[*Translating and the Computer 27*, November 2005 [London: Aslib, 2005]

EXTER: a breakthrough solution for efficient Terminology Extraction

Cyril Chantrier,
Strategic Partners Director,

TEMIS SA
Grenoble site
5, rue du tour de l'Eau
38 400 Saint Martin d'Heres
France

cyril.chantrier@temis-group.com
Tel:+33 (0)456 38 24 02

1. Summary

In 1998, Xerox's Linguistics Business Unit, at XRCE, started to develop embedded Natural Language Processing applications and products for Computer Aided Translation markets.

TEMIS, as a dedicated text analytics company, followed in this direction by acquiring the Xerox technology and team. It was very clear to us that advanced terminology creation and extraction solution were required to bootstrap any project dealing with Content creation and Annotation, Technical authoring, & Translation. Additional application areas such as Knowledge Management or Information Access Technology are developing rapidly.

Therefore, we made the decision to dedicate some part of TEMIS R&D effort and to partner with EDF R&D to develop a new generation of extraction engine, ExTer.

The purpose of this paper is not to enter deeply into the product engineering nor its technological components but to provide an overview of the solution with also some customer case that confirm the value of the solution, both on a technical and a financial standpoint.

ASLIB community and Conference is the ideal place for TEMIS to announce such a new solution.

2. History and Background: From "Xeronyms"to "Temisonyms"

In June 1998 - Xerox Research Centre Europe (XRCE) announced the commercialization of one of its key innovations, Xerox Translation & Authoring Systems (XTRAS). These tools were brought to market by a new Xerox business unit, Multilingual & Knowledge Management Services (MKMS). This unit had been

created by Xerox Document Services Group (DSG), the corporation's worldwide and fastest growing, consultancy, systems integration and outsourcing solutions division.

Monica Beltrametti, director of XRCE Grenoble Laboratory, commented at that time: "Our researchers work alongside Xerox customers to ensure that the technologies we develop have real business applications and lead the corporation towards new market opportunities. Today, with the commercialization of the XTRAS applications, we have once again achieved our goal of bringing document technology solutions to market."

In this period, The XRCE's Grenoble Laboratory research group was led by Annie Zaenen and Laurie Karttunen.

The XTRAS applications were marketed first as "The Multilingual Suite" by MKMS and incorporated tools to support the terminology and translation processes as well as end-user comprehension. At that time, the offering encompasses a Translation Memory system using "linguistic fuzzy matching for sentences" versus character based fuzzy matching (as an example, among "cheval", "cheveu" and "chevaux" the system was able to recall the pair "cheval-chevaux" - linguistic analysis - rather than the false candidate "cheval-cheveu" - character based), and Terminology applications. These solutions were addressing the following processes:

- Automatic Terminology extraction, monolingual or bilingual. This solution is based on TermFinder, using XeLDA® as a linguistic layer to extract Noun Phrases and proposing terms candidates. TermOrganizer is the visualization and validation interface, once the extraction has been done by TermFinder. Note also that this solution enables bi-lingual terminology extraction and validation
- Authoring Process document consistency while using the correct terms at the source. The tools used are TermChecker and TermOrganizer
- Comprehension Assistance by providing, the contextual translation of a word or a complete expression in a single click. The tools here are Web@ssistant and TermOrganiser

Xerox' presence in the linguistic tools market was first established by InXight, which marketed information retrieval software based on XRCE's linguistic technology. This software had been integrated at that time into leading Internet search and retrieval engines, including those of Verity, AOL and InfoSeek. XRCE's partnership with MKMS was to extend this market presence, allowing Xerox to address global companies' multilingual document management issues.

At the beginning of 2001, MKMS decided to focus on Terminology Solutions and marketed its product offering under "Xerox Terminology Suite". In 2002, MKMS was split in two entities, one dedicated to federated search technology, named "Xerox askOnce" and one dedicated to linguistic products. In July 2003, the Linguistic Products Business Unit was acquired by TEMIS, which led to the renaming of XTS as "eXtraction Terminology Suite ™"

This terminology extraction solution has been licensed by many organizations and has been found to increase terminology extraction and validation productivity by a factor 10 compared to manual extraction.

2.1 XeLDA® as the underlying Technology

The TEMIS extraction Terminology Suite ™ uses the XeLDA® framework as an underlying linguistic technology.

XeLDA® is a multilingual linguistic engine. It models and standardizes unstructured documents in order to automatically exploit their content.

Based on a technology developed through 20 years of research and development, XeLDA® provides advanced Text Mining features enabling textual information processing in 16 languages: English, German, French, Italian, Spanish, Portuguese, Dutch, Czech, Greek, Hungarian, Polish, Russian, Danish, Finish, Swedish and Norwegian (Bokmal).

XeLDA® offers a scalable range of services based on natural language processing components that can be integrated in business applications:

- Language identification: automatically recognizes the language used by each document

- Segmentation: divides a text into sentences

- Tokenization: splits a text into basic lexical units. While in some languages punctuation and spacing often provide a good indication of word and sentence boundaries, special cases such as contractions(aren't), possessives (Bart's) and abbreviations (Inc.) can make accurate tokenization a non-trivial task.

- Morphological analysis: returns the normalized form (the lemma) and the potential grammatical categories for all the words identified during the tokenization stage. For instance, when a user inputs 'bought' the system will deliver 'buy'. This differs from non-linguistic methods, such as wildcarding or tail-chopping, which will degrade the accuracy of text-based information retrieval, e.g. a tail-chopping system would match 'dining' with 'din' instead of 'dine'. For instance the French word 'lune' (moon) will be identified as a feminine singular noun and 'swam' as a simple past tense of the verb swim.

- Part-Of-Speech disambiguation: determines the exact grammatical category of a word according to its context.

