

# CATALOGA®: a Software for Semantic and Terminological Information Retrieval

\*Annibale Elia  
+39089963017  
elia@unisa.it

\*Alberto Postiglione  
+39089962215  
ap@unisa.it

\*Mario Monteleone  
+39089963009  
mmonteleone@unisa.it

\*Johanna Monti  
+39089963009  
jmont@tin.it

\*Daniela Guglielmo  
+39089963009  
daniguglielmo@gmail.com

\*Department of Political, Social and Communication Sciences, Università degli Studi di Salerno  
Via Ponte Don Melillo s.n.c.  
84084 Fisciano (SA), ITALY

## ABSTRACT

One of the most relevant problems with Information Retrieval (IR) softwares is the correct processing of complex lexical units, today also known as multiword units. The shortcomings are mainly due to the fact that such units are often considered as extemporaneous combinations of words retrievable by means of statistical routines. On the contrary, several linguistic studies, also dating back to the '60s, show that multiword units, and mainly compound nouns, are almost always fixed meaning units, with specific formal, morphological, grammatical and semantic characteristics. Furthermore, these units can be processed as dictionary entries, thus becoming concrete lingware tools useful to achieve efficient semantic information retrieval (IR).

Therefore, in this paper we will focus on CATALOGA®, an automatic IR software which retrieves terminological information from digitized texts without any human intervention. CATALOGA® is actually configured as a stand-alone software which can be integrated in Web sites and portals to be used online. More specifically, we will describe its lingware and software characteristics, discussing their usage as a possible solution to current IR software limitations. The analytical procedure here described will prove itself appropriate for any type of digitized text, and will also represent a relevant support for the building and implementing of Semantic Web (SW) interactive platforms.

## Categories and Subject Descriptors

I.7.3 [Computing Methodologies]: Document and Text Processing – *index generation*.

## General Terms

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WIMS'11, May 25-27, 2011 Sogndal, Norway

Copyright © 2011 ACM 978-1-4503-0148-0/11/05...\$10.00

Management, Experimentation, Human Factors, Standardization.

## Keywords

Lexicon-Grammar, Semantic Web, Automatic Textual Analysis, Terminological Information Retrieval, CATALOGA®, Text Mining, Data Mining.

## 1. INTRODUCTION: A SHORT SURVEY ON SEMANTIC AND TERMINOLOGICAL INFORMATION RETRIEVAL APPROACHES

As for semantic and terminological information retrieval, two main distinct approaches are used nowadays. The first one is essentially based on lingware development and exploitation, namely on the use of electronic dictionaries; the second one is based on onomasiologically-built ontologies.

As we will show in paragraph 3, Cataloga® is an IR software based on the use of specific electronic dictionaries, which essentially provide the encoding of concepts and knowledge based on stable and non-extemporaneous ontological equivalence relations. On the contrary, other information retrieval softwares<sup>1</sup> allow users to create and structure electronic dictionaries so as to suit their specific research needs. This approach, however, has some limitations, not the least of which is the assumption that all users have a sufficient knowledge of the semantic field or fields they intend to include in their electronic dictionaries. In any case, this method leads to the creation of self-made IR electronic dictionaries, which include non-standard ontological relations, due to the fact that each user is likely to build and validate only his own ones. This implies that any electronic dictionary of this kind is efficient only for the user who has built it, i.e. that it cannot be considered a standard IR tool.

---

<sup>1</sup> For instance, this is the case with Collocate (<http://www.athel.com/colloc.html>) and WordStat 6 (<http://www.provalisresearch.com/wordstat/Wordstatkd.html>).

On the other hand, it is certainly possible that users have a good knowledge of the semantic field or fields they intend to include in their electronic dictionaries. Also, this is statistically most frequent with those areas of knowledge which are not overly specialized, and which sometimes use fairly semantically-transparent words and expressions – for instance, this is the case with social and political sciences. On the contrary, areas of knowledge such as chemistry, particle physics or molecular biology are certainly less semantically transparent, and experts in these areas generally need to use IR software equipped with an ontologically stable and reliable updated embedded lingware, in which concepts are non-extemporaneously encoded. In other words, they need to retrieve concepts they still do not know; this is the very reason why they often retrieve new meanings and expressions by consulting specialized literature and magazines, as well as updated specialized paper dictionaries.

At the same time, onomasiologically-built ontologies [23] are used to retrieve words using concept descriptions. The structuring of such ontologies is based on the observation that many users may be able to describe an object and its specific functions even not remembering its name. Therefore, onomasiological paper dictionary are structured so as to provide access to entries starting from the descriptions these entries are associated to. Recently, the spread of computerized dictionaries on CD-Rom has also increased the possibility for users to surf descriptions of objects using the same search functions as Web engines, including Boolean operators. [23] describes an experimental research which proposes the application of complex algorithms to the definitions of two different onomasiological dictionaries<sup>2</sup>. The result sought after is the homologation of all definitions in a single surfable database, without losing any of the various descriptive elements present in the two dictionaries. The results obtained are very interesting, although the application of these ontologies leaves room for some conceptual and pragmatic doubts.

