Belinda Maia  
Universidade do Porto

USING CORPORA FOR TERMINOLOGY EXTRACTION:  
Pedagogical and computational approaches.

Abstract

In this paper we look at ways of using specialised corpora for term extraction. We begin by referring to the more pedagogical approach used by teachers to train students of translation in language for special purposes, and then compare it to the methods and objectives of those in computational terminology. Although these approaches are different in scope, the problems involved in research are similar and the ultimate goals are complementary.

Keywords

Corpora  
Specialised corpora  
Terminology  
Pedagogical approach  
Computational terminology  
Term extraction  
Parallel text processing  
Subject specialists

1 Introduction

Over the years we have explored the possibilities of using electronic texts as a means of finding terminology in context. Our work originally developed out of an interest in research using corpora for linguistic analysis, and began as a teaching method with the possibility of searching texts to check correct words and phrases, collocations, and syntactic patterns in more general language, using WORDSMITH\(^1\) and a general selection of newspaper text and novels. This type of work, with the help now of on-line corpora like the British National Corpus\(^2\) and the CETEMPúblico Portuguese corpus, the bilingual COMPARA corpus\(^3\) and other general corpora, online or otherwise, is still an important part of our translation practice.

Recently we have concentrated on making and using small corpora to study specialised subjects in more depth and for term extraction. This paper will describe the path we have followed and try to explain where we might be heading in the future. It will try to put our

\( ^1\) WORDSMITH – at http://www.oup.co.uk/elt/catalogue/Multimedia/WordSmithTools3.0/\n\( ^2\) The British National Corpus (BNC) - the Oxford University Computing Service:  
- General information: http://info.ox.ac.uk/bnc/  
- Look up – 50 examples: http://sara.natcorp.ox.ac.uk/lookup.html  
\( ^3\) CETEMPÚBLICO and COMPARA - http://www.portugues.mct.pt/
more pedagogically orientated efforts into perspective in relation to current work in translation, terminology, computational linguistics, language technology and the other special subject disciplines for which terminology is so important. The sequence of events that led to the position we are in at present is similar to that experienced by other people, and it reflects a more general movement in the use of corpora for a variety of ends.

2 Terminology and corpora – the pedagogical approach

The type of terminology and corpora work we have been involved with has until recently been largely of a pedagogical nature. One of the principal objectives of our work is to convince the humanities educated translator that the serious study of specialised language – from the text to the term – is essential training for future translators.

How to teach specialised translation is a perennial problem in translator training because, as Robinson (1997: 148) points out, translators in real life often have to fake their knowledge of a subject. He quotes Kussmaul (1995: 33) as saying “Expert behaviour is acquired role-playing”. It would be idealistic to believe that translators will be able to acquire an in-depth knowledge of every subject they deal with, but the fact is that good, experienced, free-lance translators are often highly knowledgeable in a variety of areas. More commonly, translators specialise in certain areas and, once they have established a niche in the market, restrict themselves to these areas. The obvious advantage of having, or being, an in-house translator is precisely that such knowledge can be acquired and consolidated over time. Today, however, as in other areas of employment, positions as in-house translators tend to be coveted rarities, and the result is that translation is often outsourced to translators who may have no knowledge of the subject matter of the texts they are translating.

During translator training, therefore, we would argue that there are two ways in which one can educate future translators to cope with specialised language. One is to provide a good theoretical introduction to genre and text analysis, something which is often sacrificed in favour of more ‘practical’ work, but which we believe can and should be part of general theoretical courses dealing with theory of translation or contrastive studies. This theoretical approach can and should be coordinated with the practical aspects of encouraging students to translate as wide a variety of text types as possible, which is common translation teaching practice. This provides a good consciousness raising exercise in text appreciation as well as a broad view of the translation problems involved.

