Corpora and LSP Translation

Natalie Kübler[1]

0. Introduction

This paper reports on an experimental approach to the training of specialized translators through the application of corpus query tools to textual data. The use of corpora within the frame of translation and languages for special purposes (LSPs) is nothing really new. In specialized translation, translators often have to work as terminologists, as they have to deal with terms (and their translation into the target language) that are specific to a subject area that they may not know very well. People working in terminology have been using paper corpora for a long time, to search for term candidates and their phraseology. The great change in the past years has been characterized by a greater accessibility to electronic corpora and powerful personal computers.

The experiment took place at the *Department of Intercultural Studies and Applied Languages* (" *Etudes interculturelles et langues appliquées*") at University Paris 7. The students on which the experiment was carried out were undergraduates and postgraduates preparing for a diploma in specialized translation and language engineering. The approach I used took the form of projects based on group work.

The first section of this paper describes the students with whom the experiment was led, and the pedagogic objectives of this experiment. Section two describes the corpora and the tools used. The projects run during the academic year are described in section three. The use of corpora and the results obtained are detailed in section four and five.

1.1. Students Groups and Pedagogic Objectives

The students were divided into two groups: undergraduate and postgraduate. Postgraduate students have more experience than undergraduate students. For this reason, the pedagogic objectives were slightly different.

1.2. Students and student subjects

Undergraduate students are in their fourth year of university study (French degree = " Maîtrise "). They usually do not have any professional experience as translators. They are trained, as full-time students, to translate from English into French, and from German or Spanish into French. The vast majority of them have been taught basic computer skills for the humanities, i.e. word processing, spreadsheet software, database use and Web browsing. Most of them have a basic knowlege on General Linguistics, but know nothing about Natural Language Processing or Corpus Linguistics.

Postgraduate students are in their fifth year of study (French degree = " DESS ", a postgraduate professional degree that leads to work with private companies and not to a PhD).
Postgraduate students are divided into two groups :

- IL : " Industrie des Langues " (Language Industry) ;
- TS : " Traduction Spécialisée " (Specialized Translation).

The first one (IL) is more oriented towards the language industry; students are taught more computing skills, such as working with Unix, and using SGML, HTML, SQL, PHP, and perl. SGML and HTML are markup languages, the former being an international standard to format documents,the latter being the most-widely used format to build Web-pages. SQL is a language used to query databases and PHP is used to build databases on the Web. Perl is a programming language that is widely used to process natural language. Multilingual linguists or translators with skill in computer techniques and programming languages are more and more required by companies dealing with NLP, translation, localisation, etc. Students are trained to translate mostly from English into French.
The second group usually translates from English into French, and either German, Spanish or Portuguese into French. As some of them are not native speakers of French, other translation pairs are also accepted (e.g. English into German or Greek into English). The TS group learns fewer computing skills, i.e. basic skills in HTML and nothing about Unix. The two postgraduate groups work part time in private companies, i.e., they spend one week at

1 University Paris 7, kubler@ccr.jussieu.fr

university, and the other one with a company, as interns. They have to carry out various tasks depending on the company they work with : translating technical documents, post-editing, technical writing, building term bases, building electronic dictionaries for Natural Language Processing (NLP) systems, and manipulating texts using different tools.

1.2. Pedagogic objectives

There are two types of pedagogic objectives in the experimental approach described : the general objectives are the same for the undergraduate and postgraduate students ; the particular objectives depend on the knowledge and situation specific to undergraduate or postgraduate studies, which are detailed below. These differences also mean different projects.

1.2.1. General objectives

The students concerned are generally computer-literate at both undergraduate and postgraduate level. As students in specialized translation, they have to build term bases in varied areas:
- health-related subjects,
- variations in the fur of cats,
- wine-making,
- computer science,
- astronomy,
- geology,
- spiders, and so on.

Students are used to working on paper corpora and extracting potential terms manually. The general objectives of the experiment that was carried out consisted in graduating from paper to electronic corpora : the aim was to help students become familiar with electronic corpora and corpus query tools, and to use the Web as a "mega-corpus[2]", browsing it for linguistic (and encyclopaedic) information, as Maia (2000) underlined it. Linking the paper corpora which students have to collect in their terminology projects with electronic corpora and the Web resulted also in the gathering of corpora in the various subject areas students were working on . Another general objective was to help people learn to work in groups on specific projects in order to prepare them for the non-academic world, in which dealines must be met for job completion.

