



**HAL**  
open science

## Working with different corpora in translation teaching

Natalie Kübler

► **To cite this version:**

Natalie Kübler. Working with different corpora in translation teaching. Ana Frankenberg-Garcia, Lynne Flowerdew, and Guy Aston. *New Trends in Corpora and Language Learning*, Continuum, pp.62-80, 2011. hal-01134954

**HAL Id: hal-01134954**

**<https://u-paris.hal.science/hal-01134954>**

Submitted on 24 Mar 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## Working with different corpora in translation teaching

Natalie Kübler

### Introduction

Studies on corpus use for translation, translation teaching and translation studies started during the nineties. Baker (1999) investigated translators' behaviour by studying parallel corpora in order to assess the distance between the language of translation (what she called *third code*) and the language of native speaker texts. Around the same time, Aston (1999) outlined the uses of corpora in translation and translation teaching, and the first Corpus Use and Learning to Translate (CULT) conferences took place (Beeby et al. 2009; Bernardini and Zanettin (eds.) 2000; Zanettin et al. 2003). Much work has since been put into creating and enhancing corpora and corpus tools in order to study translation, and to showing how appropriate corpora can help translators find the information they need to translate.

Despite this interest in academic circles, professional translators are still not very keen on using corpora for translation. A survey carried out in 2005 by the European project MeLLANGE<sup>1</sup> (also reported in Bernardini 2006), obtained 623 completed questionnaires, the majority returned by UK professional translators, but also by professionals in France, Italy, and Germany. This survey showed that of the 40.5% who collected domain-specific-texts, most (69.4%) saved these in electronic form and used computer software (mainly word processing packages), to explore them. In contrast, 41.9% said they had never heard of corpora, even if they would be interested in learning more about them (see table 1).

Table 1. Results from the MeLLANGE questionnaire

1. Do you collect domain specific texts?	6. What do you use to search the corpora you
No                      59.5%	<u>use?</u>

Yes	40.5%	Search facility in word processor	65.9%
		Concordancer	19.0%
		Other search tools (specify: Trados, Concordance in translation memory)	14.4%
		UNIX utilities	0.7%
<u>2. How do you collect them?</u>		<u>7. If you do not use corpora, why?</u>	
In electronic form	69.4%	Never heard about them	41.9%
On paper	30.6%	Other answers?	
<u>3. How do you use them?</u>		<u>8. Would you be interested in a service which quickly provided domain- and language- specific corpora tailored to your needs?</u>	
Search them with software	53.1%	Yes	78.6%
Read them	46.9%	No	21.4%
<u>4. Do you use corpora in your translation practice?</u>		<u>9. Would you be interested in a tool for extracting terminology from a domain-specific corpus?</u>	
No	60.2%	Yes	77.9%
Yes	39.8%	No	22.1%
<u>5. If yes, do you use :</u>		<u>10. Would you be interested in learning more about the potential that corpora offer?</u>	
Corpora of the target language?	26.1%	Yes	82.4%
Corpora of the source language?	23.1%	No	17.6%
Parallel corpora?	19.7%		
Domain specific corpora?	15.3%		
Comparable corpora?	13.6%		
General language corpora?	2.3%		

Why do so many professional translators not use corpora? First, not all training syllabuses include corpus use as a skill to be taught. And where it is included, students also learn to use other tools, such as translation memories and Google, which may seem easier to approach. Second, corpus resources are not equally available in all languages and domains, and although query tools have improved considerably over the last decade, they still require specific competences and may not seem user-friendly. Third, corpus skills are never mentioned in job advertisements on the translation market (Bowker 2004). Fourth, the impact of cultural studies seems to have reduced interest in linguistics in the translation community. All of these factors may negatively influence attitudes towards the use of corpora, not just of professional and trainee translators, but also of translator trainers. There is still limited use of corpora in pedagogic settings. As Frankenberg (2010) notes as far as language teaching is concerned, insufficient attention is paid to training teachers to use corpora. This is true of translation teaching as well.

So how can we render the use of corpora more attractive? A great deal of work is being put into providing better tools for term extraction and corpus querying. We will concentrate here on the teaching situation, and ways in which trainee translators can be taught how to obtain relevant information from corpora.