Actually, the first questions that naturally arise are: can these procedure and its complex algorithms account for all concepts included in any onomasiological paper dictionary? In other words, could this IR method become universally agreed-upon and hence be used as a standard IR procedure to formalize semantic concepts, even in different languages? And also: can it be assumed that the same concept will always be expressed by all internauts in the same predetermined way, and then that all users will be capable to produce a concept research, always using a formal, predefined set of words and expressions? Almost certainly, the current state of the art does not allow us to give affirmative answers to these questions, even if, for the future, it cannot be excluded that these procedures could become a

<sup>2</sup> The concept of the onomasiological dictionary presupposes the converse relation between onomasiology and semasiology. Although that volume of a bilingual dictionary which leads from the user language to the foreign language is essentially an onomasiological dictionary, the term ‘onomasiological dictionary’ commonly designates a monolingual dictionary. Under that proviso, an onomasiological dictionary is a dictionary which is organized by semantic criteria and thus leads the user from concepts to expressions. (see also [http://www.christianlehmann.eu/ling/ling\\_meth/ling\\_description/lexicography/onomas\\_dictionary.html](http://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/onomas_dictionary.html))

standard, also thanks to the constant developments of computer science and computational linguistics.

But other aspects lead us to think that the method of onomasiologically-built ontologies is not the most straightforward to follow. Actually, from a more linguistic point of view, we must not forget that there is a clear difference in terms of polysemy between simple words such as *carta* (“paper”, but also, “card”, “map”, “chart”, “charter”, and “document”) and compound words such as the *Carta dei Diritti dell’Uomo* (only “Charter of Rights”). The first type of words always needs to be included in a sentence or in a discourse to be disambiguated, and often does not intrinsically have the semantic features to specify whether the context in which it occurs is or is not technical-scientific; the latter type, in most cases, retains its meaning regardless of the sentences and the discourses in which it occurs, and it also specifies its semantic content. For instance, the sentence:

1. *Max ha discusso della carta con Paolo* [“Max has talked with Paul about the (paper+card+map+chart+charter+document)”]

is syntactically and semantically governed by the operator *Onn discutere con qualcuno di qualcosa* (to talk with so. about st.), while as for the sentence:

2. *Max ha discusso con Paolo della Carta dei Diritti dell’Uomo* (“Max has talked with Paul about the Charter of Rights”)

apart from the syntactic characteristics of the operator, it is possible to say that it deals with law and right subjects, due to the presence of the terminological proper compound noun *Carta dei Diritti dell’Uomo*.

In cases as the one observed here, efficient onomasiologically-built ontologies should then account for all the meanings of *carta* (on the basis of all their potential occurrence contexts) and also for the meaning of *Carta dei Diritti dell’Uomo*, which can hardly be defined in an unanimously and uniquely manner, being a norm composed by several rules and subset of rules.

Besides, as for clear technical-scientific simple words such as *spinterogeno* (“distributor”), it is important to stress that they may often occur inside figured-meaning sentences as:

3. *Max è lo spinterogeno della squadra* (“?Max is the distributor of the team”)

and that in such cases it is not straightforward how the concept should be structured to define the word itself. For instance, a distributor is an inanimate object, while in the previous sentence it is used as a human agent. This difficulty seems more evident in other sentences as:

4. *Lo spinterogeno della squadra si è fratturato un braccio* (“?The distributor of the team broke one of his arms”)

in which an onomasiologically-built ontology should also account for human body parts with reference to an inanimate object.

Considering that the sentences 1-4 are all acceptable and grammatical, we may reach the conclusion that we need to study syntax in order to coherently structure concepts and concept descriptions; so, the building of onomasiological ontologies becomes a work of immeasurable extent, very difficult to achieve

in short times, and maybe achievable by means of more complete analytical linguistic procedures, as for instance Lexicon-Grammar, which will be described in the following pages.

## 2. SEMANTIC WEB AND NATURAL LANGUAGE FORMALISATION

This paper intends to be a contribution to the debate on SW in a new perspective based on the possibility to extract semantic information from texts automatically retrieving concrete and lemmatized lexical forms. In these pages, the theoretical and methodological framework which we refer is Lexicon-Grammar<sup>3</sup>

