In search of knowledge: text mining dedicated to technical translation

Johanna MONTI[1], Annibale ELIA[1], Alberto POSTIGLIONE[2]
Mario MONTELEONE[1], Federica MARANO[1]

jmonti@unisa.it, elia@unisa.it , ap@unisa.it, mmonteleone@unisa.it, fmarano@unisa.it

[1] Department of Political, Social and Communication Sciences , Via Ponte Don Melillo, University of Salerno, Italy
2 Department of Computer Science, Via Ponte Don Melillo, University of Salerno, Italy

Abstract

Although a vast amount of contents and knowledge has been made available in electronic format and on the web in recent years, translators still do not have friendly and targeted tools at their disposal for the various aspects of a translation process, i.e., the analysis phase, automatic creation and management of the linguistic resources needed and automatic updating with the relevant information generated by the computer translation tools used in the process (Machine Translation, Translation Memories, and so on).

Text mining and information retrieval are not typically connected with the translation process and no existing online translation workspace integrates text mining or information retrieval facilities that are specifically aimed at improving the documentary competence of translators in order to process unstructured (textual) information, and make the information on the web or in texts accessible to translators.

This paper explores a new approach to helping translators look for different types of information (glossaries, corpora, Wikipedia, and so on) related to the specific translation work they have to perform which can then be used to update the lexical base needed for the translation workflow (both human or machine-aided). This new approach is based on CATALOGA, a text mining tool, which can be combined with an IR application and/or an MT/TM system and used for different purposes.

1 Johanna Monti is author of the Abstract, Introduction, Sections 2 and 4.3 and Conclusions, Annibale Elia is author of Sections 3.1 and 3.4, Alberto Postiglione is author of Section 3.2. and 5, Mario Monteleone is author of Sections 3.3 and 4.1 and Federica Marano is author of Sections 4.2. and 4.4.

1. INTRODUCTION

Information acquisition is a very crucial aspect in the translation workflow that has been underlined by many scholars in the field of translation theory and practice [Lebtahi and Ibert 2004] [Hönig H. 2006] [Kuβmaul P. 2007]. However, up to now, very little help has come from the IT community to make this task faster and easier.

Information extracted from the web or from texts has a valuable role to play in the analysis phase of a translation process (whether human or computer-assisted) since it helps detect typical translation reference material (texts, glossaries, dictionaries, comparable or parallel corpora and so on), identify the subject domain of texts and relevant concepts and terminology and detect similarities between documents or how they are related to other variables of interest.

Translators spend a lot of time on the preliminary phases of translation and this is particularly true when working on technical translations.

Surprisingly, computer translation tool producers focus mainly on the post-editing of raw machine translations and completely ignore the initial phases of the translation workflow, i.e., the search for information needed for the translation task and, for technical translations, mainly terminological compound words; if done properly, this can lead to a significant improvement in the translation engines and can speed up the whole translation process from the very beginning, i.e., analysis of the source text.

This paper explores a new approach to helping translators look for different types of information (glossaries, corpora, Wikipedia, and so on) related to the specific translation work they have to perform which can then be used to update the lexical base needed for the translation workflow (both human or computer-aided). This new approach is based on the combination of CATALOGA, a text mining tool, an IR application and/or an MT/TM system.

Up to now, CATALOGA, in its monolingual configuration, has been used to analyse large and heterogeneous text corpora with outstanding results. The paper illustrates the evolution of this tool towards a bilingual version and will show the results of experiments performed on Italian texts and their translations into English in order to evaluate the feasibility and limitations of our approach. Section 2 presents (i) an overview of the latest debate on documentary search in the field of translation theory and practice, (ii) the approach that we intend to propose, based on the extraction of knowledge from source texts, and finally (iii) how it can be used in relation to the translation task. Section 3 describes the CATALOGA system, the dictionaries that are used in combination with it and the different uses of the system in the translation task. Section 4 describes how CATALOGA can be used in a scientific/technical translation process. Section 5 outlines future research perspectives and Section 6 presents some conclusions.

2. DOCUMENTARY SEARCH IN THE TRANSLATION TASK: IS TEXT MINING OF ANY HELP?

According to Wills in *Science of Translation* [1997:118], the translator « must have a SL [source-language] text-analytical competence and a corresponding TL [target-language] text-reproductive competence».

These competences are based on the knowledge acquired by translators during their studies and their professional experience. However, it is unlikely that this knowledge and experience are always sufficient to cope with the translation task. The analytical competence of the source text and the reproductive competence of the original text in a target text are mainly based on the ability to

understand and interpret the original message and the ability to search and find all the necessary information is a fundamental aspect for translators.

The search for information represents a crucial part of the translation process and many scholars have recently analysed this aspect.

Lebtahi and Ibert [2004: 223] observe that nowadays the translation profession requires «une aptitude a traduire, une aptitude à la recherche documentaire et terminologique, et enfin une aptitude à travailler vite sous la pression du temps». They analyse the interdependencies between these three different competences: in particular, with regard to the interdependence between translation competence and technological competence in using software applications for the documentary and terminological search, they underline that:

> Les aptitudes à traduire, qui impliquent des aptitudes à la recherché documentaire et terminologique, sont en fait indissociables de la maîtrise des nouvelles technologies qui investissent l'activité. Disposer du materiel informatique adequate est aujourd'hui un imperative pour tous les traducteurs (adresse e-mail, retranscription des texts effectués sur un fichier informatique et respect de la mise en page demandée). La maîtrise de l'outil informatique et des resources qui lui sont liées devient une condition pour exercer le métier de traducteur, sans parler de la traduction des supports eux-mêmes. [2004:231]

Hönig [2006:160] discusses the relationship that exists between text comprehension and the search for information and highlights how search competence is a critical aspect of the professional competence of translators:

> Den für ein ausreichendes Textverstehen nötigen Recherchierbedarf zu erkennen ist ein wichtiger Teil professioneller Kompetenz. Voraussetzung dafür ist jedoch, dass Textverstehen und Recherchieren eng aufeinander bezogen werden.

