

# The Web for Corpus and the Web as Corpus in Translator Training<sup>1</sup>

Miriam Buendía-Castro and Clara Inés López-Rodríguez

University of Granada

**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 distribution made by De Schryver (2002), there are two corpus 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 these two approaches in the context of a scientific and technical translation course at university level. We asked two groups of students in the Translation and Interpreting Degree Program at the University of Granada to carry out a technical translation assignment. One of the groups used the WaC approach, whereas the other group used traditional WfC methods. Our objective was to find out whether the Web as Corpus approach was able to compensate for the lack of subject field knowledge of one of the student groups. We wished to see if the quality of the translations of these students was more or less similar to that of the other group that had previously translated texts in the subject field. 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).

## 1. Introduction

Since 1997, Corpus Use and Learning to Translate (CULT) has been a fruitful area of research (Beeby, Rodríguez & Sánchez 2009; Bowker 1998, 2000; López 2002, López & Tercedor 2008; Zanettin 1998; Zanettin, Bernardini & Stewart 2003). Both corpora and the Internet have been 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 & 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%)<sup>2</sup>. Nonetheless, the growing number of Internet users who speak other languages is in the process of changing the situation (Fletcher, in press; *Internet World Users by Language*, 2009<sup>3</sup>). Because of the vast amount of information offered, the Internet constitutes *a fabulous linguists' playground* (Kilgarriff & Grefenstette 2003: 333), from which Translation can also benefit.

It is widely acknowledged that documentation and 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 the

area of Corpus Use and Learning to Translate (CULT), 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 lot 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).

In any case, 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), when learning foreign languages or translating. This kind of corpus compiled for a particular translation assignment has also been called *ad hoc, virtual corpus* (Ahmad et al. 1994), and *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 (Kilgarriff & Grefenstette 2003; Fletcher 2004, 2007; Baroni, Marco & Bernardini 2006).

In this study, we compare and evaluate these two approaches in the context of a scientific and technical translation course at university level. Firstly, we review the notions of Web for Corpus and Web as Corpus. Secondly, we describe an experiment carried out with two groups of 3<sup>rd</sup> year students in the Translation and Interpreting Degree Program at the University of Granada in order to test which of the two approaches was more beneficial for the completion of a specialized translation assignment on swine flu. 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.

## **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 traditionally used to compile texts in digital format for the subsequent implementation of offline corpora. In this regard, in the field of Translation, the notion of DIY (do-it-yourself) corpus (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 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 manually searching the web for valuable information. 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<sup>4</sup>.

The fact that Internet is currently the principal source of texts for corpus compilation 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 (2001: 52) 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.”

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 & 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.

## **2.2. The Web as Corpus**

There is no agreement on the meaning of the expression, *Web as Corpus (WaC)*. According to the classification of Bernardini, Baroni & Evert (2006), there are three ways of approaching the WaC from a linguistic perspective: (i) the Web as a corpus surrogate; (ii) the mega-corpus or mini Web; (iii) the Web as a Corpus supermarket.

### **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. Some of the most widely known are *WebCorp*<sup>5</sup> (Kehoe & Renouf 2002), *KWiCFinder*<sup>6</sup> (Fletcher 2001), *Linguistic's Search Engine* (Elkiss & Resnik, 2004), *WebCorpus*<sup>7</sup> (Fletcher 2007) or *Corpeus*<sup>8</sup> (Leturia et al. 2007). 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.

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 & Biever 2007: 2-3). Regarding popularity, Fletcher (in press) 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 fragile. The information on 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<sup>9</sup>, 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, though only 5% of *new* pages have new content.

Because of these limitations, it is necessary to develop a search engine for linguists (Lüdeling, Evert & Baroni 2007). 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 search engine 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 & Evert 2006: 14).

