The Web for Corpus and the Web as Corpus in Translator Training[1]

Miriam Buendía-Castro, Clara Inés López-Rodríguez
University of Granada, SPAIN

ABSTRACT
Corpora are rich information sources that can provide the translator with both linguistic and conceptual knowledge that is not found in dictionaries. The question that arises within this context is whether the web can be considered as a corpus. Following the distinction made by De Schryver (2002), there are two corpus-based approaches to the web: (i) web for corpus (WfC), in which the web is used as a source of texts in digital format for the subsequent implementation of an offline corpus; (ii) web as corpus (WaC), which uses the web directly as a corpus. In this paper, we compare and evaluate as translation aid tools, "automatically built" corpora (both general, i.e. WaCky corpora, and specialized, i.e. corpora created by the students themselves through WebBootCat, which are accessible through Sketch Engine), as opposed to the manual building of corpora. To that end, we asked two groups of students in the Translation and Interpreting Degree Program at the University of Granada (Spain) to carry out a technical translation assignment. One of the groups used automatically built corpora, whereas the other group used the web to manually extract texts for the later compilation of corpora. The results obtained showed that these two methods are complementary, and that students should decide for one or the other depending on their needs (i.e. translation assignment, novelty of the translation, directionality and specificity of the translation, time allotted, or level of analysis required).

KEYWORDS: corpora, corpus query system, specialized translation, Web as Corpus, Web for Corpus.

1. Introduction

Since 1997, Corpus Use and Learning to Translate (CULT) has been a fruitful area of research (Beeby, Rodríguez and Sánchez 2009; Bowker 1998, 2000; López 2002, López and Tercedor 2008; Zanettin 1998; Zanettin, Bernardini and Stewart 2003). Both corpora and the internet are nowadays included amongst translation tools, along with lexicographic and terminographic resources, translation memories, and other applications.

The internet gives users the possibility of accessing any type of information at any time and at any place. However, this also has its downside since it can result in an "information overload" (Jiménez Piano and Ortiz-Repiso Jiménez 2007: 18). The amount of data circulating on the internet on any given day is greater than all the information available in the nineteenth century (Austermühl 2001: 7). English continues to dominate the web, representing 45% of the total number of web pages. Other European languages with a significant percentage of webpages are German (5.9%), French (4.41%), Spanish (3.8%), Italian (2.66%), and Portuguese (1.39%)[2]. Because of the vast amount of information offered, the internet constitutes "a fabulous linguists' playground" (Kilgarriff and Grefenstette 2003: 333), from which Translation can also benefit. It is widely acknowledged that documentation and

[1] A Spanish short version of this study was published in TransKom *4* (1) (López and Buendía 2011). This English version explains the procedure in more detail, and discusses the findings more fully.

[2] These results correspond to the investigation carried out by the Union Latina (Latin Union). Available at "http://dtil.unilat.org/LI/2007/es/resultados_es.htm (accessed 13 May 2010)".

terminological extraction are amongst the most important tasks for translators, and that corpora have become essential for performing them. When translators face a new translation assignment, a new corpus is usually required. Corpora are rich information sources that can provide the translator with both linguistic and conceptual knowledge that is not generally found in traditional lexicographical repositories, such as dictionaries.

Many years ago, corpus compilation used to be an arduous process that required many hours spent in libraries. Even today, after more than a decade of research and experience in this area, many scholars and professional translators still consider that compiling a corpus is time-consuming in the short term. However, this is no longer true since on the internet, hundreds of texts can be compiled in a few minutes. Nonetheless, using corpora for translation purposes is not just a question of retrieving a large number of texts:

> The difficulty of using corpora is that they rarely provide immediate answers to a translator's problems. Unlike translation memory or machine translation systems, they do not instantly present a preferred candidate for the user to accept, modify or reject. Corpus data has to be interpreted and evaluated comparatively to reach conclusions, and this requires not only technical skill [...] but above all critical thought (Aston 2009).

Within Translation Studies, many researchers have highlighted the pedagogical advantages of using a do-it-yourself corpus (DIY corpus). This means a collection of internet documents compiled *ad hoc* as a response to a specific text to be translated (Zanettin 2002: 242). This kind of corpus compiled for a particular translation assignment has also been called "disposable corpus" (Varantola 2003) since texts are harvested for satisfying transitory needs, rather than to enrich a permanent corpus. Additionally, Varantola stresses the importance of determining the criteria for compiling and using *ad hoc* corpora:

> I would even go a step further and claim that the knowledge of how to compile and use corpora is an essential part of modern translational competence and should therefore be dealt in the training of prospective professional translators (Varantola 2003: 56).

The main question that arises within this context is whether the web should be considered as a corpus itself. In this regard, there are two approaches to the web (De Schryver 2002): (i) *web for corpus (WfC)*, in which the web is used as a source of texts in digital format for the subsequent implementation of an offline corpus; (ii) *web as corpus (WaC)*, which uses the web directly as a corpus[3]. Supporters of the WfC approach include authors such as Sinclair, whereas Kilgarriff and Grefenstette (2003); Fletcher (2004, 2007), and Baroni and Bernardini (2006) follow a WaC approach.

In this study, we compare, and evaluate as translation aid tools, "automatically built" corpora, as opposed to the manual building of corpora in the context of a scientific and technical translation course at university level. To this end, firstly, we review the notions of Web for Corpus and Web as Corpus. Secondly, we describe an experiment carried out with two groups of 3rd year students in the Translation and Interpreting Degree Program at the University of Granada (Spain). The objective was to test whether the use of a corpus query system to search

[3] De Schryver (2002: 272) makes a distinction between the Web for Corpus, that is, the Web "as a provider of data 'for' the creation of corpora" and the Web as Corpus, in which the focus in "on the potential of the Web 'as' a corpus in itself".

online corpora (*Sketch Engine)* available from http://www.sketchengine.co.uk/ was more beneficial for the completion of a specialized translation assignment on swine flu than the usual tools (paper and electronic dictionaries, and access to internet search engines). Special attention was given to the design of the study, characteristics of the sample population reflected in questionnaire data, and the assessment of the quality of their translations. Finally, we discuss what the average quality of the translations of each group can tell us about the advantages and disadvantages of both approaches. A summary of the results of this experiment in Spanish is available in López and Buendía (2011).

