Using the Web as a Linguistic Tool in Translation Practice

Snezhina Gileva

Sofia University
Master's Program in
Computational Linguistics

1. Introduction

Internet, in its capacity of a global information environment, represents a unique source of linguistic information, unfortunately, still not fully utilized by translators. The wide use of the net in everyday translation practice makes it possible not only to solve multiple linguistic problems but also to improve significantly the quality of translation. Precisely for that reason we can be positive that in the near future the ability to use the linguistic capabilities of the net will become such an essential skill as the use of computers is nowadays.

There are several main areas in which Internet is invaluable to the translator:

- Quick access to a vast range of reference information materials: electronic dictionaries, encyclopaedias, glossaries, and various terminological resources.
- Use of the net in its capacity of a universal multilingual corpus for extracting diverse linguistic information.
- Acquisition of background information about the source and the target texts
- Operational connection: internet is a medium for communication that facilitates the fast exchange of information with customers, making the translator quite independent and globalizing the range of translation services.

The main topic of this paper shall be the investigation of the possibilities of the web as a multilingual corpus. First, we shall go over the potentials of the web as a reference guide offering multiple tools for linguistic references. Then we shall try to find out why dictionaries are not a sufficient source of information and make a short overview of the types of corpora used in translation. In addition, we shall try to prove that the web is actually a corpus and find out what its potentials as such are with regard to translation. Finally, we shall perform a small field test consisting of a translation problem and an attempt to find its answer in the web using first Google and then BNC.

2. The web as a linguistic reference guide

Undoubtedly, the main advantage of the net as a reference guide is the wealth of specialised glossaries and dictionaries from all possible fields of knowledge. Practically all well-known publishing houses offer electronic versions of their dictionaries and encyclopaedias on a CD-ROM, and many of them like Merriam Webster, Encyclopaedia Britannica, Larousse, Hachette, Meyers, Brockhaus, Garzanti, etc. also have free online versions of their products. Most often the access to the big dictionaries is in online mode but the majority of the specialised glossaries can be downloaded and used offline. Of special interest are dictionaries of slang, idioms, differences between British and American English, reference guides of grammar, style and many other.

The diverse choice of dictionaries and encyclopaedias is yet not the most valuable thing Internet has to offer to the translator. The modern search engines are also widely used as front-end solutions for linguistic queries on the net. Practically we can look at the whole set of pages present online as a colossal corpus covering all imaginable fields of knowledge.

Here it is worthwhile mentioning how the search engines actually work. Each such system is basically a huge database containing copies of websites scattered all over Internet. A robot program (called web crawler) travels around the web, collecting hypertext links and feeding them into this database. All pages are automatically indexed and when a user runs a query in the search engine, the program searches not across all pages on the net but looks up the key words in an alphabetic index and returns a link to the page on which the keyword is present.

The practice of using the web in translation practice has allowed to increase radically the quality of the end product, especially when translating from one's mother tongue into a foreign language and in fields where the terminology is updated virtually every day like telecommunications, business and finance, international relations etc. Moreover, even the most experienced translators inevitably come across unknown terms, neologisms, professional slang etc. that cannot be found even in the most recent dictionaries. In situations like that Internet is invaluable. Typing a few words in the search engine window is often worth several trips to the library and long hours of consultations with experts.

3. Why dictionaries are not enough?

In this line of thought the question arises why dictionaries are not sufficient resources for lexical information and why should translators complicate matters further by introducing additional information sources. The reasons can be briefly summarised in the following way (Varantola 200:172):

- dictionary-makers usually aim at context-free descriptions of word use, whereas dictionary users resort to dictionaries to solve a context-dependent problem.
- translators certainly need equivalents, but they also need reassurance: for this reason translators do not like to find equivalents that they do not recognize
- translators often need information relating to longer stretches of text rather than a single lexical item
- translators try to find non-dictionary type information in dictionaries because it is not readily and systematically available in other sources

It has been estimated that up to 50 % of the total time for performing a particular translation task may be spent while trying to find relevant lexical information. We can therefore argue that the more **varied tools** language professionals have at their disposal, the better their decisions and work will be (Varantola 2000:173).