The ideal method would be to compile a corpus directly from the web without having to trust a search engine to automatically download documents. 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; Kilgarriff 2003; Fletcher 2004, 2007). Various research groups are currently working on the implementation of this type of system (e.g. Webcorp project, GlossaNet<sup>10</sup> project, and the Wacky<sup>11</sup> project from the University of Bologna-Forli).

We shall now briefly describe the Wacky project since some of its corpora were accessed by our subject population, who used the Sketch Engine interface as part of our experiment. The main objectives of Wacky (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). They are currently working on the Spanish corpus.

### **2.2.3. The Web as a corpus supermarket**

Finally, the web can be perceived as a *supermarket* where a corpus can be selected and acquired. Internet users go to the web to search for texts on a search engine. Users select their texts and download them to create a corpus. This approach, which is often adopted by translators, has much in common with traditional corpus compilation methods (Web for Corpus). However, in the case of the Web as a corpus supermarket, the initial stages of the process are done semi-automatically.

There are valuable tools for translation that permit users to quickly and automatically compile corpora from the web. For example, the BootCat (*bootstrapping corpora and terms from the web*) toolkit quickly provides translators with knowledge of the terminology of a given specialized domain (Baroni & Bernardini 2004). Such information is often not found in general or specialized dictionaries. In fact, even if specialized dictionaries do exist for a particular domain, they are difficult to find, and by the time they are published, likely to be out of date (Baroni et al. 2006).

WebBootCat (Baroni et al. 2006) is a version of the BootCat tools. It is a web service to aid translators by quickly producing corpora for specialist areas, in any language, from the web. The application does not have to be downloaded, but can be easily accessed with the corpus analysis tool, Sketch Engine<sup>12</sup> (cf. Kilgarriff et al. 2004). 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 incorporating word sketches<sup>13</sup>, grammatical relations, and a distributional thesaurus. As can be seen in Fig. 1, 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 provided by Sketch Engine in order to work with it, 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, and work with them from a linguistic perspective.

The screenshot shows the Sketch Engine interface. On the left is a navigation sidebar with the LEXCOM logo and 'Corpus Architect' branding. It includes user information (Miriam Buendía, 1,000,000 tokens free, 16 days left), navigation links (Corpora, Create corpus, WebBootCaT, Configuration templates, Sketch grammars, User groups, Settings, Log out), and a Support section (Help, Report a bug). The main area is titled 'Corpora' and contains a table of available corpora:

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

Below the table is a link 'Show 33 more corpora'. Underneath is a section 'My corpora' with a table showing 'no corpora' and buttons for 'Create corpus' and 'WebBootCaT'.

Figure 1. Interface of Sketch Engine showing its main facilities

Castagnoli (2006) shows the advantages and limitations of the use of BootCat in a course in Terminology and LSP. She 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 (ibid: 171).

### 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 & Tercedor 2008). However, we designed an experiment that would test whether the WaC or the WfC was more effective within the context of an actual Scientific and Technical Translation course. Our aim was to discover whether the use of Sketch Engine (Web as Corpus approach) offered more advantages to

translation students than the traditional Web for Corpus approach, involving the manual selection of a corpus from the web.

### **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 worse. In our case, the initial assumption was that students who had previously translated two texts on swine flu, and who used Google 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 the 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 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 improved the quality of their translations.

### **3.2. Design of the study**

As part of our study, we asked two groups of 3<sup>rd</sup> 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*<sup>14</sup> under exam conditions. In 2 hours they had to translate 350 words. This was done with access to Internet and to reference materials such as

electronic and paper dictionaries, and corpora. The differences between the groups were the following:

- Students in the experimental group had never translated texts on the 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 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.

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

### **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 would evidently 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.

Finally, Section 3 elicited the opinion of students in the experimental group about the usefulness of 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, 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 mode 21.