Kuβmaul, in his recent *Verstehen und Übersetzen* [2007:76], focuses his attention on the relationship between the understanding of the source text, the search for information and text analysis and underlines the complexity of search activity. He highlights its importance in relation to the extension of the different types of knowledge relevant to translators, i.e., the mental lexicon and world knowledge. Furthermore, this activity must not be considered an autonomous activity but part of the understanding and reformulating process.

Search activity plays a relevant role throughout the whole translation process: in the initial phase, during the analysis, comprehension and interpretation phase, afterwards, when reformulating the message and choosing the appropriate translations in the target text and finally, during the revision phase in order to guarantee the correctness and accuracy of the choices. Schmitt P.A. [2006:186] underlines how the search for information is a critical aspect of the translation process and consequently how search tools are an essential part of a translator's equipment:

> Da Übersetzen per definitionem etwas mit schriftlicher Textproduktion zu tun hat, handelt es sich bei den technischen Arbeitsmitteln des Übersetzers zunächst um Schreibwerkzeug; anderseits spielt auch die Recherche, der Zugriff auf externe Wissensbestände, eine zentrale Rolle, so daβ auch Recherchewerkzeuge zum typischen Instrumentarium des Übersetzers gehören.

Translators should therefore be aware of the main search applications since they are vital to their professional expertise. M. Pinto [2001:298] also underlines that translators should receive adequate training in the use of *documentary tools*:

> The translator is the nucleus of translating operations. Given his importance in the context of a functional approach to such operations, the documentary support of his activity should be given special care: firstly, by supplying more and better documentary tools (original documents, references, glossaries, encyclopedia,

> dictionaries), and secondly, by improving his degree of documentary formation, not only as a processor of information, but also and above all, as a user of a practically unlimited documentary orbit.

The mastery in using these instruments is considered a specific expertise and is defined by M. Pinto as *documentary competence* [2001:294] which is used by translators throughout the whole translation process with positive fallouts on the quality of the final product. Nowadays, traditional documentary tools (dictionaries, glossaries, thesauri, reference material in general) are all available in electronic format and can be used both off-line and on-line. Many tools and resources for linguistic applications are available on the Internet and professional translators have many more possibilities to access information compared with a few years ago: a great deal of information can be found on the Internet and therefore the abilities that translators have to develop are mainly search abilities but also the ability to identify and choose relevant information.

In "A professional Approach to Translator Training", Olvera Lobo and her colleagues [2007:518] observe that:

> […] today's translators must develop research strategies and evaluate the quality of information, tasks previously carried out with the help of other professionals such as librarians, information scientists, and subject matter specialists who are so vital for the success of their work.

Translators have to be able to look for information in an effective way by using software applications and the web and, at the same time, they have to be able to evaluate the quality of the results. It is a combination of two different competences: the technological one and the ability to effectively search for information.

At present, translation environments specifically address the translation task itself by means of Machine Translation (MT) and/or Translation Memory (TM) applications and offer additional translation support tools for terminology management, text alignment and text pre-editing. However, up to now, little attention has been devoted to the time-consuming task of analysing the translation text, identifying the main concepts, finding the relevant information and the appropriate translation in the various documentary tools available to translators both off-line and on-line and building and automatically updating the knowledge base.

The approach that we intend to propose here is based on the integration of text mining technologies in the translation workflow. Since text mining aims to automatize tasks such as text analysis, text categorization, information extraction, text summarisation and text understanding in general, it represents, in our opinion, a useful tool for easing and speeding up translators' work in the early stages of the translation cycle.

Text mining can have a significant role in scientific and technical translations where translators have to cope with the problem of finding adequate linguistic resources (dictionaries, corpora, thesauri) in order to identify and properly translate multi-word units with a specific terminological meaning.

In most languages, there is a close and necessary relation between terminology and a specific subset of multi-word units, i.e., the one formed by compound words. This is proved by the fact that specialized lexica consist mainly of compounds: recent theoretical estimations show that specialized lexica may contain between 50% and 70% of multi-word units [Sag et al. 2002]. Lately, these estimations were confirmed by Ramisch et al. [2010] who found that 56.7% of the terms annotated in the *Genia* corpus[2] are composed of two or more words, and this is an underestimation since it does not include general-purpose multi-word units such as phrasal verbs and fixed expressions. This

[2] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi

means that in any given scientific and/or technical text, a large part of cognition is conveyed by multi-word units which may be ontologically classified on the basis of the knowledge domain(s) in which they have been created and for which they express precise, unambiguous meanings. Non-compositionality, numerousness and morpho-syntactic variations of multi-word units cause many problems in the choice of the correct translation, both during a human or a computer-assisted translation process.

Although multi-word units are so numerous and have a critical role in scientific and technical translations, state-of-the-art CAT tools still do not have dedicated functionalities to handle them [Fernández-Parra M. & ten Hacken P. 2010].

In the following section, we describe our approach to the handling (identification and extraction) of multi-word units, with a terminological meaning, i.e., compound words, in a text mining tool developed at the Laboratory of Computational Linguistics "Maurice Gross" (Department of Social, Political and Communication Sciences of the University of Salerno, Italy) and its possible applications in a translation process.

3. CATALOGA

3.1 General features of CATALOGA®

CATALOGA® is a text mining software based on matching digitised texts and electronic dictionaries of terminological compound words developed according to the Lexicon-Grammar (LG) lexical formalization method[3]. The monolingual version was designed and developed by Elia, Postiglione and Monteleone and the bilingual version is under development.

At present, it is configured as a stand-alone software which can be integrated in web sites and portals. Its main goal is to extract terminological compound words from a given scientific or technical text and to automatically determine – without human intervention – the main knowledge domains it deals with and detect the terminological compound words in the analysed texts.