  Morphological analysis is crucial with languages other than English, where words are affected by inflections and features such as gender, number, case, person or tense. The system has to identify all the variations of a word associated with plural or singular, feminine, masculine or neutral, or with different tenses and persons in the case of verbs, as well as the exceptions to the rule. For example, the word 'ground' can have different grammatical categories in each of the following context situations:

• Falafel is made from ground (adjective) chickpeas
• It is safer to stand on the ground (noun) than on the table
• I ground (verb) the coffee beans

- Extraction of noun phrases: extracts sequences of words that form noun phrases

- Dictionary lookup: identifies the context of a word to find the corresponding dictionary entry

- Recognition of idiomatic expressions: recognizes the expressions found in a text

2.2 TEMIS overview

TEMIS was founded in September 2000 by a team of managers, researchers and consultants from IBM who saw a market requirement for innovative Text Mining Solutions. This skilled team had earlier developed and distributed IBM Text Mining solutions worldwide.

This team had developed IBM linguistic solutions such as Text Knowledge Miner and IBM Technology Watch. TEMIS management signed an initial licensing agreement with Xerox for XeLDA®, their acclaimed linguistic engine. This agreement later led TEMIS management to acquire Xerox linguistics operations.

With this acquisition, TEMIS complemented its development team with renowned linguistic experts and the existing TEMIS solutions with XeLDA® and extraction Terminology Suite™.

Currently, TEMIS employs more than 50 people and operates subsidiaries in France, Germany, Italy, the UK and the US. TEMIS products are distributed world-wide through its partner network.

TEMIS is a software company designing, developing and distributing corporate Text Mining solutions. The company is the European leader for this technology in size, international presence and revenue.

TEMIS is the first company to have packaged its products both by business needs(Competitive Intelligence, Customer Relationship Management, and Human Resources) and vertical markets (Life Sciences, Publishing, Automotive, and Homeland Security). TEMIS offers breakthrough solutions for indexing and organizing collections of documents and extracting information. These solutions enable corporations or organizations to tackle information overload and unlock the business value of large collections of textual data so far unexploited.

The product offering and the value proposition are based upon:

- Information Extraction engines: XeLDA® and Insight Discoverer™ Extractor (IDE)

- the Skill Cartridge™ Library, such as ExTer
  Information Classification engines: Insight Discoverer Categorizer (IDK) and Insight Discoverer Clusterer (IDC)
- Corporate application solutions: Online Miner™ and extraction Terminology Suite™.

2.3 IDE

Insight Discoverer™ Extractor is an information extraction server dedicated to analysis of text documents. It detects pieces of information that are the most relevant to users: a merger announcement in an article for an analyst, a sales opportunity in an e-mail message for a customer relationship manager, a specific skill in a resume for a recruiter, for example. It reads electronic documents in over fifty different formats and covers 16 languages: English, German, French, Italian, Spanish, Portuguese, Dutch, Czech, Greek, Hungarian, Polish, Russian, Danish, Finish, Swedish and Norwegian (Bokmal).

Insight Discoverer™ Extractor performs a sequence of linguistic analyses:

- Corpus recognition: reads 50 formats with automatic language identification
- Morpho-syntactic analysis:
  - Assignation of grammatical categories to each word in a document (noun, adjective, verb, etc.)
  - Lemmatization: returns each word to its base form (singular for a plural, infinitive for a conjugated verb) so that it can be recognized independently of its inflected form
- Knowledge extraction (runs extraction rules):
  - Recognition of entities (names of companies or people, numerical data, dates, places, etc.)
  - Identification of relationships between the entities (mergers and acquisitions between companies, interactions between proteins and small molecules, etc.)
  - Information extraction is the process that enables identification of relevant information. The relevance criteria are expressed in the form of knowledge components grouped in a Skill Cartridge™.
- The flexibility of its Skill Cartridge™ system means that Insight Discoverer™ Extractor can extract sophisticated information from written language:
  - Identification of negative or positive trends
  - Differentiation between rumors and actions
  - Resolution of anaphora
  - Acronym management
  - Enhancement of ontology

2.4 The Skill Cartridge™ Library

TEMIS has developed a unique technology to address the diversity of issues its clients regularly face. A Skill Cartridge™ is a linguistic plug-in which adapts the extraction of information to the areas of specific interest to each client. They consist of both specific terminologies (taxonomies, thesauri, ontologies) and linguistic rules. The TEMIS Skill Cartridge™ system enables businesses to adapt extraction rules to their

business systems. A Skill Cartridge™ can be either vertical (e.g. Life Sciences, Automotive, Publishing, etc.) or functional (Competitive Intelligence, Bioinformatics, Intellectual Asset Management, etc.) TEMIS has developed a Skill Cartridge™ Library in different languages (among which English, French, German, Spanish or Italian). Skill Cartridges™ can be packaged and distributed by TEMIS or Certified Partners, or they can be developed directly by TEMIS customers and then remain their sole intellectual property.

TEMIS provides its customers with the Skill Cartridge™ Development Kit. This package includes the Skill Cartridge™ Studio, a Skill Cartridge™ compiler and a testing environment, which enable TEMIS customers to customize the standard Skill Cartridges™ (specific rules, specific vocabulary) or to create their own Skill Cartridges™. The standard Skill Cartridge™ Library currently comprises nine off-the-shelf Skill Cartridges™ and their vertical adaptations.