---

<sup>3</sup> Lexicon-Grammar is one of the most consistent methods for natural language formalization, automatic textual analysis and parsing. Its main goal is to describe all mechanisms of word combinations closely related to concrete lexical units and sentence creation, and to give an exhaustive description of lexical and syntactic structures of natural language. LG was set up by the French linguist Maurice Gross during the '60s, and subsequently developed for and applied to Italian by Annibale Elia, Emilio D'Agostino and Maurizio Martinelli. Its theoretical approach is prevalently based on Zelig Harris' Operator-Argument Grammar, which assumes that each human language is a self-organizing system, and that the syntactic and semantic properties of a given word may be calculated on the basis of the relationships this word has with all other co-occurring words inside given sentence contexts. For LG, the simple sentence [9] is the minimal linguistic meaning context which can be analyzed and on the basis of which concrete studies on natural language may be achieved. A simple sentence is a context formed by a unique predicative element (a verb, but also a name or an adjective) and all the necessary arguments selected by this element in order to obtain an acceptable and grammatical sentence. The study of simple sentences is achieved by analyzing the rules of co-occurrence and selection restriction, i.e. distributional and transformational rules based on predicate syntactic-semantic properties. Today LG methodology is used all over the world by RELEX, a network formed by several groups of researchers working in different universities, laboratories and international structures. The Department of Communication Sciences of Salerno University is part of RELEX and has had a long scientific collaboration with LADL (Laboratoire d'Automatique Documentaire et Linguistique) of Paris 7 University, directed for over twenty years by Maurice Gross. This close scientific collaboration aimed at refining lingware in order to adapt it to NLP software. In the last twenty years, LG has also reached important results in the domain of automatic textual analysis and parsing, creating software and lingware fully oriented towards NLP, such as INTEX (<http://intex.univ-fcomte.fr/>), UNITEX (<http://www-igm.univ-mlv.fr/~unitex/>), and more recently NOOJ (<http://www.nooj4nlp.net/pages/nooj.html>). For more information about LG, refer also to <http://it.wikipedia.org/wiki/Lessico-grammatica> and <http://fr.wikipedia.org/wiki/Lexique-grammaire>. Furthermore, [28] is an updated presentation both of Italian Lexicon-grammar and of the linguistic and computational tools developed at the Department of Communication Sciences of Salerno University (Italy). For other books and publications about multilingual

(LG), which gives an exhaustive description of lexical and syntactic structures of languages.

### 2.1 Web Engines, Ontologies and Lexical Meaning Units

As for Web search engines, it is possible to state that faulty information retrieval results are often caused by inaccurate handling of the linguistic data analysed. In other words, queries that do not retrieve the information required, or that give results which are partially correct, often indicate that Web documents have not been read "carefully", regardless their semantic content and/or textual structure.

Also, it is well-known that a standardized implementation of ontologies would create concrete and universal linguistic bridging among different knowledge domains, therefore allowing the achievement of Web queries which correlate different types of texts and refer to more than a single URL. For instance, as for the domains of "music", "musician", "music composer" and "opera", well-structured ontologies could allow the synchronous retrieval of information concerning a given composer, the operas he composed, all theatres in which his operas have been or are performed, all cities where the theatres are located, and so on. Consequently, queries like "In which city was/is performed 'Il Barbiere di Siviglia?'", "In which store in Naples can I find a CD of 'Il Barbiere di Siviglia?'" would be given unique, unambiguous and complex answers.

Against this background, from a linguistic point of view, some observations seem necessary on the type of semantics involved in the most spread SW approaches. First of all, it is a fact that specific linguistic complex meanings are always conveyed by expressions that are equally complex, both from a lexical and from a morphological point of view. For instance, this is the case with *acceleratori di particelle* ("particle accelerators"), which is a lexical meaning unit composed by three words,  *fusione fredda* ("cold fusion"), which is a lexical meaning unit composed by two words, or  *particelle di antimateria* ("particles of antimatter") which is a lexical meaning unit as well composed by three words. Secondly, it is also a fact that such complex expressions always tend to maintain unaltered their formal aspect, mainly in order to constantly express the same unambiguous meanings.

These two observations lead to think that in the production of written texts, many combinations of words, and consequently the meanings that such combinations produce, almost never have the characteristics of a Markov chain. Besides, it is worth stressing that such combinations of words are often registered by specialized paper dictionaries, and that they can be formalized to create specific lingware for automatic textual analysis, data mining and semantic information retrieval.

These considerations let us assume the existence of pre-coded lexical ontologies, i.e. ontologies which may be observed between fixed lexical meaning units and their correspondent immutable and non ambiguous meanings. At the same time, all this indicates that meaning extraction from texts can also be completed bybypassing the application of rule-based procedures, i.e. that

---

Lexicon-Grammar, see the link "Bibliographie" on <http://infolingu.univ-mlv.fr/>.

meaning extraction can be more rapidly and easily achieved embedding lexical data bases inside automatic textual analysis software routines.

Therefore, in this paper, we will illustrate the linguistic basis on which to structure efficient semantic and terminological information retrieval software. In section 3 we will describe the main principles for the formalization of complex lexical meaning units, today also known as multiword units. In section 4 we will describe CATALOGA®, a software based on lexical ontologies and which can determine – without human reading – the main topic a specific digitized document deals with. We will close this paper with section 5, in which we will expose our perspectives for future research.

