Published in LSP & Professional Communication Volume 5, Number 1, April 2005 - ISSN 1601-1929

REPORT:

The CIBLSP Project: Using Electronic Corpora to Investigate Specialised Bilingual Terminology

Nathalie Arlin, Amélie Depierre, Pascaline Dury,
Amélie Josselin, Susanne Lervad and Claire Rougemont
Research Centre for Terminology and Translation (CRTT)
Université Lumière Lyon 2, France

Introduction

Although the compiling and analysing of general language corpora has been common practice for some time now, specialised language corpora are still scarce. The CIBLSP (*Corpus Informatisés Bilingues de Langues de Spécialités*) Project presented in this article has been started at the CRTT by a researcher team of six, in order to investigate scientific English and French, and thus provide much needed information on specialised translation.

It is based on compiling a bilingual (French and English) comparable electronic corpus, in five specialised fields of knowledge: volcanology, medicine, pharmacology, drugs and ecology. Each specialised field represents a sub-corpus of the overall project, and each researcher of the team compiles a different sub-corpus.

As explained below, this project is based on common compiling criteria and a common methodology regarding the sampling and the analysis of the documents. The ultimate objective of the overall project is to give a better picture of terminological links across specialised fields, and to design better tools for investigating and teaching ESP (English for Special Purposes), and specialised translation.

However, although the project is currently underway, CIBLSP is still an in-house corpus at the University Lyon 2. This explains why this paper provides detailed information on the creation of the project and the objectives it follows in the long run, but only gives glances at the work achieved until now, and at the first results obtained so far (the emphasis has been put on the results achieved in the field of volcanology and medicine, the other fields of the project being only briefly presented here).

1. Objectives of the Project

Above all, the project consists in compiling a set of large comparable computerised corpora in English and in French, in the five above-mentioned sub-fields (henceforth called *sub-corpora*).

As explained in the introduction, although each member of the team pursues her own specific goals in the project, the global corpus has been based on common pragmatic and theoretical objectives, which are the following:

1.1. Glossary and Dictionary Making

The most obvious purpose of building a corpus is to extract terminological and terminographic information, i.e. to establish which terms are actually in use and therefore suitable for recording in glossaries and dictionaries.
In addition, analysing the compiled documents from a diachronic point of view should make it possible to study the evolution of terms and concepts, from the time when they appear in a given language to the moment when some of them possibly disappear, together with the semantic or conceptual changes that might affect them.

1.2. Teaching and Translation

The sub-corpora will also be carefully exploited in order to improve specialised languages and terminology teaching methods, by providing genuine examples of terms in context or lexical statistics.
In the long run, the sub-corpora can be made part of translation classes and used as translation tools as detailed below.

1.3. Semantic and Theoretical Objectives

Cross-field analyses of the sub-corpora included in the project should help to detect either common or field-specific phenomena, and hence give a deeper insight into the phraseology of specialised languages.

Moreover, an etymology-based method for calculating the level of specialization of terms and texts (Depierre 2004) should make it possible first to compare the sub-corpora of the project at similar levels of specialisation, (for example comparing specialised papers on the one hand, and popular science articles on the other), and secondly to point out constant or shifting patterns in the use of more or less specialised terms.

Thanks to information extracted from each sub-corpus, various issues will be tackled, such as trying to back up the theoretical assumption that concepts are mobile entities, which can be borrowed and used in different fields and in different communication contexts (non-specialised / specialised), or to prove that translations need to be conceptually accurate, and that it takes a good knowledge of concepts to translate terms properly.

2. Methodology

2.1. Constitution of the Sub-Corpora

The main asset of the CIBLSP project is the common methodology. The coherence and the quality of the project are ensured by the methodology chosen in order to constitute each sub-corpus.

Indeed, the sub-corpora are being compiled according to the following methodological principles: They are bilingual (English/French) and comparable for all the fields included in the project.

A comparable corpus consists of sets of texts in different languages that are not translations of each other. Moreover, the word *comparable* is used in this paper in order to indicate that the texts in both languages have been selected because they have some characteristics in common. According to Altenberg and Granger (2002: 7-8) “comparable corpora consist of original texts in each language, matched as far as possible in terms of text type, subject matter and communicative function”.