Generic Skill Cartridges™

- Analytics Skill Cartridge™ that targets the extraction of key linguistic categories including terminology for a broad range of languages. This Skill Cartridge™ automatically identifies major semantic categories that provide a content snapshot of a document. This knowledge is used by the Temis Insight Discoverer™ Clusterer and Insight Discoverer™ Categorizer for an enhanced classification of unstructured documents. The Analytics Skill Cartridge™ is an integral part of the Insight Discoverer™ Extractor.
- Text Mining 360° Skill Cartridge™ is an advanced multilingual text analysis tool for automatic content exploration of all types of textual data. This Skill Cartridge™ identifies key named entities, such as names of people, companies, products, locations, establishments and organizations, as well as numerical data, such as time and monetary expressions, measurements, addresses, phone numbers.
- Competitive Intelligence Skill Cartridge™ that extracts strategic and critical business information. This Skill Cartridge™ extracts financial information (revenue, growth, sales) commercial information (market share, number of customers), stock information (capitalization, trends) and all the information related to merger and acquisition, joint-ventures, partnerships, research strategies, etc. The Competitive Intelligence Skill Cartridge™ can be applied to news feeds, competitors' websites, analyst reports, scientific publications or legal information sources.
- **ExTer Skill Cartridge™** for Extraction of Terminology, which is dedicated to terminology extraction. See below.

Domain Specific Skill Cartridge™

- Human Resources Skill Cartridge™ to support businesses in their recruiting process. This Skill Cartridge™ automatically extracts competencies and know-how from applications and resumes. Its extraction rules can be combined with recruitment criteria to only extract relevant information according to recruiting strategies.
- Life Sciences Skill Cartridges™ embedding Competitive Intelligence Skill Cartridges™ special edition for Life sciences, Biological Entity Relationships Skill Cartridge™, Medical Entity Relationships Skill Cartridge™ or Chemical Entity Relationships Skill Cartridge.

3. ExTer Project

3.5 Background

The EDF Group is among the key players in the field of electricity generation, distribution and supply in Europe. Managing a generation mix with a capacity of 125.4 GWe, it provides energies and services to 42.1 million customers throughout the world, including 36.2 million in Europe.

The EDF Group is made up of Electricité de France, parent company (EDF SA), and a network of 75 affiliates and investments established in Europe and around the world.

EDF Research and Development Division has written and maintained LEXTER, an application for French language terminology extraction, for many years. LEXTER was considered a reference application by the experts and many universities which were using it.

EDF R&D was keen to commercialise this software and was looking for an industrial partnership to make it happen. As TEMIS and EDF had a previous, successful commercial and cooperation relationship, TEMIS has been selected as the appointed partner.

TEMIS was very interested in this project as the new terminology solution would complement the existing one based on eXtraction Terminology Suite™.

ExTer is a new generation of terminology extraction tool based on Lexter prototype and TEMIS extraction technology,

3.6 General architecture of the proposed solution

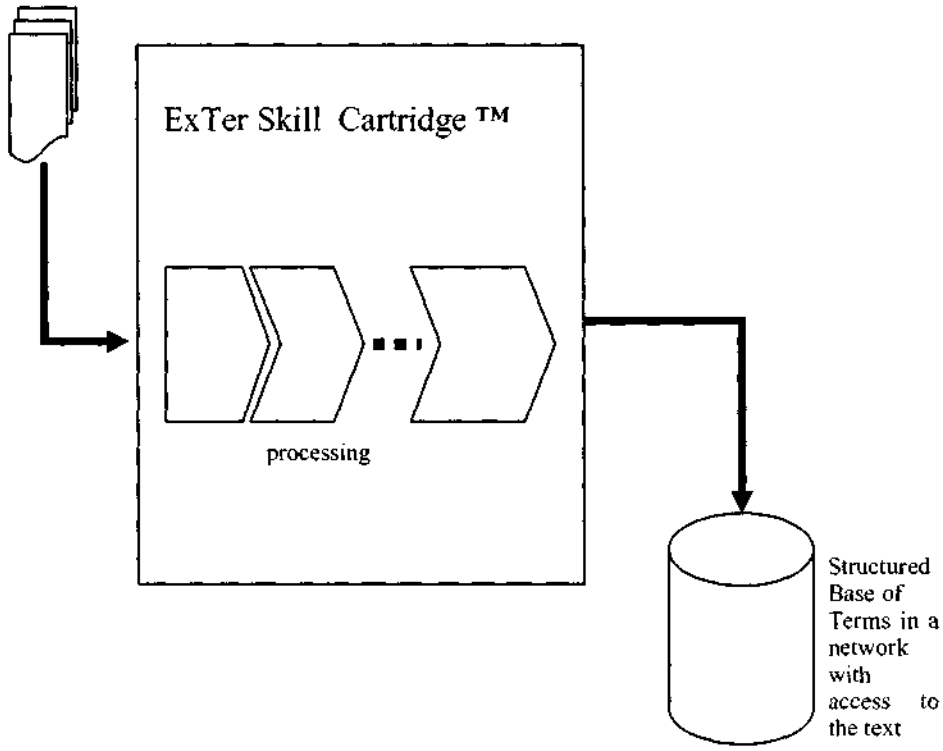
The overall architecture - see Figure 1 below- of the solution is based on:

1. Technical Architecture of IDE :
   Insight Discovery™ Extractor (IDE) V2.0, as the TEMIS generic knowledge extraction tool that runs:
   - Formatting: Transformation of input documents- XML, MS Word, TXT, ASCII, PDF into a pivot format - HTML - easier to use for other processing
   - Morpho syntactic tagging of the text by XeLDA®
   - Extraction of Noun Phrase of Maximum size (NP Max) associated with their left context.

Other kind of specific processes (plug -ins) can also be incorporated in the process, to address (e.g.) filtering, cleaning, enrichment among those concepts which have been extracted by a linguistic Skill Cartridge ™ (SC).

This plug-in mechanism will allow the ExTer system to disambiguate and structure the NP max.