Beeby et al. (2009) propose that the conjunction of corpora and translation in teaching can be seen from two perspectives:

- *Learning to use corpora to translate*, i.e. using corpora as tools and corpus linguistics as a method to find linguistic information useful in the translation process;
- *Learning to translate using corpora*, i.e. studying the process of translating using corpora, as in Castagnoli et al. (2010) who show how using a learner translator corpus in the classroom can lead to raising students understanding of different translation strategies.

This paper illustrates how different types of corpora can be used from the first of these perspectives.

## 1. The activity of translation

We are concerned here with pragmatic specialised translation. Pragmatic translation, defined by Newmark (1988: 133) as a practice taking into account the *reader's or the readership's reception of the translation*, is the type of translation in which corpus use is at its best. A pragmatic translator should know that the source text may not be perfect, and that the most important thing to take into account is its communicative intent (Froeliger 2004). This means that the translator must take the real world into account, and have a global general knowledge of the source and target cultures. Specialized translation will be viewed as the translation of texts written in languages for specific purposes (LSPs), not just as conventionally defined, such as science, medicine, or law, but also in the broader senses of general academic language or general business language. Pragmatically translating LSPs requires not only knowledge of the source and target cultures in general, but also knowledge of very specific areas. Even a very well-educated translator may not know the terminology, phraseology, or even grammar of a particular specialist domain. So what is the solution? Calling on the specialists for help? Becoming specialized in a specific domain? Experts often disagree about interpretations in their field, and for a translator to become specialized can be a frustrating experience, as they will never be as specialized as the true specialists. Using corpora can help the translator to acquire specialized knowledge in a subject area, to discover a specific domain, and to find linguistic information, which enables them to convey the intent of the source text.

The issues of intent, genre and register are central to the translation task, but common linguistic difficulties should not be put aside, or cultural gaps. Students must be trained to think 'bilingually', which means being able to understand the source text (meaning and intent) in the context of a specialised domain, and to formulate what they have understood in the target language. This means finding out about terminology, phraseology, and more delicate questions such as semantic prosody. Training future translators means raising their linguistic awareness, both in the source and target language. They must be taught to avoid literal translations, distortions of meaning, use of the wrong register, and so on. These are all areas where corpora can help, as one of the tools

available in the translation process. As Bernardini (2006) notes, translation is in many ways an ideal field for corpus applications.

This paper focuses on specialized translation, trying to show in what ways corpus use is ideal in this case. Examples from practical situations in the translation classroom will show some advantages and drawbacks of corpus use for the translation process.

## **2. Corpus Types and Corpus Tools**

Depending on the type of task to be performed, different types of corpora may be called for. However, not all may be equally available. What seems the ideal tool for the translator is a parallel corpus, in which source texts are aligned with their translations in the target language. An appropriate parallel corpus can provide the terminology and phraseology necessary for the translation, as well as examples of alternative translation strategies. However, parallel corpora do not exist for many language pairs and domains, and to compile one requires specific competences and is very time-consuming. Comparable corpora, i.e. collections of texts dealing with the same subject in the source language and the target language, *can help overcome problems of 'artificiality' in parallel corpora* (Kübler 2003: 41) and can largely make up for the lack of those (Frankenberg-Garcia 2009). These are much easier to compile, even if the task of extracting bilingual terminology and phraseology from them is more complex.

Varantola (2002) calls small comparable corpora compiled from the Web *disposable corpora*. They may never be re-used, but are nonetheless well-suited for a specific task. Such corpora are collections of texts in a specific subject area, which may belong to different genres. For example, a specialized corpus in physics can consist of papers from specialized journals, textbook materials, and popular science articles. These can be backed up by reference monolingual corpora, such as the BNC for British English, the COCA for American English, or Kosmas for German, to provide information about more general linguistic features.

While all these types of corpora can be used in the translation classroom (and also by

translation professionals), their limited availability represents one of the reasons why professional translators use them so little. English is well served as far as reference corpora are concerned. But French does not have a reference corpus like the BNC. The Frantext corpus, which contains 4000 texts from the end of the 16<sup>th</sup> century up to the 21<sup>st</sup> in arts, literature, sciences and technologies, is hardly a reference corpus for contemporary French. While currently being completed with more contemporary texts, it is not a balanced corpus, even if it contains 210 million running words, and at most may be useful to literary translators. Some newspaper-based tools are available: Thus Glossanet<sup>2</sup> allows users to create and query their own on-line specialized corpora from French and Belgian newspapers; concordances are sent to the user by e-mail. *Les Voisins de le Monde*<sup>3</sup> allows users to search ten years of *Le Monde* and find collocates for arguments which are governed by verbs, nouns, or adjectives, or predicates, i.e. PoS which govern argument collocates, but also find the distributional neighbours of arguments and predicates (i.e. arguments or predicates which share the same collocates).