2. The Web for Corpus and the Web as Corpus Approaches

2.1. The Web for Corpus

As previously mentioned, the Web for Corpus (WfC) is the approach that has been used in the longer term to download texts in digital format for the subsequent implementation of offline corpora. In this regard, in the field of Translation, the notion of do-it-yourself corpus (DIY) (Zanettin 2002: 242) has been used to describe the collection of internet documents compiled *ad hoc* as a response to a specific text to be translated. Authors such as Sinclair (2005) are clearly in favour of this traditional approach. Although Sinclair admits the internet's usefulness for linguists, he underlines the fact that the WWW is not a corpus because it has not been defined from a linguistic perspective.

The WfC approach involves searching the web for valuable information non-automatically in order to select and download appropriate texts. Users thus enter a list of keywords in a Search Engine (SE) or a particular URL, which leads them to other websites. They then select texts to download and process in a corpus analysis program, such as Wordsmith Tools (Scott 2008). Wordsmith Tools is a corpus analysis tool developed by Mike Scott at the University of Liverpool. The software includes three tools: *Wordlist, Concord,* and *Keywords*.[4] Corpus analysis tools are tools designed to help linguists to exploit large quantities of texts (i.e. in order to look for specific terminology or for phraseology) more rapidly and systematically than by using what translators have traditionally called "parallel texts".

It is well-known among corpus linguists that the internet is currently the principal source of texts for corpus compilation. This means that corpus quality is directly related to the quality of websites. For this reason, text selection is crucial for the development of a representative corpus with reliable data. As Austermühl points out, "Finding data on the World Wide Web is no problem at all. But finding reliable information is rather a difficult task. And finding the information you really need can be very time-consuming and often frustrating (2001: 52)."

Much has been written about the parameters for determining the quality of digital resources, but as yet there is no broad consensus of opinion. Buendía and Ureña (2009) reviewed the literature on such parameters, and established an evaluation protocol composed of three parameters: (i) *authority*, which evaluates mainly the reputation and expertise of the authors; (ii) *content*, which includes coverage, accuracy, objectivity, currency and audience; (iii) *design*, composed of navigational aids, accessibility and presentation and management.

[4] Its first version was released in 1996, and by April 2013 its 6th version was already on the market (Scott 2008, updated by the authors).

2.2. The Web as Corpus

There is no consensus of agreement on the meaning of the expression *Web as Corpus (WaC).* According to the classification proposed by Bernardini, Baroni and Evert (2006), there are three ways of approaching the WaC from a methodological point: (i) *the Web as a corpus surrogate*; (ii) the *mega-corpus or mini Web*; (iii) *the Web as a corpus shop*.

2.2.1. The Web as a Corpus Surrogate

This first approach to the Web as Corpus regards the web itself as one huge corpus. Systems that implement this approach generally have an interface in which the search words are entered. Results are then displayed as concordances, in the same way as if a corpus had been entered in a corpus analysis tool on the user's computer, but with the difference that the "corpus" is online.

These systems are rather different from conventional search engines such as Google or Altavista in that they pre-process the questions before sending them to the search engines and then post-process the results and present them in such a way as to facilitate linguistic studies. In other words, these systems provide a layer of pre- and post-processing which redirect queries entered by the user to existing search engines, and then filter and present the results in appropriate format. Some of the most widely known are *WebCorp* (http://wse1.webcorp.org.uk/); *KWiCFinder* (http://www.kwicfinder.com/KWiCFinder.html); and *Corpeus* (http://www.corpeus.org). However, these systems of pre-post processing have certain limitations, which coincide to a great extent with the limitations of search engines.

Firstly, the quantity of web text searched is limited by time constraints, and thus the recall can be poor. Since search engines offer a limited number of results for a particular query, these systems cannot retrieve more results than the search engines because they depend on them. As a result, WaC systems will normally offer fewer results since they have to filter the results that do not satisfy the user's search query. Additionally, if there is information unavailable on the search engine, it is almost impossible for these web corpus systems to provide it. Usually each system relies on a specific search engine. However, WebCorp allows the user to select from a group of search engines such as Google, Bing or/and FAROO.

Secondly, the proportion of potentially relevant web texts is limited by the search criteria of search engines. Systems such as WebCorp do not have any control over Google ranking. When making a query, the system should ideally offer a random sample of reliable webpages. However, search engines return a list of pages according to specific criteria, such as popularity or geographical proximity, something which is less interesting for linguists. Thus, when the same query is entered in the same search engine, the results will be different, depending on whether the query is made in the United Kingdom or the United States, for example (Hundt, Nesselhauf and Biever 2007: 2-3). Regarding popularity, Fletcher (2012) states that search engine hits are very different from corpus frequencies, and that "most widespread does not necessarily mean 'preferred' in linguistic terms".

Thirdly, search engines are inherently limited. Information on the internet updates so rapidly that experiments can never be replicated. Fletcher (2007: 37) talks about the volatility of the web, and states that "not only do hit counts vary widely due to non-linguistic factors, but the same query on the same search site can return different sets of SERPs (search engine report pages), not only from different places at different times, but even during a single user session". Ntoulas et al. (2004) also studied the dynamicity and volatility of the web, based on the analysis of 154 webpages. The results of their analysis concluded that new webpages appear at a rate of 8% a week. However, *new* does not necessarily mean *additional* or *novel*. The study concluded that the total number and size of webpages remained relatively constant, since *old* pages disappear, and only 5% of *new* pages have new content. Because of these limitations, it is necessary to develop a search engine for linguists and translators (Lüdeling, Evert and Baroni 2007). Translators' needs are mainly focused on the linguistic features of the target language and on the search for equivalent terms and concepts in their working languages. This is the approach followed by those who regard the web as a mega corpus.