Figure 1: Functional Schema:



2. ExTer process scheme:

   In this terminology extraction process, IDE use the ExTer Skill Cartridge ™ which handles various tasks:

   - Extraction of Candidate Max NP and their context
   - Gathering of statistical measures on NP constituents: association scores (mutual information)
   - Handling of simple cases of Prepositional Attachments
   - Decompounding of NP Max in a tree structure of sub NP organized in "head-expansion"
   - Structuring: organization of the list of extracted NP in a network of terms that have relations like “Has for Head”, “Has for Expansion”, “is a variant of”.

One of the requirements of the system is to be able to keep the Key Word In Context accessible, alongside the extraction processes, which is critical for the terminologist, the translator or the knowledge officer.

Figure 2 shows the details of the various processes enabled by the ExTer Skill Cartridge ™

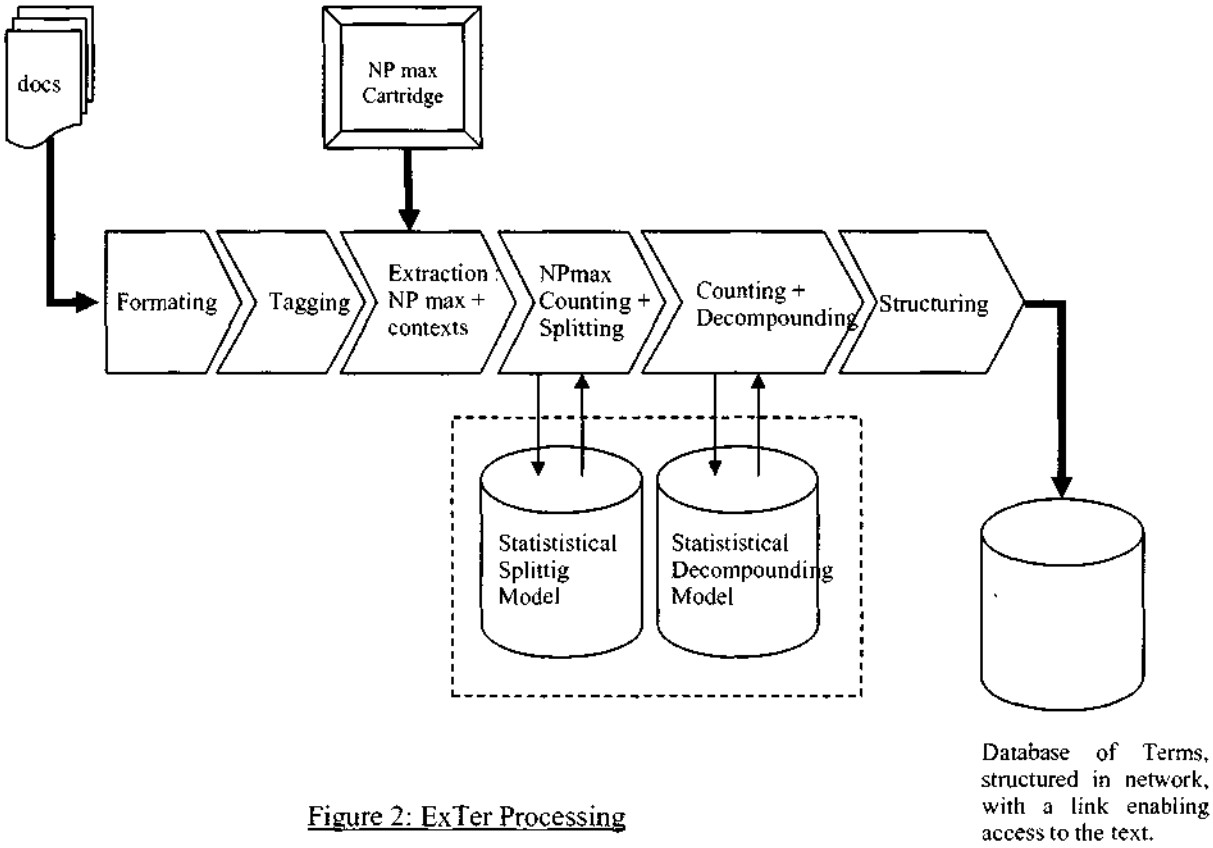


Figure 2: ExTer Processing

General Principles behind ExTer are driven by:

- A minimal set of language dependant syntactical rules to handle non ambiguous decomposition patterns
- A statistical decision procedure based on association scores for ambiguous NPs. max.

3.7 NP max Extraction Module

This module is based on a linguistic cartridge which is based on the already existing Temis NP extraction cartridge. This cartridge was "extended" so that it could fit as much as possible to the NP coverage enabled by Lexter. To do that match, we had to consider the following cases:

- Simple coordination cases handling **(chauffage urbain *et* collectif)**
- Handling of adverbial modification **(orbite *quasi* elliptique)**
- Handling of preposed adjectives ***(nouveau* système de satellite)**
- Handling of preposition (***à***, ***de,*** étendu à *dans, en, par, pour, sur, avec, chez*)

This cartridge enables to extract both « simple NP » for which structuring « head - expansion » is not ambiguous.

This is the case for all NP max of a length of 2 which are very present in corpus with the following patterns:

- (ADV) ADJ NOUN *très belle voiture*
- NOUN (ADV) ADJ *orbite quasi elliptique*
- NOUN1 NOUN2 *frein moteur*
- NOUN1 PREP (DET) NOUN2 *satellite de communication*

Note: NP length is related only to names, adjectives or verbs.

