[*Translating and the Computer 22*: Proceedings of the Twenty-second international conference…16-17 November 2000. (London: Aslib, 2000)]

AUTOMATIC BILINGUAL TERMINOLOGY EXTRACTION

A Practical Approach

David Chambers.
WIPO. Geneva
david.chambers@wipo.int

Abstract

Faced with large and steadily increasing work volumes, the Patent Cooperation Treaty Translation Section at the World Intellectual Property Organization, Geneva, is looking for ways to improve the efficiency of its translation process. A terminology problem has been identified, and attention has turned to automatic bilingual terminology extraction as a possible means of solving that problem. A project has been defined and evaluation tests implemented with the aims of automatically capturing bilingual terminology from existing technical texts and their translations, validating the candidate term pairs generated, defining an appropriate database structure and generating terminological records in an automatic or semi-automatic manner. Benefits of this approach are becoming apparent and, as work progresses, the potential for extending the scope of the project to other related applications offers interesting prospects for the future.

1. Introduction

At the World Intellectual Property Organization (WIPO) in Geneva, Switzerland, a specialised agency of the United Nations, the Office of the Patent Cooperation Treaty (OPCT) is responsible for processing international patent applications received from all Contracting States. Within the OPCT, translation is an important aspect of operations, helping to ensure the widest possible and easiest access to information contained in documents describing and evaluating new inventions.

Given the rapid increase in the number of patent applications filed under the Patent Cooperation Treaty (PCT), and the ensuing increase in the translation work load, it is important to ensure a continued improvement in the efficiency and cost effectiveness of the translation process. This paper will look at development work undertaken in PCT Translation with a view to facilitating the translation operation.

We shall first briefly examine operations in PCT Translation and identify the problems which have to be faced. Subsequently, the solution adopted for terminology development and the associated terminology extraction and database creation project will be examined in detail. Problems encountered in the software evaluation and term validation phases of the project will be indicated, methodologies discussed, and a brief outline of future activities given. A short section has also been included to give an insight into the underlying principles of automatic term extraction.

2. Translation in the Office of the PCT

2.1 Translation Work

In the course of the PCT international patent application procedure, the Office of the PCT is required to translate two main types of documents:

a) Abstracts of international patent applications.

These are half-page summaries of descriptions of inventions and are written in dense technical language, often with poor syntax as in the example.

> Vehicle window and method of making the same
>
> A vehicle window comprising a relatively thin sheet of clear plastic material having opposed surfaces, an electrically operable defrosting grid adhered to one surface of the relatively thin sheet, and a relatively thick substrate layer of clear plastic material having opposed surfaces curved into a vehicle window configuration. The relatively thick substrate layer is adhered to the one surface of the relatively thin sheet and the electrically operable defrosting grid adhered thereto while in contact therewith in a molten state under heat and pressure within a cavity defined by two generally parallel curved die surfaces of cooperating injection moulding dies so that upon solidification the surfaces of the relatively thin sheet are retained in a curved configuration in generally parallel coextensive relation to corresponding curved surfaces of the relatively thick substrate layer, and a method of making the window wherein the electrically operable defrosting is formed by silk screening onto one surface of the relatively thin sheet while in a substantially planar condition a curable electrically conductive ink in the form of a defrosting grid and then curing the curable electrically conductive ink on the one surface of said relatively thin sheet so that the defrosting grid is stably adhered thereto. *(PCT Gazette 20/1996, WIPO, Geneva)*

b) International Preliminary Examination Reports

These are technical reports drawn up by engineers qualified in the field of the invention concerned. They give an opinion on the technical content of the invention with respect to novelty and inventiveness.

The majority of texts received for translation are written in English, French, German or Japanese, with some in Russian and Spanish, and translated from these languages into English and/or French. All translations are done internally at WIPO, with the exception of Japanese abstracts and all Chinese texts.

2.2 Development Work

Our activities are not limited to translation. Development work has been carried out on an ongoing basis in PCT Translation for many years, but it is only recently that the need for this work has been officially recognised and a mandate given to develop translation tools and computer assisted translation systems.

Various lines of approach are being followed:

- Internet search facilities
- Terminology database
- Electronic dictionaries
- Translation memory
- Voice recognition systems.

The most positive results to date have been achieved with the Internet search facilities, which have been implemented in the form of our TermLinks targeted search interface *(Gomez, J. ITI Bulletin, August 2000).* Terminology database creation has required extensive groundwork, but is

advancing well. We will see below how we now hope to move ahead rapidly. Electronic dictionaries on CD-ROM or loaded from CD-ROM onto our local area network have been introduced, and others are available on-line over the Internet. Translation memory is a slow starter since we do not believe our texts to be adequately repetitive for it to be effective, although as we shall see, the use of a workbench environment will necessitate its development. Little has been done with voice recognition, although progress may be more apparent once a collection of specialised technical terms has been created.

2.3 Constraints

As in all translation operations, a number of problems and constraints have to be faced.

2.3.1 High work load

Work load is currently around 70,000 pages, or over 20 million words, translated per year. The growing popularity of the international patent application procedure, whereby an inventor can file a single patent application which has effect in any or all of 108 countries, has meant that our work volumes have increased at a rate of between 15 and 20 % annually.

2.3.2 Critical, short deadlines

In the course of the patent application procedure, certain document publication and preparation dates are legally binding. Frequently, texts for translation are received only a few days before those deadlines.

2.3.3 Highly technical texts in an exceptionally wide range of technical subject areas.

It is inevitable when dealing with patents that the subject matter is highly technical, frequently making use of terminology not yet to be found in any dictionary or employing existing terms with modified meanings. The vast range of subjects covered means that it is difficult for the translators to specialise in any one field. They are therefore heavily dependent on reference material and up-to-date glossaries.

2.4 Solutions

As in all business processes, increased output and reduced costs are permanent objectives. At the same time, work quality must be maintained, and indeed improved. In the OPCT, these objectives are shared by both management and the translators themselves. If the translators are able to achieve greater efficiency and improve the quality of their work, they attain a higher level of job satisfaction, a key factor often overlooked when working towards improved productivity.