On a scale of 1 to 5, their level of English was 4 (Advanced/C1/CAE). All the students, except one, were native Spanish speakers. 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 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. In fact, the majority of them normally consulted on-line lexicographic materials along with electronic dictionaries installed in their computers rather than paper dictionaries.<sup>15</sup>

For this translation assignment, the most popular documentation resources were the following (Table 1):

Table 1. Most popular documentation resources used by the students

<b>Online dictionaries:</b>
<ul style="list-style-type: none"> <li>- WordReference.com: <a href="http://www.wordreference.com/es/">http://www.wordreference.com/es/</a></li> <li>- Reverso Online Dictionary: <a href="http://dictionary.reverso.net/">http://dictionary.reverso.net/</a></li> <li>- The Free Dictionary: <a href="http://www.thefreedictionary.com/">http://www.thefreedictionary.com/</a></li> <li>- Oxford Dictionary of English (OED) Online: subscription of the University of Granada</li> <li>- Merriam Webster Online: <a href="http://www.merriam-webster.com/">http://www.merriam-webster.com/</a></li> <li>- Diccionario de la Real Academia Española [Dictionary of the Royal Academy of the Spanish Language]: <a href="http://www.rae.es">http://www.rae.es</a></li> </ul>
<b>Electronic dictionaries installed on their computers:</b>
<ul style="list-style-type: none"> <li>- Oxford Spanish dictionary (bilingual)</li> <li>- Collins bilingual dictionary</li> </ul>
<b>Bibliographical databases:</b>
<ul style="list-style-type: none"> <li>- Medline</li> </ul>
<b>Paper dictionaries:</b>
<ul style="list-style-type: none"> <li>- Diccionario crítico de dudas inglés-español de medicina by Fernando Navarro.</li> </ul>

### 3.2.3. Use of Sketch Engine

The experimental group attended two sessions of 2 and 1 hours respectively. In these sessions, they learned the differences between the Web for Corpus and the Web as Corpus, and how to

use Sketch Engine, the Corpus Query System described in section 2.2.3. In the first two-hour session, students registered in Sketch Engine<sup>16</sup>, 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 give for improving their use of Sketch Engine, following Castagnoli (2006). 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 the swine flu, one in English and one in Spanish. We told them that they were going to use 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 get acquainted 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 & 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 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), adding the punctuations 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 & Tercedor 2006)

	<b>A. Content</b>	<b>B. Register, vocabulary, terminology</b>
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 the 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

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

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

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

However, when the data were analyzed 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.

This brief description provides us with specific terminology that comes up in the source text, such as *antígeno* (antigen), *hemagglutinina* (hemagglutinin), *neuraminidase* (neuraminidase), *escindir* (cleave), or *residuo de ácido siálico* (sialic acid moieties).

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 Sketch Engine and good specialized dictionaries, such as the one in Table 4, 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:

- 52% of the students stated that they rarely used paper dictionaries.
- 81% of students stated 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.
- Regarding the time used to finish the translation assignment, 30% concluded Sketch Engine helped them to finish it in less time, whereas the rest said that there was no time difference. No one stated that it took longer.
- 50% of the students said that SkE was more useful for English than Spanish. One reason given was that it contained more corpora in English with more tokens. Furthermore,

when using WebBootCat, the number of texts retrieved by SkE was greater in English than in Spanish because there are more webpages in English.

- Regarding the advantages of Sketch Engine, 100% of the students highlighted that they were able to retrieve more reliable texts than the ones offered by 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. Other disadvantages were the impossibility of comparing equivalents<sup>17</sup> as well as the fact that sometimes Sketch Engine does not offer all the results needed. It often excludes texts that could be useful, whereas Google generally offers more hits.

## **5. Conclusions**

Documentation and terminological extraction are among the most important phases of the translation process. In this respect, corpora are essential for performing these tasks. The fact that the Internet is currently the main source for corpus compilation means that the web can be considered as a corpus itself. Nevertheless, there is 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 these two approaches in the context of a scientific and technical translation course at university level. We asked two groups of 3<sup>rd</sup> 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 the WaC (Web as Corpus) approach (with Sketch Engine), whereas the other group used conventional methods of WfC (Web for Corpus). Our main objective was to test whether the WaC approach 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 the 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 Web as Corpus approach, 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.