The « complex »NP composed of single NP with more than one prepositional modificator (Prepositional Phrase or PP) or adjectival (Adjectival Phrase or AP). These ones are ambiguous on a "head-expansion" structuring point of view, as multiple concurrent decompoundings are possible:

A typical case is given by the pattern :

- NOUN1 PREP NOUN2 ADJ *satellite de communication géostationnaire*
  combustion de fuel liquide

Possible decompositions :

1. NOUN1 PREP [NOUN2 ADJ] *combustion de [fuel liquide]*
2. [NOUN1 PREP NOUN2] ADJ *[satellite de communication] géostationnaire*

As a conclusion, the simple morpho-syntactic features are insufficient to choose a valid decompounding and to disambiguate. This decision has to be taken using other kind of information and algorithms.

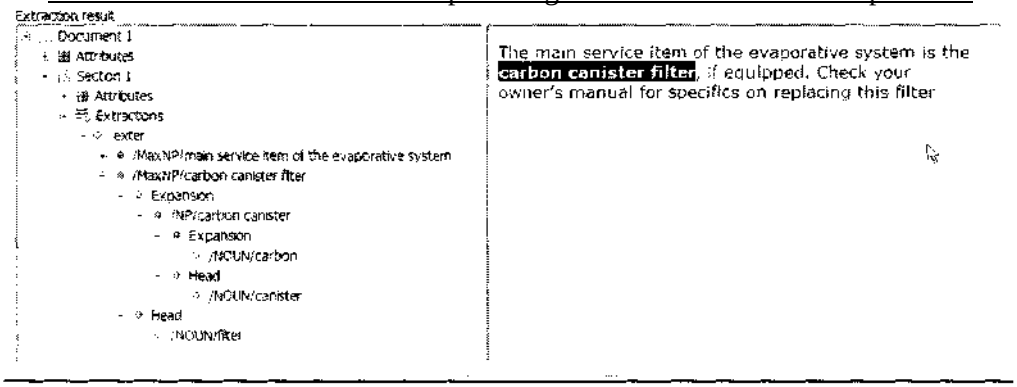
It was obvious that:

- The Lexter rules were useful and provided good quality but TEMIS also found out that the model was very language dependant. In addition, the number of rules was making the system difficult to maintain.
- The pure statistical model was also possible with advantage of being virtually "portable" to any language but we had the feeling that some quality increase could be leveraged using other method.

As a result of these investigations, the ExTer solution is an hybrid solution combining both statistics and linguistic rules:

- Rules are stable and easy to maintain, especially as they are a small number
- Statistical calculation is based on non ambiguous NP (length 2) on the totality of the corpus.

This screen shot shows the decompounding of a NP max in "Head-Expansion"



3.8 Structuring Module

We use here the decompounding of a term in « HEAD/EXPANSION » to build a network of terms. In this network, each tern is linked to the other terms with whom this term possesses a relation, such as "HAS FOR HEAD", "HAS FOR EXPANSION", "IS A VARIANT OF".

The goal of this structuring module is to store in a Database or XML File, all the information required to feed either the EDF R&D WorldTrek validation module or to export the content to any Terminology Management tool.

This information: is related to:

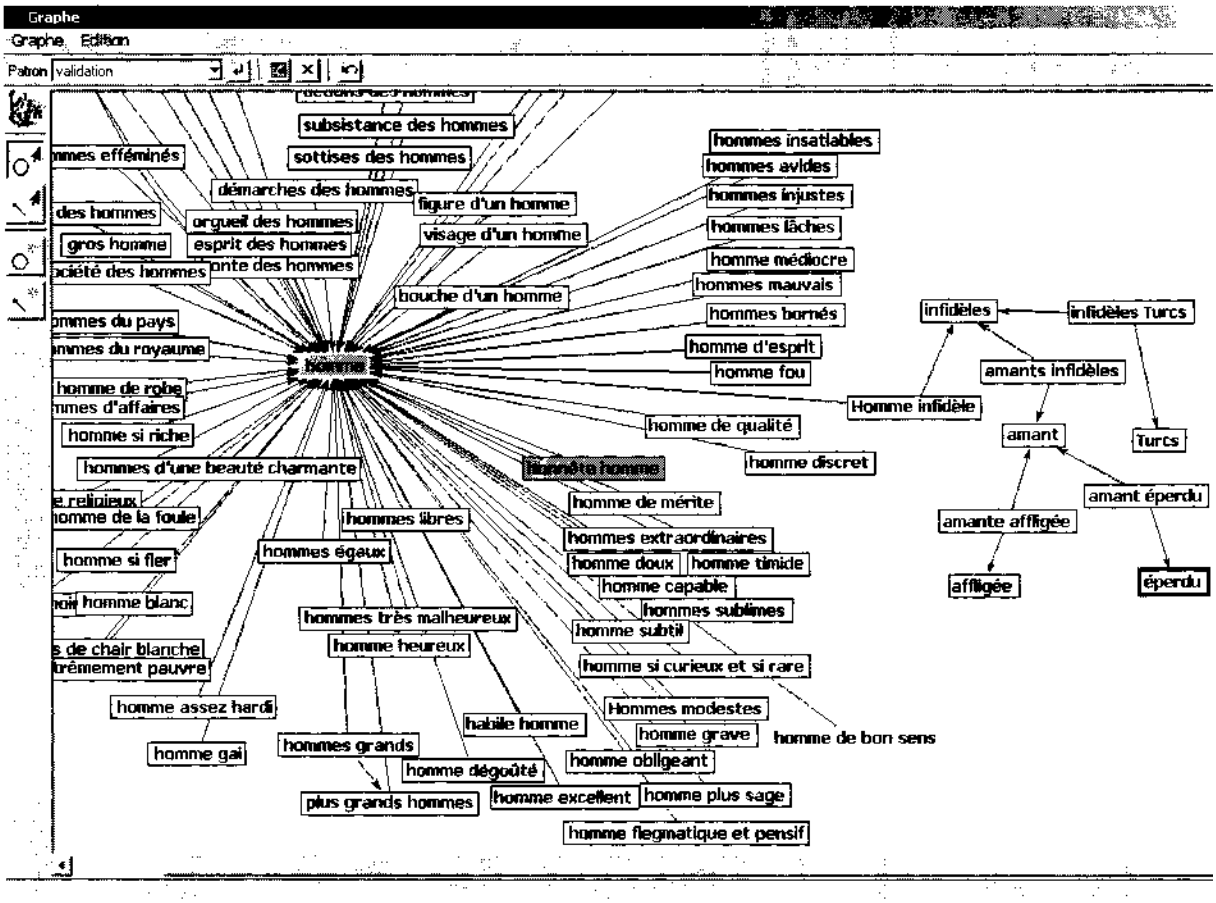
- Terms and their links between one another
- Identifiers /addresses-labels - enabling the access to the text.
- Frequencies