While a simple answer to the high volumes/short deadlines dilemma may appear to be to hire more staff and apply high, rigid output requirements, it has been seen that this does not constitute a satisfactory or indeed a realistic long-term solution. The effects of such a policy tend to be negative for in practice an imposed excessively high output leads to a rise in stress levels and dissatisfaction. Staff skills and expertise do not develop. Motivation is lost. Quality and productivity are likely to suffer as a result.

The approach adopted has therefore been first to analyse the translation process and identity those elements of the work most suited to the application of language technologies offering the best answers to the needs of the translators. A Feasibility Study was undertaken and provided a number of recommendations for improving the translation process.

2.5 Terminology Problem

That Study identified the development of terminology resources as an essential step towards improved efficiency in translation.

In view of the wide range of subject matter, and given the need to find the latest terminology in any subject field, translators can spend over 40% of their time on searching for highly technical, specialised and new terminology, and in many cases the same search is repeated by other translators. An enormous amount of highly technical terminology transits through PCT Translation without it being possible either to capture this terminology in a retrievable form, or make it easily accessible to translators.

In order that terminology found by one person can be stored and made available to all, work has been started on adapting dedicated terminology database software to the specific needs encountered when translating technical information in international applications. Our aim is to provide desktop access to translations, definitions and explanations of technical terms, and where appropriate also provide links in the terminological records to external reference material.

The problem faced at present is how to create terminological records rapidly in the database in order to achieve a critical mass of data that will make use of the database effective. The manual input of terms either by a terminologist or by individual translators is a slow process. The direct importation of existing glossaries would be possible, but terms are generally devoid of context. Furthermore, copyright issues could arise. Consequently, an alternative, innovative approach has had to be found.

Our attention was drawn to automatic term extraction, and in particular to the automatic extraction of bilingual terminology from parallel corpora with a view to capturing existing terminology and generating a large number of bilingual term pairs to be incorporated into the terminology database. We also had available a large corpus of parallel texts in two and sometimes three languages (English and French plus some German), and expertise in the form of a team of experienced translators for validating the candidate terms extracted.

Consequently, a project was defined for terminology extraction and database creation.

3. Terminology Extraction Project

A project, entitled "Terminology Extraction and Terminology Database Creation - Integration in the Translation Process" was launched in June 1999.

3.1 Project objectives

The aims of the project are as follows.

- Automatic capture of bilingual terminology from PCT texts and their translations.
- Manual validation of bilingual term pairs.
- Definition and creation of an appropriate database structure.
- Automatic generation of terminological records.
- Integration of database into the translation process.

3.2 Project Software Components

During our exploratory work it became apparent that although developers of computer assisted translation systems may claim to provide global solutions to translation problems, none of the systems offered met all our needs. Specific tools were needed for specific tasks.

Consequently, the software adopted falls into two main categories.

a) Bilingual terminology extraction software

For us, this type of product was something completely new and represented a departure from the known translation tools such as translation memory systems. As an international organisation, it is not our role to be involved with software development or research work. We therefore had to find a product that was commercially available and not in alpha or beta testing phases. Our investigations revealed that a number of products existed for monolingual term extraction, for example Lexter belonging to Electricité de France or SystemQuirk developed at Surrey University in the UK. However, the only product we found that would provide results automatically in the form of bilingual term pairs was the Xerox Xtras Terminology Suite, and in particular the TermFinder and TermOrganiser modules.

b) Database management and integration software

We are on reasonably familiar ground here and had a choice of terminology database software from suppliers such as Atril, IBM, Star, Trados and Xerox. Our requirements were that the product should:

- be tried and tested,
- integrate terminology and translation memory functions,
- offer integration with MSWord.

Only Trados was able to offer a product, Translator's Workbench, which met all these requirements.

3.3 Project Partners

Our expertise was insufficient for us to be able to implement the project alone. We therefore called on external services to assist with various aspects of the work.

- **PCT Translation Section I** acted as project leader and provided translation expertise for term pair validation.

- **Xerox Multilingual Knowledge Management Services,** Xerox Research Centre Europe, provided consultancy for use of the Xerox software. Assistance was also received from the Xerox research team

- **ISSCO,** the natural language research institute of the University of Geneva provided technical assistance and metrics. Translation technology students were also available to assist with processing.

3.4 Source Material

We have already mentioned the nature of texts at our disposal, i.e. abstracts and examination reports relating to international patent applications. Only the abstracts are available electronically in both English and French, and in some cases in German. Electronic storage in a readily accessible form (SGML) has been in place for approximately five years, which means that around 130 000 pages of text are available at least in English and French.

4. Automatic Capture of Bilingual Term Pairs

Before moving on to look at project implementation, it would be useful to examine briefly some of the principles involved in automatic term extraction. This paper does not aim to give a detailed explanation of the linguistic or theoretical aspects, which are covered in the specialised literature, nor are we able to discuss the actual algorithms used which are proprietary to the software developer.

A number of definitions should serve to differentiate first between various levels of extraction.

A Term is defined in ISO standard 1087 as "A designation of a defined concept in a special language by a linguistic expression.". A note states that terms can consist of single words or be composed of multiword strings. The distinguishing characteristic of a term is that it is assigned to a single concept.

Term Extraction has been defined as "selecting those elements out of the corpus that are considered terms of the special subject field which is the object of the terminology".
(Cabré, M.T. Terminology theory, methods and applications. Trad. DeCesaris, J.A. 1998: John Benjamins).

Automatic Terminology Extraction can be defined as the application of dedicated software to the Term Extraction process to identify candidate terms in a text or set of texts. Extraction is performed monolingually. The software may be language independent, relying purely on statistical methods for identifying terms, or language dependent, taking into account linguistic information for the language concerned, or apply a combination of statistical and linguistic methods.

Automatic Bilingual Terminology Extraction takes the above definition a step further to cover the extraction of terms and then: translations across two different languages from a set of parallel texts, thereby generating a set of bilingual term pairs.

In the case of the algorithms used for our project, the process involves monolingual term extraction in each of the languages concerned, followed by pairing of the terms bilingually. The monolingual extraction process uses language-specific term identification methods, whereas the bilingual matching process relies entirely on statistical methods to pair off terms in source and target languages.

Statistical processing may take the form of measuring the frequency of occurrence of words and word associations. Probability theory is also applied.