No geographical variety of English (British, American, etc.) has been given preference, but the documents selected are texts written by native speakers or texts written by non-native speakers but reviewed by international selection committees. Although each researcher pursues her own goals and choice of documents for each field and for each language, we all follow Bowker and Pearson’s basic rules (2002) about how to constitute and analyse a corpus. The compiling is therefore based on "choice but not chance". In other words, we collect the texts in order to follow the global objectives of the project, instead of trying to find a conducting line in a set of various documents accumulated by chance.

The sub-corpora will be of approximately the same size and as homogenous as possible. Each sub-corpus in each language and in each sub-field will comprise approximately 500,000 words.

The texts collected in the two languages represent different levels of specialisation (from very specialised to non-specialised texts, but explicit enough to inform about the structure of a field and its basic principles).

Therefore the user will be provided with a whole range of texts thanks to which it will be possible to study phenomena such as science popularisation and to analyse the terminology used in the various fields represented in the project.

2.2. Tools Used for the Project

In order to have the whole corpus in electronic form, we digitised (i.e. scanned and re-read) those documents that were not already available in electronic form using optical character recognition software such as Omnipage (version 10 and 11) and HP PrecisionScan Pro 2.0.

The automatic terminology extraction from the sub-corpora and the concordance analyses are being made with tools such as Hyperbase, Syntex, Wordsmith Tools, TERMplus and Termwatch.

3. What has been Achieved in Each Field

3.1. Pharmacology

As far as pharmacology is concerned, the building of a comparable electronic sub-corpus is linked to the so-called DIBPHARM project, an English/French pharmacological dictionary project, started as a collaborative terminological activity between linguists and subject field experts, based on a printed corpus.

DIBPHARM can be of great help not only to translators, but also to specialised writers and even to specialists. It will be an electronic tool providing information concerning about 4,000 terms and their use, through definitions, contexts and notes, all developed by the various work teams.

The sub-corpus will be composed of a large variety of reference texts. First, digitising texts from a collection of documents compiled for the initial corpus will allow us to check the validity of the terms manually extracted. The initial sub-corpus consists of specialized books, didactic references, summaries of product characteristics, good manufacturing practices or European procedures for marketing authorisations.

In addition, downloaded Web documents identified as using specific criteria can be used to help to find new terms in the field of pharmacology and to find new contexts for new meanings, or to reject some hypotheses.

Due to the fact that pharmacology is a very wide field of research, our sub-corpus analysis will first be tested on 2 sub-fields: *pharmacokinetics* (i.e. Study of drug disposition in a body) and *pharmacovigilance* (i.e. Post-marketing surveillance).

The sub-corpus of pharmacology currently comprises approximately 200,000 words per language.

3.2. Ecology

The sub-corpus is made of texts pertaining to the field of ecology. But since ecology is a very large and fast-expanding field, the collection of documents has been narrowed down to the following subjects: terrestrial ecosystems, ecological successions, niches, habitats and guilds, species communities and their interactions (especially predation and parasitism). Texts relating to aquatic ecosystems and the ecology of waters are not included, and the political aspect of environmentalism has been left out as well.
The sub-corpus of ecology has been devised in order to study the diachronic evolution of terms and concepts of the field, in French and English. The period

chosen for the sub-corpus covers the 20th century[1], in both languages. The oldest document included in the English part of the sub-corpus dates back to 1903.

The sub-corpus will hopefully provide valuable information on the major concepts of the field (and on the extent to which these concepts have evolved over time), and could be used in order to improve the definitions contained in databases accessible to terminologists and translators. It can also prove to be a valuable tool to observe the migration over time of terms and concepts from specialised to general communication.

The total word count is at present a little under 700,000 words for the English part and as the same approximate number is hoped to be achieved in French, the final sub-corpus of ecology will be around 1.5 million words.

3.3. Drugs

The drug terminology in the CIBLSP project is based on the multilingual terminology work of the AVENTINUS project presented below.

3.3.1. The AVENTINUS Project

The aim of the AVENTINUS Project funded by the European Union in the Linguistic Engineering Program (LE-2238) was to provide drug enforcement departments within the European national police and intelligence organisations with linguistic tools that will help the users overcome cross-language problems. The idea is that they should be able to use their own language when searching documents and databases in foreign languages. Thus five languages are supported: English, German, Spanish French and Swedish. AVENTINUS is not a full-fledged information system, but provides the users with the linguistic tools to be integrated into their existing operational environments. Modularity and integrating capacity are the most prominent features proposed. The main partner of the project is the Europol Drug Unit. The department of SRAAKDATA at the University of Gothenburg is in charge for the website (http://scrooge.spraakdata.gu.se).