4. Corpora in translation

Speaking about linguistic tools, it is worth mentioning that an increasing number of scholars in translation studies have begun to consider seriously the corpus-based approach as a feasible perspective for studying translation in an original and systematic way. Contrastive linguists have also recognized the value of translation corpora as resources for the study of languages, and translator trainers have begun to design general and specialized corpora to aid the comprehension of source language texts and improve production skills (Laviosa 1998:1).

There are several types of corpora as we know them but the ones which attracts gratest attention in translation are **multilingual** and **aligned parallel corpora**.

4.1 Annotated vs. Unannotated

If corpora is said to be **unannotated** it appears in its existing raw state of plain text, whereas **annotated** corpora has been enhanced with various types of linguistic information. Naturally, the usefulness of a corpus is increased when it is annotated. For example, the form "gives" contains the implicit part-of-speech information "third person singular present tense verb". However, in an annotated corpus the form "gives" might appear as "gives_VVZ", with the code VVZ indicating that it is a third person singular present tense (Z) form of a lexical verb (VV). Such annotation makes it quicker and easier to retrieve and analyse information about the language contained in the corpus. (McEnery, Wilson 2004, Encoding and annotation, para. 2)

4.2 Sample vs. Monitor corpus

When talking about corpora, in most cases we mean bodies of text having **finite** size, for example, 1,000,000 words. However, there is another type of corpora called **monitor corpora,** which are constantly being updated and extended, like John Sinclair's Collins COBUILD corpus. According to McEnery and Wilson the main advantages of monitor corpora are:

- They are not static - new texts can always be added, unlike the synchronic "snapshot" provided by finite corpora.
- Their scope - they provide for a large and broad sample of language.

Their main disadvantage is:

- They are *not* such a reliable source of **quantitative** data (as opposed to qualitative data) because they are constantly changing in size and are less rigorously sampled than finite corpora.

Apart from monitor corpora, most often the case is that a corpus consists of a finite number of word. Usually this figure is determined at the beginning of a corpus-building project. An exception is the London-Lund corpus, which was increased in the mid-1970s to cover a wider variety of genres.

4.3 Monolingual vs. Multilingual corpora

Apart from monolingual, there are also corpora containing texts from several different languages which are called **multilingual** and an increasing amount of work in being done on the building of such corpora.

First we must make a distinction between two types of multilingual corpora: the first can be described as **small collections of individual monolingual corpora** - they contain completely different texts in those several languages, which are **not** translations of each other.

The second type of multilingual corpora attracts most attention and is known as **parallel corpora**. This refers to corpora which hold the same texts in more than one language. The parallel corpus dates back to the famous Rosetta Stone and mediaeval times when "polyglot bibles" were produced which contained the biblical texts side by side in Hebrew, Latin and Greek etc.

In order for such corpora to be useful, it is necessary to indicate which sentences in the source language are translations of which sentences in the target language, possibly even which words are translations of each other. A corpus which has this additional information is known as aligned corpus. If we take the sentence “The boy loves the girl” in such a corpus it should be aligned next to “Der Junge liebt ein Mädchen”, at a higher level of organisation even “the” can be aligned next to “der”. This is not always a simple task because due to the specific features of the languages often one word in the source text may be translated with more than one word in the target text (e.g. “raucht” in German and “is smoking” in English), not to mention the stylistic and syntactic differences between languages which may complicate matters even further.

Annotated parallel corpora are rare and those which exist tend to be bilingual rather than multilingual. However, two EU-funded projects (CRATER and MULTEXT) are aiming to produce genuinely multilingual parallel corpora. The Canadian Hansard corpus is annotated, and contains parallel texts in French and English, but it only covers a restricted range of text types (proceedings of the Canadian Parliament). However, this is an area of growth, and the situation is likely to change dramatically in the near future. (McEnery&Wilson, 2004, Multilingual corpora, para. 4)

5. Is the web really a corpus?

Due to the versatile nature of texts that need translation often even the most voluminous corpora cannot offer sufficient information to solve a translation problem. The web is far from the ideal of an orderly annotated corpus but has one undeniable advantage over other collection of texts - it is huge. I use the term “collection of texts” because in order to name the web a “corpus” we first need to try to prove its status as such. McEnery and Wilson outline four main characteristics of a corpus: **finite size**, **sampling and representativeness**, **machine readable form** and **a standard reference**.