Linguistic processing methods rely on identification of words and part of speech patterns. Various steps are applied in linguistic processing, for example tokenization, morphological analysis, part of speech disambiguation, phrase identification and delimitation, shallow parsing.

As can be seen from the above definitions of **Term and Term Extraction,** the idea of a defined concept or special subject field is prevalent. In human term extraction, it is relatively easy for a terminologist or an expert in the subject field to identify terms relating to a given concept. In machine or automatic term extraction, however, the software is not necessarily able to exclude general language terms or unrelated subject fields from the list of candidate terms generated.

Applying linguistic technology algorithms to one of our English texts:

> "Laser energy is directed between the beam director and the transmitters/receivers by the active monolithic structure solely as collimated light.",

the following results are obtained *(Xerox website http://www.xrce.xerox.com/research/mltt/demos):*

Tokenization identifies the individual words in the text and identifies them as tokens.

```
Laser
energy
is
directed
between
the
beam
director
```

etc.

Morphological analysis

```
Laser       laser+Noun+Sg
energy      energy+Noun+Sg
is          be+Verb+Pres+3sg
directed    direct+Verb+PastBoth+123SP
directed    directed+Adj
between     between+Adv
between     between+Prep
the         the+Det+Def+SP
beam        beam+Noun+Sg
beam       beam+Verb+Pres+Non3sg
```

Words such as "directed" or "beam" are identified as various possible parts of speech.

Part of Speech Disambiguation will then identify the relevant part of speech for the word in the sentence.

```
Laser laser     +NOUN
energy          energy +NOUN
is    be        +VBPRES
directed        direct +VPAP
between         between      +PREP
the   the       +DET
beam  beam      +NOUN
director        director     +NOUN
```

An example of incorrect part of speech disambiguation which recurred frequently in our initial tests on the French language was the proposed French term "invention porte". This is not as one may suppose a translation of the English "invention of a door", but came from the recurring phrase "l'invention porte sur ..." (the invention concerns...) where "porte" should have been classed as a verb.

The text is then tagged with this information for further processing, in our case for extraction of nouns and noun phrases.

Multiword terms can be identified on the basis of certain predefined sequences such as adjective+noun (AN), noun+noun (NN), or in French noun+preposition+noun (NPN). Also certain signs such as quotation marks, parentheses or punctuation marks, or certain non-term words such as a verb, or "e.g.", can serve as phrase delimiters.

Some examples of patterns for describing terms are:

AN	Acoustic transducer
ANN	Spectral energy deviation
NN	Data transfer
AAN	Public cellular network

Or for French

NA	Transducteur acoustique
NPN	Transfert de données
NPNA	Satellite à orbite basse
NPNPN	Réseau de communication par paquets

Shallow Parsing rules may also be applied as in the sequence:
Nounl+adjective+preposition+noun2 → nounl+adjective
As in the example
"Applications intégrées en bande" to give "Applications intégrées".
Albeit with possible ambiguities as in nounl+preposition+noun2+adjective giving
Poste de traitement central
Poste de traitement
Poste central
Traitement central

5. Project Implementation

The project was launched with a view primarily to evaluating the term extraction software. Although planned initially as a single test, this evaluation actually took place in two phases which we have designated Evaluation Phase I and Evaluation Phase II in the following description. Each phase involved the following steps:

- Data preparation
- Sentence alignment
- Term Extraction and Bilingual Matching
- Bilingual Term Pair Validation
- Generation of Terminological Records

6. Evaluation Phase I

6.1 Data Preparation - Evaluation Phase I

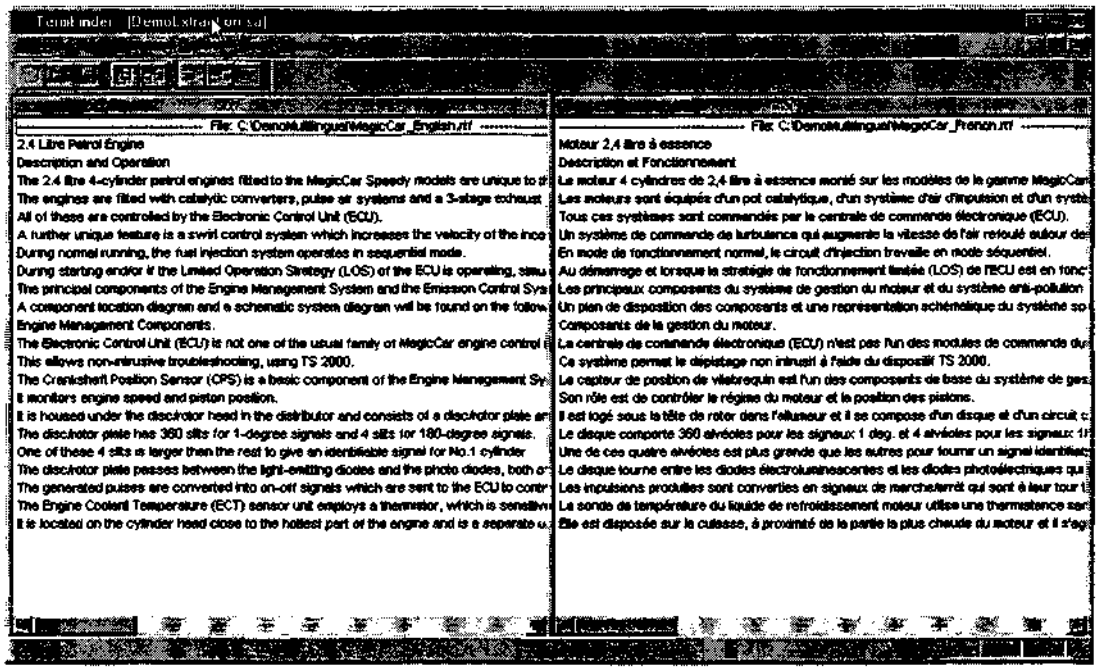
The source data (patent abstracts) are held on archive tapes in SGML format. Our IT department was able to provide a script for downloading these data to the PC. Each file downloaded contained an original text, a translation, document number, date, name of patent applicant and, of particular importance, the subject field code in accordance with the International Patent Classification *(WIPO, Geneva)* which classifies inventions into main and subsidiary subject areas. Using that code, we were able to select a data set relating to a given technical subject field. Some additional processing was needed to split the initial file into separate files for each language.