The second teaching technique is to encourage students to do in depth individual or group work on specialised subjects. This trains them in the theory and methodology of how to find out as much as possible about any subject, and teaches them how to make finding information and developing their attitude to knowledge a goal in itself. Both these techniques are needed if translator training institutions are to provide professionals who can be considered a reasonable alternative to the subject expert with a knowledge of languages.
At an earlier PALC conference in Maia (1997), we described the ‘do-it-yourself’ corpora with which we had been experimenting. This work arose out of the need to develop our restricted curriculum – which only allows for ‘general’ translation - in a way that allowed our students to find their way into the world of specialised language. In pre-Internet days, the idea of teaching only general translation was a solution to the problems posed by committing the university to supplying specialised translation courses within the structure of a modern language course. In this way, a teacher who wished to venture into a specialised field would have to do so at their own risk, choosing texts on subjects for which they could provide the necessary vocabulary, with perhaps a few extra items related to the topic under discussion. The idea that the student might need to do more than consult a general dictionary, or one of the technical dictionaries available, was hardly an issue, since finding such information involved time and, essentially, money.4

Despite this, several teachers did encourage students to go beyond the dictionaries and search for information by either arranging the necessary material themselves, through friends and family, consulting specialists, or organising special visits to places of interest. Needless to say, we still use these methods, but the Internet has opened up possibilities that allow us to go way beyond the local environment. The project work we develop with undergraduates, involving the collection of small, specialised corpora and glossary work, has also been described in Maia (2000).

This year our Master’s course in Terminology and Translation has allowed us to take this all to another level. Most of the students involved are trained translators with some years’ experience and are well aware of what the professional world expects of them. Several of them had already worked with us at undergraduate level and therefore had a good idea of what was expected of them.

The first semester’s programme provided a theoretical background in the relevant areas of linguistics, terminology, lexicography, translation and genre and text analysis5, as well as a short introductory course designed to put the whole project into focus against a background of international endeavour and research in the area of human language technology. During the second semester, they had the opportunity of working with subject experts, who were all members of the teaching staff at the university, on specialised areas of Engineering, Geography and History on terminology projects. Apart from the usual academic report on their activities, the end-result of each project included a description of the conceptual framework of the area, a terminology database, and specialised parallel and comparable corpora. This experiment in inter-disciplinary study has provided some interesting insights that will be useful for the future development of their master’s dissertations and further versions of this post-graduate qualification.

3 Terminology research
One lesson that has been learnt by our attempts to coordinate the worlds of terminology and translation is the urgent need for more interdisciplinary communication, and I am not referring only to the humanities and the subject specialists but also to people in the

5 The programme can be found at http://www.letras.up.pt/translat/i_mepr.html
different areas of language specialisation. It would seem that people interested in translation, terminology, corpora and computational linguists are not always as aware of each other’s work as they should be. For example, we admit that, although Pearson’s (1998) *Terms in Context* was always a set book for the Master’s, we were less aware of the developments described by Bourigault et al (2001) and Veronis (Ed. 2000) and the recent publication of these books meant they were not part of the original programme. However, others too are guilty of lack of contact with neighbouring disciplines. Computational linguistics regards terminology work as only tangential to its interests, translation specialists pay too little attention to specialised languages, and not everyone in the terminology world seems aware of the complexity of full text translation. There is also some misunderstanding or disagreement over what constitutes a ‘parallel’ text.

An important point to remember is that, although terminology in all its forms, from the special language glossaries and dictionaries to the big databases available, is of considerable interest to the translator, the world of terminology is not always primarily interested in the translator. It may come as a surprise to the translator, but monolingual terminology is often the starting point for many terminological projects, and is seen as an end in itself.

### 3.1 Monolingual terminology research

Monolingual terminology research has applications in quite local environments. This is particularly true of the American context where, as Wright (2001: 468-479) points out, terminology management is an integral part of information management in efficiently run companies, and is essential to Quality Control and Total Quality Management. She points out (ibid: 468) that ‘…terminological products were primarily used (and sometimes even generated) at the very end of the production process’ often ‘just before applying the shipping label’. She then goes on to say that ‘forward-looking enterprises have now integrated terminology generation and documentation into the document and the product generation process’. Much of her teaching practice is orientated towards preparing people to work with this process (personal communication).