The tasks performed by the monolingual version of CATALOGA® can be summarised as follows:

1. automatic reading of a text;
2. computation of all the occurring terminological multi-word units, i.e., location and computation of all the occurrences of any of a finite number of compound words;
3. statistical computation of the ratio between terminological and non-terminological occurrences;
4. statistics-based listing of all the terminological occurrences in decreasing order. classed on the basis of the relevant knowledge domains.

In the following paragraphs we will provide detailed information on (i) the computational aspects of the tool, (ii) the lingware, used by the automatic routines of CATALOGA® (iii) the bilingual version of the tool which is currently under development and finally (iv) the possible applications of the tool in a scientific and technical translation process.

3.2 Computational aspects of CATALOGA®

The technology used to assemble this software is "Borland Developer Studio 2006" (i.e. Delphi 10 or Delphi 2006). Delphi is a powerful RAD (Rapid Application Development) visual software development tool, based on an Object Programming Language. No specific hardware architecture is

[3] For more information about the Lexicon-Grammar approach refer to http://infolingu.univ-mlv.fr/english/Bibliographie/biblieng.html

explicitly required since the software is normally installed and used on both house-desktop and laptop standard Windows computers.

If we consider a Windows XP computer with 3 GB of RAM and a 1.66 GHz dual-core CPU, the software loads and pre-processes a whole Cataloga® electronic dictionary (approximately 500,000 entries) in less than 10 seconds during the start-up phase. This pre-processing step is performed only once for each software session whereas the text processing step is achieved in real time.

The time complexity of the CATALOGA® dictionary pre-processing algorithm is O(n), i.e., is linearly proportional to the sum of compound word lengths. At the same time, the matching algorithm has an O(m)-time complexity, i.e., it takes $m<X<2m$ state transitions to process a text string of length m. All terminological compound words can be simultaneously recognised in one pass. During the analysis procedure, the terminological electronic dictionaries are completely allocated in the RAM so that the effective software execution time is very short and does not depend on the dictionary or the text size.

3.3 CATALOGA Lingware

The electronic dictionaries used by CATALOGA® are part of the DELA system, developed according to the Lexicon Grammar (LG) approach. This system is formed by Simple-Word Electronic Dictionaries (DELAS-DELAF), and Compound-Word Electronic Dictionaries (DELAC-DELACF). CATALOGA® uses the DELAC-DELACF dictionary which includes mainly terminological compound nouns. Each entry of the dictionaries is given a consistent ontological description, being coherently tagged with reference to the knowledge domain(s) in which it is commonly used (i.e., in which it has a terminological unambiguous meaning). For instance, the compound word *acconto dividendo* ("interim dividend") is marked with the tag ECON which stands for "Economics". As a further example, the compound *massimizzazione del gettito fiscale* ("revenue maximization") is marked with two different tags: ECON and FISC (Tax Regulations), due to the fact that it is used in both knowledge domains.

The development and management of an electronic dictionary consist of three main steps:

1. *Lexical acquisition*. During this on-going phase, MWUs are extracted from corpora and/or certified glossaries and continuously updated.

1. *Morpho-grammatical, syntactic and domain tagging*. Each lexical entry is given a coherent linguistic description consisting of:
   - a morpho-grammatical and inflectional paradigm,
   - the internal structure of the compound word,
   - the domain.

   The same information is given to the corresponding translation, together with the syntactic function of the terminological compound word (both in the source and the target language).

   In the following entry extracted from the English-Italian bilingual dictionary:

   macchia/bianca,.NA:fs-+;MED/ =white/spot,.AN:s+/N

   the Italian compound noun "macchia bianca" is followed by the tag "NA:fs-+" which indicates the morphologic and grammatical pattern of the compound noun, i.e., the compound consists of a noun (N) followed by an adjective (A), it is feminine singular (fs), it does not have a masculine form (-) but a feminine plural form (+); the tag "MED" (for Medicine) refers to the domain that the entry belongs to. The English translation "white

spot" which follows after the equal sign is given the same consistent ontological description. Finally, at the end of the string, the tag "N" indicates the syntactic function of the compound noun, both in Italian and in English.

Examples of different possible morpho-syntactic subcategories are provided in the following table.

N° of constituents in the lexical unit	POS tags	Example
bi-gram	NA NN …	aborto spontaneo (MED) interfaccia utente (INF) …
tri-gram	NPN NPN NPN …	capacità del disco (INF) cassa di risparmio (ECON) morbo di Crohn (MED) …
fourth-gram	NAPN …	disturbo respiratorio del sonno (MED) …
fifth-gram	NPNPN …	disturbo da deficit di attenzione (MED) …
…	…	…

Table 1: Morpho-syntactic subcategories of MWU

2. *Testing on corpora*. The dictionary is used to automatically analyse and process large corpora.

As a sample, we provide a short excerpt of the Italian Electronic Dictionary of Medicine:

macchia/bianca,.NA:fs-+;MED/ =white/spot,.AN:s+/N
macchia/blu,.NA:fs-+;MED/ =blue/spot,.AN:s+/N
macchia/blu,.NA:fs-+;MED/ =macula/cerulea,.NA:s+/N
macchia/corneale,.NA:fs-+;MED/ =aglia,.N:s+/N
macchia/cribrosa,.NA:fs-+;MED/ =lamina/cribrosa,.NA:s+/N
macchia/di/Bier,.NPN:fs-+;MED/ =Bier/spots,.NN:s+/N
macchia/di/Bitot,.NPN:fs-+;MED/ =Bitot's/spots,.NPN:s+/N
macchia/di/Brushfield,.NPN:fs-+;MED/ =Brushfield's/spots,.NPN:s+/N
macchia/di/De Morgan,.NPN:fs-+;MED/ =De Morgan's/spot,.NPN:s+/N
macchia/di/Filatov,.NPN:fs-+;MED/ =Filatov's/spots,.NPN:s+/N
macchia/di/Flindt,.NPN:fs-+;MED/ =Flindt's/spots,.NPN:s+/N
macchia/di/Koplik,.NPN:fs-+;MED/ =Koplik's/sign,.NPN:s+/N
macchia/di/Koplik,.NPN:fs-+;MED/ =Koplik's/spots,.NPN:s+/N
macchia/di/Maurer,.NPN:fs-+;MED/ =Maurer's/cleft,.NPN:s+/N
macchia/di/Maurer,.NPN:fs-+;MED/ =Maurer's/doft,.NPN:s+/N
macchia/di/Michel,.NPN:fs-+;MED/ =Michel's/flecks,.NPN:s+/N
macchia/di/Michel,.NPN:fs-+;MED/ =Michel's/spots,.NPN:s+/N
macchia/di/Mueller,.NPN:fs-+;MED/ =Mueller's/spots,.NPN:s+/N
macchia/di/Mueller,.NPN:fs-+;MED/ =vitiligo/iridis,.NN:s+/N
macchia/di/Roth,.NPN:fs-+;MED/ =Roth's/spots,.NPN:s+/N
macchia/di/Soemmering,.NPN:fs-+;MED/ =Soemmering's/spot,.NPN:s+/N