They write:

> *„The term "corpus" also implies a body of text of finite size, for example, 1,000,000 words. This is not universally so - for example, at Birmingham University, John Sinclair's COBUILD team have been engaged in the construction and analysis of a* ***monitor corpus****. This "collection of texts" as Sinclair's team prefer to call them, is an open-ended entity - texts are constantly being added to it, so it gets bigger and bigger.“ (Mc Enery and Wilson 2004, Definition of a corpus, para. 1)*

If we apply that definition to the nature of the web, it turns out that it is basically what we call a **monitor corpus** – one that has no finite size but grows constantly. If we look it that way, the web is actually the biggest possible monitor corpus.

Undoubtedly, all texts on the web are **machine readable**. Moreover, their primary existence is in electronic form and although many of the web documents have hard copies, most of them are available only on the hard disks of computers.

According to Mc Enery and Wilson “a corpus represents a **standard reference** to the language variety it represents” and “provides a yardstick by which successive studies can be measured”.

However, outside very specialised domains we do not really know what existing corpora might be representative of (Kaligarriff, Grefenstette 2003:340) and although the web cannot really be called a yardstick it may be a very lucrative source of information, which structured in an appropriate way, may present a linguistic playground not worse than that offered by other well-known corpora.

As far as **sampling and representativeness** are concerned, I would allow myself to quote Adam Kaligarriff and Gregory Grefenstette, who in their paper „Introduction to the Special Issue on the Web as Corpus“(2003), adopt a very straightforward approach to this question. They disagree with the definition of Mc Enery and Wilson claiming that these authors mix the question “What is a corpus?” with the question “What is a good corpus?”. It is indeed true that many of the corpora used for literary, linguistic or language-technology studies do not fit into the McEnery-Wilson definition, especially in the part “sampling and representativeness”. Kaligarriff and Grefenstette give an example with a corpus consisting of the complete works of Jane Austen, which is neither a sample, nor representative of anything else. Finally they come up with the following definition “*A corpus is a collection of texts when considered as an object of language and literary study*”. So far this definition is the most suitable one establishing the status of the web as a corpus and also fitting best into the concept of this paper.

6. Potential uses of the web in translation

In recent years the web has become a most useful tool for translators as a place where they can easily look up how a certain word or phrase is used. Since queries to standard search engines allow for restrictions to a particular language and, via the URL domain, to a particular country, it has become easy to obtain usage information which has been buried in books and papers prior to the advent of the web. In addition to simply browsing through usage examples one may exploit the frequency information (Volk 2002:6). Here we will summarize several examples how a translator may profit from the web.

6.1 Decision-making process in translation

One of the most common cases in translation practice is **checking the translation variant** one already has in mind. For example, can we translate “Spontanprämie” as “special bonus” or “key ratio” as “Schlüsselzahlen”. The simplest way is to type the corresponding keyword in the Google site (after enclosing the search items in quotation marks “”) to get a quick answer. The decision of the translator in this case is based predominantly on frequency information. If there are enough occurrences of the word or phrase in question in Google and if they come from reliable sources one may safely conclude that this is a good candidate for the target translation term. Of course, on the web everything is relative, and what we call “enough occurrences” and “reliable sources” may vary greatly from one search item to another. In addition, the number of hits in Google only show if the word or combination of words actually exists in language, only after carefully examining the context in which the word occurs (e.g. Shlüsselzahlen very often appears in contexts like “Verwendung von **Schlüsselzahlen** für Eintragungen im Führerschein”) we can decide if this is the correct term or not.

This method allows to also check the **correct translation of proper nouns and names of institutions.** For example, if one does not know how to translate the name of Sofia University “Св. Климент Охридски” and hesitates whether to use the actual name of the patron saint *St. Climent of Ochrid* or transliterate it as *St. Kliment Ohridski*, one can always use Google to find out that the latter is much more frequent. Names of institutions can also pose a problem but it is in the web where we can find out that the German “Direktion für Hochschulbildung“ and the French „Direction des enseignements supérieurs“ are equivalent to the English “Department for Higher Education” (as a structure within a ministry).