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 SkE 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.

Finally, from this study we can conclude that computer applications for terminology in the new degree programs in Translation and Interpreting must be continuously updated since these new tools are extremely useful for translators and terminologists.

## Acknowledgements

The authors thank Dr. Pamela Faber for proofreading the article.

## References

- Ahmad, K., P. Holmes-Higgin and S. Raza Abidi (1994). "A description of texts in a corpus: 'Virtual' and 'real' corpora". In W. Martin, W. Mejis, M. Moerland, E. ten Pas, P. van Sterkenburg and P. Vossen (eds) *Proceedings of EURALEX*. Amsterdam: Vrije Universiteit, 390-402.
- Aston, G. (2009). "Foreword". In A. Beeby, P. Rodríguez Inés and P. Sánchez Gijón (eds) *Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate*. Amsterdam/Philadelphia: Benjamins Translation Library 82. John Benjamins, ix-x.
- Austermühl, F. (2001). *Electronic tools for translators*. Manchester: St. Jerome.
- Baroni, M. and S. Bernardini (eds) (2006). *Wacky! Working papers on the Web as Corpus*. Bologna: GEDIT.
- Baroni, M., A Kilgarriff, J. Pomikálek and P. 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, M. and S. Bernardini (2004). "BootCaT: Bootstrapping corpora and terms from the web". In *Proceedings of LREC*, Lisbon (Portugal).
- Available at: [http://sslmit.unibo.it/~baroni/publications/lrec2004/bootcat\\_lrec\\_2004.pdf](http://sslmit.unibo.it/~baroni/publications/lrec2004/bootcat_lrec_2004.pdf) (accessed 13 May 2010)
- Beeby, A., P. Rodríguez Inés and P. Sánchez Gijón (eds) (2009). *Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate*. Amsterdam/Philadelphia: Benjamins Translation Library 82. John Benjamins.

- Bernardini, S., M. Baroni and S. Evert (2006). "A WaCky introduction". In M. Baroni and S. Bernardini (eds) *WaCky! working papers on the web as corpus*. Bologna: GEDIT, 1-32.
- Bowker, L. (2000). "Towards a methodology for exploiting specialized target language corpora as translation resources". *International Journal of Corpus Linguistics*, 5(1), 17-52.
- Bowker, L. (1998). "Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study". *Meta* 43(4), 631-651.
- Buendía Castro, M. and J.M. Ureña Gómez-Moreno (2009). "Parameters of evaluation for corpus design". *International Journal of Translation*, 21, 73-88.
- Castagnoli, S. (2006). "Using the Web as a Source of LSP Corpora in the Terminology Classroom". In M. Baroni and S. Bernardini (eds) *Wacky! Working papers on the Web as Corpus*. Bologna: GEDIT, 159-172.
- De Schryver, G-M. (2002). "Web for / as corpus: a perspective for the African languages". *Nordic Journal of African Studies*, 11(2), 266-282.  
Available at: <http://tshwanedje.com/publications/webtocorpus.pdf> (accessed: 2 May 2010)
- Elkiss, A. and P. Resnik (2004). "The Linguist's Search Engine User's Guide".  
Available at: <http://lse.umiacs.umd.edu/lseuser.pdf> (accessed: 2 April 2010).
- Fletcher, W. H. (2011) (in press). "Corpus Analysis of the World Wide Web". In C.A. Chapelle (ed.) *Encyclopedia of Applied Linguistics*. Wiley-Blackwell.  
<<http://www.encyclopediaofappliedlinguistics.com/>>.
- Fletcher, W. H. (2007). "Concordancing the web: promise and problems, tools and techniques". In M. Hundt, N. Nesselhauf and C. Biewer (eds) *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 25-45.
- Fletcher, W. H. (2004). "Facilitating the compilation and dissemination of ad-hoc web corpora". In G. Aston, S. Bernardini and D. Stewart (eds) *Corpora and Language Learners*. Amsterdam: Benjamins, 275-302.