When working in specialized translation with French as a target language, parallel or comparable specialized corpora are also necessary. All information concerning LSPs, such as terminology and different text types, has to be found in specialized corpora. However, it is not always easy to build a comparable French/English corpus in a specialized domain, as French documents belonging to the same genres as the English ones may be rare. For instance, scientific research articles are rarely written or translated into French, as most French scientists publish directly in English. Textbook materials and popular science articles in French are easier to find. So working with comparable corpora in these areas raises issues of how comparable the corpora really are. In technical domains, comparable French texts may be easier to find, but they are not always devoid of mistakes. There may also be confidentiality issues, such as proprietary information of a company. This is especially a problem for terminology and specialized phraseology.

Zanettin (2002) reported the advantages of compiling do-it-yourself corpora to translate specific documents. His hope that corpora would 'find their place in the translator workstation

together with other corpus resources and computer-assisted tools' (2002: 8) does not yet seem to have been fulfilled, however. Bernardini (2006) called for work in three different areas: *the role of corpus work for awareness-raising*, the construction of *translator-oriented (e-)learning material*, and the fact that *corpus construction and corpus searching tools should be more user-friendly*. But Aston and Kübler (2010) note that the situation has still not changed much. So part of translator education for corpus use must still deal with corpus creation and corpus query tools. Various tools can be used in the classroom, which are equally applicable in professional situations, such as *Wordsmith*<sup>4</sup>, *AntConc*<sup>5</sup>, *ParaConc*<sup>6</sup>, *Xaira*<sup>7</sup>. Let us see how these issues can be faced in the classroom.

### **3. Translation tasks and basic exercises**

Corpora can play different roles at different stages of the translation process: (a) during the documentation phase, in which translators look for initial information on content, terminology and phraseology in the source and target language; (b) during the translation phase, in which translators look for solutions to specific terminology and phraseology problems; (c) during the revision phase, in which they investigate other alternative strategies. Each of these phases can be seen as involving a series of tasks. For instance, the first phase requires:

- identifying the genre and register of the document to be translated;
- collecting a corpus;
- exploring the domain and understanding difficult or unknown concepts;
- acquiring useful information on linguistic points, particularly terminology and phraseology, in both source and target languages.

To perform these tasks, the translator or learner translator must have already acquired a number of concepts in linguistics, and have attained a certain level of linguistic awareness. This is one reason

why corpus linguistics should form part of translator training curricula, as a precondition to using corpora in the translation process. Students must first understand what a corpus is, what types of corpora exist, and what is in them. Frankenberg-Garcia (2010) proposes tasks to raise student awareness of these issues, such as *understanding different corpora*, *formulating corpus queries*, and *interpreting corpus output*. These can all be adapted for translator training, and need not necessarily be linked to translation tasks as such.

### 3.1. Translation tasks

In the two tasks which are described below have been carried out by French-speaking students at the University Paris Diderot in the frame of a Master's in specialised translation for several years (since 2004). Therefore all the examples mentioned here come from real-life classroom situations. , Students have thus access to a variety of corpora and tools that are listed here: the corpora and tools used by French speaking students in translating from English into French are:

## LAST BEFORE PREPRINTS

- a ten-million word corpus of *Le Monde*<sup>8</sup> (which represents one year) and a home-made concordancer using Perl regular expressions;
- the *Les Voisins de Le Monde* web interface to the 200-million-word *Le Monde* corpus ;
- the English/French *Europarl*<sup>9</sup> corpus (Koehn 2005) with *ParaConc*; and
- a series of English/French and DIY comparable corpora in Earth Science<sup>8</sup> and in digital camera technology, which have been compiled by students over the years, and queried either with the home-made online concordancer or with [AntConc](#).

*Task 1: Group translation of a research article in an imposed specific domain: Earth Science*

First-year master's students are assigned specialised research articles in two or three sub-domains of

Earth Science, such as volcanoes, the birth and evolution of mountains, plate tectonics, hydrology, ice, climatology, and mud volcanoes. Each sub-domain and each article are assigned to a group of students. The articles are then divided into sections of about one thousand words, and each student is assigned one section. The aim of the project is to achieve a complete translation of the articles, with a consistent terminology for the sub-domain. The pedagogic objective is to lead students to discover the use of corpora in the process of translating a specialized text in a group translation context.