At present, 180 different tags are included in the data-base of CATALOGA® knowledge domains; this means that this software can analyse texts using 180 different electronic dictionaries. The most

important dictionaries are: Computing/IT (approx. 54,000 entries), Medicine (approx. 46,000 entries), Law (approx. 21,000 entries) and Engineering (approx. 19,000 entries). Subset tags are also provided for domains that include specific subsectors. This is the case with Engineering for which a generic tag ING is used whereas nine more explicit tags are used for Acoustic Engineering (ING ACUS), Aeronautics and Aerospace Engineering (ING AER), Chemical Engineering (ING CHIM), Civil Engineering (ING CIV), Mechanical Engineering (ING MECC), Mining Engineering (ING MIN), Naval Engineering (ING NAV), Nuclear Engineering (ING NUCL) and Oil Engineering (ING PETROL). The same formalization method was used for Physics which has been given a generic tag FIS plus more specific tags for Atomic Physics (FIS ATOM), Nuclear Physics (FIS NUCL), Physics of Plasma (FIS PLASMA), Solid-State Physics (FIS SOL) and Subnuclear Physics (FIS SUBNUCL). Each dictionary has been created and verified under the supervision of domain experts.

3.4 The bilingual version of CATALOGA®

The initial phases of a scientific or technical translation process imply several tasks that have to be performed by translators i.e., reading of the source text, identification of the main concepts and relevant terminology, documentary search using traditional documentary tools (paper dictionaries, thesauri, etc.) or web pages on the Internet, use of general, and specialized, monolingual, bilingual and multilingual electronic dictionaries on the Internet or on CD-ROM, consulting reference material provided by the customer or text corpora on the Internet or on CD-ROM, looking up information in a personal text corpus by means of text analysis or concordance software programs and updating and tailoring the linguistic resources or the translation tools according to the specific translation task that has to be performed. No tools are available on the market that speed up these complex and time-consuming activities.

The approach we would like to propose here is to introduce a higher degree of automation and integration for this crucial phase of the translation cycle which could also be beneficial to the subsequent translation phase.

An ideal documentary tool, in this respect, should contain a text mining and information extraction facility from corpora which enables:

- document classification (identification of domain and extraction of relevant concepts) and automatic indexing based on linguistic information;
- retrieval of useful reference material by users such as appropriate terminology resources, parallel corpora, etc. which are automatically assigned to a specific translation project;
- pre-translation of the source text and /or updating of the translation tools (both MT and TM) with the relevant information found during the query phase.

This tool would allow users to semi-automate the translation analysis phase with regard to the retrieval of reference material (documents, terminology, corpora) for a particular translation project. Unlike state-of-the art collaborative translation workspaces, this would provide an advanced and indispensable feature based on linguistic knowledge within a typical translation workflow.

With this idea in mind, we are developing the bilingual version of CATALOGA®. In addition to the features described for the monolingual version, the main features of the bilingual version of Cataloga are:

1. listing of all the terminological occurrences in decreasing order classed on the basis of the relevant knowledge domains with their translation;
2. tagging of all the terminological compound words with their translation in the source

text in XML format;
3. automatic replacement of the translations in the target text.

These features are very useful for different purposes:

1. the list of words obtained at the end of the text analysis process can be used in a specific crawling tool, such as BootCat[4] for instance, to automatically retrieve useful reference material such as parallel or comparable corpora;
2. the tagged text can be used for training purposes in conjunction with MT, in specific SMT and TM applications, in order to identify and pre-translate linguistically significant phrases, with the aim of improving the computer-assisted translation results;
3. the pre-translated target text can be used as a basis during a traditional human-based translation cycle.

In the following section, we provide the results of some experiments performed with CATALOGA both on Italian texts and their translations and some examples of the possible use of CATALOGA in the initial stages of a scientific or technical translation process.

4. USE OF CATALOGA® IN A SCIENTIFIC/TECHNICAL TRANSLATION PROCESS

4.1 Text analysis performed by CATALOGA®: results

In order to provide a concrete example of how CATALOGA® processes texts and automatically extracts meanings, we will consider the following short passage which a human reader with an average cultural level could straightforwardly define as dealing with the field of medicine:

> La vitamina A (Retinolo) svolge un'azione protettiva delle mucose e degli epiteti. Inoltre ha un ruolo nella crescita, favorendo lo sviluppo scheletrico. La carenza di vitamina A è una delle più comuni carenze vitaminiche. È comune soprattutto nei Paesi in via di sviluppo, rappresentando una della principali cause di cecità. La carenza di vitamina A è spesso dovuta a malassorbimento lipidico, ad alcolismo, e si osserva più comunemente negli anziani. Un sintomo precoce di carenza di vitamina A è la cecità notturna, seguita da secchezza della congiuntiva, macchie di Bitot (macchie biancastre della sclera). Questa risposta fatta da me su altro sito le fa capire a che cosa è dovuta la macchia di Bitot e di che colore è ovvero biancastro. La sua sembra più o un piccolo nevo nevocellulare piano oppure una zona di assottigliamento sclerale, completamente innocua e sine materia dal punto di vista patologico, che lascia intravedere la componente bluastra sottostante. (Available on http://www.medicitalia.it/consulti/Oculistica/65819/Macchianella- sclera)[5]