### 3. MULTIWORD UNITS IN LEXICON-GRAMMAR

Multiword units have been an increasingly important concern for natural language processing scholars, being considered as a *pain in the neck for NLP* [20], because of the many difficulties they raise. To begin with, there is no universally agreed definition or term for the concept of multiword unit. Concurrent terms of multiword unit are multiword, multiword expression, fixed expression, idiom, compound word, and collocation used by many authors of different theoretical schools or following distinct natural language processing approaches.

In earlier Lexicon Grammar Theoretical Framework, the most essential features of what we call multiword units were non-compositionality and semantic opaqueness. Maurice Gross [13] uses the term compound word to refer to a string composed of several words the meaning of which cannot be computed from its elements. De Mauro [5] describes it as a group of words with a single meaning, which cannot be inferred from the meanings of the individual words that are part of it, both in common use and in specialized language. Recently the significance of compositionality has changed and the term multiword unit has evolved in such a way that it can also be referred to non-idiomatic units, being now used to refer to various types of linguistic entities, including idioms, compounds, phrasal verbs, light or support verb constructions, lexical bundles, etc. Lexicon-Grammar scholars have long been studying multiword units, and the practical analytical formalization has been done for several languages ([1] and [19] for Portuguese; [15] and [23] for French; [3], [8] and [27] for Italian, among others). Lexicon-Grammar scholar Morris Salkoff has dedicated attention to multiword units in his contrastive studies, particularly between English and French ([21], [22]).

Besides, in [3] multiword units are considered as part of a continuum in which combinations can vary from a high degree of variability of co-occurrence of words (combinations with free distribution), to the absence of variability of co-occurrence. They identify four different types of combinations of phrases or sentences, namely:

1. with a high degree of word co-occurrence variability, i.e. with free internal distribution, compositional and denotative meaning;
2. with a limited degree of word co-occurrence variability, i.e. combinations with restricted internal distribution;

3. with no or almost no word co-occurrence variability, i.e. combinations with fixed internal distribution;
4. without any word co-occurrence variability.

Relations between these mentioned classes can be interpreted not only as relations between distinct classes, but also as relations between poles of the continuum. We give here some examples of these combination classes:

- a) (for combinations at point 1.)
  - Verbal structures: (Max, Ugo, your nephew,...) looks at (a book, the river, Eva,...)
  - Nominal structures: (clean, dirty,...) water
  - Adverbial structures: with (elegance, love, devotion,...)
- b) (for combinations at point 2.)
  - Verbal structures: (Max, Ugo, your nephew,...) dries (the clothes, the laundry,...)
  - Nominal structures: (mineral, sparkling, natural,...) water
  - Adverbial structures: from one (moment, day, year,...) to the other
- c) (for combination at point 3.)
  - Verbal structures: (Max, Ugo, your nephew,...) bends his elbow
  - Nominal structures: heavy water, arsenic water
  - Adverbial structures: in no uncertain terms
- d) (for combination at point 4.)
  - Proverbs: Walls have ears

From a semantic point of view, and with reference to communication processes, we observe that types (c) and (d) may also have “idiomatic” interpretations, or rather interpretations that are not semantically compositional (i.e. not coming from a compositional computation of the meanings of each lexical element). Probably, some of these fixed and idiomatic combinations are the result of metaphoric and metonymic drifts which have been lexicalized.

Starting from these assumptions, we may deduce that the use of the four mentioned combination types originates from the need for incisive and immediate communication processes rather than for ordinary ones. While metaphor and metonymy, as any figure of speech, involve an additional operation of decoding and interpretation, fixed and idiomatic combinations are used as a single block: they are semantic shortcuts, and it is not necessary to know the meaning of each element of the linguistic sequences they are conveyed by. But it is important to stress that in LG, all these types of lexical entries can be formalized, coherently inserted inside linguistic databases (i.e. electronic dictionaries), and used within NLP routines, such as for instance information retrieval and parsing.

#### 3.1 Multiword units Formalization

Each type of multiword unit may need to follow a different formalization method. There is the morphological aspect of multiword units (i.e., the morphology of composition) that weights considerable for morphologically-rich languages and remains a highly challenging task. From a lexicographical point of view, multiword units with a specific grammatical function and an autonomous meaning need to be registered in dictionaries in a systematic way, i.e. as autonomous lemmata and not, as often is the case in traditional dictionaries, as examples of use of head

nouns or adjectives. As far as lemmatization<sup>4</sup> is concerned, a clear distinction between multiword units with a high degree of variability of co-occurrence among words and those with a limited or no variability of co-occurrence among words (compound words, idiomatic expressions, and proverbs) should be made. This is one of the most critical issues in the description of natural languages. For example, there is a relevant difference in Italian between *colletto bianco* and *colletto celeste* (which only has the meaning of “blue collar”). The first has to be lemmatized since it has also the specific meaning of “white collar worker”, and has distinctive formal, morpho-grammatical and lexical properties, i.e.:

- it is invariable, as it does not accept any insertion or addition, for instance *\*colletto molto bianco* (“\*very white collar worker”);
- is a singular masculine compound noun only referring to a “human being”, with *colletti bianchi* as its masculine plural form.

