4. Concordance lines for the word "*globale*" in the specialised corpus, sorted by left co-text

More crucially perhaps, 27 out of 52 concordance lines for "*globale*" instantiate the collocation "*variabile globale*" (i.e. "*global variable*").[11] It would be a safe bet therefore to explicitate the noun "*variabile*" and translate "*global*" as "*variabile globale*".

4. Conclusion

Drawing on the literature produced within "web-as-corpus" research, in this paper different approaches to the use of the web for translation purposes were presented, and their advantages and limitations highlighted. In particular, it was argued that the most productive approaches are those whereby web data are used to compile corpora, offering translators more flexible and reliable resources, compared to simply using commercial web search engines. Finally, a simple yet hopefully convincing example was made of some ways in which web-derived corpora, both general and specialised, can be profitably applied to a translation task.

References

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. (to appear). "The WaCky Wide Web: a collection of very large linguistically processed web-crawled corpora". *Language Resources and Evaluation Journal*.

Baroni, M. and Bernardini, S. (2004). "BootCaT: Bootstrapping corpora and terms from the web". *Proceedings of LREC 2004*. 1313-1316.

Baroni, M. and Ueyama, M. (2006). "Building general- and special-purpose corpora by Web crawling". *Proceedings of the 13th NIJL International Symposium*. 31-40.

Bernardini, S. (2006). "Corpora for translator education and translation practice. Achievements and challenges". In *Proceedings of LREC 2006*.

Gatto, M. (2009). *From body to web. An introduction to the web as corpus*. Bari: Laterza.

Hundt, M., Nesselhauf, N., Biewer, C. (eds.) (2007). *Corpus linguistics and the web*. Amsterdam: Rodopi.

Kilgarriff, A. (2007). "Googleology is bad science". *Computational linguistics*, 33(1): 147-151.

Kilgarriff, A. and Grefenstette, G. (2003). "Introduction to the special issue on the web as corpus". *Computational linguistics*, 29(3): 333-347.

Lüdeling, A., Evert, S. and Baroni, M. (2007). "Using Web data for linguistic purposes". In Hundt et al. 7-24.

Renouf, A., Kehoe, A. and Banerjee, J. (2007). "WebCorp: an integrated system for web text search". In Hundt et al. 47-67.

Sánchez-Gijón, P. (2003). "És la web pública la nova biblioteca del traductor?" *Revista tradumàtica*, 2. < http://www.fti.uab.es/tradumatica/revista/num2/articles/07/07art.htm > Page consulted on date: 01.09.09

Sharoff, S. (2006). "Creating general-purpose corpora using automated search engine queries". In Baroni, M. and Bernardini, S. (eds.) *WaCky! Working Papers on the Web as Corpus*. Bologna: GEDIT. 63-98.

Thelwall, M. (2005). "Creating and using web corpora" *International journal of corpus linguistics*, 10(4): 517-541.

[1] http://mellange.eila.univ-paris-diderot.fr/index.en.shtml Page consulted on date: 03.08.09.
[2] http://www.kwicfinder.com/KWiCFinder.html Page consulted on date: 03.08.09.
[3] http://www.webcorp.org.uk/ Page consulted on date: 03.08.09.
[4] http://corpus.leeds.ac.uk/internet.html Page consulted on date: 01.09.09.
[5] A similar effort, though on a larger scale, was carried out by the *WaCky* group, who built reference corpora of nearly 2 billion words for Italian, German, and English (Baroni et al. to appear). At the moment, the WaCky corpora are only downloadable (http://wacky.sslmit.unibo.it/) or consultable through the (commercial) SketchEngine (http://sketchengine.co.uk/). A free web-based interface, as well as similar corpora for French and Spanish are in the pipeline. Pages consulted on date: 01.09.09.
[6] http://cpansearch.perl.org/src/JHI/perl-5.7.0/pod/perltooc.pod Page consulted on date: 30.07.09
[7] http://www.natcorp.ox.ac.uk/ Page consulted on date: 01.09.09
[8] In fact, ideally one would need a specialised comparable corpus, i.e. one that also includes a source language component for purposes of text comprehension, but if one is pressed for time, the odds are usually in favour of the target language corpus, which provides information about appropriate rendering as well as content (Bernardini 2006).
[9] The SketchEngine offers a 30-day trial license. A free, UNIX-based version of BootCaT is available here (http://sslmit.unibo.it/~baroni/bootcat.html), but, in order to use it, more advanced computational skills are required. Page consulted on date: 01.09.09
[10] http://www.antlab.sci.waseda.ac.jp/software.html Page consulted on date: 01.09.09
[11] If more evidence were needed, one could equally use a very simple regular expression like "*variabil.*", with a full stop standing in for any character, thus searching for both the singular and the plural forms of the adjective.