Subject specialists, regardless of their language, will often draw up monolingual glossaries of terms for use in a variety of circumstances, of which the academic environment is only one. If one looks at the enormous number of on-line glossaries, – largely in English – one will get an idea of the breadth of incentives and interests behind the construction of such glossaries. There is everything from the databases of multi-national organisations to the short, pedagogically orientated glossaries Internet-aware secondary school teachers put up for their students.

---

6 A leading terminology school has published an important article (in a volume that shall remain nameless), which was translated into English and revised, according to the footnotes, by non-native speakers or native speakers with no training in translation - or even writing originals. The syntax is frequently confused, and confusing, and the punctuation appalling – but the terminology, no doubt, is correct!

7 For example, Veronis (Ed. 2000) on *Parallel Text Processing* - begins with an apology for a lack of awareness of the fact that ‘parallel’ texts are understood differently by some people in the translation world, who would prefer the term ‘aligned’ texts. But at least the apology exists.
A lot of the research into computational terminology is also monolingual, and the objectives are usually more related to knowledge engineering than translation or terminology databases. The objective of providing quick computational access to terms is not restricted to the collection of terms *per se*, although this is an obvious use for this type of research. There is also a need to be able to discover appropriate texts easily and to find terms that act as keywords for automatic abstracting, see Oakes and Paice (2001).

### 3.2 Technical and financial aspects

When terminological databases are the main objective of research, the big problems, computational and otherwise, are related to making such databases compatible and accessible to as many people as possible. There are the problems of incorporating different databases prepared in different software and according to different conceptions of what is required of the database. Since 1999, the MARTIF standard (ISO 12200) now exists for terminological databases, but this will help more with cajoling compatibility out of databases based on it, than with producing instant compatibility out of older systems.

A major problem of well-researched terminology is its accessibility. In order to be effective as a tool to subject specialists in all levels of government, industry and commerce, it must be minimally accessible. The usual practice in the past was to produce glossaries and dictionaries, which those interested either commissioned or bought, often at considerable expense. The attempts to put all this on CD-ROM, as has happened with both traditional general dictionaries and specialised ones, have not yet shown great strides in theory or technique. However, a lot of work has been done in electronic format and, as Schmitz (forthcoming) points out, when arguing in favour of the project Webterm, some effort should be made among academic institutions to cooperate in putting all the terminology work done by students - some of it quite valuable – into a format that could be accessible on-line.

All this gives rise to a variety of problems of which copyright is only one. The bottom line is really who pays for what. While some terminology is still paid for by standardising organisations or commercial interests, the rest is subsidised – if at all – by projects, or is done for free by enthusiasts, or as a curricular obligation by students. Under these conditions, all terminology work will be subject to the needs and convictions of those in charge, and the requirements of the final database will also affect the availability of the research raw material - corpora.

### 3.3 Multilingual terminology research and corpora

Computational research on multilingual terminology nearly always begins with the alignment of the texts being used, and this process is the subject of Veronis (2000). Alignment of texts presupposes users who will be interested in bi- or multilingual

---

8 See Cimino (2001) for an example of combining medical databases of this sort.
9 It is no coincidence that two of the leading schools of terminology, in Canada and Catalonia, are driven, and paid for, by the political need to make sure that the two languages used in these areas are equally well represented in official terminologies.
language analysis, whether the objective is terminology extraction, translation or more
general linguistic analysis. The articles here reflect a variety of problems in alignment –
from the normal problems of aligning even quite similar languages at lower than sentence
level (Veronis, 1-24), attempts at parsing and matching both languages, Chinese and
English, (Wu, 139-167), trying to match very different languages like English and
Hebrew (Choeka et al, 69-96), to using knowledge about translation acquired from the
results of alignment to improve the tool itself (Santos, 169-186). There is also an
appreciation of different systems (Veronis & Langlais, 369-398), and an article by Melby
(347-368) that looks at the whole problem of alignment methods, explains clearly the
differences between aligned texts, translation memories and terminological databases,
and goes on to show the connections between them and statistically based machine
translation. He even suggests that as technology progresses and more and more textual
material goes on-line, that these corpora will eventually take over from traditional
lexicography and terminology. It is very possible that, apart from lexicographical
traditions, the one thing that is holding us all back from proceeding in that direction is the
Big Problem – Copyright.