1.2.2. Particular objectives for undergraduate students

The full-time students in their fourth year at university have never translated a whole document. There is no project until the fourth year. As they have always translated text samples, they only have a vague idea of the processes involved in the translation of a document outside the university. Our aim was to show them the tasks that must be done in the " real " world and the steps that must be taken beyond the process of translating itself, such as documentation, terminology, working with experts, proofreading etc.

1.2.3. Particular objectives for postgraduate students

The postgraduate students already know of the processes associated with translation. However, they are still quite wary of computers, especially machine translation (MT) systems. The pedagogic objective here lies in convincing students, especially the TS group, that MT systems can be useful tools (and are no threat) for human translators. They therefore learn how to use a Web-based MT system (Systran) as a tool to help them in the translation process. A side-effect of the argumentation leads them to understand how corpora can help enhance MT systems, which is a bonus, as some of our students work in NLP companies.

2. Corpora and Tools

The corpora students use are accessible on a Web site via internally-grown tools. Our tools were developed at the University of Paris 13 (Foucou et al. 2000) and have migrated to the University Paris 7. They can be accessed on

2 The Web can help find linguistic information, but is of course not as balanced and liable as a carefully chosen collection of texts.

the working Web site of the DESS[3].

2.1. Corpora

The corpora that are accessible to the students consist of monolingual, comparable, and parallel corpora in English and French. There are two type of corpora: general language and specialized language corpora (specialized corpora).

2.1.1. Specialized corpora

We define specialized corpora as a collection of texts dealing with a particular subject area, and written by experts to various types of audience (experts to experts, experts to students, experts to laymen). For the time being, the on-line available[4] specialized corpora deal with computer science, digital camera, gene therapy, and video editing. The underlying philosophy of our corpus collecting principles is to take advantage of the existing resources that are accessible on the Web. For the projects described below, the specialized corpora that were used were the Computer Science and the Digital Camera corpora.

Computer Science: Our computer science corpus has been collected to teach computational English to French-speaking students in computer science (Foucou & Kübler 2000). The corpus must therefore be representative of the different genres computers science students are confronted with: OS manuals, Internet manuals, newsgroups, specialized dictionary, computer science jargon. This corpus was also used by our postgraduate students in the project that will be described below. The available corpora in this domain are the following :

- the Free On-Line Dictionary Of Computing: 500'000 words,
- the Internet Request For Comments (RFCs): 8,5 million words,
- the Unix manual (man): 1,6 million words,
- articles from *Wired*: 100'000 words,
- mails from computer science newsgroups: 100'000 words,
- the Linux HOWTOs: 500'000 words. The HOWTOs have been aligned with their French translation and thus are considered as a translation corpus.

Digital Camera : All the documentation used comes from around twenty Web sites either in French or in English, and from user manuals. They can be accessed separately, although not all are accessible outside the university, for copyright reasons. The size of this corpus is around 400 000 words.

2.1.2. "General" corpora

Our aim is not to collect a general corpus of English or French, but a "general" comparable corpus built of newspaper is available to check the degree of specialization of a term. It is clear that using the British National Corpus or The Bank of English will provide more information about general English. It is though quite convenient to have a "general" corpus at hand. The available newspapers in French and English are the following:

- *The Times* : 3,5 million words,
- *The Herald Tribune*: 1,5 million words
- *Le Monde*: 1 million words

A whole year of each of those newspapers has been collected.

2.2. Tools[5]

The concordancer to which students have access is based on perl-like regular expressions and allows queries containing POS tags. Although the concordancer uses POS tags on plain text, the words are not disambiguated, following thus Sinclair's (Sinclair 1991) view of using corpora. This allows the user to begin with as wide a search

3 The URL is the following: http://wall.jussieu.fr.

4 Some of the corpora that are used are not available outside university for confidentiality reasons with the companies we work with.

5 All the examples in the " Tools " section come from our digital cameras corpora in French and in English.

as possible and to narrow it little by little looking thoroughly at the results.