French was only included in the second phase of the project and because of a lack of time, less extensively so than the other languages with only around 900 terms extracted from internal communication texts like police reports. The CRTT was therefore later asked to correct and complete the French part of the multilingual term database (GOT) of the AVENTINUS project. Moreover, drug terminology differs from “normal” terminology in a substantial way as it is used not to make communication easier, but rather to hide acts which are illegal and considered as criminal. This difference is even more emphasised by the fact that many of the terms are slang words or argot. Normally, one can expect a terminological environment to cover a rather specific, well-defined field, and to be rather consistent with ambiguity and stability in meaning and also often in growth. The field of drugs, however, includes such opposite areas as street slang, police and customs vocabulary, drug legislation, medical treatment, and complex chemical compounds. New products based upon new chemical compounds (referred to as

designer drugs) are constantly being developed to keep the trade ahead of the legislation since a drug is not prohibited in our society until it is explicitly put on the list of illegal drugs, i.e. classified as narcotic.

3.3.2. Ontological References

GOT contains some 14,000 terms alltogether (mostly English terms -roughly 7,000-, but only around 2,000 Swedish, German and Spanish terms and less than 1,000 French terms). The objective is to reach approximately 2,000 French terms in the completion phase.

Furthermore, the GOT drug term base is connected to a world model ontology, containing seven classes of concepts and their sub-classes. In this dimension, the terms operate on a conceptual level, as each term is immediately linked to the ontology through a restricted set of concepts.

1) DRUG (*substance* and *tool*: 142 FR terms vs. 3,927 EN terms)
2) PERSON (*dealer, user, official, smuggler* and *producer:* 18 FR terms vs .401 EN terms)
3) SOCIAL LOCATION (*hotel, house:* 3 FR terms vs. 71 EN terms)
4) ORGANISATION (*criminal, company, government:* 4 FR terms vs. 108 EN terms)
5) GEOGRAPHICAL LOCATION (*city, province, country, region:* 0 FR terms vs. 71 EN terms)
6) ROUTE (*trade and smuggling geographical patterns*: 0 FR terms vs. 23 EN terms)
7) OTHER

We found an urgent need to complete the drug category with French terms and this work is therefore still ongoing.

3.3.3. Strict and Soft Terminology

The terminology in the AVENTINUS database is twofold: *strict terminology* on the one hand,, i.e. the kind of well-defined and unambiguous terms which are traditionally associated with certain fields like generic and chemical names for substances (like *diacethylmorphine og H* for heroin), and *soft terminology* on the other hand, characterised by metaphors, and street names.

Examples of soft terminology in the field of drug substances and ecstasy are French collocations like: *soleil avec visage souriant, le ya ba, croissant de lune sans visage,* and simple noun terms like *papillon, coeur, Adam, Eve.*

3.3.4. Semantic Relations

As GOT is a relational database, it contains a semantic hierarchy which is mainly constituted by two levels: terms, their *synonyms* (equal senses – same level) and hyperonyms[2] (superior concepts – superior level). Synonym relations in English amount to 21,926 while the French relations amount to 205 only. The synonym group of the English term *marijuana cigarette*, is, to mention one, rather extensive,

with hundreds of synonyms. The French completion of synonyms for marijuana is expected to be around 100.

3.3.5. Overview of the French Sub-Corpus

The data collected for the French sub-corpus come from texts from open sources. The French sub-corpus consists mostly of popular and semi-popular science articles and reports of the governement site (http://www.gouv.fr) and the CNDT (*Centre National de Documentation Sur la Toxicomanie*) in Lyon. Using a report of re-transcribed texts from different professionals, we have included material from the TREND project undertaken by specialist Anne Fontaine, a sociologist of the field. Documentalists working with the information database TOXIBASE at the CNDT helped to find and validate the texts for the sub-corpus.

3.3.6. Automatic Terminology Extraction

Using the extraction tool TERMplusExtract, we have, until now, only extracted terms and related information from the French sub-corpus. Because TERMplusExtract gave too many responses, it extracted around 9,000 potential terms. We are currently establishing criteria to make our selection of terms with a specialist from the TREND project and the person in charge of the GOT database.

The frequency of some terms (like *cannabis* and *ecstacy)* is taken into account along with other criteria such as variation, in order to show neologisms. Frequent verbs are: *consommer*, *risquer*, *pratiquer*, *opiacer*, *troubler* and nouns and adjectives are *espaces festif et thérapeutique.*

The same procedure will be used in a next step for the English part of the sub-corpus.