4. Problems for computational and pedagogical approaches to terminology
When reading reports on research into computational terminology, we were struck by the
fact that many of their problems reflected aspects of our own, more traditional, work.
However, since their research is more theoretical, and concentrates more on finding ways
of doing things in the future, rather than tackling more immediate real-life terminology
work in the present, their solutions can only help us in limited ways. These problems are
those related to the (often missing) links between language specialists and subject experts
and their specialised languages, established attitudes to terminology compilation, already
available material - both traditional reference works and corpora, and translation at a
linguistic and socio-cognitive level.

4.1 Language specialists and subject experts – attitudes
Perhaps the most important lessons to be learnt in any exercise in terminology which
involves people from the humanities with subject specialists is on the conflict that arises
from two very different attitudes to specialised language. To the linguist coming from
the humanities, lexicography is a discipline that, in any case, is thoroughly under-rated by
mainstream linguistics. The reasoning has been that the real world lexicon is so enormous
in scope and so dependent on context that it can obscure the real object of linguistic study
- syntax. Most people interested in practical uses of corpora are dedicated to the
empirical study of how people use language in normal circumstances. However, when we
move on to study texts in specialised areas, we tend to take our methods of work with us,
in order to search such texts for naturally occurring terminology.

To the specialist, the terminology of his/her discipline is an essential key to the
understanding of the subject. Introductory literature to specialised areas concentrates on
explaining the main concepts, and the terms used to represent them. At another level, the
objective of terminology is to provide normalised terms, within a semi-legal framework,
so that confusion generated by different people using different terms for the same thing,
or vice versa, can be avoided. The specialist is therefore interested in the concepts, and
in arriving at some linguistic solution that will represent these concepts satisfactorily. Although specialists will sometimes accept the offer of terminologists to help them sort out their material, they expect to make the final choices. The resulting problems are often a question of the competition among specialists as to whose terminology is best.

Subject specialists are, without doubt, essential to the terminology process. Their help can be requested at the beginning of the process, as when computational terminology starts with a set of keywords or basic terms provided by experts, or when the experts provide terminologists with basic literature and other material. It can come during the process, as when computational terminologists request help when their system produces too much ‘noise’ – i.e. too many structurally similar but irrelevant examples, or when the terminologists need guidance in conceptualisation and planning as they flounder in an unstructured sea of information. The expert will always be required at the end of the process, when validation of the terms is called for. However, the circle does not necessary close there – the terminologist will still have to cope with changes made after the experts have argued about everything and, hopefully, agreed to compromise and agree on the final version.

The attitudes of both linguists and specialists to translators and terminologists are also an interesting clue to the way terminology work has developed, and make one reflect on the position of the translator as a transformer of texts from one language to another, and the real role of the terminologist. Even the more functional-systemic linguists, with their more holistic understanding of how different levels of language interact, will be more interested in how the term is integrated into the text or syntax than in the term itself. On the other hand, subject specialists often express considerable reservations over allowing anyone but one of themselves - with a good knowledge of another language - to do top class scientific translation\(^\text{10}\). No doubt, this attitude is partly a result of experience with translation done by translators who have either been inadequately trained in the techniques of scientific translation, and/or have no access to properly elaborated terminology.