On the contrary, *colletto celeste* does not possess these characteristics, being a free nominal group which does not have to be lemmatized.

This is quite a simple example of the difference between opposite poles in the continuum. Sometimes, however, multiword units are much more difficult to classify and describe. For example, the Italian multiword unit *editto bulgaro*<sup>5</sup> (“Bulgarian edict”) and *elezione bulgara* (“Bulgarian elections”) are on the edge between the status of compound words and that of free nominal groups. This is a problem that occurs most frequently with compound words. According to (Silberstein, 2004:117), an accurate identification of compound words must be based on the following criteria:

1. **semantic atomicity:** if the exact meaning of a nominal group cannot be deduced from the meaning of its components, the nominal group must be lemmatized (=> it is therefore treated as a compound noun; this happens with the already mentioned *colletto bianco*, but also with *teste di cuoio* (members of a special anti-terrorist police team), *casa chiusa* (“brothel”) *Guerra Fredda* (the proper noun “Cold War”), in which each element of the compound participate in the construction of a complete and non-literal meaning;
2. **distributional restriction:** if certain constituents of the nominal group, which by the way, belong to certain natural distributional classes, cannot be freely replaced, then this distributional restriction must be acknowledged by classifying the series of nominal groups in a lexicon, which again, amounts to treating it as a compound noun. For instance, the above-mentioned examples of *colletto bianco* and *colletto celeste* follow this criterion;

<sup>4</sup> The word “lemmatization” is here used with reference to electronic dictionaries, i.e. homogeneous lexical-data bases created to be embedded inside information retrieval software routines.

<sup>5</sup> This compound noun comes from political language and refers to a 2002 speech of the Italian Prime Minister Berlusconi pronounced against some journalists and their banishment from the Italian Broadcasting Service,

3. **institutionalization of the usage:** certain nominal groups, even those that are semantically and distributionally “free”, are used in a quasi-obligatory manner, to the detriment of other potential syntactic constructions that are just as valid, but are never used. The Italian expression *in tempo reale* (a loan translation of the English “in real time”) is an example for this criterion, which use in Italian seems to be unmotivated if we take into consideration that the antonym *\*in tempo irreale* (“\*in unreal time”) is not used at all. These criteria allow identifying a larger group of compound words than it is normally and traditionally assumed for a language.

Concerning information retrieval, we will show in the following pages how crucial is the correct lemmatization of compound words belonging to terminological and technical knowledge domains. In the first place, we will observe that a consistent and successful information retrieval system must necessarily be based on a coherent codification of reality, i.e. on a correct ontological formalization of compounds. CATALOGA®, i.e. the software which will be presented in the following pages, is based on all the theoretical and formal criteria mentioned before.

#### 4. CATALOGA®: A SOFTWARE FOR SEMANTIC INFORMATION RETRIEVAL

CATALOGA® is an information retrieval software developed by Annibale Elia, Alberto Postiglione and Mario Monteleone; it is based on the matching between digitized texts and electronic dictionaries of terminological compound words developed according to the LG lexical formalization method. It is actually configured as a stand-alone software which can be integrated in Web sites and portals to be used online. As already stated, the main linguistic goal of this software is to extract terminology from a given scientific or technological text and to automatically determine – without human reading – the main knowledge domains it deals with.

From a computational point of view, the tasks performed by CATALOGA® can be summarized as follows:

- automatic reading of a text;
- computing of all the occurring terminological compound words, i.e. location and computing of all the occurrences of any of a finite number of compound words;
- statistical computing of the ratio between terminological and non-terminological occurrences;
- statistics-based listing of all the terminological occurrences, in decreasing order and classed on the basis of the pertaining knowledge domains.

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. On the contrary, no specific hardware architecture is explicitly required, given the fact that the software is normally installed and used on both house-desktop and laptop standard Windows computers.

During the start-up phase, considering a Windows XP computer equipped with 3 GB of RAM and a 1.66 GHz dual-core CPU, the software loads and preprocess a whole electronic dictionary

(approximately 500,000 entries) in less than 10 seconds. This preprocessing step is performed only once for each software session. On the contrary, texts processing step is achieved in real time.

The time complexity of CATALOGA® dictionary preprocessing 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 recognized in one pass. During the analysis procedure, the terminological electronic dictionaries are completely allocated in RAM, so that the effective software execution time is very short and does not depend on the dictionary or the text size.