The task is divided into a series of subtasks, some of which also relate to other courses. These are:

- Defining the genre;
- Collecting a corpus;
- Exploring the domain and understanding difficult or unknown concepts;
- Acquiring bilingual information on domain-specific terminology and phraseology, and on the phraseology of scientific argumentation;
- Conveying information appropriately in the target (native) language;
- Working together to agree on terminology and phraseology; and
- Revising the translation.

#### *Task 2: Individual translation of a specialized text in any domain*

This is a year-long individual project, achieved by second-year master's students, and that deals with a larger text, usually about 5000 words. It must be in a specialized domain, but may be of any genre. The translation process is again divided into a series of sub-tasks.

### **3.2. Identifying genre and register**

Understanding genre is vital to the translation process, because the same genre can present very different linguistic features according to the language. For example, the genre of user manuals uses

a much more formal register in French than in English. In English, giving instructions is done using the second person (*you*), whereas French tends to use the infinitive or a third person form:

Example 1: extracts from a comparable EN/FR corpus on digital cameras

EN

***You** must put these file back after the firmware update has been completed.*

***make** sure you put time aside to learn it properly.*

*The SSFDCs are keyed, so **you** can't insert them backwards.*

***You** can adjust the brightness of the LCD at any time by holding down the DISP button.*

FR

*Les utilisateurs expérimentés **peuvent** contrôler et ajuster la sensibilité de l'appareil.*

*Nikon Capture permet d'**ajuster** la taille de sortie.*

***Mettre** le cordon série sur l'ordinateur et **brancher** la fiche sur l'appareil photo.*

As the translator's first task is to define the genre of the text to be translated, emphasis is placed on raising students' awareness of what genre is and how to distinguish different genres. The main questions to be asked are:

- What is the speaker's purpose and topic?
- Who is the intended audience?

In answering these questions, the student should relate them to linguistic and rhetorical features. Is it a didactic text, in which definitions and explanations are given? If it is a scientific article, then there will probably be few definitions and explanations, as the author will take it for granted that the reader already knows a lot about the subject. If it is a manual for a washing-machine, then the reader will not be presumed to know anything technical about the field, and the translation should be kept as clear and simple as possible. Genre and domain are also important from a terminological point of view: a highly specialized scientific text will require a lot of research on terminology, while an article in a newspaper may require less such work, but more on other points, such as general and

cultural knowledge. Once the genre of the text has been determined, the internal characteristics of the text can be studied. Simple exercises with corpora can give students insights into these characteristics and lead them to understand more about internal linguistic characteristics, which are associated with the genre, such as the difference in addressing the reader in French and English user's manuals, or the use of *nous* for the author in French scientific articles, whereas English tends to use the passive.

Take the word *hypothesis*, which is supposed to occur quite often in scientific articles. Even before collecting a comparable corpus for the specific domain, a general corpus can provide important information. The examples below come from the BYU BNC and the BYU COCA<sup>10</sup>. Asking students to search for the collocates of *hypothesis* in the fiction and the academic subcorpora of BNC and COCA very easily led them to see differences between the two genres. There are far more significant collocates for *hypothesis* in the academic subcorpus than in the fiction one (100 against 24 in COCA, 100 against 12 in BNC). In the fiction subcorpus of COCA, only one of these collocates is not a grammatical word, namely *test*. In the academic subcorpus, almost 50% of the top 24 significant collocates are non-grammatical words, as shown in table 2.

Table 2. The significant 24 collocates of *hypothesis* in the fiction subcorpus of COCA, and the top 24 collocates in the academic subcorpus. The collocates are ordered by significance.