2.2.2. The Mega-Corpus or Mini Web

Some linguists have attempted to create a new object, namely, a kind of mini-web or mega-corpus adapted to linguistic research. This new tool for linguists could benefit users that wish to study aspects of language through the Web, and also those users who wish to investigate aspects of the Web through language (Bernardini, Baroni and Evert 2006: 14). The ideal method would be to compile a corpus directly from the web without having to trust a search engine to look for available documents, and then to manually download them one by one. If it were possible to access the corpus obtained through the web from an interface offering sophisticated search options (linguistic annotation, metadata, *inter alia*), this would be a real "search engine for linguists" (Volk 2002; Fletcher 2004, 2007). Various research groups are currently working on the implementation of this type of system, e.g. the Webcorp project (http://www.webcorp.org.uk), the GlossaNet project (http://glossa.fltr.ucl.ac.be), and the Wacky project from the University of Bologna at Forlì (http://wacky.sslmit.unibo.it).

We shall now briefly describe the Wacky project since some of its corpora were accessed by our subject population by means of the Sketch Engine interface as part of our experiment. The main objectives of Wacky, which stands for *Web as Corpus kool ynitiative*, are to compile huge corpora (more than two billion words) extracted from the web, and to offer tools to process and exploit them. Within this initiative, there are currently some corpora already available: *deWaC* (for German), *itWaC* (for Italian), *ukWaC* (for English), and *frWaC* (for French). Our students sought the *uKWaC* and the *Spanish Web Corpus*.

For example, the ukWaC is a 2 billion word corpus constructed from the Web constraining the crawl to the .uk domain and using medium-frequency words from the BNC as seeds, i.e. as input words for the search. The corpus was POS-tagged, i.e. every word in the corpus was tagged according to its part of speech (noun, verb, adjective, adverb, etc.), and lemmatized, i.e. the different forms of a word were grouped together, with the software TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/). As shall be seen, what distinguishes this approach from the Web as a corpus shop is that the process is done automatically.

2.2.3. The Web as a Corpus Shop

Finally, the Web can be perceived as a *shop* where a corpus can be selected and acquired. Internet users go to the web to search for texts through a search engine. Users select their texts and download them to create a corpus. The process normally implies both manual and (semi)automatic compilation methods.

There are valuable tools for translation that allow users to quickly and automatically compile corpora from the web. For example, the BootCat toolkit can help speed up the corpus creation process of a given specialized domain (Baroni and Bernardini 2004). Users just have to download the program from http://bootcat.sslmit.unibo.it/ and follow the instructions of a wizard through the process of creating a simple web corpus. There is an online version of the

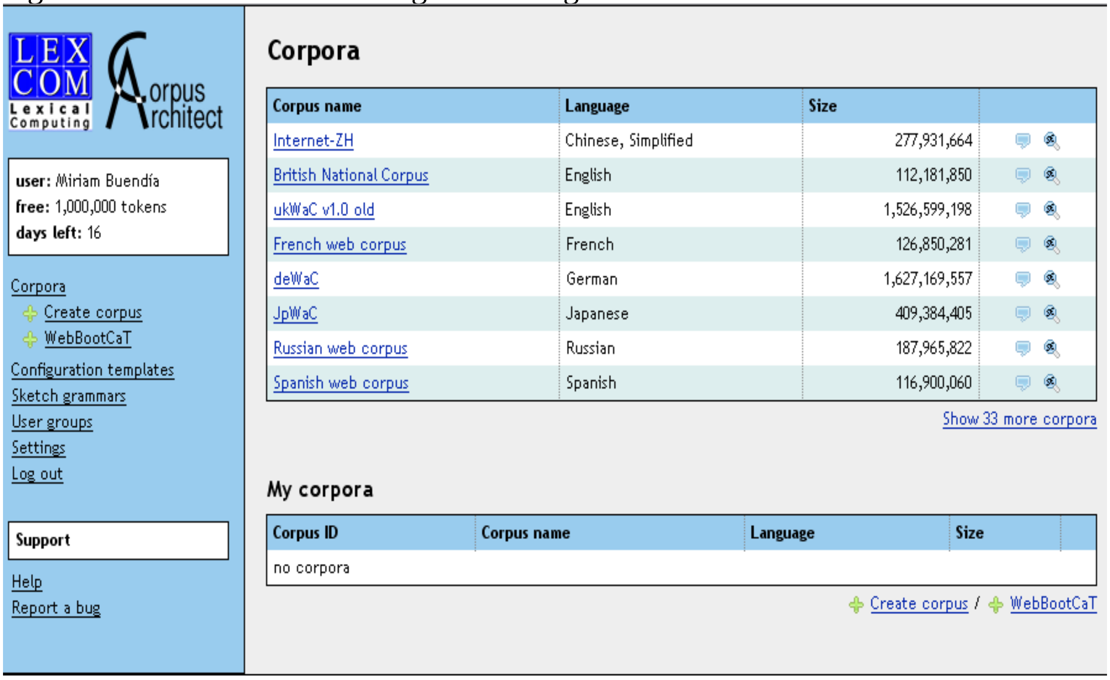
BootCat tools called WebBootCat (Baroni et al. 2006). It is a web service aimed at assisting translators by quickly producing corpora for specialist areas, in any language (including English, Spanish, Italian), from the web. The application does not have to be downloaded, but can be easily accessed from within a famous corpus analysis tool: Sketch Engine (http://www.sketchengine.co.uk) (cf. Kilgarriff et al. 2004). Briefly speaking the process to build an automatic corpus with WebBootCat is the following: (1) specify the language; and (2) select the seed words, i.e. words that are specific of your domain and that the system would use to launch the query to Google. In this study, our translation students took advantage of Sketch Engine as a translation tool.