A common word in the field of digital cameras in French is the term *exposition* (“ exposure ”). Instead of looking up only *exposition*, the first step consists in finding all the sequences including *expos* in French. The search string is the following : \w**expos*\w**, which will retrieve occurrences such as those shown in the following short concordance sample :

1) La synchro lente permet au film d être **exposé** à l éclairage ambiant de l arrière plan et la synchro retard sur le
par exemple quand le sujet est sous- **exposé** (voir le Guide Rapide). Cette section décrit la manière de composer
Trop sombre, une photographie est sous- **exposée** . Trop claire, elle est surexposée. Ouverture L'ouverture des
d obturateur au-dessus de 1/8.000 pour **exposer** correctement. La vitesse de synchro la plus élevée d instantané
expliquez-vous que je n arrive jamais à **exposer** correctement un groupe de personnes ? Si vous constatez que
fréquemment trop claires et apparaissent **surexposées** . Le mode de prise de vue nocturne du modèle PowerShot
400 ISO et un flash intégré assure une **exposition** correcte dans toutes les situations. Une utilisation

This search result is teeming with terms and linguistic information, such as the adjectives *sous-exposé* and *surexposé*. The extracts …*le sujet est sous-exposé*… and …*exposer correctement un groupe de personnes*… reveal a specialized use of the verb *exposer* and its compounds. In general French, the argument in the position of the direct object cannot either be animate nor human with the specific meaning it has here.

The next two search strings describe structures of compound nouns in French, that are common processes used to coin new terms in LSPs : *&N/\w+ de &N/\w+* *&N/\w+ &A/\w+*

The first one defines a multi-word noun composed of a noun followed by the preposition *de* (“ of ”) and followed again by a noun, and the second one a multi-word noun composed of a noun and an adjective. As there is no POS disambiguation, the first results of this kind of search are a little too wide, as shown in example (3) :

2) La grande majorité des **amateurs de photo** numérique pense que la résolution de l
200-1200), vérification et **analyse de l** image en temps réel sur écran couleur à
la focale utilisée, et donc l **angle de champ** , ainsi que les accessoires montés) **
d un contraste élevé et d un **angle de vision** latéral étendu, l appareil photo
bonne qualité en A4. Si l **augmentation de la** taille d image est un plus, la nouvelle
incluse qui raine dans la **base de l** appareil-photo dans l' ensemble (verrou
Effect d images Livré : **cable de connexion** , drivers Le MAVICA MVC-FD 51 Prix :

This problem is readily solved by defining the minimum number of characters a word must contain as: *&N/\w{3,} de &N/\w{4,}*. In this case, words such as *la*, which can be either a definite article or a noun (a key in music), are discarded, and many terms of the domain can be found using this method :

3) couleur 3D, la balance des blancs TTL*, et la **compensation de tons** **. * Le Nikon D1 est le premier appareil
Ce chapitre fournit des informations sur la **composition de photographies** , en utilisant l autofocus, les
Timer. Le dessus du contrôleur sont les **configurations de foyer** , d instantané, continues et d' encadrer. Avec l
du réglage automatique de distance et **confirmation de charge** du flash. fonctionnement multimode (auto, off,)

A tokenizer using perl-like regular expressions is also available. It sorts single word units frequencies, but also all the words containing specific sequences. The sequence below describes all the words ending with the suffix *ible* : *\w+ible*. It results in a list containing words such as “ accessible ”, “ compatible ”, “ flexible ”, “ impossible ” etc.

3. Project Description

The undergraduate and postgraduate students had to carry out and undertake different projects, as their background was different, as were some of the pedagogic objectives.

3.1. Undergraduate project and procedure

The undergraduate project consisted in translating a Web Site on digital cameras, collecting and using comparable corpora in English and French in this specific subject area. General language corpora were also available in English

and French, as they are useful to test the degree of specialization of a term. The Web had to be used as a huge corpus when more linguistic or encyclopaedic information was needed.

The Web site that was chosen consisted of a series of reviews on digital cameras. The class was divided into groups of three, each group being responsible for the translation of one review and one of the following tasks :

- Downloading their review in HTML and plain text format
- Collecting documentation from the Web, user manuals
- Checking the documentation with an expert
- Collecting French and English corpora on digital cameras
- Completing a term base with terms in the two languages
- Translating their review
- Proofreading
- Creating the French Web site.