For Phase I we prepared data sets comprising about 1000 pages medical texts, 700 pages chemistry and 800 pages electrical and telecommunications. Each subject field was to be processed separately to give an appreciation of the term capture in each field.

6.2 Sentence Alignment - Evaluation Phase I

The Xerox TermFinder module provided the sentence alignment, which follows the principles known from translation memory software. Source and target files were first associated, and the alignment process was then run. The results of the alignment were good and the program allowed for correcting the alignment manually with functions for merging and splitting text segments.

Unfortunately, the user interface proved unergonomic and difficult to use. Each line is the start of a segment or sentence, the source text being on the left and the target text on the right.



There is no way of seeing the full segment in both languages at the same time. By clicking on a line, the full segment will be displayed for that language, but this box must then be closed before the corresponding segment in the other language can be opened. Manual correction of the alignment was hence extremely tedious.

6.3 Term Extraction and Bilingual Matching - Evaluation Phase I

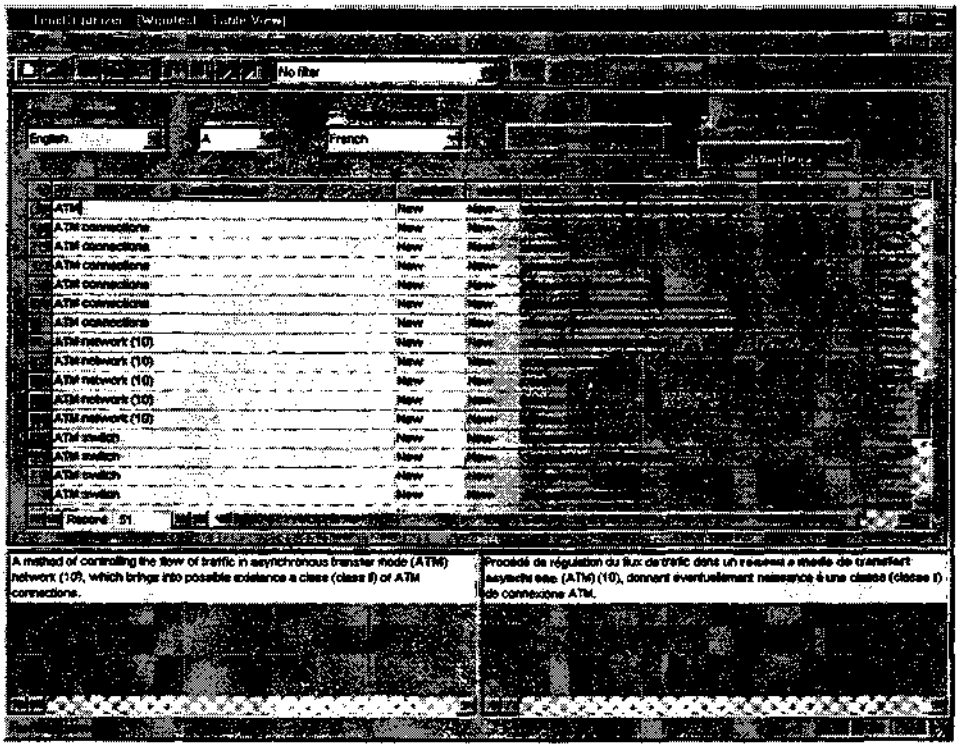
The aligned segments are then processed by the term extraction function of the program. This runs quickly in the background and extracts the nouns and noun phrases from the texts in accordance with the principles described earlier. There are in fact three operations: monolingual extraction in the source text, monolingual extraction in the target text, bilingual matching to produce candidate term pairs.

6.4 Bilingual Term Pair Validation - Evaluation Phase I

There is no system that can automatically validate the accuracy of the candidate term pairs generated. We are totally dependent on human expertise and must call on experienced translators and/or terminologists for validation.

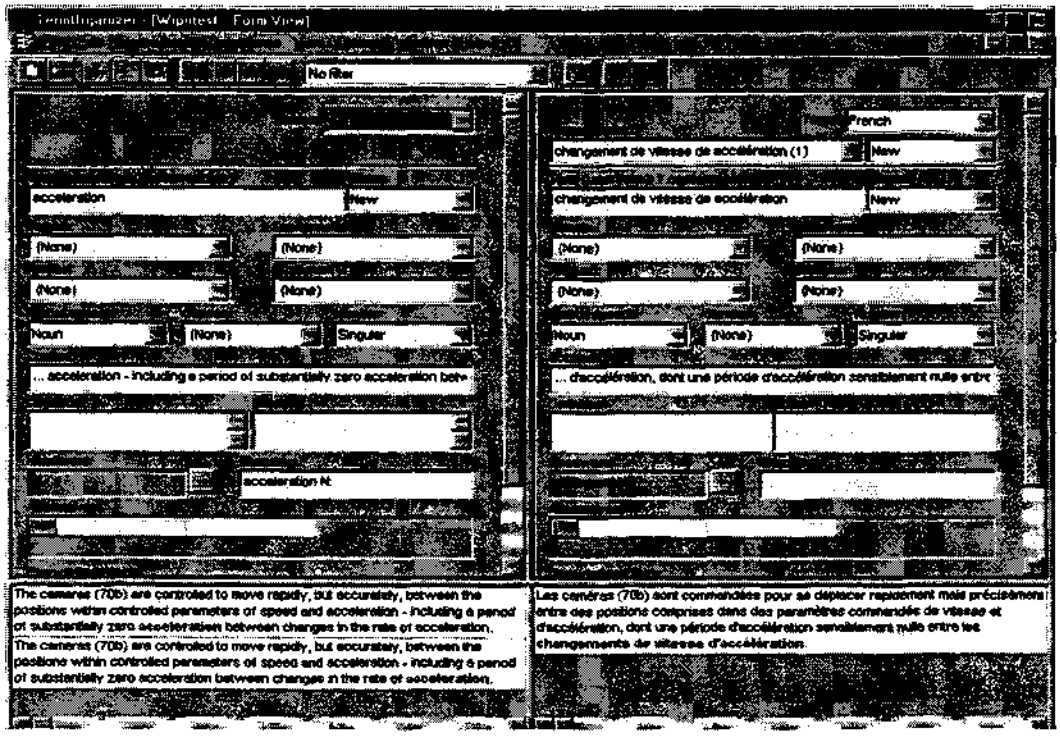
The Xerox TermOrganiser module provides the necessary tools and interfaces for carrying out this validation work. A series of screen shots will show the different work modes available.