The lingware embedded in CATALOGA® is developed also taking into consideration two specific formal and linguistic considerations:

- in any given scientific and/or technological text, the large part of cognition is conveyed by a small number of terminological words, 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 and non-ambiguous meanings;
- in most languages, there is a close and necessary relation between terminology and a specific subset of multiword expressions, i.e. the one formed by compound words<sup>6</sup>. This is proved by the fact that specialized lexica are mainly formed by compounds, in an average that is often higher than the 80% of the whole registered lexical set. Therefore it is possible to state that from a formal and semantic point of view, terminology fully exploits the syntagmatic procedures of compound word structuring and formation, in which a starting lexical element – for instance a noun with a generic meaning such as *carta* (“card”) – can be specified by adding other lexical elements, as in *carta di credito* (credit card, debit card).

<sup>6</sup> As already stated, within the complex set of multiword units, compound words have specific formal, morphological, grammatical and semantic characteristics which push towards a clear differentiation from other multiword units. Actually, the definition of compound words is based on two different aspects. The first is at the same time morpho-grammatical and semantic, and refers to compounds as to lexical meaning units, i.e. sequences of words different from free groups or phrases; each compound word, and more specifically each compound noun, is in fact a single unit, with a precise grammatical function, and a specific meaning which cannot almost ever be inferred from the words that compose it. The second is essentially formal: it introduces a distinction between uninterrupted sequences of letters limited by blanks and forming simple words such as *sedia* (“chair”) and sequences of words separated by blanks or other diacritics elements, such as *sedia a dondolo* (“rocking chair”). For more specifications on such topics see especially [4], [13], [16], [23], [24] and [27].

In the following paragraphs we will give detailed information on the electronic dictionaries used by the automatic routines of CATALOGA®, more specifically on the matching of entries with knowledge domain tags, which actually represent the bulk of DELAF-DELACF lexical ontologies.

## 4.1 Cataloga® Lingware

The electronic dictionaries used by CATALOGA® are part of the DELA system, originally established by LG. This system is formed by Simple-Word Electronic Dictionaries (DELAS-DELAF), and Compound-Word Electronic Dictionaries (DELAC-DELACF). CATALOGA® uses DELAC-DELACF dictionary which mainly include terminological compounds. Each entry of DELAC-DELACF dictionaries is given a homogeneous ontological description, as it is coherently tagged with reference to the knowledge domain(s) in which it is commonly used (i.e. in which it has a terminological non-ambiguous meaning). For instance, the compound word *acconto dividendo* (“interim dividend”) is marked with the tag ECON, which stands for “Economics”. Also, 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.

These electronic dictionaries are also bidirectionally bilingual, which means that their target-language section is formalised in the same way as the source one.

Also, for each entry a formal and morphological description is given with:

- the internal structure of each compound. For instance, the tag “NPN” indicates that a given compound is formed by a Noun, followed by a Preposition, followed by a Noun. At the same time, the tag “NA”, indicates that a given compound is formed by a Noun followed by an Adjective;
- the gender and the number of a given compound, together with all potential plural forms, both masculine and feminine. For instance, the tag “ms-+” indicates that a given compound is masculine singular (ms), does not have any feminine correspondent singular form (-), and has a plural masculine form (+). At the same time, the tag “fp--“ indicates that a given compound is feminine plural (fp), does not have any masculine correspondent singular or plural form, neither has a correspondent singular feminine form.

As a sample, we give here 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/cleft,.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 inside the data-base of CATALOGA® knowledge domains; this means that this software may analyze texts using 180 different electronic dictionaries. The most important dictionaries are those of Informatics (approx. 54,000 entries), Medicine (approx. 46,000 entries), Law (approx. 21,000 entries) and Engineering (approx. 19,000 entries). Subset tags are also previewed for those domain sectors which include specific subsectors. This is the case with Engineering, for which a generic tag ING is used, while 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). A same formalization 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.

## 4.2 Cataloga® Analysis Procedure

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

*La vitamina A (Retinolo) svolge un'azione protettiva delle mucose e degli epiteli. 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 delle 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<sup>7</sup>.*

After reading and analyzing it, CATALOGA® gives as output a table with the results of the text processing:

**Table 1. CATALOGA® Analysis Results**

|                                                                                 |      |        |      |      |
|---------------------------------------------------------------------------------|------|--------|------|------|
| 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.                                          |      |        |      |      |
| *****                                                                           |      |        |      |      |
| ANALISYS                                                                        |      |        |      |      |
| *****                                                                           |      |        |      |      |
| 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  |
| *****                                                                           |      |        |      |      |
| *                                                                               |      |        |      |      |
| assottigliamento sclerale                                                       | 1    | N+NA   | ms+  | MED  |
| carenze vitaminiche                                                             | 1    | N+NA   | fs+  | MED  |
| cecità notturna                                                                 | 1    | N+NA   | fs-  | MED  |
| macchia di Bitot                                                                | 1    | N+NPN  | fs+  | MED  |
| macchie biancastre della sclera                                                 | 1    | N+NAPA | fp+  | MED  |
| macchie di bitot                                                                | 1    | N+NPN  | fp+  | MED  |
| malassorbimento lipidico                                                        | 1    | N+NA   | ms+  | MED  |
| nevo nevocellulare piano                                                        | 1    | N+NAA  | ms+  | MED  |
| punto di vista                                                                  | 1    | N+NPN  | ms+  | DIGE |
| secchezza della congiuntiva                                                     | 1    | N+NPN  | fs+  | MED  |
| sviluppo scheletrico                                                            | 1    | N+NA   | ms+  | ANAT |
| vitamina a                                                                      | 4    | N+NN   | fs-  | MED  |