Another case is **when we do not have a translation variant** (e.g. how to translate the German word “Pistenflitzer” in English) **or our variant has not been confirmed** (*special bonus* does not express well the implicit meaning of *Spontanprämie* as a term opposed to annual, performance-related bonus). In that case the translator needs to find texts from the corresponding thematic domain where he or she stands a good chance of finding the necessary term (key words may be “bonus”, “reward” and “compensation”). Practically all search engines allow advanced search options where you can limit the search scope to a particular language (e.g. only sites in German) or to one domain area (e.g. sites ending in .bg are hosted in Bulgaria, .ch in Switzerland etc.)

6.2 Translation of compound nouns

In 1999 Grefenstette used the web to show how one can locate the correct translations of German compound nouns if the potential translations of their constituent parts were known. He selected a number of German compounds from a machine-readable German-English dictionary. The requirement was that every compound had to be decomposable into two German words found in the dictionary and the English translation also had to consist of two words. However, for each segment of the words in question there was more than one translation possible. For example, the German noun Aktienkurs (share price) can be segmented into Aktie (share, stock) and Kurs (course, price, rate) both of which have multiple possible translations. By generating all possible translations (share course, share price, share rate, share course,…) and submitting them to Alta Vista queries, Grefenstette obtained frequency results for all possible translations. He tested the hypothesis that the most frequent translation is the correct one. He extracted 724 German compounds according to the above criteria and found that his method predicted the correct translations for 631 of these compounds (87%) which is an impressive result given the simplicity of the method (Volk 2002:6).

6.3 Mining the web for parallel texts

Translation memory systems have become an indispensable tool for translators in recent years. They store parallel texts in a database and can retrieve a unit (typically a sentence) with its previous translation equivalents when it needs to be translated again. Such systems come to their full use when a database of the correct subject domain and text type is already stored. They are of no use when few or no entries have been made. However, very often previous translations exist and have been published on the web (e.g. documents in the administrative sphere, or common EU documents in multiple languages). The idea is to find these translation pairs, evaluate them, download and align them and, finally, feed them into a translation memory.

In 1999 Philip Resnik developed a method for automatically finding parallel texts in the web. Initially he ran queries on Alta Vista by asking for parent pages containing the string “English” and “German” in anchor text within a fixed distance of each other. This generated many good pairs of pages such as those reading “Click here for English version” and “Click here for German version” but of course also many bad pairs.

Then he made use of the fact that the translations and the originals are very similarly arranged in terms of HTML structure. He used a statistical language identification system to discover if the documents are in the suspected language. Then he submitted 192 pairs to human judgement and 92% of the pages judged as good by the human experts were judged as good by his system as well.

In the second phase of the experiment he expanded his scope of research by looking not only for parent pages but also for sibling pages (linked pages which are translation of each other). For the language pair English-French he obtained more than 16 000 page pairs.

7. Web and other corpora in the translation process

A corpus of finite size cannot always offer sufficient information for disambiguating complicated translation questions. Even the British National Corpus with its 100 million words is often ill-equipped for that purpose. On the other hand, the web has a potential to be invaluable for linguistic research due to its width, breadth, up-to-dateness and universal availability.

A simple search conducted in Google and BNC was meant to illustrate the suggestion that relatively simple search techniques for querying the web can be used for solving quite complex translation problems. The translation pair came from an academic certificate written in Bulgarian. The source text was "издържал изпити по" (= [the student] has passed exams in/on [several subjects]) followed by a list of subjects attended and marks of the student. The two translation variants that offered themselves were "passed exams on" and "passed exams in". Preposition usage is one of the trickiest things in translation and since it was not quite clear which preposition was more appropriate in that case a search on both Google and BNC was designed to solve the problem.

Initially a full-text search in Google was done for the two phrases. The results were:

Query	Results (Hits in Google)
„passed exams in"	735
„passed exams on"	224

However, there was a lot of "noise" in the results like:

...The number of **passed exams in** nine rural schools in the area where...