	COCA fiction subcorpus 78,752,154 tokens		COCA academic subcorpus 79,292,295 tokens
	WORD		
1	N'T	1	SUPPORTED
2	YOUR	2	SUPPORT
3	YOU	3	THIS
4	HE	4	TESTING

5	MY	5	(
6	?	6	RESEARCH
7	I	7	NULL
8	"	8	FOLLOWING
9	IT	9	OUR
10	'S	10	FOR
11	A	11	HYPOTHESIS
12	AS	12	BE
13	WAS	13	THESE
14	,	14	SUPPORTS
15	IS	15	WITH
16	.	16	MAY
17	TEST	17	)
18	AND	18	PREDICTS
19	OF	19	IN
20	THAT	20	;
21	TO	21	WHEN
22	THE	22	FROM
23	:	23	:
24	IN	24	THE

The collocates for *hypothesis\_in* in the two subcorpora show a number of differences in register. The academic subcorpus has no pronouns among the top collocates, whereas four pronouns are near the top of the list for the fiction corpus. The degree of formality seems lower in fiction, witness also the presence of contracted forms. These simple results can then lead to a discussion of

other linguistic differences.

### **3.3. Compiling a comparable specialised corpus**

DIY corpora represent a useful documentation resource in specialised translation. Over the years, our students have compiled comparable corpora in a number of specialised domains (Kübler 2003), using the keywords found in the research article on which they are working. As mentioned above, for some domains, especially scientific ones, this poses problems insofar as there are very few research articles written in French. (A parallel corpus is also impossible because research articles are almost never translated into French.)<sup>11</sup> In such cases students are guided to look for PhDs<sup>12</sup> didactic texts, and websites on popular science.

Students are made aware that their corpora may not be fully comparable, and that this can present drawbacks, particularly as far as terminology is concerned. Young French researchers in hard sciences read scientific literature in English but write their PhDs in French. Even though there has been a French scientific terminology for a long time, most new concepts are coined in English and then translated into French almost literally. As PhDs are not produced for publication, this terminology may not be fully reliable. Some didactic texts can be found on university websites, but most of the French texts that staff use in their university teaching are very difficult to get hold of. There is, on the other hand, plenty of popular science in French available on the web, but again this may not be a reliable source of terminology.

### **3.4. Learning more about the domain**

In understanding the source text, not only are language problems at stake, but also cultural ones. In texts that deal with specialized subjects it is necessary to get acquainted with the domain. An expert in the domain can help the translator understand it, but it is not always possible to have an expert at hand, and here corpora can play an important role (see, for example, Maia 2003). In the Earth Science translation task described above, students are required to write definitions for a series of

terms before they start to translate, in order to improve their understanding of the domain. They are then encouraged to search for definitions in their corpora, using linguistic markers to find these (Pearson 1998). In the examples below, the markers *is a* and *i.e.* in English, and *est un(e)* and *c'est-à-dire* in French, were used to find term definitions.

Example 2: results of corpus searches for definitions, using linguistic markers

*is a*

*Gelifluction is a thaw-related solifluction (Matsuoka, 2001) that is controlled by elasto-plastic soil deformation.*

*est un*

*Un bassin avant arc est un bassin océanique situé entre la subduction et la terre, au contraire d'une fosse océanique qui est le bassin provoqué par la subduction elle-même.*

*i.e.*

*an extrusive (i.e. mud-volcanic) rather than an intrusive (i.e. diapiric) mechanism builds up the mud domes...*

*c'est-à-dire*

*Les roches sédimentaires, c'est-à-dire ces roches qui proviennent de la transformation de sédiments comme les sables et les boues...*

### 3.4. Terminology and phraseology

The use of corpora to find term equivalents has been abundantly described in the literature (e.g. Bowker and Pearson 2002; Kübler 2003; L'Homme 2004; Maia 2003). While much current research is focused on the automatic retrieval of term equivalents (see Zweigenbaum 2008, for example), there are as yet no available tools for this. So translators must learn how to query comparable

corpora to locate term equivalents. This methodology is nowadays well-documented and can also be applied to look for phraseological equivalents. Showing students that one of the prepositional collocates for *hypothesis* is *with* led to an exercise aimed at raising their awareness of the different definitions of a collocation (two words with or without a grammatical relationship, which co-occur in a statistically significant manner), and of the existence of phraseological units that appear more often in academic than in fictional texts. Example 3 shows concordances from the BNC academic and fiction subcorpora. A simple gap-filling exercise, in which students have to fill in the blanks using the words in boldface, shows that there are a number of phraseological units associated with *hypothesis* (*help with*, *compatible with*, etc.), and that *hypothesis* + *with* collocation reported by the tool does not necessarily imply a syntactic relationship between node and collocate, since, in the last example from the BNC academic subcorpus, *with* occurs in a different sentence from *hypothesis*.