Sketch Engine (SkE, also known as Word Sketch Engine) is a corpus query system such as Wordsmith Tools, previously described. Apart from the conventional functions offered by WordSmith (concordances, keywords, and wordlists), it integrates grammatical relations, a distributional thesaurus and word sketches. Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour (Kilgarriff et al. 2004). As can be seen in Figure 1, once registered[5], a Sketch Engine account offers the user:

- Pre-loaded corpora (60 million - 2 billion words) in a wide range of languages (i.e. English, French, German, Japanese, Russian, Italian and Spanish, and for other languages such as Arabic, Chinese, Dutch, Croatian, Greek, Hebrew, Hindi, Persian, Polish, Portuguese, Romanian, Serbian, Slovenian, Swedish and Vietnamese).
- Access to WebBootCaT. This allows users to compile a corpus of thousands of tokens in a few minutes from the 'seed terms' entered by the users. Additionally, it permits users to download the corpus to their computer; add new documents to their corpus from the web or from the hard disk; extract keywords of the domain; view the different texts in plain format or vertical format (i.e., annotated morphologically and by lemmas); and open the corpus with a lexical analysis program, and do things like generate concordances, wordlist, frequency lists, collocations, and word sketches.
- A CorpusBuilder, which permits users to upload and set up their own corpora from the hard drive.

[5] Free access is only available for 30 days.

Figure 1: Interface of Sketch Engine showing its main facilities[6]



user: Miriam Buendía
free: 1,000,000 tokens
days left: 16

- Corpora
  - Create corpus
  - WebBootCaT
- Configuration templates
- Sketch grammars
- User groups
- Settings
- Log out

Support

- Help
- Report a bug

Corpora

Corpus name	Language	Size	
Internet-ZH	Chinese, Simplified	277,931,664	
British National Corpus	English	112,181,850	
ukWaC v1.0 old	English	1,526,599,198	
French web corpus	French	126,850,281	
deWaC	German	1,627,169,557	
JpWaC	Japanese	409,384,405	
Russian web corpus	Russian	187,965,822	
Spanish web corpus	Spanish	116,900,060	

Show 33 more corpora

My corpora

Corpus ID	Corpus name	Language	Size	
no corpora				

Create corpus / WebBootCaT

3. Using the Web for Corpus, and the Web as Corpus in a Scientific and Technical Translation Course

There have been many studies on the application of corpus linguistics to translation teaching since the use of corpora helps students to find the appropriate words for a specific context and text type, thus increasing their learning autonomy (López and Tercedor 2008). However, we designed an experiment that would test whether the use of corpora when compared to translating without corpora was more effective within the context of an actual Scientific and Technical Translation course. Our aim was to discover whether the use of automatically compiled corpora, more specifically, WaCky corpora and WebBootCat accessed through the Sketch Engine interface offered more advantages to translation students than the usual approach followed in the translation classroom, involving the use of paper and online lexicographic and terminographic resources, and the selection of parallel texts from the web for documentation, without using lexical analysis tools.

3.1. Research Hypotheses

When translating a text, previous knowledge of the subject field, as well as the experience of having translated texts on the same topic facilitate the translation process. In addition, when translating in a new environment with unfamiliar computer tools, the result of the translation is usually of lower quality. In our case, the initial assumption was that students who had previously translated two texts on swine flu, and who used the Google search engine to translate a text on this subject in exam conditions (control group), would perform better than students without access to Google and who had not previously translated texts about swine flu (experimental group).

Nonetheless, this presupposition might not hold if the students without Google access or previous translation experience in the subject field (experimental group) were provided with a tool that was capable of compensating for their lack of experience in the translation of swine

[6] Reproduced by kind permission of Adam Kilgarriff, Sketch Engine's founder and owner.

flu texts, as well as their lack of access to Google. Furthermore, if the quality of their performance turned out to be similar to that of the other group (control group), then this would mean that the resource was a valuable tool for translators since it helped them produce a good quality translation.

3.2. Design of the Study

As part of our study, we asked two groups of third year students in the Translation and Interpreting Degree Program at the University of Granada (Spain) to translate a fragment of a research article entitled 'Insights from investigating the interaction of oseltamivir (Tamiflu) with neuraminidase of the 2009 H1N1 swine flu virus'[7] under exam conditions. In 2 hours they had to translate 350 words. This was done with access to the internet and to reference materials such as electronic and paper dictionaries (see Table 1) in the case of both the experimental and the control groups. The differences between the groups were the following:

- Students in the experimental group had never translated texts on swine flu, as opposed to students in the control group, who had previously translated two texts.
- Students in the experimental group were asked to translate without Google, and thus could not use Google to verify terminology or for documentation purposes.
- Students in the experimental group used corpora as a translation aid. These corpora were accessed by means of the Sketch Engine interface.
- Students in the experimental group had received training in the use of Sketch Engine (two sessions), and had compiled a corpus on swine flu using this resource one week before the exam (see details in section 3.2.3.). Therefore, they had already sought expressions in English and Spanish in WaCky corpora such as the British National Corpus, the ukWaC, and the Spanish Web Corpus.

A questionnaire was given to both groups to gather information about their background and to receive feedback about the experiment.

3.2.1. The Questionnaire

The questionnaire was divided into three sections, and was composed of 34 questions. The experimental group had to answer all three sections, whereas the control group only answered the first two sections (see Appendix).

Section 1 elicited background data from our subjects that could influence the results of the exam. These included age, mother tongue, knowledge of foreign languages, level of English, keyboard skills, previous higher education (for example, a student with a BSc in Biology, and good command in English could be expected to perform the task better than a student with no background in science).

Section 2 included nine questions about their previous knowledge of swine flu as well as their documentation skills and habits. We asked respondents whether they normally compiled a corpus for their translation assignments. If that was the case, we asked them to explain the usual stages for this process, and whether they analysed the corpus manually or with lexical analysis software. There were two questions about the electronic and paper dictionaries used for this translation exam.

[7] S.-Q. Wang et al. *Biochemical and Biophysical Research Communications* 386 (2009) 432–436. "http://download.thelancet.com/flatcontentassets/H1N1-flu/virology/virology-36.pdf (accessed 20 September 2011)".

Finally, Section 3 elicited the opinion of students in the experimental group about the usefulness of automatically compiled corpora through Sketch Engine both for this translation assignment, and for documentation and translation purposes in general. They also had to describe the advantages and disadvantages of using Sketch Engine.