... l) average score of the **passed exams** (**in** words and in figures); ...

...In 1999 I **passed exams in** Moscow State University...

...have **passed exams in** one sitting to become a Memeber of...

AND

... (Both **passed exams on** November 6, callsigns pending)...

...a student who took and **passed exams on** the same day...

...270 candidates who have **passed exams on** paper...

...graduated from Bergen school in 1707; MA in 1727. Religion, **passed exams on** Oct 22, 1714...

which clearly did not fit into the formula:

passed + exams + <preposition> + <name(s) of subjects>

Despite that everything pointed to the fact that the first variant was the more common of the two.

To confirm the results, I decided to do a second search using the main form of the verb "to pass" in order to collect more results. Again a full-phrase search in Google was used. The results were:

Query	Results
“pass exams in”	1460 hits
“pass exams on”	1550 hits

Once again only frequency information was not enough and not quite representative to demonstrate clearly that either one or the other of the prepositions was the correct one in that case.

For that reason a third search on Google was conducted including the word “subjects” to the above-mentioned search phrase. The reason to do that was to find web pages that mimic the original academic certificate, i.e. pages containing the phrase “passed exams **in** OR **on**” followed by a list of subjects. The results were:

Query	Results
“pass exams in” + subjects	199
“pass exams on” + subjects	30

With the search query structured in that way, the word “subjects” may appear anywhere on the page, not necessarily in close proximity to the search phrase. What turned out was that in the first case “subjects” was much closer to “pass exams in”: e.g.

“pass exams”:

- in a number of subjects
- in four core subjects
- in the following national subjects
- in all subjects of the curriculum.

In the second case the results looked like that:

- **...** Candidates must **pass exams on** Windows architecture and services plus programming languages of **...** the person who trains others in these **subjects** through Microsoft **...**
- **...** They will be required to **pass exams on** driving and the Rukhnama in order to get **...** academic year, students can choose to sit three exams in the **subjects** of their **...**
- **...** Where only a few **subjects** in a JAR-66 module are covered by the ratings **...** Existing licence holders need not **pass exams on** Human Factors (module 9) and Aviation **...**

Bearing in mind the frequency of occurrence and the proximity search results, the conclusion imposed itself that in our case the preposition IN was the more appropriate of the two.

However, after closer observation of the results, it turned out that there was a fine separation between the use of IN and ON in that particular case. ON appeared mainly in combinations with nouns or phrases denoting **specific areas of knowledge** and **fields of expertise** like:

- to pass exams **on the US Constitution**
- to pass exams **on the latest Microsoft Technologies**
- to pass exams **on two different instruments**
- to pass exams **on earlier release versions of Oracle**

while IN, in most cases, collocated with names of subjects as they are defined in the school curriculum (**Mathematics, Physics, Biology, Chemistry** etc.) and although there were examples like “to pass an exam on Maths” they were not very common. Despite the simplicity of the search method the results were convincing enough to show that both prepositions had their domain of usage.

What were the results in BNC? For that particular query were used both the online version of BNC and the downloadable SARA client. Unfortunately, only 1 single entry in the corpus matched the queries which were conducted. The results were the following:

Query	Results
„passed exams in“	K5J **2150** Michael Wittet: CA who passed exams in German prison camp
„passed exams on“	No matches
“pass exams in”	No matches
“pass exams on”	No matches

The following search queries have been tested as well:

- **`pass=VVB + {exam[s]} + PRP`** – to which no results were returned
- **`Pass + {exam[s]}`** returned 8 matches but they excluded exactly the part of speech I was interested in – namely the preposition.

The results from BNC were insufficient to solve the problem at hand but there is no doubt that BNC and other well-known corpora offer options for linguistic research incomparable to any online search engine. There are, of course, projects like WebCorp, KwiCFinder, Gsearch etc., which try to facilitate the process of searching the web for grammatical constructions, still this ultimate linguistic goal has not be fully achieved. However, when faced with no time and, in most cases limited budget, the most obvious place to search for an answer to a translation problem is the net.