Table View is perhaps the most useful way of working initially. The source terms are listed on one side of the screen and the target terms on the other. The context(s) for the term



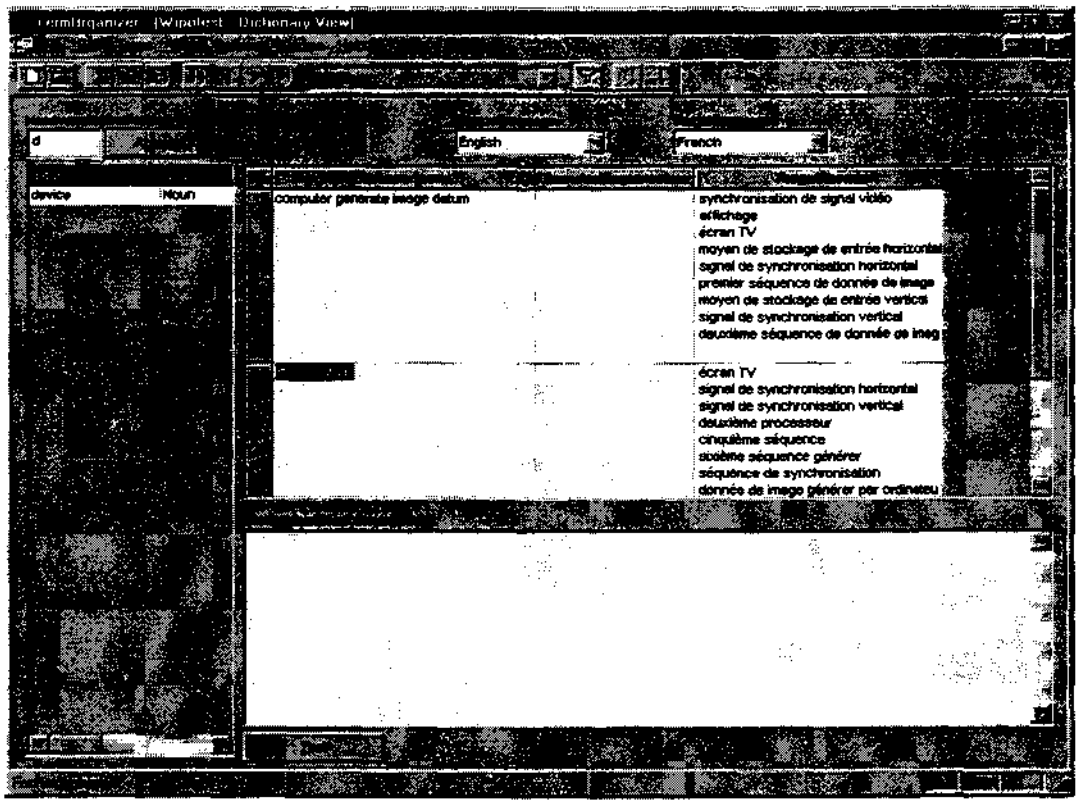
selected in the list will be displayed in the boxes at the bottom. These contexts are in fact the segments defined at the alignment stage. Various administrative columns can be included such as a status column for each of the terms. The status can thus be set for example to indicate that the term is valid as a monolingual entry, in either of the languages, or as a term pair if English and French terms have been correctly matched and the translation is acceptable.

Form View allows one term to be viewed at a time with the corresponding terminological record. The source language and source term appear at the top left; the target language and



target term(s) at the top right. The boxes contain data making up the terminological record such as domain, sub-domain, name, etc. The bottom boxes contain the context(s) for the terms in each of the languages. A status box at the top right indicates the current status of the term, and allows that status to be modified manually, for example from new to valid once a candidate term has been validated. The various boxes can be filled in manually to complete the terminological record.

Dictionary View enables a headword to be selected and displayed at the top left. The box on the right will then show the various target terms that the software has generated for the source



term or phrase. In this way, the various candidate translations for the source term can be seen at a glance.

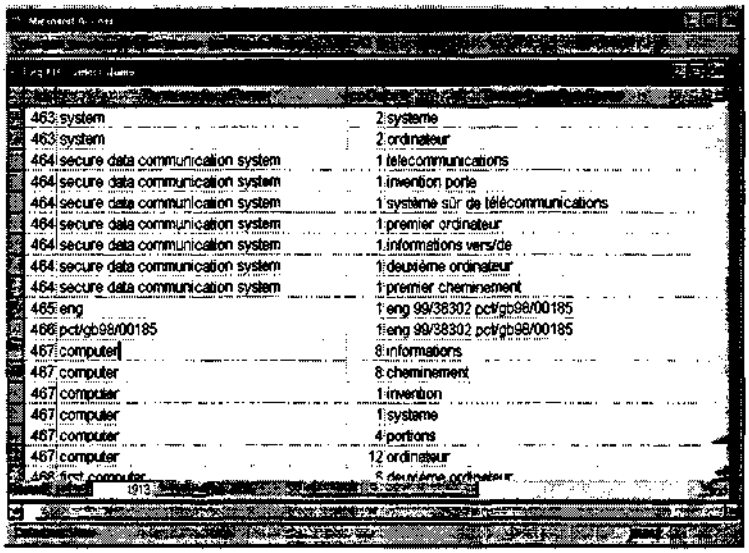
6.5 Automatic Generation of Terminological Records

The software allows most of the data in the original SGML file to be retained for possible incorporation in the terminological record. In our case we should be able to enter automatically the source and target terms and contexts, the subject area, document date, document originator and document reference number. Certain grammatical data such as part of speech, gender and number are also generated by the software and can be included automatically in the record.

7. Results — Evaluation Phase I

Initial extraction results were marked by an excessive amount of noise present in the term pairs generated. Although terms had been identified monolingually, they had not been correctly matched to form a correct candidate term pair.