3.4. Volcanology

3.4.1. Goal of the Study

3.4.1.1. General Goal

As part of a research project which aims at improving the treatment of terms in general-purpose (monolingual and bilingual) dictionaries, two sub-corpora dealing with the field of volcanology have been built to see to what extent the comparison of corpus data with dictionary data can improve the content of general dictionaries in order to meet the needs of users, particularly translators[3].

This field caught our attention especially because it is a good example of a popularised field and because some volcanologists have noticed that its terminology is poorly treated in existing dictionaries.

3.4.1.2. Particular Objectives

Our goal is to improve the treatment of terms both at the *macrostructure*[4] and the *microstructure* level of dictionaries. As far as the *macrostructure* is concerned, we want to see what types of terms should be included in the dictionary (e.g. simple

terms (*lava*, *crater*, *lapilli*) or complex terms (*shield volcano*, *bread-crust bomb*) and where they should be included (something of a problem regarding the complex terms). As far as the *microstructure* is concerned, we are particularly interested in definitions, phraseology (collocations and compounds), examples, and cross-references in *monolingual* dictionaries, and sense indications, translation equivalents (number and accuracy), phraseology (collocations and compounds) and examples in *bilingual* dictionaries.

As a consequence, our research objectives that are corpus-related are the following: (i) extract a list of terms and compare it to the nomenclature of six existing general-purpose dictionaries (two English and two French monolinguals, two English / French bilinguals), (ii) retrieve phraseological units (collocations and compounds), (iii) retrieve defining contexts to help to improve definitions (monolingual dictionaries) and sense indications (bilingual dictionaries), (iv) identify semantic relationships between terms (synonymy, hyponymy, etc.), (v) identify interesting examples, and (vi) identify or verify the equivalent(s) of a term.

3.4.2. Methodology

3.4.2.1. Overview of the Two Corpora

The two sub-corpora we have designed are a comparable corpus, as defined earlier, and a translation corpus, which consists of original texts in one language and their translation in another language, and which, moreover, is bi-directional (translations are in both directions: from English to French and from French to English).
The comparable corpus is used in order to attain all six objectives, while the translation corpus is used mainly for the sixth objective. The main characteristics of the two corpora are summed up in the table joined below :

		COMPARABLE CORPUS	**TRANSLATION CORPUS**
Subject-field		Volcanology	
Languages	**Names**	English and French	
	Authors	Native language speakers	Native language speakers for source language (SL) in all cases ; possibly also for Target Language (TL)
	Geographic Variety	French : FR, (CD) English : US, CD, GB	French: FR *(SL & TL)* English: US *(SL)*, GB *(TL)*
Size		400,000 words / language => total 800,000	100,000 words / language => total 200,000
Time-Period covered		Circa 20 years (1977 - 2002)	Circa 20 years (1979 - 2002)
Type of Texts		- Written texts - Whole texts - Reliable texts - Popular-Science texts	

Table 1. Main Features of the Volcanology Corpora

Several elements of the preceding table are worth commenting upon, but here we will focus on only one aspect - the choice of popular science texts-, and refer the reader to Josselin forthcoming for more details on the other elements as well as the problems encountered when compiling the comparable sub-corpus.

Three main reasons account for our choice of *popular science* texts: (*i*) according to Delavigne (2001), even popular science texts include terms; (*ii*) since a popular science corpus is by definition aimed at non-specialists, terms found in such a corpus should logically be included in general dictionaries aimed at the general public (as opposed to specialised dictionaries); (*iii*) the typical explanatory style of popular science texts provides term definitions that can be used by lexicographers.

These popular-science texts are classified according to two main criteria– *discourse* and *genre*. The corpus represents the following three discourse levels (based on Pearson 1998, and Meyer & Mackintosh 1996): (*i*) "semi-popularised" discourse, written by experts for those with some knowledge of the field (*e.g. Scientific American*– US; *Pour la Science*– FR) ; (*ii*) popularised discourse, written by relative experts for the uninitiated (*e.g. New Scientist*– GB; *Discover*– US; *Science et Vie*– FR) ; (*iii*) instructional discourse, written by teachers for students (*e.g. A Teacher's Guide to the Geology of Hawaii Volcanoes National Park*).

The comparable sub-corpus also contains texts from various genres. First, it covers both running texts and glossaries. The running texts are subdivided into the following categories: journalistic (newspapers and magazines) and non-journalistic (textbooks, books, exhibition texts, Web documents). For more details about the structure of the comparable sub-corpus, see Josselin & Frérot (2004).