[4] Bootcat (http://bootcat.sslmit.unibo.it/?section=home) is an open source crawling tool that creates random tuples from a seed term list and runs a query for each tuple (on the Bing search engine). It constructs a URL list on the basis of the first 10 results obtained from the query and downloads the corresponding web pages. Bootcat is also available on the Sketchengine webpage (http://www.sketchengine.co.uk/?page=Website/SketchEngine) and as BootCat front-end, a web service front-end and a graphical user interface to the core tool, respectively.

5 Vitamin A (Retinol) exerts a protective action on the mucous membranes and epithets. It also has a role in growth, supporting skeletal development. Lack of vitamin A is one of the most common vitamin deficiencies. It is especially common in developing countries, representing one of the main causes of blindness. Vitamin A deficiency is usually due to fat malabsorption and alcoholism and is most commonly seen in elderly people. An early sign of vitamin A deficiency is night blindness, followed by dryness of the conjunctiva and Bitot's spots (white spots on the sclera). This answer I gave on another site helps you understand the origins of Bitot's spots, and their colour, i.e. whitish. Your spot looks more like a small melanocytic nevus or a scleral thinning area that is completely harmless and sine materia from a pathological point of view, with a bluish part underneath. (Available on http://www.medicitalia.it/consulti/Oculistica/65819/Macchianella-sclera ; English translation by Mario Monteleone)

After reading and analysing it, CATALOGA® automatically produces a table with the results of the text processing:

CATALOGA® - Rel. 4.8 del 9 mar 2010 - 11:00
Global number of knowledge domains in the database: 180
****

Total Number of lines in the input text:	1
Total Number of words in the input text:	154
Total Number of chars in the input text:	972
Longest line in the input text:	972
Average sentence length (in words):	9.6
Average word length (in syllables):	2.2
Flesh index for this paper:	62.0

****

Generic Dictionary Occurrences:	1	
Thematic dictionaries occurrences:	14	

Therefore, the input text is thematic.

********** ANALYSIS **********
The input Text deals with (in frequency order): MED (Medicine), ANAT (Anatomy).
**********ORDERED FREQUENCIES **********

MED (MEDICINE)	12	92.9%
ANAT (ANATOMY)	1	7.1%
DIGE (GENERIC DICTIONARY)	1	7.1%

****
File name: Medicina.txt
Number of different compound words: 12

COMPOUNDS	OCC.	MORPH	INFL	DOM	ENG	MORPH	INFL
assottigliamento sclerale	1	N+NA	ms-+	MED	scleral thinning	N+AN	s+
carenze vitaminiche	1	N+NA	fs-+	MED	vitamin deficiencies	N+NN	p+
cecità notturna	1	N+NA	fs--	MED	night blindness	N+NN	s+
macchia di Bitot	1	N+NPN	fs-+	MED	Bitot's spot	N+NPN	s+
macchie biancastre della sclera	1	N+NAPA	fp-+	MED	white spots of the sclera	N+ANPDETN	p+
macchie di Bitot	1	N+NPN	fp-+	MED	Bitot's spots	N+NPN	p+
malassorbimento lipidico	1	N+NA	ms-+	MED	fat malabsorption	N+NN	s+
nevo nevocellulare piano	1	N+NAA	ms-+	MED	small melanocytic nevus	N+AAN	s+
punto di vista	1	N+NPN	ms-+	DIGE	point of view	N+NPN	s+
secchezza della congiuntiva	1	N+NPN	fs-+	MED	dryness of the conjunctiva	N+NPDETN	s+
sviluppo scheletrico	1	N+NA	ms-+	ANAT	sketelal development	N+AN	s+
vitamina A	4	N+NN	fs--	MED	Vitamin A	N+NN	s-

Table 1 - CATALOGA® Analysis Results

This table shows that CATALOGA® has inferred that the input text deals with medicine as a result of the analysis and computation of the terminological compound words in it, i.e., it has made the same conclusions as the human reader.

Similar precise results have been obtained with the analysis of a corpus consisting of 1,070 text files (approx. 10 MB) extracted from Italian online newspapers. The results of these analyses are shown in the following table:

Analysis Results (1,070 files tested)		
Correct	Partially faulty	Faulty
71%	29%	0%

Table 2 - CATALOGA® Analysis Results on a corpus of 1, 070 files (approx. 10MB)

It is worth noting that partially faulty results depend on terminological entries of the dictionaries that have not yet been updated and that CATALOGA® also achieves detailed and successful analyses with very short text files.

We have shown that CATALOGA® analytic routine can extract terminology, and more generally, semantic information from given texts in a precise way. In addition, these results confirm that:

- the information given in any terminological text is mainly conveyed by terminological compound words;
- the automatic retrieval of terminological compound words from texts allows automatic retrieval of its general meaning in the form of lexical ontologies made up of one compound word matched with one non-ambiguous knowledge domain tag.

Therefore, some more specific considerations can be made based on the results obtained so far, i.e. CATALOGA®:

- achieves more efficient and less "noisy" automatic terminological information retrieval compared with statistical approaches to IR;
- may allow automatic information-based data storage and bypass human reading;
- can support the creation of information data bases in which items (i.e. digitised texts) may be linked/grouped according to the terminological compound keywords and/or the non-ambiguous knowledge domain tags they share;
- can offer more rapid management and updating procedures as far as textual terminological information is concerned;
- allows the automatic creation of bilingual lists of terminological compound words which can be used for web crawling purposes or as training set for MT and TM applications.