- Fletcher, W. H. (2001). "Concordancing the web with KWICFinder". In *Proceedings of the 3<sup>rd</sup> North American Symposium on Corpus Linguistics and Language Teaching*. Boston.  
Available at: <http://kwicfinder.com/FletcherCLLT2001.pdf> (accessed: 22 April 2010)
- Hundt, M., N. Nesselhauf and C. Biewer (2007). "Corpus linguistics and the web". In M. Hundt, N. Nesselhauf and C. Biewer (eds) *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 1-5.
- Internet World Statistics (2009). "Top ten languages". Available at: <http://www.internetworldstats.com/stats7.htm> (accessed: 1 April 2010).
- Jiménez Piano, M. and V. Ortiz-Repiso Jiménez (2007). *Evaluación y calidad de sedes web*. Gijón: Ediciones Trea, S.L.
- Kehoe, A. and A. Renouf (2002). "WebCorp: applying the web to linguistics and linguistics to the web". In *Proceedings of the WWW 2002 Conference*. Honolulu.
- Kilgarriff, A., P. Rychly, P. Smrz and D. Tugwell (2004). "The Sketch Engine". In *Proceedings Euralex*. Lorient (France), 105-116.
- Kilgarriff, A. (2003). "Linguistic search engine. Abstract". In *Proceedings of the Workshop on Shallow Processing of Large Corpora*. Lancaster.
- Kilgarriff, A. and G. Grefenstette (2003). "Introduction to the special issue on the web as corpus". *Computational Linguistics*, 29(3), 333-347.
- Leturia, I., A. Gurrutxaga, I. Alegria and A. Ezeiza (2007). "CorpEus, a 'web as corpus' tool designed for the agglutinative nature of basque". In C. Fairon, A. Kilgarriff, H. Naets and G. M. De Schryver (eds) *Building and exploring web corpora*. Louvain-la-Neuve: Cahiers du Central, 69-82.
- López-Rodríguez C.I. and M. Tercedor-Sánchez (2008). "Corpora and students' autonomy in scientific and technical translation training". *Jostrans (Journal of Specialised Translation)*, 9.

Available at: [http://www.jostrans.org/issue09/art\\_lopez\\_tercedor.php](http://www.jostrans.org/issue09/art_lopez_tercedor.php) (accessed: 3 June 2010)

López-Rodríguez, C.I. (2002). "Training translators to learn from news report corpora: the case of Anglo-American cultural references". In B. Maia, J. Haller and M. Ulrych (eds) *Training the Language Services Provider for the New Millenium*. Oporto: Faculdade de Letras Universidade do Porto, 213-222.

Lüdeling, A., S. Evert and M. Baroni (2007). "Using web data for linguistic purposes". In M. Hundt, N. Nesselhauf and C. Biewer (eds) *Corpus Linguistics and the Web*. Amsterdam/New York: Rodopi, 7-24.

Ntoulas, A., J. Cho and C. Olston (2004). "What's new on the web? The evolution of the web from a search engine perspective". In *Proceedings of the 13<sup>th</sup> International Conference on World Wide Web*. New York: ACM Press, 1-12.

Robinson, B., C.I. López-Rodríguez and M. Tercedor-Sánchez (2006). "Self-assessment in translator training". *Perspectives: Studies in Translatology*, 14(2), 115–138.

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

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

Tercedor-Sánchez, M., López-Rodríguez, C.I. and P. Faber (in press). "Working with words: research approaches to translation-oriented lexicographic practice". *TTR: traduction, terminologie, rédaction*, 23.