We can see that, even though a correct term and translation may have been obtained, the number of incorrect results presented is excessive. It appeared as though, for a given English term, the software was picking out all noun phrases in the corresponding segment in the French text and presenting them as candidates.



On occasions, for a given English term, the correct term in French was present in the list of French terms, but it had not been paired with the corresponding English term.

For about five pages of source text, therefore, we were obtaining some 40 pages of lists of terms. It was calculated that a year's translations would generate some 500 000 pages of term lists.

Furthermore, the TermOrganiser screens we have seen above proved inadequate for handling this volume of unwanted terms. Each screen view contained essentially unwanted data and this, combined with problems of ergonomy when a basic operation for each term called for several mouse clicks, rendered the validation operation unpractical.

The TermFinder and TermOrganiser modules had been introduced successfully by Xerox to other clients, particularly in the automobile field where terminology is more specific than in our case. It was clear, however, that our texts offered a special challenge which called for an unproved version of the software.

The outcome of this first evaluation phase was that work was stopped at an early stage prior to any final term validation, and no terminology was retained. We believed that Xerox had the know-how to offer a better solution and therefore arranged further discussions with the Xerox Research team in Grenoble, France.

Xerox were extremely receptive to our needs and, following detailed discussions of our problems and requirements, offered to run further tests using algorithms which were still in the research stage and had not yet been integrated into a commercial product. We were therefore able to envisage further evaluation in the hope of greater success.

8. Evaluation Phase II

Thanks to Evaluation Phase I we were able to fully master the extraction procedure, from data preparation through to terminological record generation. Each step undertaken during that first phase had been totally new to us, and we did not have a real world model to follow for guidance in bilingual extraction and validation.

Evaluation Phase II allowed us to draw on experience gained. We took the opportunity of carefully examining and consolidating each stage of the process.

The key objective of this second phase was to reach a decision on the validity of the new extraction algorithms offered by Xerox. These algorithms could not at this stage be incorporated in the commercial product without further development of the TermOrganiser module. It was therefore decided to run the new algorithms independently in the Xerox research laboratory, and devise our own simple tools for carrying out term validation.

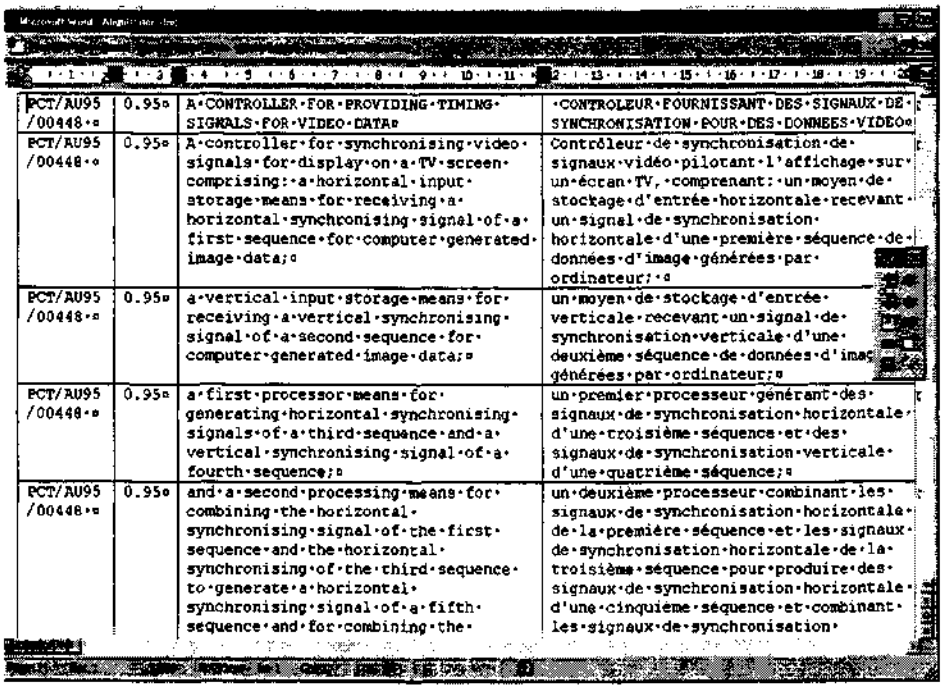
8.1 Data Preparation - Evaluation Phase II

The initial data processing scripts for loading the data from archive tapes and preparing the required SGML files were reviewed and considerably streamlined. The resulting files were fully adapted to the Xerox input requirements and document type definition (DTD). Selection according to subject area on the basis of IPC codes was also improved. Reliable routines are now in place for future use.

For the Phase II tests, we decided to create a new, large, homogeneous corpus which theoretically would give better extraction results due to greater term repetition and a higher number of term pair co-occurrences. A total of 4436 English abstracts, with their translations into French, constituting a corpus of 1.34 million words, were selected in a single subject field, that of "Communication technique" corresponding to IPC code H04.

8.2 Sentence Alignment - Evaluation Phase II

This data set was aligned by Xerox. We had the task of checking the alignment. For this



PCT/AU95 /00448	0.95	A CONTROLLER FOR PROVIDING TIMING SIGNALS FOR VIDEO DATA	CONTROLEUR FOURNISSANT DES SIGNAUX DE SYNCHRONISATION POUR DES DONNEES VIDEO
PCT/AU95 /00448	0.95	A controller for synchronising video signals for display on a TV screen comprising: a horizontal input storage means for receiving a horizontal synchronising signal of a first sequence for computer generated image data;	Contrôleur de synchronisation de signaux vidéo pilotant l'affichage sur un écran TV, comprenant: un moyen de stockage d'entrée horizontale recevant un signal de synchronisation horizontale d'une première séquence de données d'image générées par ordinateur;
PCT/AU95 /00448	0.95	a vertical input storage means for receiving a vertical synchronising signal of a second sequence for computer generated image data;	un moyen de stockage d'entrée verticale recevant un signal de synchronisation verticale d'une deuxième séquence de données d'image générées par ordinateur;
PCT/AU95 /00448	0.95	a first processor means for generating horizontal synchronising signals of a third sequence and a vertical synchronising signal of a fourth sequence;	un premier processeur générant des signaux de synchronisation horizontale d'une troisième séquence et des signaux de synchronisation verticale d'une quatrième séquence;
PCT/AU95 /00448	0.95	and a second processing means for combining the horizontal synchronising signal of the first sequence and the horizontal synchronising of the third sequence to generate a horizontal synchronising signal of a fifth sequence and for combining the	un deuxième processeur combinant les signaux de synchronisation horizontale de la première séquence et les signaux de synchronisation horizontale de la troisième séquence pour produire des signaux de synchronisation horizontale d'une cinquième séquence et combinant les signaux de synchronisation

purpose we imported the alignment results into a tabular format in MSWord and created a number of macro-commands to provide the necessary functionalities. The screen shot shows our layout with the full contexts being visible in each language. Segments can be merged, or

split at the cursor position, by clicking on the relevant up or down arrows in the dedicated toolbar. Lines can be created to accommodate new segments resulting from a split.