However, the ideal of the subject expert with a good knowledge of languages and, possibly, training in translation, still remains very largely an ideal, and there is a wide spectrum between this ideal and the reality of humanities educated translators doing their best with the help of dictionaries, reference books and, now, everything that the Internet provides. The reality of the language specialist being expected to cope is largely due to the structuralist perception of language in which syntax merely provides a framework on which to hang words or terms.\(^\text{11}\) This same structuralist view still pervades much terminology work and the relationship of translation to terminology. It is also prevalent in much of the computational work being done on term extraction. Any system that has

\(^{10}\) The Gulbenkian Foundation in Portugal contracts university lecturers in a discipline for translations of highly regarded scientific work into Portuguese. One of these professors once expressed horror at the idea of a humanities educated translator attempting to handle such a text – although he had the grace to admit that the translation was proving far more difficult than he had expected, with correct terminology being the least of his problems!

\(^{11}\) People too easily presume that the subject specialist is fully aware of and competent in the conventions of syntax, text, register and other aspects of both the source and target languages.
to rely on using either lexical items or the tags produced by parsing, must by its nature rely on a structure of some kind. Even semantic tagging of the kind proposed by Oakes and Paice (2001) is structured.

There are, of course, a lot of advantages to be gained from a sound structural analysis of syntax and the lexicon, and the work of many centuries in this direction underlies a lot of our assumptions about knowledge in general. However, this view should not obscure the need to look at the functions and relations between elements of the structure, or prevent the flexibility and creativity necessary for knowledge to advance, and be seen to be advancing, in hitherto unmapped directions.

4.2 Terminology compilation – processes and attitudes

Our experience of working in a variety of areas and with different kinds of experts has been interesting in its diversity, and the results have been both positive and negative. On the more negative side is the project work that attracts our weaker, shyer undergraduates - taxonomies in the natural sciences. They tend to take refuge from the world in the type of book produced on animals, flowers or herbs by Reader’s Digest or Dorian Kindersley and their translations. The negative aspect of this work is not in the possible usefulness of the terminology produced, but in the failure of the pedagogical element involved. The project becomes ‘cut-and-paste’ dictionary work, in which corpora play little part as, although explanatory definitions and images are to be found in these publications, the short, very repetitive texts from which they are taken are of limited value as corpora. Besides this, the conceptual frameworks involved are stable and often fossilized, and contact with the outside world of experts is largely avoided12. The taxonomies themselves reflect the earliest kinds of terminology practice.

Linguists often use the word ‘nomenclature’ negatively to describe the terminological systematisation of concepts and terms in the sciences, often as a way of dismissing the whole problem of terminology as a series of fixed word choices with no interest to general language. However, the need of the subject specialists for these nomenclatures is reflected in the thesauri, standardised terminology, official glossaries and other carefully organised structures with which they try to impose some sort of order on the profusion of terms that are used and abused in their professions. However, science progresses at such a rate that this type of terminology is far less stable than the taxonomies just described. Scientific philosophy since Popper has been only too aware of the restrictions placed on innovation by over-dependence on the fixed reality of a concept/term, and recognises that science can only progress if this supposed reality is constantly challenged. This understanding does not, of course, mean that all previous work in categorisation and terminology has to be simply abandoned, but it does mean that terminologists, especially when asked to review and up-date terminology, very often have to work within pre-defined structures and according to accepted norms.

Our experience of working with engineers, for example, is that they already have an enormous amount of this normative material. They recognise that a lot of it is

12 The honourable exception was a student who produced an excellent project on endangered species with the help of a zoologist.
inadequate, outdated, and far less easy to systematize than they would like. They thrill to the idea that they could have updateable and explanatory databases to consult instead of twenty-year-old printed glossaries. However, their expectations are of something better than what they have, rather than something different. On the positive side, however, we are learning how to cope with their need for prescriptive terminology, and they are coming to appreciate the more descriptive approach.

4.3 Existing terminology resources – possibilities and limitations
Quite obviously, if one is going to work in terminology, one needs to make use of resources that already exist. Terms often have a long history and there will be documentation referring to wide variety of information.