Each group was also responsible for coordinating one of the different tasks, and for making the information available to the whole group :

- The corpora groups were in charge of merging the corpora collected by the different groups, deleting the duplicate corpora, checking for possible mistakes and submitting the completed corpora to me so that I could integrate those into a concordancer.
- The terminology group set up a term base under ACCESS that other groups added to ; they also had to check for consistency.
- The documentation group collected and commented on the various glossaries found by the other groups, and took charge of copyright problems.
- Two groups were responsible for proofreading the translations.
- The last group was responsible for creating the French Web site, linking the files and checking the HTML tagging (it should be noted however that this task was done in a parallel class in which students were taught the basics of HTML).

Everybody had to do a little of everything and each group was responsible for one task on behalf of the whole class.

3.2. Postgraduate project and procedure

As the aim for the postgraduate students was different, the project consisted in translating a part of the Free On-Line Dictionary Of Computing (FOLDOC[6]) into French using :

- an on-line MT system (SYSTRAN) ;
- comparable corpora in general, and computer science (CS), English and French ;
- parallel corpora (also called translation corpora) in which the source text was in CS English and the target text in CS French ;
- the Web as a corpus and source of linguistic information.

Working in small groups, students elected to translate several entries in the same subject area, such as programming languages, networks, games, e-mail, the Web, operating systems and so on. The first step was to carry out a rough translation using SYSTRAN MT system. The next step consisted in analyzing SYSTRAN's translation problems at all linguistic and non linguistic levels : format, lexicon, terminology, lexicon-grammar, syntax, semantics, and pragmatics. They then had to correct the translations. Working in a parallel course on HTML, each group had to publish their project on the Web.

3.3. A Short Presentation of SYSTRAN

The SYSTRAN machine translation system is based on a transformer architecture : texts are translated sentence by sentence and input sentences are transformed into output sentences with the simplest possible parse. There is no complete parsing, thus no complete representation, of the sentence. A package of lexical and grammatical translation

6 http://www.foldoc.org

rules transform the source sentence into a target sentence, re-ordering words and taking into account phenomena such as agreement.

This system has the advantage of being quite robust, carrying out a translation in any case, even when sentences are not gramatically correct. Obviously, the drawback of this type of system lies in the results which are never sure to be reliable. Some translations are surprisingly good, others have nothing to do with the source text. All well-known and difficult to parse phenomena, such as conjunctions, disjunctions, long-distance dependencies, and global ambiguity pose problems. Simpler issues, such as the position of noun modifiers in French and in English, are not always well managed :

4) EN. The IETF is a **large, open international community** of network designers, operators, vendors and researchers whose purpose is to coordinate the operation, management and evolution of the Internet and to resolve **short- and mid-range protocol and architectural issues.**

   FR. L'IETF est une **grande, ouverte communauté internationale** des créateurs de réseau, des opérateurs, des constructeurs et des chercheurs dont le but est coordonner l'exécution, la gestion et l'évolution de l'Internet et de les résoudre **protocole sous peu et de mi-portée et issues architecturales**.

Apart from the difficulties related to syntactic analysis, issues such as anaphora resolution or those connected with pragmatics and world knowledge are not taken into account in this MT system. As will be shown below, issues raised by translating LSPs can be dealt with using various types of corpora.

4. Working with Comparable Corpora

The first part of this paper dealt with specific projects and tools used as background information, before explaining the role of corpora. Corpora sustain the whole work done in the two projects. I am now going to describe how they were used in those, which shall reveal the experimental approach that was adopted.

4.1. Understanding terms

When reading a text to be translated, translators are liable to find terms they will not understand because they are too specialized. The first use of corpora, here, is to help translators find definitions of specific terms that cannot be found in specialized dictionaries or glossaries. Pearson's (Pearson 1998) approach to finding terms can be adapted in this case.