4.2 Use of CATALOGA with a crawling tool

The bilingual list of terminological compound words produced by CATALOGA® can be used to automatically produce a precise and specific list of 'seed terms', both in the source and the target language, tailored on the source text, to be used in queries on the web with a crawling tool. For our experiment we used the BootCat toolkit [Baroni et al., 2004], a well-known suite of Perl scripts for bootstrapping specialized language corpora from the web.

Taking as input the key terms extracted by means of the automatic text analysis procedure performed by CATALOGA®, BootCat draws upon web data to automatically build a specialised corpus for the domain of interest and tailored on the specific text to be translated. In this way, the most relevant web pages which specifically refer to the subject matter of the text to be translated can be collected.

For instance, if we take the list of English terminological compound words (refer to Table 1) produced during the text analysis phase illustrated in the previous section and we use it as 'seed terms' in Bootcat, we obtain the following list of web sites:

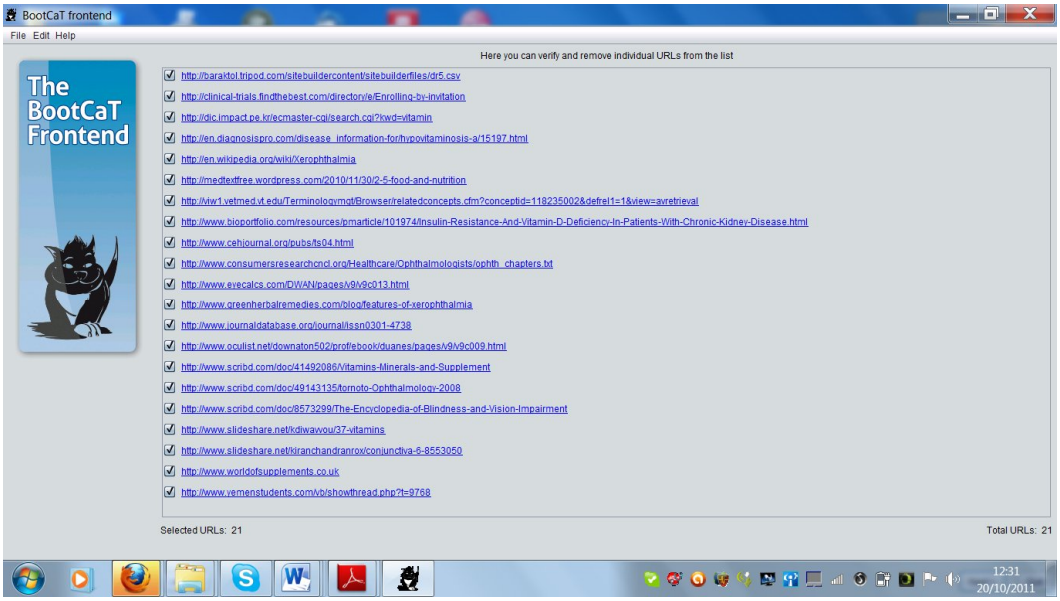

Figure 1 - URL list generated on the basis of the CATALOGA list of compound words

The list of web sites contains relevant information sources such as medical texts, glossaries, thesauri and text corpora related to the subject matter of the analysed text.

4.3 Use of CATALOGA with MT

The handling of multi-word units in MT is a well-known problem. The importance of a correct processing of multi-word units in Machine Translation (MT) and Computer Aided Translation (CAT) has been highlighted by several authors including Thurmair [2004], Villavicenzio et al. [2005], Diakonescu [2004], Váradi [2006], Hurskainen [2008], Rayson et al. [2009], Moszczyński [2010], Barreiro et al. [2010], Monti et al. [2011]. For MT, Piao et al. [2005] have pointed out that the issue of MWU identification and accurate translation has remained an unsolved problem for current MT systems.

One of the proposed solutions for overcoming translation problems in MT and in SMT in particular is based on the idea that multi-word units should be identified and bilingual multi-word units should be grouped prior to statistical alignment [Lambert and Banchs, 2006]. More recently, Zhixiang Ren et al. [2009] have underlined that experiments show that the integration of bilingual domain MWUs in SMT could significantly improve translation performance.

Since terminological compound words are multi-word units with limited or no variability of co-occurrence among words, they have to be considered as a single lexical unit with specific semantic and syntactic features and therefore treated as one single token rather than a sum of different tokens. In this way, translation problems can be substantially solved since terminological compound words are mono-referential and unambiguous translations can be assigned prior to an MT process.

In the wake of Wu et al. [2008], who proposed a method for constructing a phrase table using a manually-made translation dictionary in order to improve SMT performance, especially when translating domain texts, we intend to experiment on a large scale how the results of text analysis performed by CATALOGA can be used in a translation process based on SMT and phrase-based MT in particular. The bilingual list of compound words provided by this text mining tool represents the specific dictionary of the analysed text and can therefore be used for domain adaptation purposes, adding it as a bilingual phrase table to an SMT system, which seems to outperform the

use of dictionaries as a training corpus [Wu et al. 2008], to improve the quality of a baseline SMT system.

4.4 Use of CATALOGA to pre-translate source text

In addition to the above mentioned options for integrating CATALOGA in a translation environment, a further possibility, already available, is to use the list of compound words generated during the analysis of the source text performed by CATALOGA to pre-translate the source text, thereby ensuring coherent use of terminology throughout the whole target text.