The use of corpora in terminology would seem to produce reactions similar to those produced by the original proposals of the COBUILD project in lexicography, and there are still people who expect terminology to conform to practices of some decades ago. For example, they expect definitions to conform to the expert’s expectation of a concise, fine-grained definition using other specialised words, rather than the coarser grained explanation using more general language found in explanatory texts. Yet, the latter are those that are most easily found in corpora and are arguably of more general interest. Besides this, using up-to-date corpora to extract terms, which by their very nature are often less stable than words in general language, would seem to make very good sense, and Sager (1992:130) refers to this as ‘established practice’. Yet Cabré (1992/1998:118-20) barely touches on the possibilities of corpora, listing instead all the traditional sources – thesauri, dictionaries, encyclopaedias, databases etc – in detail.

We have found that subject specialists often provide us with thesauri, glossaries, normative documents and other terminology work – more often than not in English, French or German - but do not find it that easy to understand why we want texts, particularly in electronic form. They do not always seem to appreciate that terminologists are not simply word collecting machines, and that they need to understand the concepts behind the words they are dealing with. They expect the terminologist to ‘fake’ knowledge, and never fail to be amazed when the (humanities trained) terminologist actually shows a capacity to understand what they are talking about. This ability to amaze is what makes the difference between good and bad terminologists and translators, and teaching our students to be able to amaze is of maximum importance.

Apart from the assumptions of the subject specialists, there is another factor that militates against our pedagogical objectives. Surprising though it may seem, the on-line glossaries and databases already referred to can have a negative effect on the type of undergraduate project work we do. When they are good, they can be God’s gift to the translator, but when they are bad, they are worse than useless, as they give false expectations to those who use them. Besides this, the majority are monolingual, and usually in English.

Students are well aware of them and often use them as a basis for their own glossary work. This can be positive. Taking a glossary in English and working towards the other language using proper research techniques and consulting experts can produce good
results. However, it is pedagogically limiting because it continues to promote the idea that one can depend on word lists to fill in the lexical items in the structure of a text, while remaining ignorant of the content. The weaker students are perfectly happy to rely on glossaries and often allege that they cannot find suitable texts with which to make corpora – i.e. having found the ‘right word’, why worry about finding out more about the subject? Although such a process may be unavoidable when one is translating under pressure to earn one’s living, it is hardly one we should be encouraging at a training level. Therefore, since we cannot make them go away, the best thing to do with these glossaries is concentrate on getting students to discriminate between good and bad material.

3.4 Finding Corpora

Our Master’s students still need the lessons of the pedagogical approach, but they are also progressing towards real-life terminology, and our biggest problem is finding appropriate corpora to work with. The Internet is awash with documents in English, and French and German do not do too badly but, after that, the more minor a language, the more difficult it is to find reliable texts, outside the corpora already available through the EC pages, ELRA and other sources.

One must also remember that many people are still reluctant to publish top-quality research on-line – or are forbidden to do so by the copyright agreements with their publishers. Finding texts on a subject dictated by the circumstances of a particular terminology project, rather than by more abstract research, is sometimes difficult even in English, and it is far more difficult to find appropriate material in minor languages. This is also true of normal written texts. Too often experts – or even the university library – can only present us with English texts, and parallel texts of this specialised nature are rare.

In our case, when texts do exist, the chances are that they are in Brazilian Portuguese, something to which continental Portuguese-speaking people often react negatively. The vocabulary and text structure of Brazilian Portuguese – particularly in scientific and technical domains – suffers from considerable influence from American English. Terms are often adopted from English with barely a nod in the direction of Portuguese morphology, and sentence structure in such texts tends to follow the dominant SVO order of English rather than the more hedged style of natural Portuguese. The dominance of international English is an insidious influence that many of us will recognise.