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

- Volk, M. (2002). "Using the web as corpus for linguistic research". In R. Pajusalu and T. Hennoste (eds) *Tähendusepüüdja. Hatcher of the Meaning. A Festschrift for Professor Haldur Oim*. Tartu: University of Tartu.
- Zanettin, F., S. Bernardini and D. Stewart (eds) (2003). *Corpora in Translator Education*. Manchester-Northampton: St Jerome Publishing.
- Zanettin, F. (2002). "DIY Corpora: The WWW and the Translator". In B. Maia, J. Haller and M. Urlrych (eds) *Training the Language Services Provider for the New Millennium*. Porto: Faculdade de Letras, Universidade do Porto, 239-248.
- Zanettin, F. (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: <input type="checkbox"/> Level 2 / Low intermediate level of English / B1 / PET <input type="checkbox"/> Level 3 / High intermediate level of English / B2 / FCE <input type="checkbox"/> Level 4 / Advanced level of English / C1 / CAE <input type="checkbox"/> Level 5 / Proficient in English / C2 / CPE
8. Keyboard speed/precision: <input type="checkbox"/> Very good <input type="checkbox"/> Good <input type="checkbox"/> Average <input type="checkbox"/> 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 <u>manually</u> to document myself about the subject or do I use <u>lexical/ corpus analysis software</u> instead?
18. Electronic dictionaries consulted
19. Paper dictionaries consulted

### C. USE OF SKETCH ENGINE TO TRANSLATE THE TEXT PROVIDED

20. <b>Corpus in English:</b> time to compile it and number of words
21. <b>Corpus in Spanish:</b> time to compile it and number of words
22. <i>Seeds</i> 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 <i>Seeds</i>
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? <input type="checkbox"/> Concordance <input type="checkbox"/> Word List <input type="checkbox"/> Word Sketch <input type="checkbox"/> Thesaurus <input type="checkbox"/> Sketch-Diff
34. In what sense have I found Sketch Engine useful? <input type="checkbox"/> To retrieve more reliable texts than the ones offered by Google. <input type="checkbox"/> To document myself about the subject and to acquire expert knowledge <input type="checkbox"/> To decide upon which term is the most appropriate one <input type="checkbox"/> To check out complementation patterns <input type="checkbox"/> Other (specify)
35. Advantages of Sketch Engine
36. Drawbacks of Sketch Engine
37. Other observations I would like to highlight

### Notes

<sup>1</sup> This research was funded by a grant received from the Spanish Ministry of Science and Innovation for the project EcoSystem: Single Information Space for Frame-based Environmental Data and Thesaurus (FFI2008-06080-C03-01/FILO).

<sup>2</sup> These results correspond to the investigation carried out by the Union Latina (Latin Union). Available at: [http://dtiil.unilat.org/LI/2007/es/resultados\\_es.htm](http://dtiil.unilat.org/LI/2007/es/resultados_es.htm) (accessed 13 May 2010).

<sup>3</sup> <<http://www.internetworldstats.com/stats7.htm>>.

<sup>4</sup> <<http://www.lexically.net/wordsmith>>.

---

<sup>5</sup> <<http://www.webcorp.org.uk/>>.

<sup>6</sup> <<http://www.kwicfinder.com/KWiCFinder.html>>.

<sup>7</sup> <<http://webascopus.org/searchwac.html>>.

<sup>8</sup> <<http://www.corpeus.org>>.

<sup>9</sup> SERPs stands for *search engine report pages*.

<sup>10</sup> <<http://glossa.fltr.ucl.ac.be>>.

<sup>11</sup> <<http://wacky.sslmit.unibo.it>>.

<sup>12</sup> <<http://www.sketchengine.co.uk/>>. Now Sketch Engine is going to be integrated in a new system called *new Corpus Architect corpus management system*.

<sup>13</sup> Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour (Kilgarriff et al., 2004).

<sup>14</sup> 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>>.

<sup>15</sup> This tendency and the new relations of students towards traditional paper dictionaries has been described in Tercedor, López & Faber (in press).

<sup>16</sup> Free access is only available for 30 days.

<sup>17</sup> In Google, for example, you can search a Spanish term into an English site, by adding, for instance, the domain *.uk*, so that you are able to compare equivalents.