In comparison with XTS, where the content of extraction was "flat", ExTer provides a network of terms which can be considered as the basic layer to build a semantic

network and later on, ontology.

The figure below shows the EDF R&D WorldTrek user interface dedicated terminology validation.

(Copyright EDF R&D)



ExTer XML output format

```
- <entity id="2" typeref="1" freq="1">
    <l>valeur de=le train roulant</l>
    <c>VALEURS DES TRAINS ROULANTS</c>
    <attribute id="4" typeref="2" typename="freqIsol" value="1" />
    <attribute id="5" typeref="3" typename="prod" value="0" />
    <attribute id="6" typeref="1" typename="cat" value="MaxNP" />
  - <relation id="2" typeref="1" typename="te">
    - <member id="3" roleref="2" rolename="Head">
      - <entityref ref="3">
          <l>valeur</l>
        </entityref>
      </member>
    - <member id="4" roleref="1" rolename="Connector">
      - <entityref ref="4">
          <l>de=le</l>
        </entityref>
      </member>
    - <member id="5" roleref="3" rolename="Expansion">
      - <entityref ref="5">
          <l>train roulant</l>
        </entityref>
      </member>
    </relation>
  </entity>
</entities>
<relations>
- <relation id="4292" typeref="2" typename="VariantClass">
  - <member id="38012" roleref="5" rolename="Representative">
    - <entityref ref="6156">
        <l>interrupteur démarrage</l>
      </entityref>
    </member>
  - <member id="38013" roleref="4" rolename="Variants">
    - <entityref ref="5913">
        <l>interrupteur de démarrage</l>
      </entityref>
```

4. Customer case: TEMIS- Automotive Industry Customer- STAR Paris

4.9 Background

TEMIS and STAR Paris, the French entity of STAR Group, the Language Service Provider, have developed business relations for a couple of years now. Some typical

solutions which were delivered to customers were for example the eXtraction Terminology Suite™, with TermFinder to run the terminology extraction process and STAR WebTerm software to handle and diffuse terminology. STAR was also using TermFinder in-house for Global Translation projects where terminology extraction was required.

In spring 2005, STAR had to face a major terminology extraction project from one of its automotive industry customer. The decision has been jointly taken, after a first validation prototype to assess the quality and speed of extraction, to run the project with ExTer.

We will now describe the whole extraction process with formatting, conversion, selection and validation steps. A Final paragraph will capture early “business benefits” and “Return on Investment” as calculated by STAR.

4.10 Process

The Corpus was in French language and made of 4,099,268 words from eight different repair manuals, each divided in three parts (Bodywork, Mechanical parts and Diagnostic)

STAR converted the corpus in a standard format for Terminology Exchange and send this TMX file to TEMIS who hosted the ExTer extraction engine and who processed the extraction.

As a result, TEMIS delivered to STAR Paris an XML format with a number of candidates of 3,470,679 (i.e. every preposition, noun or noun phrase in the corpus)

Then STAR made the conversion of XML in CSV, for further processing in MS Access. At that stage, the number of candidates which were kept for selection was 2,055,648 (i.e. only noun and noun phrases from the previous total)

A Selection step had to be done independently for each part of manual and this leads to a number of candidates, without duplicates, of 43,730

The details of the selection step (i.e. list of candidates respecting client’s criteria, to be validated by experts) is as follow:
- First step: delete the noise (all obvious false records - like "$", "152D#",.... kept by the extraction engine)
- Second step: delete less obvious noise (errors in extraction due to language specificities like “aide de 1’outil” for “à 1’aide de 1’outil”)
- Third step: review of all terms to assign either a selected or a rejected status (all terms capable of being inserted in an automotive termbase)

At the end of this selection process, the number of terms sent for validation to the car manufacturer experts - mainly technical writers - was of 22,621 (with potential duplicates as the client wanted to see each term in the part of manuals where it appeared)

The validation step consisted of:
- the results of the selection were sent to the client to be reviewed by experts.
- the review by experts had the objective to assign 3 kind of status for each term validated if term has to be kept, refused if not, and erroneous, if term has to be mentioned in the base as false (with hyperlink to the corresponding validated term)

The finalization step showed that the number of terms validated was about 3,000 (only tools and process as that was all the client wanted to insert in the termbase). A consolidation as several experts have worked on the terms enabled this final selection. The terms were then exported in MS Access and Excel for diffusion and implementation and experts can now check whether a term is validated or erroneous when they write manuals. This project opened the door for Terminology Checking as an additional way to increase the consistency of document authoring and to reduce the translation costs.

4.11 Benefits

The STAR terminology extraction project had numerous targets, one of those being the implementation of a terminology-controlled technical writing environment in the customer authoring department.

In order to ensure the terminology consistency of the customer Repair Manuals, STAR proposed to install a XeLDA® based, dynamic terminology control feature in the existing customer authoring environment, namely “TermChecker”.. Before going ahead with this project, the customer wanted to get clear figures on the project return on investment (ROI).