### Example 3: phraseological units associated with *hypothesis*

#### BNC academic subcorpus

*only that can NAEP not **help with** research hypothesis formation or,*

*evidence which was **inconsistent with** the working hypothesis*

*fasten onto those that **agree with** his hypothesis and overlook those*

*behaviour which is **inconsistent with** the hypothesis of budget.*

*This is **compatible with** the hypothesis that prenatal nutrition affects*

*Our results support this hypothesis. **Reactions with** carcinoembryonic*

*formation in these patients is also **consistent with** this hypothesis*

*proposed hypothesis that Mozart suffered **with** Tourette's syndrome.*

#### BNC fiction subcorpus

*it dishonesty; comparing facts **with a hypothesis**. However, even if*

The next step consists in finding equivalents for these phraseological units, using in this case the corpus in Earth Science. Looking for *hypothèse(s)* preceded by *avec* (as a translation equivalent of *with*) gives the following results, which provide insights as to French phraseological equivalents:

Example 4: *avec....hypothèse* in a hydrology corpus

*compatible avec cette hypothèse Chapitre III. Etude géochimique en corrélation avec cette hypothèse. Ils permettent de mettre en toluène est compatible avec cette hypothèse. Les thiols à courte avec les Mammifères. Cette hypothèse essaye d'expliquer est en accord avec une hypothèse de réajustement tectonique du en accord avec cette hypothèse. sur ce volcan est également en accord avec cette hypothèse qui, si elle s'avère applicable a été réalisée avec une hypothèse de calcul plus réaliste que*

LAST BEFORE PREPRINTS  
Where equivalents are not found in specialized corpora, general corpora can often provide answers

for specialized translations. The following example relates to the translation of *aggressively* in a computer science article entitled 'Index poisoning attacks in peer-to-peer file-sharing systems'. In his comment on translation problems, the student, who had not been able to compile an English/French comparable corpus in this domain because of the lack of research articles in French, noted that *aggressively* was used in two different contexts in the source text:

Example 5: extracts of *aggressively* in two different contexts in a computer science article

*the "copyright industry" (including the music, film, television, gaming, and book publishing industries) is aggressively attempting to curtail the unauthorized distribution of content in P2P file sharing systems*

*Attackers have discovered this vulnerability and are now aggressively index poisoning popular file-sharing systems.*

His translation hypothesis was to use *agressivement*. However, during the revision phase, this was felt to sound incorrect in French, so it was checked against *Le Monde*:

Example 6: *agressivement* in *Le Monde*

*Mussolini a rompu avec la ligne **agressivement** antibourgeoise des premiers faisceaux, au milieu de cette foule **agressivement** banalisée, on trouve quelques par la recherche d'un son cru, **agressivement** dépouillé, son travail de producteur à l'énoncé ce qu'il pourrait avoir d **agressivement** masochiste. N'empêche : même*

In this newspaper corpus, *agressivement* seems to modify adjectives rather than verbs. As its contexts differ substantially from those of *aggressively* in the source text, there is no support for *agressivement* being an appropriate translation. Looking for *aggressively* in the *Europarl* parallel corpus, the interesting equivalent *de manière offensive* was noted, which comes from the domain of war. *Offensif* collocates with *missile, armée, guerre, stratégie* and appears in many metaphors. *Les Voisins de Le Monde* gave the following results:

Example 7: neighbours of *agressivement* in *Les voisins de le Monde*

*conforme autoritaire xénophobe **virulent** radical irresponsable brutal inacceptable spontané opposer neutre efficace cibler intelligent muscler inadmissible arbitraire contraire généreux provocateur souple violent concerter semblable volontariste préventif juger unilatéral pervers provocant dangereux digne contradictoire inhabituel consensuel énergique défensif civiliser confus ferme similaire scandaleux courageux mesurer prudent identique répressif archaïque transparent paradoxal habile maladroit cohérent suicidaire **offensif** discriminatoire polémique destructeur menaçant injuste pragmatique délibérer rationnel coordonner guerrier nuancer inhumain indigne répréhensible conciliant revendicatif*

Among these neighbours, two have a very close meaning to *agressif*, namely *offensif* (already noticed in the *Europarl* corpus) and *virulent*. Looking for their collocates in *Les Voisins de le Monde* showed that the first is used in war metaphors, the second in other metaphors in the

discourse. This finally led to the following French translations of the source text sentences containing *aggressively*:

Example 8a: text extract of *avec virulence*

*L'industrie du copyright (qui regroupe l'industrie du film et celle du jeu vidéo, les télévisions ainsi que les maisons d'édition musicales et littéraires) les attaque avec virulence.*

Example 8b : text extract of *de manière offensive*

*Les attaquants ont découvert cette faille et s'en prennent de manière offensive aux index des réseaux P2P les plus populaires afin de les empoisonner.*

The expression *avec virulence* can only be used with speech verbs, whereas *de manière offensive*, is found with verbs describing a physical or virtual attack, but not in the discourse.