3.4.2.2. Use of Corpora

We use the comparable sub-corpus as a starting-point. Thanks to the corpus-based parser Syntex (developed by D. Bourigault, cf. Fabre & Bourigault 2001), we extract terms and related information from the corpus and analyse the corpus data, which we then compare with dictionary data. Then we turn to the translation corpus for research in equivalents. A return to the comparable corpus is often required for further information. Not only is there constant to and fro between the two corpora, but also between the corpora and the dictionaries.

3.4.3. The Findings so Far

As far as objective 1 is concerned, we are extracting a list of 110 English terms and 110 French terms from the comparable sub-corpus (52 simple nouns, 36 noun phrases or compounds, 16 adjectives and 6 verbs for each language). Because Syntex gave too many responses (e.g. it gave us 7,095 potential simple noun terms!), we had to establish a number of criteria to make our selection. For example, we rejected proper nouns and acronyms. Of course, we took into account the frequency of the term in the sub-corpus, along with some other criteria such as distribution in the various sources of the sub- corpus. Since the manual validation of the potential terms suggested by Syntex is time-consuming, the process of term

selection is still ongoing, and comparison of the selected terms with the nomenclatures of existing dictionaries is yet to be done.

Objectives 2, 3, 4 and 6 have been worked upon to some extent in Josselin & Frérot (2004), and Josselin & Roberts (2004). In the former, we focused on bilingual English-French dictionaries[5], in the latter, on monolingual English and monolingual French dictionaries[6].

By studying the treatment of two sets of collocations (the conceptual series *active / dormant / extinct volcano* and its French equivalent; and the collocation *volcano / lava + erupt* and its French equivalent) in two bilingual dictionaries and comparing the dictionary data to corpus data, we found that corpus use can improve the content of general bilingual dictionaries both in terms of quantity and quality: for example, we discovered in the corpus that the verb *erupt* is used in a transitive manner in approximately 10% of the occurrences of the corpus, which is recorded in neither bilingual dictionary under study; we also found that some equivalents recorded in the dictionaries did not appear in the corpus (e.g. *volcan dormant, volcan au repos*).

By studying the definitions of two conceptual series (again, the series *active / dormant / extinct volcano* and its equivalent in French, and the series relating to some volcanic products: *bomb, block, lapilli, ash and dust* and its equivalent in French) provided in monolingual English and French dictionaries, and comparing them to defining contexts found in the sub-corpus, we discovered that the information extracted from the corpus could help to solve some of the inconsistencies contained in the dictionaries. We also found that, although the definitions present in the sub-corpus were rather different from those of the dictionaries, they were not necessarily incompatible**.** A happy medium can indeed be found between the terminological and the lexicographical approaches to defining strategies; for instance, a generic term can be used systematically in order to make the semantic relationships more explicit (thus meeting the terminographic requirements) but can be paraphrased with some defining elements found in the sub-corpus therefore meeting the general-lexicography needs): thus, a paraphrase of the hyperonomic term *pyroclastic* – "ejected lava fragment"- found in the corpus can be used to define *lapilli*.

3.5. Medicine

3.5.1. Medical Terminology for LSP and Translation Teaching

In the field of medicine, a comparable sub-corpus is also being built. In its final stage, it will comprise texts of various levels of specialisation, from those aimed at the general public to very specialised research papers, in a wide range of medical sub-fields.

This corpus is going to be analysed while bearing in mind two main objectives:

From a teacher's viewpoint, introducing Applied Languages (Langues Etrangères Appliquées) fourth-year students at Lyon 2 University to the techniques and practices of corpus linguistics as a part of terminology and translation lectures and tutorials.

From a terminologist's viewpoint, studying synonymy and suppletion of terms and their use in the various contexts represented in the compiled corpora, plus comparing and cross-analysing the different sub-corpora of the CIBLSP project, and the different levels of specialisation.

3.5.2. A Long-Term Experiment with Students

Each year, students are asked to observe what has been done in the previous years by other students and to continue the work by developing one particular aspect of corpus analysis. In the first year of this experiment, each student chose a part of the human body and compiled a mini-corpus of specialised texts using as key words the chosen body part and the diseases that might affect it. The students' work is obviously to be checked through before being exploited as a basis of further research.