This arrangement constituted a very flexible alignment correction tool, making on-screen work quite efficient. In correcting the segmentation, the choice arose as to the extent to which segments should be broken down into smaller portions. For monolingual extraction the longer segments in the form of complete sentences give better results due to more effective parsing, while for bilingual matching, the shorter the segment the better. It would be useful to do comparative tests with short and long segmentation of a text. Unfortunately, this goes beyond the scope of our activities.

The data set aligned into 25 000 segments, 85 % of the segments being aligned with 100 % accuracy (as determined by software). Manual correction of the alignment took 20 person-hours.

8.3 Term Extraction and Bilingual Matching - Evaluation Phase II

This operation was performed entirely by the Xerox Research team using their algorithms. The raw output data in the form of a list *of 277* 000 term pairs were then sent to us in text format.

8.4 Bilingual Term Pair Validation - Evaluation Phase II

Given the sheer volume of data, this operation is inevitably time-consuming, but essential. To rationalise the approach, it was decided to tackle validation in three stages.

a) **A clean-up stage,** aimed at reducing the corpus to a manageable size for validation of terms and term pairs.

b) **A first validation phase,** for identifying those terms and contexts that could be imported directly into the terminology database. It had become clear during the clean-up stage that the expertise of terminologists and experienced translators, with knowledge of patent translation, was essential for this operation.

c) **A final validation phase,** for processing remaining terms requiring further clarification or verification. Again, this work must be done by experienced staff.

Consideration was also given to the methodology to be applied for selecting terms during the clean-up and validation stages. Three possibilities were envisaged.

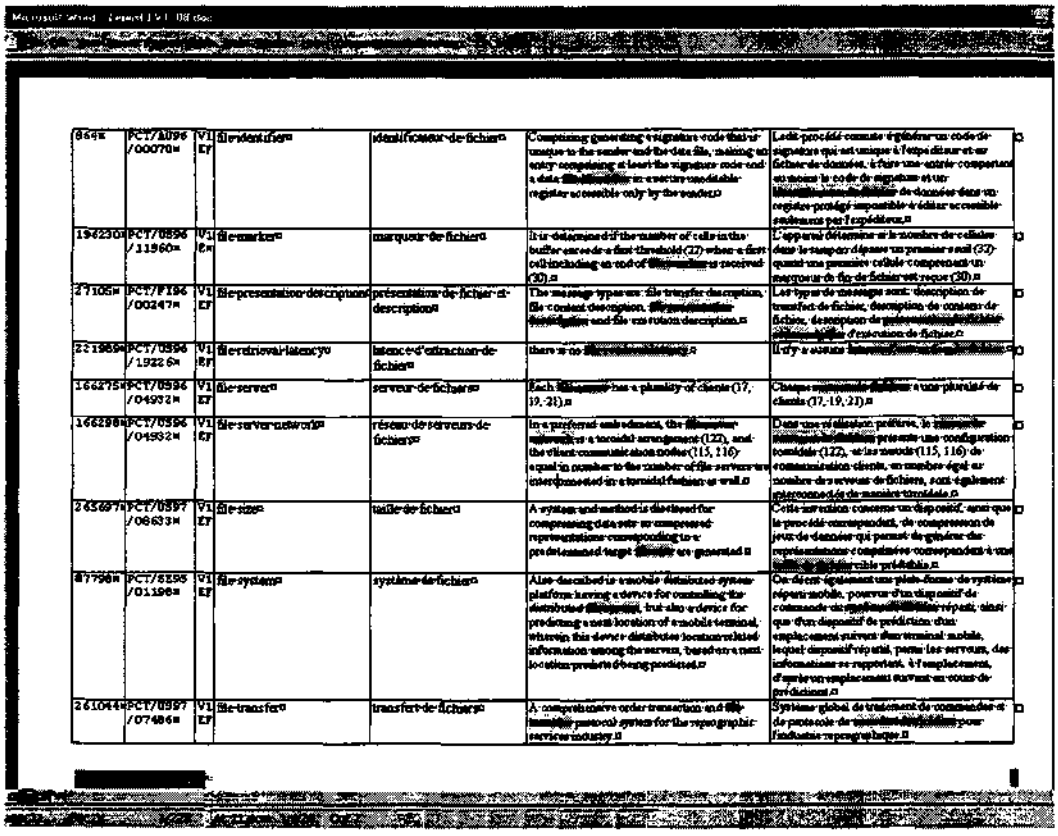
a) **Monolingual validation** of the list of terms in each language, followed by bilingual validation and then contextual validation. We could go through the list of extracted English terms, deleting unwanted terms and their translations. Various problems arose. If a term were repeated several times, we would simply select the first occurrence. However, there was no means of knowing if either the translation or the context were correct. Perhaps the second or third, or even thirtieth, occurrence would have been the most appropriate to retain. Nor would it be possible to retain several occurrences of a given source term with different translations. Furthermore, a term which appeared to be an incorrect extraction, and hence would be deleted, might in fact have been part of a

longer noun phrase which it would be useful to reconstitute. This form of "silence" would not be visible in monolingual checking.

b) **Bilingual validation** involving checking the list of term pairs bilingually, without context. This would have overcome some of the problems encountered with monolingual validation, but still left the difficulty of deciding which of several proposed translations for a given source term was indeed correct. A typical example was the English term "access port" which had three translations: "port d'acces", "porte d'acces" and "orifice d'acces". In this subject field, the temptation would have been to select the first of the translations as being the most appropriate. However, closer examination showed them all to be correct in their specific contexts.

c) **Complete in-context validation** appeared at first sight to be the most complex and most demanding approach. However, in view of the points raised above, it proved to be the most effective. The nature of the context was also important for the final terminological record since, by selecting a context which served also as a definition or explanation of the term, or in particular gave the full form of an initialism or acronym, we were able to increase the information content of the record.