In the subject area of digital cameras, the term " white balance " denotes a concept that is not accessible to non experts. Following Pearson's method of looking for terms, it is possible to search the following sequence in the digital camera corpora : The sequence *called .{0,30} white balance* allows the user to look up " called " followed by 0 up to 20 characters, spaces, symbols, punctuation marks, etc., followed by " white balance " in lower or upper case. It hits the following definition :

5) These little marvels can automatically balance the color of light electronically so that nothing comes out looking too hot or too cold. It's **called "White Balance"** (WB, for short), which simply means the camera tries to keep white objects fairly white, so they don't take on extreme color casts.

Another definition can be found using the same type of method :

6) White balance : function that allows you to have natural colors by **adapting the whites** to the light.

The following useful remarks can be drawn from these two definitions :

- " white balance " can be abbreviated (" WB ") ;
- the adjective " white " can be nominalized (see " adapting the whites ").

4.2. Linguistic information

Obtaining more linguistic information about this term entails finding concordances for the left- and right-hand side contexts of “ white balance ”, “WB ” and “ whites ”. Here is an extract of selected concordances :

(SHQ-TIFF, SHQ, HQ, SQ-HIGH, SQ)	**White Balance**	(Auto, sunlight, cloudy, tungsten,
Gray Card (18%) and use the camera in	**White Balance**	Hold mode. When ever you move from one
the menu, as well as adjusting ISO and	**White Balance**	settings. Some other items of note
rundown of everything you can change:	**White balance**	(Auto, manual, sunlight, incandescent,
continuous shooting with exposure and	**white balance**	adjusted for every shot,
capabilities, exposure compensation,	**white balance**	adjustments, stitch assist, in-camera
Only the more advanced features like	**white balance**	and continuous modes need the menuing
on this in a second) and more. The	**white balance**	controls have some cool features,
Condition under which Preset	**white balance**	data is reset changed With Ver. 1.2,
away. It actually did better in the	**white balance**	department than my CP950 usually does!
with moderate compression I found the	**white balance**	feature to be a little strange. Instead
now onto another nice feature -- manual	**white balance**	. In addition to auto, and presets for
The only way I've found to get accurate	**white balance**	in this room is to use manual white
shutter, infinite focus, and daylight	**white balance**	. It even tells you to use a tripod --
see quite a few examples of the cloudy	**white balance**	mode.) And that's all the manual

The sequence (white balance)|(WB)|(whites) sorted by the right-hand side context gives a list of possible multi-word units including “ white balance ”:

- potential terms : “ white balance setting ”, “ white balance mode ”, “ white balance control ”, “ white balance compensation ”, “ white balance system ”, “ white balance feature ”, “ white balance department ”, “ white balance thing ”.

In this case, the method consists in first checking whether other terms can be followed by “ setting ”, “ mode ”, “ control ” etc. If this is the case, it leads to a new list of terms. Then possible uses of verbs such as “ to set ”, “ to control ”, “ to compensate ” must be looked up to define the verb structures in which the term can be an argument, and which argumental position it takes. Linguistic intuition tells us that “ white balance department ” and “ white balance thing ” are not terms, but just idioms that can be found in the general language. However, the non-native speaker of English, translating from English into French, can compare the use of <term> “ department ” and <term> “ thing ” in the general English corpus ; the idiomatic structures “ in the glamour department ” or “ in the speed department ” are found and present the same use as “ in the white balance department ”. About the <term> “ thing ” occurrences in general English corpora, such as : “ Mr. Kissinger said **the petition thing** never happened ” confirms the hypothesis that “ white balance thing ” is not a term.

Examining the left-hand side context of “ white balance ” allows the translator to find collocations such as “ accurate ”, “ daylight ”, “ cloudy ”, “ manual ”, “ automatic ”, or “ auto white balance ”. A commonly applied process in English leads to the shifting of the POS of a word from noun to verb :

7) But it will look white to you when you take the pictures because your brain will automatically **white balance** it.

More information can be extracted from a monolingual corpus about only one term. Once translators have listed possible terms, phraseology, derivational processes, etc., the time comes to find the equivalents in the target language (here : French).

4.3.1. Equivalents in the target language

The single general term “ balance ” is usually translated into French by *équilibre*. Searching for *équilibre* in the French corpus on digital cameras does not yield any result. The solution lies in looking up for the other component of the term, i.e. “ white ” and searching for *blanc* in the French corpus. Examining the context in which *blanc* appears, indicates that the equivalent of “ white balance ” is *balance des blancs*. In French *balance* usually means “ scales ”. The same search as for the English term must be done on the French term to find the French verbs that are used, or

the adjectives and their position in French.