In formulating the target language text so that it is adapted to its culture, a general monolingual corpus is the student's best friend, as it helps them to find correct collocations, colligations, semantic preferences and prosodies.

### **3.6. Discovering the importance of semantic prosody in the translation process**

Specialized texts in scientific domains are usually thought to have very limited semantic prosodies. It is commonly held that connotations, be they positive or negative, should be banned from scientific articles, which should be purely factual. Semantic prosody is however an important feature of scientific writing, which is linked to evaluation. Students often do not realise the importance of semantic prosody in their own language, and this leads them to awkward translations. Stubbs (2001) describes the English verb *cause* as having a negative semantic prosody: the things that are caused are generally undesirable. This is also the case of the French *causer*. In the 200-million-word *Le Monde* corpus, nearly all of the objects of forms of *causer* have negative connotations:

### Example 9: direct objects of the verb *causer* in French

*dommage irréversible dégât irréparable dommage irréparable dégâts tracas important dégât frayeur dégât matériel r dégât tort ravage dommage désagrément préjudice traumatisme gêne émoi perturbation déception ennui embarras lésion remous trouble embouteillage désastre nuisance chagrin souffrance malheur déboire souci blessure perte infection trou choc désordre décès mort pollution plaie sorbonne destruction inondation scandale douleur émotion mal ruine incendie fracture accident cancer pénurie université surprise chute inquiétude disparition bruit maladie catastrophe préoccupation maximum danger malaise peur retard commune drame problème difficulté tension victime effet risque crise million*

In scientific English, on the other hand, *to cause* does not always have a negative semantic prosody. This is probably the reason why more and more cases of *causer* in scientific French have unannotated objects, which at times seems awkward:

### Example 10: *causer* followed by unannotated objects

*migration de joints de grains, ?causant une évolution progressive (better: générant)*  
*partie distale peut ?causer la superposition de différents régimes (better: provoquant)*  
*surface du sédiment. Le flux ?causé par l'action des marées a été (better: engendré)*

Pinpointing the differences between the general and the specialized corpus helps students become aware of the problem, and avoid using *causer* as the translation equivalent of those English verbs of causation that do not have a negative semantic prosody.

## 4. Conclusions

I hope to have shown how available and do-it-yourself corpora can be used in specialized translation training for all phases of the translation process. The examples illustrate the need to raise students' awareness in using corpora, rather than merely looking for translation equivalents. A

complete translator education cannot avoid linguistic concepts. This can be done simultaneously with discussion of how to find translation equivalents, which is made much easier by using corpora. If corpus use is to become more popular among professional translators, it is essential not only to provide them with appropriate user-friendly, integrated tools, but also to teach them the rudiments of corpus linguistics as part of their training. In the future it should be an obligatory part of all translation schools' curricula. Choosing very recent scientific articles in which students have almost no prior knowledge helps them grasp very quickly how useful corpora are and how important linguistic analysis is. This choice allows translation students to appropriate the tools and methods corpora and corpus linguistics provide. They can then apply those to any domain and genre of text.

## Notes

1. *MeLLANGE Corpora and e-Learning Questionnaire. Results Summary* (Internal Report, 20.06.05). The report is available at <http://mellange.eila.univ-paris-diderot.fr>.
2. Glossanet is available at <http://cental.fltr.ucl.ac.be/projects/glossanet/>.
3. *Les Voisins de le Monde* is available at <http://www.irit.fr:8080/voisinsdelemonde/>.
4. See Scott (2004) for details of the WordSmith tools.
5. *AntConc3.2.1* is freely available software at <http://www.antlab.sci.waseda.ac.jp/software.html>.
6. *ParaConc* is available online at <http://www.paraconc.com/>.
7. *Xaira 1.24* is available online at <http://sourceforge.net/projects/xaira>.
8. The specialised earth science corpora and the small *Le Monde* corpus are available at <http://wall.eila.univ-paris-diderot.fr>.
9. The *Europarl* (European Parliament) parallel corpus is available at <http://www.statmt.org/europarl>.
10. The *BYU*, devised by Mark Davies, is an interface for the BNC and COCA, freely available at <http://corpus.byu.edu/>.
11. Some students also have to compile comparable corpora for Spanish, where they are confronted with similar issues.
12. In France it is compulsory to write PhDs in French, and they are increasingly available on the

Web.