3.2.2. Study Population: age, language background, documentation skills, and previous knowledge about the subject field

The population consisted of two groups of 12 students each, all of whom were enrolled in our course on Scientific and Technical Translation (English to Spanish), and had English as their first foreign language. Their ages ranged from 20 to 38. In both groups, the mean age was 27, and the most common age among the group was 21. On a scale of 1 to 5, their level of English was 4 (Advanced/C1/CAE). All the students, except one, were native speakers of Spanish. As part of the degree program, our subjects also had acquired skills in a second foreign language (French, German, Chinese, Greek, Arabic and Russian), which possibly improved their ability to interpret texts.

Since our subjects were in a Translation degree program, 75 per cent were accustomed to compiling corpora in electronic format for documentation and terminological purposes. They were also experienced in searching the internet and using freely available on-line lexicographic resources, such as websites offering dictionaries, feedback from language forums, and social networks. The majority of them normally consulted on-line lexicographic materials along with electronic dictionaries installed in their computers rather than paper dictionaries. For this translation assignment, as revealed by their answers to questions 18 and 19 of the questionnaire, the most popular documentation resources were the following (Table 1):

Table 1: Most popular documentation resources used by the students

Online dictionaries:
- WordReference.com: http://www.wordreference.com/es/ - Reverso Online Dictionary: http://dictionary.reverso.net/ - The Free Dictionary: http://www.thefreedictionary.com/ - Oxford Dictionary of English (OED) Online: subscription of the University of Granada - Merriam Webster Online: http://www.merriam-webster.com/ - Diccionario de la Real Academia Española [Dictionary of the Royal Academy of the Spanish Language]: http://www.rae.es
Electronic dictionaries installed on their computers:
- Oxford Spanish dictionary (bilingual) - Collins bilingual dictionary
Bibliographical databases:
- Medline
Paper dictionaries:
- Diccionario crítico de dudas inglés-español de medicina by Fernando Navarro.

3.2.3. Use of Sketch Engine

The experimental group attended two training sessions of three hours in total. In these sessions, they learned the differences between the Web for Corpus and the Web as Corpus approaches, and how to use Sketch Engine, the Corpus Query System described in section 2.2.3. In the first two-hour session, students created a personal Sketch Engine account, and started to compile one corpus of their own choice, so as to become familiar with the interface. There was a debate afterwards in which students asked questions, and recommendations were given for improving their use of Sketch Engine, following Castagnoli (2006), who shows the advantages and limitations of the use of BootCat in a course in Terminology and LSP. As homework, they were asked to practice what they had learned.

In the second one-hour session, we first clarified doubts about the use of Sketch Engine, and then asked students to compile two corpora about swine flu, one in English and one in Spanish. We informed them that they would be using these corpora to translate a text on swine flu under exam conditions, but would not be allowed to see the text until the day of the exam. We only provided them with some general keywords that they could use as initial seeds to compile their English corpus: *flu, influenza A, influenza virus, H1N1, drugs, oseltamivir*. Students were told to write down the number of words in their corpora, and the seeds that each of them proposed for the Spanish corpus. We also recommended that they should read a few texts to acquaint themselves with the subject field.

3.2.4. Assessing the Quality of the Translations

We assessed the 24 exams following the holistic approach of Robinson (1998), and Robinson, López and Tercedor (2006). Robinson's criterion descriptors allow the identification of the main areas where translation errors occur, relating them to the main phases of the translation process: decoding the source text (meaning errors) and encoding the target text. These areas are the following: (1) content/sense; (2) register, vocabulary, terminology: (3) translation brief and orientation to target text type; (4) written expression.

We assumed that using automatically assembled corpora through Sketch Engine would improve the quality of their translation in relation to content and choice of vocabulary/terminology (areas 1 and 2). Mistakes affecting content would include not only changes in meaning (wrong sense), but also instances where cohesion was not achieved or where the data of the source text had been changed. Therefore, we marked each translation (0-10), using the marks in column A (maximum of 5 points) and B (maximum of 5 points) on the following scale:

Table 2: Criteria to assess translation quality regarding Content and Register/Terminology (adapted from Robinson, López and Tercedor 2006)

	A. Content	**B. Register, vocabulary, terminology**
0	The text fails to meet minimum requirements	The text fails to meet minimum requirements
1	Comprehension limited. Major content errors. Major omissions of ST content.	Choice of register inappropriate or inconsistent. Vocabulary limited with some basic errors. Limited awareness of appropriate terminology.
2	Comprehension adequate. Minor content errors. Some omissions of ST content.	Choice of register occasionally inappropriate or inconsistent. Occasional mistakes of basic vocabulary. Clear awareness of appropriate terminology although some errors.
3	Comprehension good. Minor omissions of less relevant ST content. Over- or under-translation distorts ST content or results in ambiguity	Choice of register mostly appropriate and consistent. Vocabulary effective despite mistakes. Terminology appropriate despite occasional errors.
4	Comprehension very good. Over- or under-translation does not distort ST content or result in ambiguity.	Choice of register appropriate and consistent. Vocabulary effective despite occasional mistakes. Terminology appropriate despite mistakes.
5	Comprehension excellent. ST content, including subtle detail, fully understood.	Choice of register consistently effective and appropriate. Sophisticated, highly effective choice of vocabulary. Terminology appropriate and wholly accurate.

4. Results

After correcting the exams and analysing the questionnaires, the following results were obtained. Despite the fact that the control group had previously translated two texts about swine flu, and the experimental group had not, there was only a very slight difference in the quality of the translations. In general terms, the percentage of errors corresponding to Content and Lexis was very similar.

Table 3: Average marks in both groups as regards Content and Vocabulary/Terminology

	CONTENT (average marks 0-5)	**VOCABULARY / TERMINOLOGY (average marks 0-5)**
Control group	3.7	3.8
Experimental group	4.0	3.4

However, when the data were analysed further, the results were surprising. Firstly, regardless of the approach followed, students who used paper dictionaries obtained better results than students who only used electronic resources. For example, just by looking up the term *neuraminidase* in the *Diccionario crítico de dudas inglés-español de medicina* by Fernando Navarro (Table 4), comprehension of the text was facilitated since this dictionary gives valuable clues about using certain terminology and collocations. The dictionary definition provides us with specific terminology that comes up in the source text, such as *antígeno* (antigen), *hemaglutinina* (hemagglutinin), *neuraminidase* (neuraminidase), *escindir* (cleave), or *residuo de ácido siálico* (sialic acid moieties).