The single term *kidney* was studied as an example and a 200,000-word corpus was built, consisting of 62 specialised articles in the field of nephrology published between 1996 and 2003 and taken from the archives of the *New England Journal of Medicine* (*NEJM*). This corpus was given the name KRN62.

This year, which has been year three of the experiment, the students were asked to investigate groups of suppletive synonyms and more particularly their use in context.

Suppletive synonyms are first of all synonyms, i.e. terms which have different forms ("signifiants") but (almost) the same meaning ("signifié"). For example, *postinfectious glomerulonephritis* and *postinfective glomerulonephritis* are synonyms, but they are not suppletive.

A group of two or more synonyms can be considered as suppletive if they are of different etymological origin, native (Anglo-Saxon) or learned (Latin or Greek); they consequently pertain to a different level of specialisation. For example: *kidney stone, renal calculus, nephrolith,* or *stroke, heart failure, cerebrovascular neurologic disease,* or else *skin disease, cutaneous disease, dermatologic disease, dermatological disease, dermatology disease* are suppletive synonyms.

Terms naturally tend to vary in context, and sometimes synonyms are mistaken for variants of terms and vice versa. Several types of variation can be studied from corpora (Depierre, forthcoming).

3.5.3. Towards a Method for Calculating the Level of Specialisation of Terms and Texts

Another concern of corpus analysis is investigating how terms are used in context. The ultimate goal of such a study, along with the global objectives of the CIBLSP project, is to make it possible to compare different sub-fields in different languages, and to highlight similarities, as well as differences, in the use of specialised terms.
In order to do so, we have devised a method for working out the level of specialisation, first of the terms themselves, then of a given text or corpus.

The idea of such a method has its roots in from the observation of several groups of numerous synonyms. Some questions have arisen, such as : Why are there so many synonyms in the medical field? Are they interchangeable? How often and in what sorts of texts are they actually used?

In a group of suppletive synonyms such as the above-mentioned *kidney stone*, *renal calculus, nephrolith*, it is clear that *kidney stone* is the least specialised of the three, and that *nephrolith* is the most specialised. This intuitive statement is consistent with the etymological origins of the morphemes of which the terms are composed. Luckily enough, the terms in this example are homogenous as far as their etymology is concerned, which is not always the case; *renal stone* and *kidney calculus* are hybrids. A more scientific method is necessary, should one need to go beyond intuition.

Step 1: The first step of the method proposed here consists in assigning each morpheme a morphology-related coefficient of 0, 1 or 2. At first, sight this is time-consuming and requires a good command of etymology. Therefore it is difficult, even improbable, to believe that the process might be successfully automated in the near future. However, the calculations are simplified, as only specialised (and therefore relatively infrequent) morphemes are rated higher than 0.

Highly specialised morphemes of Greek origin, such as *nephr-*, *-lith*, *cyt-*, *h(a)em(at)-*, etc., as well as words directly borrowed from Latin, such as *calculus, vena cava*, etc., are rated 2.

Somewhat less specialised morphemes of Latin or Greek origin, which have become part of the English language thanks to a suffix, such as *renal, syndrome, chronic*, etc., are rated 1.

Finally, native morphemes (“vernaculaires”) such as *kidney, stone, blood,* etc., are rated 0, just as all the remaining non-field-specific words are.

Step 2: Once all the morphemes have been rated, the absolute level of specialisation (ALS) of a term can be calculated simply by adding up the coefficients; in order to compare terms irrespective of the number of morphemes, the relative level of specialisation (RLS) can be calculated by dividing ALS by the number of morphemes. To simplify even

further, only root morphemes can be taken into account, excluding affixes. Thus, for *kidney stone* ALS = 0, RLS = 0, for *renal stone* ALS = 1, RLS = 0.5, for *kidney calculus* ALS = 2, RLS = 1, for *renal calculus* ALS = 3, RLS = 1.5, for *nephrolith* ALS = 4, RLS = 2.

The RLS values are represented graphically in table 2:

Table 2: Graph showing the RLS of the suppletive synonyms of *kidney stone*.

------- **0** ---------- **0.5** ---------- **1** ---------- **1.5** ---------- **2** ---------->
kidney stone *renal stone* *kidney calculus* *renal calculus* *nephrolith*

Step 3: Last, but not least, the ALS of a text or a corpus can be calculated as the sum of the coefficients (higher than 0) assigned to the specialised morphemes (Mc) multiplied by the frequency of occurrence of each morpheme (f), according to the formula: $ALS = \Sigma_1^i (Mc_i \times f_i)$.