Where such impact analysis is quite easy when switching from a non-structured, DTP based environment towards a highly granular, XML-based environment with sentence retrieval and terminology control features, it is a little more tricky when only terminology control is added.

Since it was nearly impossible to estimate the productivity gains at the authoring level, STAR proposed to calculate the impact of terminology control on the translation steps of the supply chain, especially in terms of leverage improvement following Transit pre-translation. The impact would, of course, be indirect, since terminology consistency is a word-level parameter where translation memory engines work at the sentence level.

In summary, the hypothesis was that controlling terminology would increase the statistical probability that two technical writers would use the same sentence (or a similar one) to express the same technical concept.

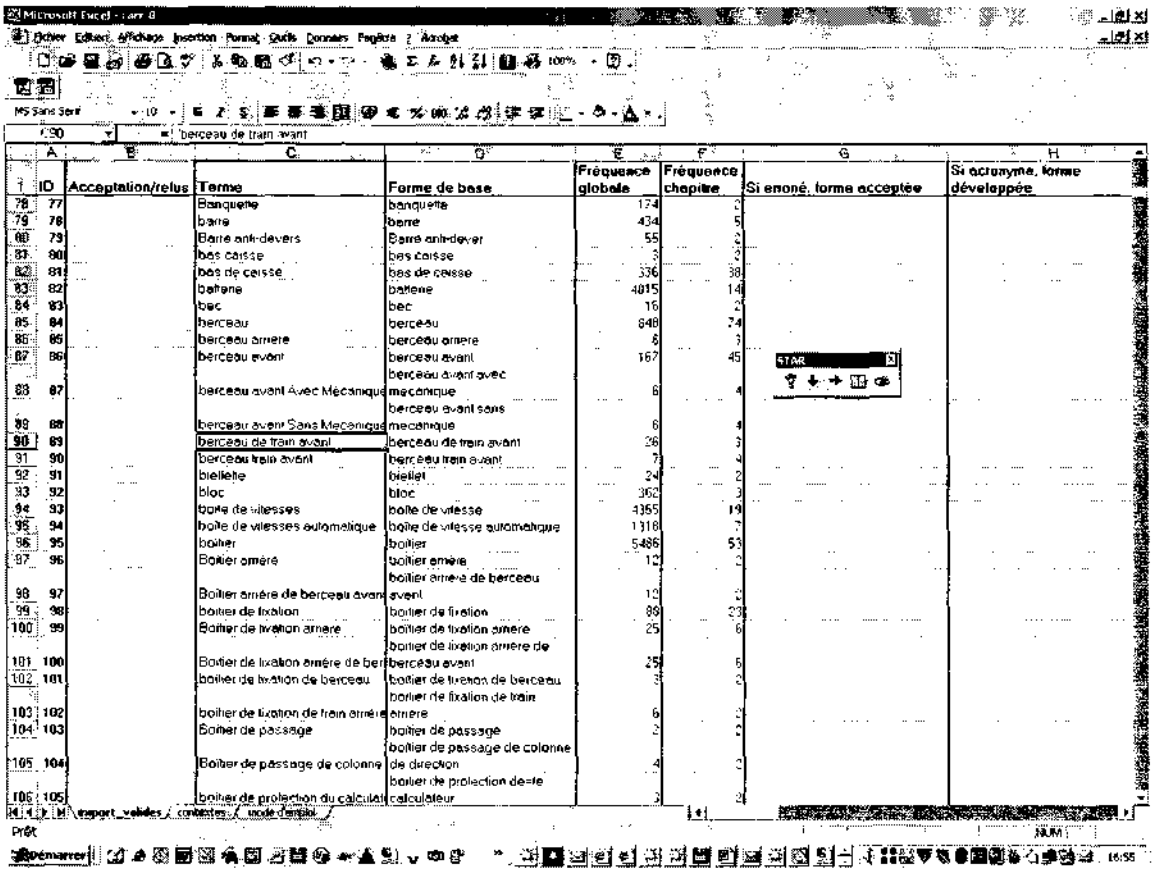
In order to estimate the potential savings without spending weeks in complex simulations, STAR used a simple, sample-based approach, replacing the different variants of four significant terms with their standardized, lemmatized forms in one translatable corpora and one reference corpora. After comparing the pre-translation results before and after this operation, STAR concluded that the increase in the number of Perfect and Fuzzy matches in the terminology-controlled sample would amount to a 5,5 % saving on the global yearly translation budget.

To this estimated ROI, STAR added potential incidental savings. For example, hectic source terminology is one major historical barrier that prevents lighter Project Management control without quality decrease; better source terminology control would allow going faster towards selective Project Management proofreading.

Amongst other incidental benefits, STAR mentions lower number of translator's requests regarding new technical concepts routed to the authors and validators, since the new terminology could be identified upstream and processed even before the translation actually starts.