These considerations lead us to a further disadvantage of many texts found on the Internet and that is both the quality of originals and the quality of the translated, or parallel, text. In pre-Internet days, translator trainers chose texts for translation carefully, and used parallel texts for teaching purposes either as examples for students to follow, or to demonstrate what should be avoided. Besides this, in professional situations, it was usually possible to find out who had done the translation and form some opinion as to its reliability. Now, we should be teaching students to recognise quality in both texts and translations more explicitly, since finding out about the original writers or translators is increasingly difficult.
Despite everything that has been said, however, there is no doubt that the Internet provides us with information on an unprecedented scale and has changed many of the perspectives from which we view translation and translator training. Translating web pages is one of the more promising areas of work for our translators, and one can safely forecast that the quantity of material on-line will increase, and that either the quality will improve, or steps will be taken to provide guidance on quality.13

The work for the Master’s degree has proved an interesting experiment in using corpora for terminology extraction. The subject specialists in Engineering, Geography and History differed considerably in their approach to providing texts for corpora. The engineers, although they eventually understood our proposed objectives with corpora, did not find it easy to supply appropriate texts. The specialists from the Humanities Faculty were quicker to appreciate what was needed, and, indeed, our History colleague would love to do diachronic terminology with the help of corpora.

Our Geography colleague, however, has come closest to realising our full objective of finding electronic texts and other documentation for term extraction. In the first place, she chose a subject well represented on the Internet – the area where demography relates to problems of the environment – and proposes to publish a dictionary with the results. Each member of her group is searching for material in a different language – English, French, German and Spanish – as well as Portuguese. Official terminology is welcome and collected carefully, and she has undertaken to validate the Portuguese terminology in the end. The interesting point for this paper is that their corpora collecting work has been extensive. They have found texts of all kinds – legal, expository, educational, general informative, etc., – besides terminology rich texts like academic abstracts, and sites that contain official keywords. In other words, this project promises to be the best example of the work so many of us would like to see done in this area.

Our colleagues in computational terminology are luckier than we are in that either they have been asked, or, even better, contracted, to find solutions to terminology problems using particular corpora with the help of subject experts, as with the LEXTER system (see Bourigault, 1994) and others, or they have been free to choose a corpus on which to practice their computational methods. They tend to choose corpora that suit their purposes. For example, Blank (2001: 239) describes the tri-lingual corpus of technical texts used in their work as ‘particularly suitable’ because:

- it is structured in a very concise, homogenous and uniform manner
- it is sufficiently big to be statistically relevant and
- the texts have been written in a legal context, i.e. the translations are of good quality.

When working from a purely pedagogical point of view, these are not characteristics that are particularly relevant to our objectives. Our interest is in a variety of texts types that are related to the subject being discussed and, therefore, a reasonable breadth of register

13 For example, there are people who advocate that academic articles on-line should be refereed in the same way as paper journals.
and style are an advantage, since terminology extraction is merely an equal partner to translation and corpora collection in these projects.

Ahmad and Rogers (2001:726), on the subject of automatic term extraction, point out that ‘Human beings still play a role in all methods proposed to date’, and Cabré et al (2001:53-88) end their appreciation of the various terminology extractors/detectors reviewed by saying:

In the future, we believe that any current terminology extractor, apart from accounting for the morphological, syntactic and structural aspects of terminological units, has to necessarily include semantic aspects if the efficiency of the system is to be improved with regard to the existing ones.

In other words, there is still a long way to go, since semantics and the influence of real world context will continue to be the main obstacles to the mechanisation of language for some time to come. However, this does not mean that current research is not valuable, or that it does not have its uses. As with other types of language technology, like machine translation and computer-assisted translation, from which it has developed and to which it will contribute, (semi-) automatic terminology extraction can only help the translator.

5 Conclusions
Using corpora for term extraction is an ideal that, if we persevere, will produce results that probably go beyond the expectations of current notions of terminology work, and become an essential adjunct to information retrieval and the advance of language technologies. It will also be particularly useful in areas that are still poorly represented in terminology databases, and it should go a long way to improving, updating and correcting anomalies in more traditional work. It will also, no doubt, be used for the correction and refinement of the very texts themselves. The projects for speeding up this process are very welcome, but full or even partial automation is still the dream of our computational colleagues – for the time being, at least.

REFERENCES