Given the volume of data and the need to scan a large amount of text, we decided to do the



validation by working on paper rather than on screen. A suitable MS Word format was again used, after first processing the data in MS Access and then applying dedicated Word macros. The terms are listed individually and highlighted in the corresponding context making them easy to locate.

The main disadvantage of working on paper was that the ID codes (left-hand column) for terms selected had to be entered manually in the MS Access database to update the records. Working in this way, the initial clean-up of the extraction results took around 200 hours, reducing the 277 000 initial candidate term pairs to around 35 000.

These results indicate the presence of a large amount of noise. Note that the figure of 277 000 terms includes all repeated extractions of a given term from the different segments. The number of different term pairs extracted was around 100 000, and the number of different source English terms was around 50 000.

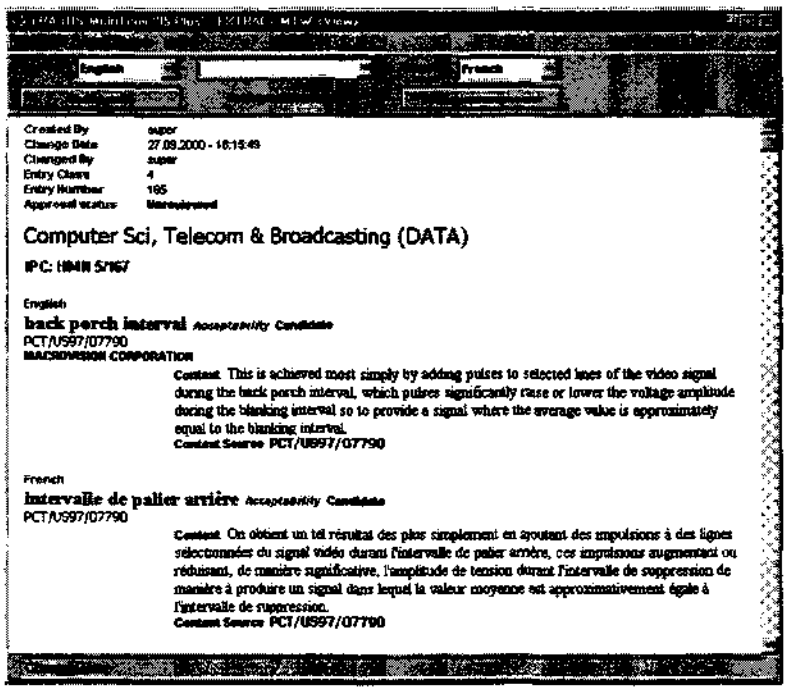
Noise took the form of incorrectly matched terms, irrelevant terms, incomplete terms, and incorrect translations.

Silence, where terms exist but have not been extracted or correctly matched, is difficult to determine. The only method of checking for silence seems to be to identify terms manually and then compare with the results of automatic extraction. This is of course extremely tedious and time-consuming.

Of interest is the nature of the terms extracted. We have the normal single and two-word combinations readily identifiable as technical terms, but also many multiple word phrases that extend beyond the normal definition of a technical term and enter the general language category. A conflict now arises between the terminologist's approach and that of the translator, the former wishing to keep only the highly technical terms and the latter wishing to retain also the less technical phrases which often provide a solution to a known translation difficulty. We may need to rethink the structure or mode of application of our terminology database.

8.5 Automatic Generation of Terminological Records - Evaluation Phase II

This step has been briefly experimented. Working from the MS Word document file, it is a



fairly simple matter to transform the selected terms and contexts automatically into an input file for direct importation into Trados MultiTerm with the result shown in the screen-shot. Additional information such as patent applicant and subject area can be added automatically from the original data files.

9. Next Steps

Having obtained what appear to be very useful results, we are now considering how best to use these results and determining the purposes they may serve. A database of technical terminology specific to our operation can be built up, with evident benefits for our translators. The extraction process can be extended to other key subject areas and to other language combinations.

Other benefits have become apparent. In particular, it has been possible to identify differing or incorrect translations of certain terms or phrases and propose correct solutions in the database with a view to ensuring greater accuracy and future harmonization of the terminology used in our translations.

It is now envisaged to investigate integration of the database into the translation process and move towards semi-automated use of the base by submitting new texts in electronic format to a translation workbench type of environment. The software would identify terms occurring in the new text for translation and propose translations from the database. Conversely, if a term in the new text is not detected in the database, it will not be highlighted in the text and the translator will know not to look for it in the base. Time wasted on fruitless searches is eliminated.

Further applications of this terminology extraction and terminology database technology within the patent application processing procedure can include:

- Contribution to sub-sentence level alignment in translation memory.
- Enhanced browsing and querying possibilities when performing searches for reference material.
- Facilitation of cross-language retrieval of technical information: enter a search term in one language and retrieve relevant documents in several languages.
- Comprehension aid for reading technical documents in languages other than one's own.
- Terminology expansion to enrich term sets used to provide designations such as in International Patent Classification.
- Controlled authoring to provide assistance and improve accuracy when drafting new texts (with consequent benefits for subsequent translation).

10. Conclusion

Automatic Bilingual Terminology Extraction provides an interesting new approach to our terminology problem. Results to date are encouraging, and we are confident that, by persevering in our search for innovative solutions, we will see a continued improvement in the extraction and validation processes. To that effect it is imperative to be able to call upon

highly qualified staff to carry out the validation phase. The "automatic" part of the process is of course vital, but in itself does not provide a useable end product. The participation of experienced translators and terminologists is therefore essential.

Close attention will now be given to the methodology to be applied in selecting and validating the candidate term pairs generated. At the same time we will be investigating methods of integrating the resulting terminology database into our translation operation.

An important place is being given to the development of computer aids in PCT Translation with the aim of enhancing present working methods and achieving improvements in translation efficiency and quality. We feel that automatic bilingual terminology extraction is an extremely promising field of research and will provide a valuable contribution to these aims.