To calculate the RLS of a text or a corpus, its ALS should be divided by the number of morphemes (Nm), according to the formula: $RLS = \Sigma_1^i (Mc_i \times f_i) / Nm$.

However, these calculations require a clear recognition of the boundaries between morphemes, which is almost impossible to do automatically; a computer program will not recognize the components of *electro-myogram, glomerulopathy, nephropathy, nephritis*, etc., unless a full list of the morphemes is incorporated as a personalised dictionary, complete with their allomorphs, for example: ***abdomen*** */* ***abdomin****al;* ***muscl****e /* *fibro****muscul****ar;* ***haemato****logy*, ***haemat****oma /* ***haemo****dialysis*, ***haem****e /* *an****aem****ia* (***hemato****logy*, ***hemat****oma /* ***hemo****dialysis*, ***hem****e /* *an****em****ia*).

The whole process would be simpler, though less precise, if graphical units are considered instead of morphemes, as computers count words, i.e. graphical units between two blanks.

In this case the ALS' of a text or a corpus can be calculated as the sum of the coefficients assigned to the specialised terms (Tc) multiplied by the frequency of occurrence of each term (f), according to the formula: $ALS' = \Sigma_1^i (Tc_i \times f_i)$. To calculate the RLS' of a text or a corpus, its ALS should be divided by the number of terms (Nt), according to the formula: $RLS' = \Sigma_1^i (Tc_i \times f_i) / Nt$.

For an even simpler calculation, the number of types instead of tokens can be considered, irrespective of their frequency of occurrence, so the ALS'' of a text or a corpus equals the number of specialised terms $\Sigma_1^i Tc_i$. To calculate the RLS'' of a text or a corpus, its ALS'' should be divided by the number of terms (Nt), according to the formula: $RLS'' = \Sigma_1^i Tc_i / Nt$. For our KRN62 corpus, all the above calculations converge to 25% (93% confidence interval), which is much higher than the usual 2% for non specialised texts.

Conclusion

A lot of work remains to be done on the CIBLSP Project before reaching the overall objectives assigned to it, and before yielding the necessary information to analyse in detail the working of the different sub-fields compiled in the global corpus. Nevertheless, carrying out this project has shown that there is still a lot of research to do in the field of terminology and specialised languages that could benefit translators, especially regarding concepts.

We strongly believe that a good translation rests on correct understanding of the concepts involved and how they are "translated" from one language to another by the translator. There is still a need to enhance the current knowledge of how concepts are formed, how they evolve, how they migrate from one field to another and from one culture to another. Then, there is also a strong need to improve the tools that could help the translator to know concepts better (dictionary definitions for instance, as shows the work on volcanology and pharmacology in the CIBLSP Project) before they can translate them properly.

Acknowledgments

We would like to acknowledge gratefully the help received from Alex Laube for reading this paper with special attention to the English language.

Bibliographical references

Ahmad, Khurshid & Margaret Rogers (2001). "Corpus Linguistics and Terminology Extraction". G. Budin & S.E. Wright (eds) (2001). *Handbook of Terminology Management* (2), 725-73.

Altenberg, B. & S, Granger (eds) (2002). "Recent Trends in Cross-Linguistic Lexical Studies". B. Altenberg & S. Granger (eds) (2002). *Lexis in Contrast. Corpus-Based Approaches*. Amsterdam/Philadelphia: John Benjamins.

Arlin, Nathalie, Depierre, Amélie, Dury, Pascaline, Josselin, Amélie, Lervad, Suzanne & Claire Rougemont (forthcoming). "Projet CIBLSP, Corpus Informatisés Bilingues de Langues de Spécialités". To appear in *Applications et Implications en Sciences du Langage, Actes des journées jeunes chercheurs 2002 et 2003*. Paris.

Bowker, Lynne & Jennifer Pearson (2002). *Working with Specialized Language - A Practical Guide to Using Corpora*. London/New York: Routledge.

Delavigne, Valérie (2001). *Les Mots du nucléaire. Contribution socioterminologique à une analyse des discours de vulgarisation*. PhD thesis. Université de Rouen, France.

Dury, Pascaline (2004). "Building a Bilingual Diachronic Corpus of Ecology: The Long Road to Completion", *Icame Journal* (28), 5-16.