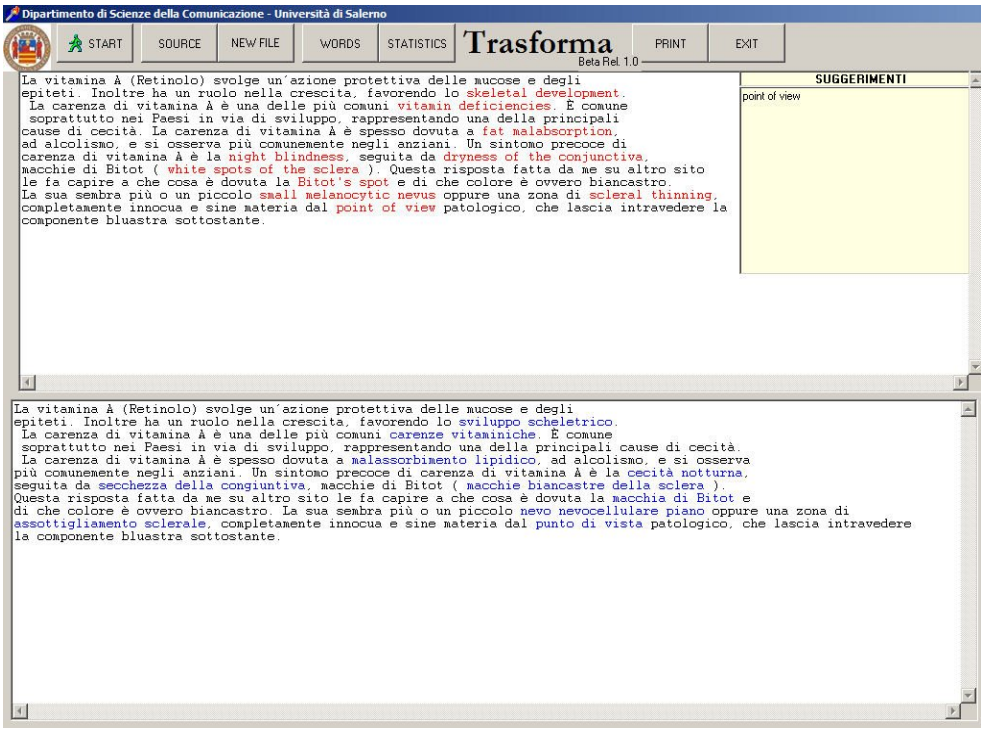

Figure 2 - Pre-translation of the source text using the CATALOGA bilingual compound word list

5. FUTURE RESEARCH PERSPECTIVES

CATALOGA and the linguistic resources used by it are based on the adoption of a well-founded, empirically coherent and solid linguistic formalisation method, i.e. the Lexicon-Grammar approach. Future research perspectives concern both the text mining tool and the DELAF/DELACF dictionaries and will include:

1. the creation of terminological electronic dictionaries for newly created knowledge domains such as e-government, bioethics, biomedicine, and so on;
2. the implementation of multilingual semantic-based terminological analysis;
3. the construction of an ontology-based query interface for information retrieval;
4. the testing and validation of the lingware on large corpora;
5. the integration of CATALOGA® in web sites and portals in order to let internet surfers test the whole system.

6. the creation of bilingual and multilingual electronic dictionaries in order to use CATALOGA® in automatic and semi-automatic machine translation routines as far as terminology is concerned;
7. the creation of an automatic smart text storing system, by means of which files can be automatically read and categorized on the basis of the main knowledge domain(s) they belong to and the terminological compound words they include. Information retrieval from textual relational data bases of this type will be achieved by means of queries structured both on knowledge domain tags and on terminological compound words.

These steps will undoubtedly be a launching pad towards new forms of experimentation of the system in translation environments.

6. CONCLUDING REMARKS

We have proposed a new approach to scientific and technical translation which is based on the combination of CATALOGA, a text mining tool, an IR application and/or an MT/TM system. The contribution of this paper lies in the following points:

- Description of the CATALOGA text mining tool, developed at the Laboratory of Computational Linguistics "Maurice Gross" of the University of Salerno (Italy), and the linguistic resources used by the system;
- Results of a pilot study using CATALOGA together with hand-crafted linguistic resources, consisting of monolingual and bilingual terminological multi-word units, for extracting information from texts to be used in combination with IR tools and MT or TM;
- Description of the possible ways of integrating text mining based on linguistic resources in a translation process.

We are convinced that this approach will definitely help translators improve their documentary competence in terms of efficiency, speed and accuracy.

REFERENCES

Baroni M. & Bernardini S. (2004). Boot-CaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004*, Lisbon: ELDA. (2004): 1313–1316.

Barreiro A., Elia A., Monteleone M., Monti J. (2010). Mixed up with Machine Translation: Multi-word Units Disambiguation Challenge. *Translating and the Computer 32 ASLIB 2010* (London, England, 18-19 November 2010).

D'Agostino E. & Elia A. (1998). Il significato delle frasi: un continuum dalle frasi semplici alle forme polirematiche. Leoni, F.A., Gambarara, D., Gensini S., Lo Piparo, F., Simone, R. *Ai limiti del linguaggio*, Bari, Roma: 287-310.

Diakonescu, S. (2004). Multiword Expression Translation Using Generative Dependency Grammar. *Advances in Natural Language Processing 4th International Conference, EsTAL 2004*, Alicante, Spain, October 20-22: 243-254.

Elia A., Postiglione A. Monteleone M. (2010). CATALOGA. Sistema informatico per la catalogazione automatica di testi, Release 4.8., Software.

Elia A., Postiglione A., Monteleone M. (2011) CATALOGA: a software for semantic-based terminological data mining. *Proceedings of the 1st International Conference on Data Compression, Communication and Processing (CCP 2011)* (Palinuro, Italy 21-24 June 2011).

Elia A., Postiglione A., Monteleone M., Monti J., Guglielmo D. (2011), CATALOGA®: a Software for Semantic and Terminological Information Retrieval. *WIMS '11 Proceedings of the International Conference on Web Intelligence, Mining and Semantics* (Sogndal, Norway. 25-27 May 2011).

Elia A., De Bueriis G. (eds.) (2008), *Lessici elettronici e descrizioni lessicali, sintattiche morfologiche ed ortografiche. Risultati del Progetto PRIN 2005 Atlanti Tematici Informatici – ALTI*, Collana "Lessici & Combinatorie", n. 2, Dipartimento di Scienze della Comunicazione dell'Università degli Studi di Salerno, Plectica, Salerno.