8. Conclusion

Corpora have been a part of the translation practice and training for a good number of years. Their use has proved beneficiary irrespective of whether the text sources are monolingual, bilingual or even aligned parallel corpora. As a new medium for the existence of information Internet provides an additional stimulus to the development of translation. It offers a much larger number of opportunities for corpus research since it exceeds in size any previously compiled corpora. Unfortunately, online search engines are suited not for linguistic but for general knowledge queries. Yet, currently the web is the first and the most obvious place where most translators look for linguistic information, and, not quite surprisingly, in most cases they are able to find it there.

The of objective of the paper in not to make sweeping generalisations about the usefulness of the web as a corpus or juxtapose it to BNC and other corpora. The task is to show that in translation practice, by using relatively simple search techniques and in within a very limited time span, one can find an answer to quite tricky questions by using the web as a huge multilingual corpus. It is my deepest conviction that this largest and most up-to-date collection of texts should be used to its full potential in translation studies and as well as in other linguistic fields.

9. References

Тиссен, Ю. (2000). *Интернет в работе переводчика.* Мир перевода No 2 (4). Санкт Петербург.

British National Corpus. (2000). Simple search on BNC-world. Available online at http://sara.natcorp.ox.co.uk/form.html

Collins Cobuild. (2000). The Bank of English. Searchable sample available online at http://titania.cobuild.collins.co.uk/form.html

Kilgarriff, Adam. (2001). *Web as corpus.* In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie and Shereen Khoja (eds.) *Proceedings of the Corpus Linguistics 2001 conference, UCREL Technical Papers: 13.* Lancaster University, 342-344.

Kilgarriff, A., Grefenstette, G. 2003. *Introduction to the special issue on the web as corpus*. Computational Linguistics 29 (3): 333-47. Retrieved November 26, 2004, from http://www-mitpress.mit.edu/journals/pdf/coli_29_3_333_0.pdf

Laviosa, S. (2004). *Corpora and the Translator.* Retrieved December 19, 2004, from http://stp.ling.uu.se/~evapet/CAT/15-Laviosa.doc

McEnery, T. , Wilson, A. (1993). *Corpora and Translation: Uses and Future Prospects.* UCREL Technical Papers.

McEnery, T. , Wilson, A. (1996). *Corpus linguistics.* Edinburgh: Edinburgh University Press.

McEnery, T., Wilson, A. (n.d.). *Web pages to be used to supplement the book "Corpus Linguistics" published by Edinburgh University Press*. Retrieved December 23, 2004, from http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/contents.htm

Resnik, P. (1999). *Mining the web for bilingual text.* Proc. Of 37th Meeting of ACL. Maryland. 527-534.

Resnik, P., Smith, N. A. (2003). *The Web as a parallel corpus*. Computational Linguistics, v.29 n.3, p.349-380, September 2003.

Smarr, J., Grow, T. (2002). *GoogleLing: the web as a linguistic corpus.* Retrieved December 16, 2004, from http://www.stanford.edu/class/cs276a/projects/reports/jsmarr-grow.pdf

Tymoczko, M. (1998). *Computerized corpora and the future of translation studies.* "Meta". vol. 43, n. 4 (98), pp. 652-659.

Varantola, K. (2003). *Translators and disposable corpora.* In Corpora in Translator Educaion F. Zanettin, S. Bernardini and D. Stewart (eds), 55-70. Manchester: St Jerome.

Varantola, K. *(2002). Disposable corpora as intelligent tools in translation.* In: Tagnin, S. E. O. (Org.). Cadernos de Tradução: Corpora e Tradução. Florianópolis: NUT, 2002, v. 1, n. 9, p. 171-189.

Volk, M. (2002). Using the web as a corpus for linguistic research. In Pajusalu, R. & Hennoste, T. (eds): *Tähendusepüüdja. Catcher of the Meaning. A festschrift for Professor Halur Õim.* Department of General Linguistics 3, University of Tartu.

WebCorp. (2002). http://www.webcorp.org.uk

Zanettin, F. (2002). *DIY corpora: the WWW and the translator.* In Maia, B, Haier, J, Ulrych, M (eds.), pp 239-248.