developing countries, representing one of the leading causes of blindness. The vitamin A deficiency is usually due to fat malabsorption, to alcoholism, and is seen most commonly in the elderly. An early sign of vitamin A deficiency is night blindness, followed by dryness of the conjunctiva, Bitot's spots (white spots of the sclera). This answer I gave on another site makes you understand which are the origins of the Bitot's spot, and which is its colour, i.e. whitish. Your spot looks more like a small melanocytic nevus, or a scleral thinning area, completely harmless, *sine materia* from the pathological point of view, and showing its bluish component below.” (Available on <http://www.medicitalia.it/consulti/Oculistica/65819/Macchia-nella-sclera>; English translation by Mario Monteleone)

<sup>7</sup> “Vitamin A (Retinol) plays a protective action on the mucous membranes and epithets. It also has a role in growth, supporting the skeletal development. Lack of vitamin A is one of the most common vitamin deficiencies. It is especially common in

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

Similar precise results have been achieved with the analysis of a corpus composed by 1070 text files (10 MB ca.) extracted from Italian online newspapers. The results of these analyses are written out in the following table:

**Table 2 . CATALOGA® Analysis Results on a corpus of 1070 files (approx 10MB)**

| Analysis Results (1070 files tested) |                  |        |
|--------------------------------------|------------------|--------|
| Correct                              | Partially faulty | Faulty |
| 71%                                  | 29%              | 0%     |

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

### 4.3 Evaluation of CATALOGA® Analysis Results

We have shown that CATALOGA® analytic routine can concretely extract terminology, and more generally, semantic information from given texts. 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 grants the 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 on CATALOGA® results, and we can state that this software:

- achieves a more efficient and less “noisy” automatic terminological information retrieval;
- may allow automatic information-based data stocking, bypassing human reading;
- can support the creation of information data bases, in which items (i.e. digitized texts) may be linked/grouped by means of the terminological compound keywords and/or the non-ambiguous knowledge domain tags which they eventually share;
- can grant more rapid management and updating procedures as far as textual terminological information is concerned.

## 5. PERSPECTIVES FOR FUTURE RESEARCH

We have seen that in order to achieve efficient semantic information retrieval, the crucial step to take is the adoption of a well-founded, empirically coherent and solid linguistic formalization method. Great importance must also be given to the

studying of language changes, which essentially occur due to new signification necessities. Therefore, as for the topics dealt within this paper, our future research perspectives will be focused on:

- the updating of the lingware structure. This step will include:
  - the creation of terminological electronic dictionaries for new-born knowledge domains, such as E-government, Bioethics, Biomedicine, and so on;
  - the implementation of multilingual semantic-based terminological analysis;
  - the construction of an ontology-based query interface for Information Retrieval;
  - the testing and validation of the lingware described by means of its application on vaster corpora;
- the integration of CATALOGA® inside Web Sites and Portals, in order to let internauts test the whole system, and to let us receive quality and usability feedback;
- 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;
- 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) to which they belong to, and also of the terminological compound words they include. The information retrieval from such textual relational data bases will be achieved by means of queries structured both on knowledge domain tags and on terminological compound words.

These steps will surely represent a launching pad towards new forms of experimentation, including the creation of SW interactive platforms.

**Author contributions.** Annibale Elia is the author of paragraph 1. Mario Monteleone is the author of paragraphs 2 and 5. Johanna Monti is the author of paragraph 3. Alberto Postiglione is the author of paragraph 4. Daniela Guglielmo has contributed to the revision of the Italian electronic dictionary (DELACF) and to the drafting of paragraph 6.