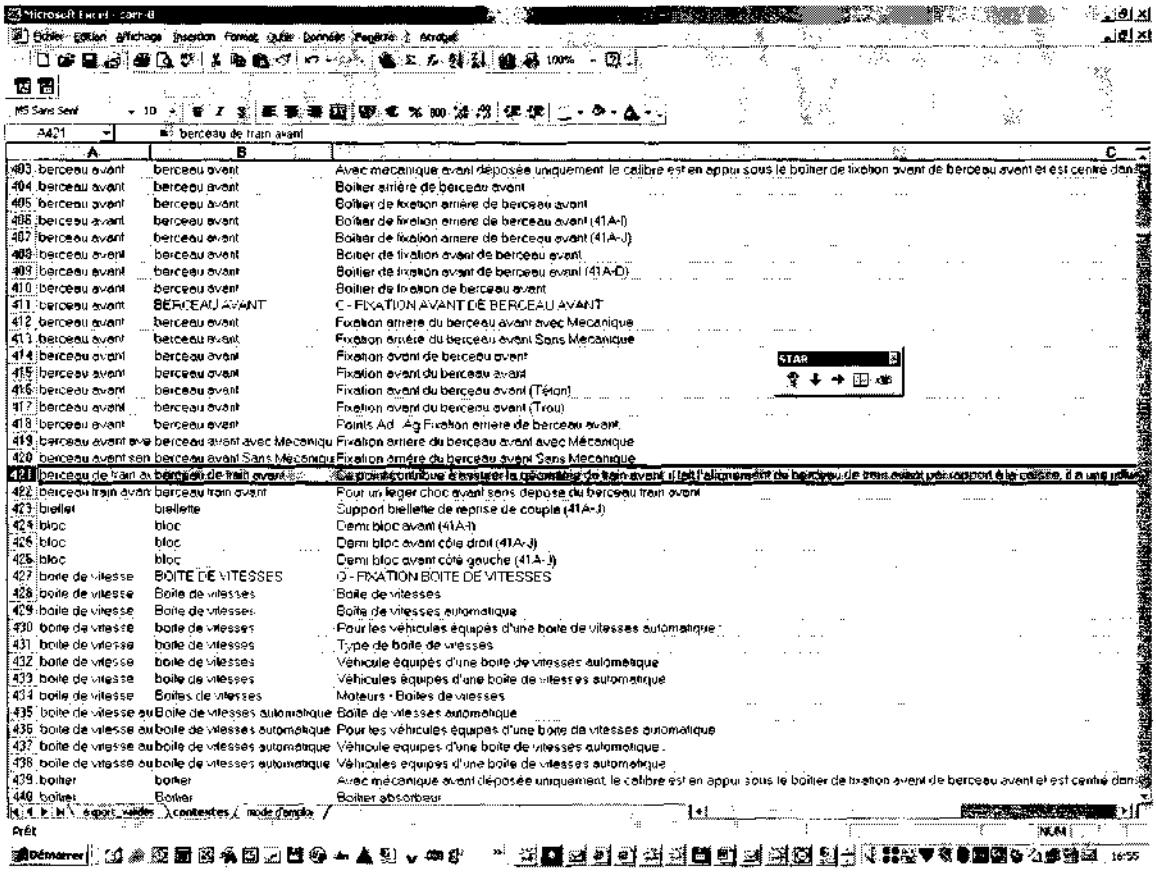
Globally, controlling terminology through the whole multilingual documentation process would definitely allow to create a direct, productive feedback between technical writers and delocalized translators : the first ones being aware of the terms that would generate interpretation problems at the translators level, they can either chose simple, homologated terms or add new terminology if a new technical or marketing concept emerges.

Validation view. Used by STAR for its Terminology Extraction project



	ID	Acceptation/relus	Terme	Forme de base	Fréquence globale	Fréquence chapitre	Si enoné, forme acceptée	Si acronyme, forme développée
78	77		Banquette	banquette	174	2		
79	78		barre	barre	434	5		
80	79		Barre anti-devers	Barre anti-dever	55	2		
81	80		bas caisse	bas caisse	3	3		
82	81		bas de caisse	bas de caisse	336	38		
83	82		batterie	batterie	4015	14		
84	83		bec	bec	16	2		
85	84		berceau	berceau	648	24		
86	85		berceau arriere	berceau arriere	6	3		
87	86		berceau avant	berceau avant	167	45		
88	87		berceau avant Avec Mécanique	berceau avant avec mecanique	6	4		
89	88		berceau avant Sans Mecanique	berceau avant sans mecanique	6	4		
90	89		berceau de train avant	berceau de train avant	26	3		
91	90		berceau train avant	berceau train avant	73	4		
92	91		biellette	biellet	24	2		
93	92		bloc	bloc	362	3		
94	93		boite de vitesses	boite de vitesse	4355	19		
95	94		boite de vitesses automatique	boite de vitesse automatique	1318	7		
96	95		boitier	boitier	5486	53		
97	96		Boitier arrière	boitier arrière	12	2		
98	97		Boitier arrière de berceau avant	boitier arrière de berceau avant	12	2		
99	98		boitier de fixation	boitier de fixation	89	23		
100	99		Boitier de fixation arriere	boitier de fixation arriere	25	6		
101	100		Boitier de fixation arrière de berceau avant	boitier de fixation arrière de berceau avant	25	6		
102	101		boitier de fixation de berceau	boitier de fixation de berceau	3	2		
103	102		boitier de fixation de train arrière	boitier de fixation de train arriere	6	2		
104	103		Boitier de passage	boitier de passage	2	2		
105	104		Boitier de passage de colonne	boitier de passage de colonne de direction	4	2		
106	105		boitier de protection du calculateur	boitier de protection de la calculateur	3	2		

Contextual view



5. Conclusion

Without ExTer such a terminology extraction project - size and quality here were critical - would even not have been possible. Both the Language Service Provider, STAR, and the customer, were extremely satisfied from the quality of the output and the relevancy of the selected and proposed terms.

This experience is very encouraging for future development like:

- Porting to new languages
- Increasing the efficiency of validation step by using external data to perform the extraction (like stop words or list of already validated terms)
- Extension to Verbal Phrase extraction
- Bridge to support Ontology building (OWL, XTM)

6. References:

- Bourigault D., « Lexter, un logiciel d'extraction de terminologie. Application à 1'acquisition des connaissances à partir de textes », Thèse de l'Ecole des Hautes Etudes en Sciences Sociales, Paris, 1994.

• Bourigault D., Gaussier E., « Etude de faisabilité de 1'intégration dans XeLDA des fonctionnalités LEXTER », Rapport du Lot 2, 1999