Table 4: Definition of Neuraminidase in the *Diccionario crítico de dudas inglés-español de medicina* (emphasis added)

NEURAMINIDASE
Esta enzima es muy conocida por ser uno de los dos ***antígenos de superficie*** *de los virus de la gripe −***hemaglutinina*** *y* ***neuraminidasa****−, que permiten clasificarlos en H1N1, H3N2, etc. (...) esta hidrolasa que* ***escinde*** *los enlaces glucosídicos entre un* ***residuo de ácido siálico*** *y uno de hexosa o hexosamina (...).*

Secondly, more problems were found in the translation of general language words used in specialized texts such as *target* and *insights* than in the translation of international terms such as *neuraminidase*, *oseltamivir*, *zanamivir*, *Tamiflu*, *NA*, *HA*, *hemagglutinin*, or *virions*. However, it was observed that~~,~~ the use of specialized corpora related to specialized dictionaries, such as the one in Table 3, helped students to better grasp the meaning of the text, thus reducing the number of mistakes relating to sense. That is the reason why the experimental group obtained slightly better results regarding content than the control group.

From the analysis of the answers in Section 3 of the questionnaire, it was ascertained that unfortunately nearly half of the students (52%) stated that they rarely use paper dictionaries. Although the Web presents a wide range of resources for use in general language queries, most specialized resources are still only available in paper format. In the case of medical resources for translation, the Web offers a variety of texts and on-line dictionaries, but these resources are normally insufficient to solve translations problems encountered when translating medical texts from English into European Spanish. Some of these problems are due to lack of expert knowledge, terminological variation, the polysemy of medical terminology, false friends, and so on (Tercedor and López 2012). In our study, some students translated the false friend "fatal" in Spanish as *fatal* instead of its correct equivalent *mortal* or *letal* (deadly). Moreover, when referring to medications, the translation of the commercial names of drugs *(Tamiflu)* and of their non-proprietary names *(oseltamivir, zanadivir)* is not normally found in medical dictionaries; even so, students who had never heard of the international non-proprietary names proposed by the World Health Organisations (WHO) provided the correct translation just by searching the corpora they had compiled with WebBootCaT.

Regarding the potential use of Sketch Engine in future translation assignments, 81% of students insisted on the fact that they would use Sketch Engine again under exam conditions, and that now that they are aware of the benefits of this tool, they no longer plan to do their translation assignments using conventional methods. In fact, 50% of the students declared they usually lost focus and wasted time querying Google. All of them agreed that the best way to make the most of Sketch Engine was to use it in combination with Google. What is really noteworthy is that 50% of the students believed that Sketch Engine was more useful for English than Spanish. One of the reasons given was that it contained more corpora in English with more tokens. Furthermore, when using WebBootCat to compile their own corpora, the number of texts retrieved by Sketch Engine was greater in English than in Spanish, possibly due to the status of English as an internet *lingua franca*. Therefore, the students found more webpages on swine flu available in English. Regarding the time used to finish the translation assignment, 30% concluded Sketch Engine helped them to finish it in less time, and no one stated that it took longer.

Commenting on the perceived advantages of corpora within the Sketch Engine and WebBootCat, 100% of students believed that they were able to retrieve and compile a specialized corpus of more reliable texts than texts retrieved using Google. They also mentioned that it allowed them to rapidly compile corpora and to analyse texts more easily through the use of concordances. In addition, 43% of the students said that they could use the texts in the corpus to acquire expert knowledge on the subject field. As for drawbacks, students highlighted the short time period of the free licence (30 days), and the novelty of the application, which meant that they did not know how to use it to its full potential.

5. Conclusions

Documentation and terminological extraction are among the most important phases of the translation process. Corpora are essential for performing these tasks. The fact that the internet is currently the main source for corpus compilation raises the question of whether the Web should be used for a corpus or as a corpus (De Schryver 2002), and which approach is best in the translation classroom. In this study, we have compared and evaluated as translation aid tools, “automatically built” corpora accessible through Sketch Engine, as opposed to the manual building of corpora in the context of a scientific and technical translation course at university level. We asked two groups of third year students in the Translation and Interpreting Degree Program at the University of Granada (Spain) to carry out a specialized translation assignment on swine flu. One group was requested to do the assignment using automatically built corpora, i.e. Wacky Corpora and corpora compiled by means of WebBootCat accessed through Sketch Engine whereas the other group used conventional methods of manual corpus compilation. Our main objective was to test whether the use of automatic corpora was able to compensate for the students’ lack of specialized knowledge of the subject field and their lack of previous experience in translating texts on this subject.

After evaluating the quality of the translations based on content and choice of the pertinent vocabulary/terminology, we concluded that despite the fact that one of the groups had already translated two texts on swine flu, in contrast to the other group, there was only a slight difference in the quality of their respective translations. Our research thus confirms that the use of automatically assembled corpora, as reflected in Sketch Engine, can compensate for limited knowledge of the subject field and its terminology, and is therefore a useful tool for translators. The results of this research also demonstrate that regardless of the approach followed, the translations of students who used paper dictionaries were generally of better quality than those of students who only used electronic resources. Additionally, we discovered that students who had achieved higher marks in previous assignments also performed better because of their translation competence and their ability to evaluate the reliability of medical sources. Our results coincide with those of Castagnoli, who concludes that the benefit of using automatically assembled corpora is in direct relation with the user's familiarity with the specialized domain, and his/her ability to critically evaluate texts/terms retrieved (2006: 171).