Elia A., Monteleone M., De Bueriis G., Di Maio F. (2008) Le poliрematiche dell'italiano. Elia, A., De Bueriis, G. (eds.), *Lessici elettronici e descrizioni semantiche, sintattiche e morfologiche. Risultati del Progetto PRIN 2005 Atlanti Tematici Informatici - ALTI*, Collana "Lessici & Combinatorie", n. 2, Dipartimento di Scienze della Comunicazione dell'Università degli Studi di Salerno, Plectica, Salerno: 11-65.

Fernández-Parra M. & ten Hacken P., (2010). Identifying Fixed Expressions: A Comparison of SDL MultiTerm Extract and Déjà Vu's Lexicon In *Translating and the Computer 32 ASLIB 2010* (London, England, 18-19 November 2010).

Gross M. & Senellart J. (1998). Nouvelles bases statistiques pour les mots du français. In *4`emes Journées internationales d'Analyse statistique des Donnees Textuelles (JADT'98)* Nice: 335–349.

Hönig H. (2006). "Textverstehen und Recherchieren", Snell-Hornby, M. et al. (eds) *Handbuch Traslation*, Stauffenburg, Tübingen.

Hurskainen, A. (2008). Multiword Expressions and Machine Translation. Technical Reports. *Language Technology Report No 1*,
[http://www.njas.helsinki.fi/salama/multiwordexpressions-and-machine-translation.pdf]

Kuβmaul P. (2007). *Verstehen und Übersetzen. Ein Lehr- und Arbeitsbuch*, Narr Studienbücher, Tübingen.

Lambert P. & Banchs R. (2006). Grouping multi-word expressions according to Part-Of-Speech in statistical machine translation. *Proceedings of the EACL Workshop on Multi-word expressions in a multilingual context*, Trento Italy.
[http://varoitus.barcelonamedia.org/rafael/Publications/Proceedings/2006_EACLWS.pdf]

Lebtahi Y. & Ibert J. (2004). Traducteurs dans la société de l'information. Évolutions et interdépendences., *Meta*, 49, 2:.221-235.

Monteleone M. (2004) *Lessicografia e dizionari elettronici. Dagli usi linguistici alle basi di dati lessicali*, Fiorentino & New Technology, Napoli,).

Monti J, Barreiro A., Elia A., Marano F., Napoli A. (2011). Taking on new challenges in multi-word unit processing for Machine Translation. F. Sanchez-Martinez, J.A. Perez-Ortiz (eds.), *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, p. 11- 19. Barcelona, Spain, January 2011. [http://hdl.handle.net/10609/5646]

Monti J. (2010). Alla ricerca della conoscenza. quali strumenti per la traduzione saggistica? Montella C. (ed.) *Tradurre Saggistica*., Franco Angeli, Milano: 143-161.

Monti J. (2010). La E-translation da Google a Second Life: le più recenti applicazioni di Traduzione Automatica online. *Atti del XLIII Congresso della Società Linguistica Italiana*. 24-26 settembre 2009. Bulzoni Editore, Roma.

Moszczyński, R. (2010). Towards a bilingual lexicon of information technology multiword units. *Proceedings of the XIV Euralex International Congress*. Leeuwarden, the Netherlands.

Olvera Lobo M. D., Robinson B., Castro Prieto R.M., Quero Gervilla E., Muñoz Martín R., Muñoz Raya E., Murillo Melero M., Senso Ruiz J. A., Vargas Quesada B. e Díez Lerma J. L: (2007), A professional approach to Translator training (PATT). *Meta*, 52 (3): 518.

Piao S.S., Rayson P., Archer D., and McEnery T. (2005). Comparing and combining a semantic tagger and a statistical tool for mwe extraction. *Computer Speech and Language*, 19(4):378– 397.

Pinto M. (2001). Quality Factors in Documentary Translation. *Meta*, 46 (2) 298.

Pym A. (2003). Redefining Translation Competence in an Electronic Age. In Defense of a Minimalist Approach. *Meta*, 48 (4): 481-497.

Ramisch C., Villavicencio A. & Boitet C. (2010). Multiword expressions in the wild? the mwetoolkit comes in handy. *Proceedings of the 23rd COLING (COLING 2010)* — Demonstrations. Beijing, China,: 57–60,

Rayson, P. Piao, S. Sharoff, S. Evert, S. and Villada Moirón B.. (2010). Multiword expressions: hard going or plain sailing? *Journal of Language Resources and Evaluation. Lang Resources & Evaluation:* 44:1–5.

Ren Z., Yajuan Cao L. J., Liu Q., and Huang Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, Singapore, August: 47–54.

Sag A., Baldwin T., Bond F., Copestake A. & Flickinger D..al. (2002). Multiword Expressions: A Pain in the Neck for NLP? *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CI-CLING 2002),* Mexico City, Mexico: 1-15 [http://lingo.stanford.edu/pubs/WP-2001-03.pdf]

Thurmair G. (2004). Multilingual content processing. *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC)* 2004, Lisbon, Portugal.

Váradi, T. (2006). Multiword Units in an MT Lexicon. *EACL-2006: 11th Conference of the European Chapter of the Association for Computational Linguistics, Workshop on Multiword expressions in a Multilingual Context,* Trento, Italy: 73-78 [http://www.aclweb.org/anthology-new/N/N03/N03-2036.pdf]

Villavicencio, A., Bond, F., Korhonen, A., McCarthy, D. (2005). Introduction to the special issue on multiword expressions: having a crack at a hard nut. *Journal of Computer Speech and Language Processing*, 19(4):365–377.

Wills W. (1977). *Übersetzungswissenschaft. Probleme und Methoden*, E. Klett, Stuttgart (trans. *The Science of Translation. Problems and methods*, Günther Narr, Tübingen),

Wu H., Wang H., Zong C. (2008). Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora". *Proceedings of Conference on Computational Linguistics (COLING*): 993–1000.