The collocates " high " or " low " are often found with the term " exposure ". Possible translations of " high " and " low " in French are usually *haut* and *bas*. In the case of *exposition* (French for " exposure "), the adjectival collocates must be *forte* (" strong ") and *faible* (" weak ").

5. Working with Comparable and Parallel (translation) Corpora

In the field of computer science, postgraduate students had access to comparable corpora, as well as to a parallel (or translation) corpus (Linux HOWTOs translated into French). As their task was to analyze SYSTRANs translation problems and find the correct translations, they already had a list of terms and structures to look for.

5.1 Literal translations

SYSTRAN translates for example the term " firewall machine " into **machine de mur pare-feu*, which is a literal and incorrect translation. Looking for the term in English leads to search for the French translation in the corresponding paragraph. Luckily the parallel corpus comprises a definition of " firewall ". Our concordancer allows the user to look for concordances, and then to have access to the English paragraph in which a selected occurrence has been found, as well as to the corresponding paragraph in French. Here is an extract of what can be found for " firewall " :

8) A firewall is a term used for a part of a car. In cars, firewalls are physical objects that separate the engine from the passengers. They are meant to protect the passenger in case the car's engine catches fire while still providing the driver access to the engine's controls.
   A firewall in computers is a device that protects a private network from the public part (the internet as a whole). The firewall computer, from now on named "firewall", (…)

9) Firewall est un terme automobile. Dans une voiture, un firewall est une pièce qui sépare le bloc-moteur du compartiment passagers. Il est prévu pour protéger les passagers en cas de feu au moteur en maintenant le controle de ce dernier par le conducteur.
   En informatique, un firewall est un péripherique qui protège la partie privée d'un réseau de la partie publique (InterNet en entier). L'ordinateur firewall, ci-après denomme "firewall", (…)

As is often the case in CS French, the English term is not translated. Further searches in French corpora show however that French equivalents for " firewall " do exist, and are used depending on the genre of the text. French computer scientists use the English word when they talk together, or when they write or translate documents for other computer scientists. In the case of a user manual that is written for a wider audience, the terms *garde-barrière* or *coupe-feu* is more widely used.

5.2.1. Terminology problems

MT systems often have trouble with LSPs. Words commonly used in general English have a very different meaning in CS English. On the other hand, translations that standardization organizations have tried to impose are not used by the experts. These two phenomena can be illustrated with the following sentence (EN) and its translation by SYSTRAN (SYS). Looking up for possible translations of " hackers " and " crackers " in the French corpora (FR) shows that the first one is usually not translated, and that the second one is translated as *pirates*, which bears the same negative connotation as " crackers " in CS English.

10) EN. Hackers create, crackers destroy
    SYS. *Les intrus créent, des biscuits détruisent.
    FR. Les hackers créent, les pirates détruisent.

5.2.2. Verb structures and their arguments

Verbs are not widely described in LSPs although they play a most important role. As described in (Kübler et al. forthcoming), specialized verbs may not exist in general language or have completely different meanings.

11) EN. Your BIOS may not allow you to boot to a Linux installed there =
FR. Votre BIOS peut ne pas vous permettre de démarrer un système Linux qui y serait installé

As said above, official standardization bodies sometimes suggest terms that are never used by the expert community. The French Commission for Computer Science Terminology tried to impose *amorcer* to translate *to boot*. A thorough search in our corpora, as well as on the Web, reveals that *amorcer* is not a possible translation.

When a specialized verb has several possible translations, i.e. possible parasynonyms or different uses, parallel corpora provide the user with the possibility of checking each French translation the other way round. The first step consists in listing the possible French translations of the English verb for all its occurrences ; syntactic structures and possible arguments must also be listed :

12) " to boot (strap) " = *lancer*, *démarrer*, *booter*, and not **amorcer*
13) When Linux boots = Quand Linux se lance

The second step requires the translator to look for all the French verbs and check whether their equivalents in English are all the same. The question is : do all the French translations of " to boot ", i.e. *lancer, démarrer, booter*, match " to boot " when looking at the French occurrence first, and then at the English corresponding paragraph ? The answer for *lancer*, which is a possible translation of " to boot ", is negative for example :

lancer = " launch" | " run ", " issue ", " type " (" a command ")

As SYSTRAN works with translation structures that are not complete, many problems arise concerning the verb structures and the types of arguments that are allowed in the different syntactic positions. Working with a parallel corpus and checking on comparable corpora gave the students a clearer picture of complex verb structures in French.