## References

- Aston, G. (1999), 'Corpus use and learning to translate'. *Textus*, 12, 289-313.
- Aston, G. and Kübler, N. (2010) 'Corpora in translators training', in M. McCarthy and A. O'Keefe (eds.). *Routledge Handbook of Corpus Linguistics*. Routledge: London.
- Baker, M. (1999), 'The role of corpora in investigating the linguistic behaviour of professional translators'. *International Journal of Corpus Linguistics*, 4, 2, 281-298.
- Bernardini, S. and Zanettin, F. (eds.) (2000), *I corpora nella didattica della traduzione/Corpus Use and Learning to Translate*. Bologna: Cooperativa Libreria Universitaria Editrice.
- Bernardini, S. (2006), 'Corpora for Translator Education and translation Practice: Achievements and Challenges'. *Proceedings of the L4Trans Workshop at LREC 2006*. [Online] <http://www.sdjt.si/bib/lrec06/>, pp. 17-22. (accessed November 4, 2009).
- Bowker, L. (2004), *Computer-aided Translation Technology*. Ottawa: University of Ottawa Press.
- Bowker, L. and Pearson, J. (2002), *Working with Specialised Language: a Guide to Using Corpora*. London: Routledge.
- Castagnoli, S., Ciobanu, D., Kunz, K., Volanschi, A. and Kübler, N. (2010), 'Designing a learner translator corpus for training purposes', in N. Kübler (ed.).
- Frankenberg-Garcia, A. (2009), 'Compiling and using a parallel corpus for research in translation'. *International Journal of Translation*, vol. XXI-1, pp. 57-71.
- Frankenberg-Garcia, A. (2010), 'Raising teachers' awareness to corpora, in N. Kübler (ed.).
- Froeliger, N. (2004), 'Felix culpa : congruence et neutralité dans la traduction des textes de réalité'. *Méta, journal des traducteurs*, volume 49, 2, juin 2004, pp 236-246.
- Koehn, P. (2005), 'Europarl: a parallel corpus for statistical machine translation'. Paper presented at the 10<sup>th</sup> Machine Translation Summit Conference. Phuket, Thailand, 12-16 September 2005.
- Kübler, N. (2003), 'Corpora and LSP translation', in F. Zanettin, S. Bernardini and D. Stewart

(eds.), pp. 25-42.

Kübler, N. (2011), *Language Corpora, Teaching, and Resources: From Theory to Practice*. Bern: Peter Lang.

L'Homme, M.C. (2004), *La terminologie: principes et techniques*. Montréal: les Presses de l'Université de Montréal.

Maia, B. (2003), 'Training Translators in Terminology and Information Retrieval using Comparable and Parallel Corpora', in F. Zanetti, S. Bernardini and D. Stewart (eds.), pp. 43-54.

Newmark, P. (1988), 'Pragmatic translation and literalism', *TTR : traduction, terminologie, rédaction*, 1/2: 133-145.

Pearson, J. (1998), *Terms in Context*. Amsterdam and Philadelphia: John Benjamins.

Scott, M. (2004), *WordSmith Tools 4.0*. Oxford: Oxford University Press.

Stubbs, M. (2001), *Words and Phrases. Corpus studies of lexical semantics*. Oxford: Blackwell.

Varantola, K. (2003). 'Translators and disposable corpora', in F. Zanettin, S. Bernard and D. Stewart (eds.), pp. 55-70.

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, pp. 239-248.

Zanettin, F., Bernardini, S. and Stewart, D. (eds.) (2003), *Corpora in Translator Education*. Manchester: St. Jerome.

Zweigenbaum, P. (2008), *Proceedings of the LREC 2008 Workshop on Comparable Corpora*, Marrakech, Morocco, 31 May 2008.