## 6. REFERENCES

- [1] Baptista, J.: “Sintaxe dos Nomes Predicativos com verbo-suporte SER DE”, in *Fundacao para a Ciencia e a Tecnologia/Fundacao Calouste Gulbenkian*, Lisboa (2005)
- [2] Barreiro, A.: *Make it simple with paraphrases. Automated paraphrasing for authoring aids and machine translation*. PhD dissertation. Faculdade de Letras da Universidade do Porto, Oporto, Portugal, (2008)
- [3] D’Agostino E., Elia A.: “Il significato delle frasi: un continuum dalle frasi semplici alle forme polirematiche”, in AA. VV, *Ai limiti del linguaggio*, Laterza, Bari, (1998), p. 287-310

- [4] De Bueriis G., Elia A. (eds.): *Lessici elettronici e descrizioni lessicali, sintattiche, morfologiche ed ortografiche*, Plectica, Salerno, (2008)
- [5] De Mauro, T.: *Dizionario della lingua italiana*, Paravia, Milano, (2000)
- [6] Elia A., Postiglione A., Monteleone M.: "CATALOGA. Sistema informatico per la catalogazione automatica di testi", Release 4.8., Software, (2010)
- [7] Elia A.: *Le verbe italien*, Schena-Nizet: Bari/ Paris, (1984)
- [8] Elia, A., Monteleone, M., De Bueriis, G., Di Maio, F.: "Le polirematiche dell'italiano", in 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, (2008), p. 11-65
- [9] Gross M.: *Grammaire transformationnelle du français. I- Syntaxe du verbe*, Cantilène, Paris, (1968)
- [10] Gross M., *Méthodes en syntaxe*, Hermann, Paris, (1975)
- [11] Gross M.: "Observations on Semantic Theories", in *Theoretical Linguistics*, n. 5:1, De Gruyter Mouton, New York/Berlin (1978), p. 3-17
- [12] Gross M.: "Les bases empiriques de la notion de prédicat sémantique", in *Langages*, n. 63, Larousse, Paris, (1981), p. 7-52
- [13] Gross M.: "Lexicon-grammar. The representation of compound words", in *Proceedings of COLING '86*, Bonn, University of Bonn, (1986), <http://acl.ldc.upenn.edu/C/C86/C86-1001.pdf>
- [14] Harris Z.: *A Grammar of English on Mathematical Principles*, Wiley, New York, (1982)
- [15] Laporte, E., Voyatzi, S.: "An electronic dictionary of French multiword adverbs." *Proceedings of the LREC Workshop - Towards a Shared Task for Multiword Expressions (MWE 2008)*, LREC. (2008), p. 31-34.
- [16] Monteleone M.: *Lessicografia e dizionari elettronici. Dagli usi linguistici alle basi di dati lessicali*, Napoli, Fiorentino & New Technology, (2004)
- [17] Monti, J.: "La E-translation da Google a Second Life: le più recenti applicazioni di Traduzione Automatica online". *Atti del XIII Congresso della Società Linguistica Italiana*, Bulzoni Editore, Roma, (forthcoming).
- [18] Monti, J.: "La traduzione automatica deve sempre essere trasparente? Spunti di riflessione su trasparenza e qualità nella traduzione", in Guarino, A., Montella, C., Silvestri, D., Vitale, M. (eds), *La Traduzione: Il Paradosso Della Trasparenza*. Liguori Editore, Napoli, (2008), p. 299- 318
- [19] Ranchhod, E., Carvalho, P.: *Expressoes Multipalavra - Questoes lexicais e sintacticas*. Material de ensino na Primeira Escola de Verao da Linguatca (UP, Porto, Portugal, 10-14 July 2006).
- [20] Sag A., Baldwin T., Bond, F., Copestake, A., Flickinger, D. et al.: "Multiword Expressions: A Pain in the Neck for NLP?", in *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CI-CLING 2002)*. Mexico City, Mexico, (2002), p. 1-15
- [21] Salkoff, M.: "Automatic translation of support verb constructions." *13<sup>th</sup> International Conference on Computational Linguistics (COLING 1990)*, University of Helsinki, Finland, 20-25, (1990a), p. 243-246.
- [22] Salkoff, M.: *A French-English Grammar: A Contrastive Grammar On Translational Principles*. John Benjamins Publishing Company, Amsterdam, Philadelphia, (1990b)
- [23] Sierra, G., McNaught, J.: "Natural Language System for Terminological Information Retrieval", in *CICLing'03 - Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing*, Springer-Verlag, Berlin – Heidelberg (2003),p. 541-552
- [24] Silberztein M.: *Dictionnaires électroniques et analyse automatique de textes*, Masson, Paris, (1993)
- [25] Silberztein M.: *Nooj Manual*, 2003, <http://www.nooj4nlp.net/NooJManual.pdf>
- [26] Vietri S., Elia A., D'agostino E.: "Lexicon-grammar, Electronic Dictionaries and Local Grammars in Italian", in Laporte E., Leclère C., Piot M., Silberztein M., (eds.), *Syntaxe, Lexique et Lexique-Grammaire. Volume dédié à Maurice Gross*, Lingvisticae Investigationes Supplementa n. 24, John Benjamins, Amsterdam/Philadelphia, (2004), p.125-136
- [27] Vietri S.: *Dizionari elettronici e grammatiche a stati finiti. Metodi di analisi formale della lingua italiana*, Plectica, Salerno, (2008)
- [28] Vietri S.: *Lessico-grammatica dell'italiano. Metodi, descrizioni e applicazioni*, UTET, Torino, (2004)