5.2.3. Using the Web

For the two projects, students were required to use the Web for different tasks. When they could not find a translation in the corpora, they had to formulate a hypothesis and try and confirm it by querying the Web. In the term " focus brackets " for example, " brackets " must not be translated by *parenthèses*. The correct translation is not in the digital cameras corpora ; the translation of " focus " however can be found : *mise au point*. Using a search engine on the Web for *mise au point* leads to the complete translation :

14) " focus brackets ": *repères de mise au point* and not *parenthèses* ...

The number of Web pages that are found containing a specific term is a criterion that helps validate a term. In CS English, a " Trojan horse " is a virus that works like the Greek Trojan Horse in Homer. The hypothesis was that the French translation was used as well in CS French. The result of a query about *cheval de Troie* on a search engine deals almost only with CS documents concerned with viruses.

The Web is thus used to complement the usual context of a term that does not have enough context in the corpora or to simply complement corpora that cannot be updated every month in fast-evolving LSPs. A last example in CS and digital camera English that was recently found on the Web, is the term " prosumer ", which is a contraction of " professional consumer " (a blend).

Conclusion and Future Prospects

This paper set out to show how combining various types of corpora and the Web can be introduced in translation training. Using comparable corpora in LSPs helps to overcome problems of " artificiality " in parallel corpora. General language corpora are also necessary to test the degree of specialization of a term. Finally the Web can be of great help in subject areas that change very quickly and in which neologisms are very common.

One of the questions that can be asked is the following : is this adequate training for future translators ? Our department works with private companies in which our students carry out various kinds of tasks that are related to corpora, and from the feedback companies give the university, it seems that the answer is positive.

Translation training however leads to various kinds of jobs, such as, terminology extraction, manipulating documents

using programming languages, building dictionaries for MT systems, localization, translating Web sites, technical writing. Learning to use corpora and corpus-query tools can give future translators the technical skills that were usually not associated with translation, but which seem to be more and more necessary, especially in technical translation.

Current work involves collecting more specialized corpora on various subjects. Postgraduate students have to collect corpora for their terminology and translation projects. They are therefore required to digitize them, when necessary, and tag them so that they can be integrated into our Web-based tools.

As the introduction of corpora in translation training radically changed the way students looked at languages, it led me to work on the development of a general methodological approach to introduce basic knowledge in linguistics and natural language processing and on how to use corpora in the fields of terminology and specialized translation.

REFERENCES

Arnold, D., L. Balkan, R. Lee Humphreys, S. Meijer, L. Sadler. (1994). *Machine Translation : An Introductory Guide*. Oxford : Blackwell.

Foucou P.-Y., N. Kübler. (2000): "A Web-based Environment for Teaching Technical English". Lou Burnard and Tony McEnery (eds.) *Rethinking Language Pedagogy: Papers from the Third International Conference on Language and Teaching.* Peter Lang GmbH : Frankfurt am Main.

Kübler, N., P.-Y. Foucou. (forthcoming) : " Teaching English Verbs With Bilingual Corpora : Examples in the Computer Science Area ". in S. Granger (ed) : *Corpus-Based Approaches to Contrastive Linguistics and Translation Studies.* Rodopi : Amsterdam.

Maia B. (2000) " Making Corpora : A Learning Process ". in S. Bernardini and F. Zanettin (eds) : *Corpus Use and Learning to Translate,* 47-60. CLUEB : Bologna.

Pearson, J. (1998). *Terms in context*. Amsterdam : John Benjamins.

Sager J. (1994). *Language Engineering and Translation : Consequences of Automation*. John Benjamins : Amsterdam.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford : Oxford University Press.

Wichmann, A., S. Fligelstone, A. McEnery and G. Knowles (eds) (1997). *Language Corpora and Teaching*. Longman : London.