The analysis of the questionnaire showed that the majority of students were pleasantly surprised by the usefulness of Sketch Engine for translation. All of them praised the reliability of texts offered by Sketch Engine, and agreed that the best way to make the most of Sketch Engine was to use it in combination with Google. In fact, 30% concluded that it took them less time to compile the corpus and do the translation, whereas the rest said it took them the same time as with more conventional approaches. Regarding directionality, 50% declared Sketch Engine was more useful for English than Spanish. Moreover, 43% of the students concluded that the resource helped them to acquire expert knowledge about the subject field. At the time of the study, the Sketch Engine platform did not allow queries in parallel texts to retrieve translation equivalents. Since then, it has become possible to search aligned sentences in English and Spanish from parallel corpora available from Sketch Engine (Opus English-Spanish or EuroParl5 English-Spanish). Although these parallel corpora are constrained by text type and topic, it seems that Sketch Engine and on-line corpora are widening their capabilities for translators in the sense of facilitating the search for equivalent segments.

Finally, from this study we can conclude that the new tools that can be used to consult corpora compiled following the WaC approach, such as Sketch Engine, can be of tremendous help for translation, translation teaching, terminology and terminology teaching. As such, in our opinion the learning of these tools should be incorporated in the new degree programs in Translation and Interpreting. In this regard, thanks to the research presented here, the use of Sketch Engine has been recently incorporated into the Terminology course currently taught in the third year of the degree of Translation and Interpreting at the University of Granada (Spain). It is our belief that the degree programs in Translation and Interpreting must be continuously updated to correspond with current trends and provide students with resources that can make their work as translators or interpreters easier. As such, in a subsequent phase, we plan to carry out a similar analysis on another specialized domain to verify whether the results obtained are similar. Nevertheless, it is our belief that, even though these new tools provide more options for the translator, specialized dictionaries on paper should not yet be set aside.

Mbuendia(a)ugr.es, clarailr(a)ugr.es

Acknowledgements

This research has been carried out within the framework of the project RECORD: Representación del Conocimiento en Redes Dinámicas [Knowledge Representation in Dynamic Networks, FFI2011-22397], funded by the Spanish Ministry for Science and Innovation.

The authors thank Dr Pamela Faber for proofreading the article.

References

Aston, Guy (2009) 'Foreword'. In Allison Beeby, Patricia Rodríguez-Inés and Pilar Sánchez-Gijón (eds) *Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate.* Amsterdam/Philadelphia: John Benjamins, ix-x.

Austermühl, Frank (2001) *Electronic tools for translators*. Manchester: St. Jerome.

Baroni, Marco and Silvia Bernardini (eds) (2006) *Wacky! Working papers on the Web as Corpus*. Bologna: GEDIT.

Baroni, Marco, Adam Kilgarriff, Jan Pomikálek and Pavel Rychlý (2006) 'WebBootCat: instant domain-specific corpora to support human translators'. In *Proceedings of EAMT 2006 - 11th Annual Conference of the European Association for Machine Translation*. Oslo, 247-252.

Baroni, Marco and Silvia Bernardini (2004) 'BootCaT: Bootstrapping corpora and terms from the web'. In *Proceedings of LREC,* Lisbon (Portugal). Available online at: [http://sslmit.unibo.it/~baroni/publications/lrec2004/bootcat_lrec_2004.pdf] (accessed 13 May 2010).

Beeby, Allison, Patricia Rodríguez-Inés and Pilar Sánchez-Gijón (2009) *Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate.* Amsterdam/Philadelphia:. John Benjamins.

Bernardini, Silvia, Marco Baroni and Stefan Evert (2006) 'A WaCky introduction'. In Marco Baroni and Silvia Bernardini (eds) *WaCky! working papers on the web as corpus*. Bologna: GEDIT, 1-32.

Bowker, Lynne (2000) 'Towards a methodology for exploiting specialized target language corpora as translation resources', *International Journal of Corpus Linguistics,* 5(1): 17-52.

Bowker, Lynne (1998) 'Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study', *Meta* 43(4): 631-651.

Buendía Castro, Miriam and José Manuel Ureña Gómez-Moreno (2009) 'Parameters of evaluation for corpus design', *International Journal of Translation,* 21: 73-88.

Castagnoli, Sara (2006) 'Using the Web as a Source of LSP Corpora in the Terminology Classroom'. In Marco Baroni and Silvia Bernardini (eds) *Wacky! Working papers on the Web as Corpus*. Bologna: GEDIT, 159-172.

De Schryver, Gilles Maurice (2002) 'Web for / as corpus: a perspective for the African languages', *Nordic Journal of African Studies,* 11(2): 266-282. Available online at: [http://tshwanedje.com/publications/webtocorpus.pdf] (accessed: 2 May 2010).

Fletcher, William H. (2012) 'Corpus Analysis of the World Wide Web'. In Caroll A. Chapelle (ed) *Encyclopedia of Applied Linguistics*. Wiley-Blackwell. Available online at: [http://www.encyclopediaofappliedlinguistics.com] (accessed 2 April 2010).

Fletcher, William H. (2007) 'Concordancing the web: promise and problems, tools and techniques'. In Marianne Hundt, Nadja Nesselhauf and Carolin Biewer (eds) *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 25-45.

Fletcher, William H. (2004) 'Facilitating the compilation and dissemination of ad-hoc web corpora'. In Guy Aston, Silvia Bernardini and Dominic Stewart (eds) *Corpora and Language Learners*. Amsterdam: Benjamins, 275-302.

Hundt, Marianne, Nadja Nesselhauf and Caroline Biever (2007) 'Corpus linguistics and the web'. In Marianne Hundt, Nadja Nesselhauf and Carolin Biewer (eds) *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 1-5.

Jiménez Piano, Marina and Virginia Ortiz-Repiso Jiménez (2007) *Evaluación y calidad de sedes web*. Gijón: Ediciones Trea, S.L.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz and David Tugwell (2004) 'The Sketch Engine'. In *Proceedings Euralex*. Lorient (France): Université de Bretagne-Sud, 105-116.

Kilgarriff, Adam and Gregory Grefenstette (2003). 'Introduction to the special issue on the web as corpus', *Computational Linguistics*, 29(3): 333-347.

López Rodríguez, Clara Inés, and Miriam Buendía (2011). 'En busca de corpus online a la carta en el aula de traducción científica y técnica'. *Trans-kom* 4, no. 1: 1-22.