Fabre, Cécile & Didier Bourigault (2001). "Linguistic Clues for Corpus-Based Acquisition of Lexical Dependencies". P. Rayson et al. (eds) (2001). *Proceedings of the Corpus Linguistics 2001 Conference*. Special issue of *UCREL Technical Papers* 13. Lancaster, 176-184.

Josselin, Amélie (forthcoming). "Constitution d'un corpus de vulgarisation dans le domaine de la volcanologie : objectifs, méthode et problèmes dans une optique de lexicographie générale". To appear in *Applications et Implications en Sciences du Langage, Actes des journées jeunes chercheurs 2002 et 2003*. Paris.

Josselin, Amélie & Cécile Frérot (2004). "Corpus-Based Terminology Extraction Applied to Lexicography: How can a Popular-Science Corpus Help Improve General Bilingual Dictionaries?". B. Lewandowska-Tomaszczyk (ed.) (2004). *Practical Applications in Language and Computers, PALC 2003*. Lodz Studies in Language, vol. 9. Frankfurt am Main: Peter Lang, 2004, 65-79.

Josselin, Amélie & Roda P. Roberts (2004). "La définition des unités techno-lectales dans les dictionnaires de langue générale : analyse de quelques exemples tirés du domaine de la volcanologie à la lumière d'un corpus de vulgarisation". Paper presented at the 72th ACFAS Conference (Université du Québec à Montréal, Canada, 13-14 mai 2004).

Lervad, Suzanne (2003). "Terminologie de la drogue. Une base de données multilingue : le projet AVENTINUS ", *Turjuman*, 12(1), 145-151.

Meyer, Ingrid & Kristen Mackintosh (1996). "The Corpus from a Terminographer's Viewpoint", *International Journal of Corpus Linguistics* 1(2), 257-285.

Pearson, Jennifer (1998). *Terms in Context*. Amsterdam: John Benjamins.

Viklund**,** Maja (1999). "Information and Information Categories in Aventinus Multilingual Drug Term Database" – *Report from the AVENTINUS project* – Maja Lindfors Viklund, GU-ISS-99-3 Reserach reports from the Department of Swedish, Göteborg University.

Viklund, Maja & Yvonne Cederholm (1999). "Chasing the Dragon – Drug related Terminology in a Multilingual Perspective", *Proceedings of the 9th International Symposium on Lexicography in Copenhagen*, Lexicography, Series Maior, Max Niemayer, Tübingen.

---

[1] Ecology, as we know it nowadays has been an independent domain (distinct from biology, botany and zoology) only since the end of the 19th century, really starting with the founding work of the German zoologist Haeckel (1899) and his coinage of the term *oekologie*.

[2] There are two semantic levels in GOT – the *standard level* (includes synonyms) and the *hyperonym* level. There are about 2,000 hyperonym relations in the database for the English language and only 20 for the French language.
For example: hyperonym 1: *person* and hyperonym 2: *addict* both lead to the term: *heroin addict*.

[3] We are currently analysing the results of a survey about dictionary use designed and carried out in late 2002-early 2003 among three different categories of potential users of terms in general dictionaries: scientists, language professionals (among whom translators are found), and the general public. This should help define more accurately the needs of these particular types of users, which could affect the lexicographer's working methods.

[4] The *macrostructure* of a dictionary is the overall wordlist of the dictionary, while the *microstructure* deals with the internal design of the dictionary by providing detailed information about the words that are included.

[5] The *Harrap's Shorter French and English Dictionary* (2000) and the *Oxford Hachette French-English, English-French Dictionary* (1996), both on Cd-Roms.

[6] The *Petit Robert* (2001) and the *Petit Larousse* (2002) for French, and the *New Oxford Dictionary of English* (2000) and the *American Heritage College Dictionary* (1996) for English. All 4 are Cd-Rom versions.

Published in LSP & Professional Communication Volume 5, Number 1, April 2005 - ISSN 1601-1929

ABSTRACT

The CIBLSP Project : Using Electronic Corpora to Investigate Specialised Bilingual Terminology

Nathalie Arlin, Amélie Depierre,
Pascaline Dury, Amélie Josselin,
Susanne Lervad and Claire Rougemont
Research Centre for Terminology and Translation (CRTT)
Université Lumière Lyon 2, France

The presentation shows the first results of a collective research project on specialised languages. The research is based on the assumption that compiling a large-scale bilingual electronic corpus in different scientific fields will provide new and more detailed information on the way specialised languages, especially English and French, work. The project is also based on the belief that such information will be used in order to improve specialised translation, as well as the way it is currently taught at university level.

***