López-Rodríguez Clara Inés and María Isabel Tercedor-Sánchez (2008) 'Corpora and students' autonomy in scientific and technical translation training'. *Jostrans* (*Journal of Specialized Translation)*, 9. Available online at: [http://www.jostrans.org/issue09/art_lopez_tercedor.php] (accessed: 3 June 2010).

López-Rodríguez, Clara Inés (2002). 'Training translators to learn from news report corpora: the case of Anglo-American cultural references'. In Belinda Maia, Johann Haller and Margherita Ulrych (eds) *Training the Language Services Provider for the New Millenium*. Oporto: Faculdade de Letras Universidade do Porto, 213-222.

Lüdeling, Anke, Stefan Evert and Marco Baroni (2007) 'Using web data for linguistic purposes'. In Marianne Hundt, Nadja Nesselhauf and Carolin Biewer (eds) *Corpus Linguistics and the Web*. Amsterdam/New York: Rodopi, 7-24.

Ntoulas, Alexandros, Junghoo Cho and Christopher Olston (2004) 'What's new on the web? The evolution of the web from a search engine perspective'. In *Proceedings of the 13th International Conference on World Wide Web*. New York: ACM Press, 1-12.

Robinson, Bryan, Clara Inés López-Rodríguez and María Isabel Tercedor-Sánchez (2006) 'Self-assessment in translator training'. *Perspectives: Studies in Translatology*, 14(2): 115–138.

Robinson, Bryan (1998) 'Traducción transparente: métodos cuantitativos y cualitativos en la evaluación de la traducción'. *Revista de Enseñanza Universitaria,* Número extraordinario, 577-89.

Scott, Mike. (2008). WordSmith Tools version 5. Liverpool: Lexical Analysis Software. Available online at: [http://www.lexically.net/wordsmith/] (accessed: 3 June 2011).

Sinclair, John (2005) 'Corpus and text- basic principles'. In Martin Wynne (ed) *Developing linguistic corpora: A guide to good practice*. Oxford: Oxford Books, 1-16.

Tercedor, Maribel, and Clara Inés López-Rodríguez (2012). 'Access to Health in an Intercultural Setting: The Role of Corpora and Images in Grasping Term Variation'. *Linguistica Antverpiensia*, 11: 153–174.

Varantola, Krista (2003) 'Translators and disposable corpora'. In Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds) *Corpora in translator education*. Manchester: St. Jerome, 55-70.

Volk, Martin (2002) 'Using the web as corpus for linguistic research'. In Renate Pajusalu and Tit Hennoste (eds) *Tähendusepüüdja. Hatcher of the Meaning. A Festschrift for Professor Haldur Oim*. Tartu: University of Tartu.

Zanettin, Federico, Silvia Bernardini and Dominic Stewart (eds) (2003) *Corpora in Translator Education*. Manchester-Northampton: St Jerome Publishing.

Zanettin, Federico (2002) 'DIY Corpora: The WWW and the Translator'. In Belinda Maia, Jonathan Haller and Margherita Urlrych (eds) *Training the Language Services Provider for the New Millennium*. Porto: Facultade de Letras, Universidade do Porto, 239-248.

Zanettin, Federico (1998) 'Bilingual Comparable Corpora and the Training of Translators'. *META*, 43(4): 616-630.

Appendix

A. Information about the Student

1. NAME
2. Age
3. Mother tongue
4. First foreign language
5. Second foreign language
6. Third foreign language
7. My level of English: □ Level 2 / Low intermediate level of English / B1 / PET □ Level 3 / High intermediate level of English / B2 / FCE □ Level 4 / Advanced level of English / C1 / CAE □ Level 5 / Proficient in English / C2 / CPE
8. Keyboard speed/precision: □ Very good □ Good □ Average □ Bad
9. I accessed the 3rd year of the degree in Translation and Interpreting, from another degree
10. Previous university-level studies

B. Previous Knowledge about the Subject and Documentation

11. I have read parallel texts about swine flu before carrying out the translation assignment
12. I have already translated texts about swine flu in another subject
13. I have used Google to check out the terminology or to document myself for this translation assignment
14. I usually compile a corpus in electronic form to carry out my translation assignments
15. What tools do I use to compile my corpora and what steps do I follow?
16. I usually analyze my corpora with a lexical analysis tool (which generates concordances, frequency lists, etc.) for documenting myself and for extracting information about terminology or phraseology:
17. Do I normally analyze my corpus manually to document myself about the subject or do I use lexical/ corpus analysis software instead?
18. Electronic dictionaries consulted
19. Paper dictionaries consulted

C. Use of Automatically Compiled Web Corpora through Sketch Engine to Translate the Text Provided

20. **Corpus in English:** time to compile it and number of words
21. **Corpus in Spanish:** time to compile it and number of words
22. *Seeds* of the Corpus in Spanish
23. Once the teacher gave me the source text, did I compile an additional corpus in Spanish? Time to compile it and number of words *Seeds*
24. Would I use again Sketch Engine under exam conditions?
25. Under exam conditions, I prefer to continue carrying out my translations as before, following a traditional methodology
26. If my Sketch Engine account did not expire, I would continue using it for my translation assignments
27. I prefer to use Sketch Engine in combination with Google
28. Do I usually lose focus in my queries with Google?
29. Regardless of the time used in compiling the corpus, did it take shorter to translate this

assignment?
30. Sketch Engine is more useful for translating assignments into a foreign language
31. Sketch Engine is useful regardless the directionality of the translation
32. Before taking the exam, I have practiced some of the advanced applications of Sketch Engine
33. What applications of Sketch Engine have I used? □ Concordance □ Word List □ Word Sketch □ Thesaurus □ Sketch-Diff
34. In what sense have I found Sketch Engine useful? □ To retrieve more reliable texts than the ones offered by Google. □ To document myself about the subject and to acquire expert knowledge □ To decide upon which term is the most appropriate one □ To check out complementation patterns □ Other (specify)
35. Advantages of Sketch Engine
36. Drawbacks of Sketch Engine
37. Other observations I would like to highlight