

Semantic Web and language resources for e-Government: linguistically motivated Data Mining

Annibale Elia*
elia@unisa.it

Daniela Vellutino*
dvellutino@unisa.it

Federica Marano*
fmarano@unisa.it

Alberto Maria Langella*
albertomarialangella@yahoo.it

Antonella Napoli*
antrapoli@unisa.it

*Department of Political, Social and Communication Sciences, Università degli Studi di Salerno

Via Ponte Don Melillo s.n.c.
84084 Fisciano (SA), ITALY
+39089963017

ABSTRACT

Language Technologies (LT) perform well when they rely on a previous Language Resources (LR) development. Hence, in this paper we illustrate how to build an efficient data mining system based on a coherent formalization of natural language and on a lingware (in Machine-Readable Form) built on the universal concepts of “lexical unit”, “meaning unit” and “morphosyntactic context”. From a linguistic and semantic point of view we get more coherent results through a linguistically motivated data mining than a statistical approach. We developed LR for Natural Language Processing (NLP) applications, composed by electronic dictionaries made of terminological multiword-expressions (Machine-Readable Form) and by local grammars (in the form of finite-state automata and transducers – FSA/FST). Both parts of this lingware were built and applied according to Lexicon-Grammar (LG) formalization principles and methods. The mentioned language resources form the basis to develop an Information Retrieval System for e-Government. This device is a Semantic Web (SW) application, which will automatically recognize a given set of frequently-asked questions (from here on, FAQs) on European Community Information, previously formalized as syntactic patterns inside local grammars.

Categories and Subject Descriptors

I.2.7 [Artificial Intelligence]: Natural Language Processing – *Language models, Language parsing and understanding, Text analysis.*

General Terms

Documentation, Experimentation, Languages, Reliability.

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

Keywords

Data mining, Semantic Web, NLP, IR, linguistic resources, multiword-expressions, electronic dictionaries, local grammars, finite state automata.

1. INTRODUCTION

In this paper we illustrate how to build an efficient data mining system applied to technical and specialized languages, starting from a coherent formalization of natural language. Technical and specialized languages, such as those one of Law, Economics, Informatics, particularly used in e-government tools are in large part formed by compounds, in an average that is often higher than the 80% of the whole lexical set registered, as it exists a strict necessity relation between terminology and a specific set of multiword expressions (i.e. those one formed by compound words¹).

In order to analyze technical and specialized languages we refer to a coherent formalization of natural language stated by LG theory, which describes words' relationship not through statistical rules, but identifying compound words treated as single “meaning unit”. By this approach we get more complete linguistic and semantic results than using a merely statistical one.

¹ As we will see in the following pages, the definition of compound words is based on two different aspects. The first one is at the same time morpho-grammatical and semantic, and refers to compounds as to “lexical” and “meaning units”, i.e. sequences of words different from free groups or phrases and having an overall (and sometimes non compositional) meaning. In this sense, we can make precise distinctions between free sequences of words such as *colletto celeste* (blue collar) and “lexical” and “meaning units” such as *colletto bianco* (white-collar worker). The second one 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 [1], [2], [3], [4] and [5].

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*).

In order to coherently describe this and other creation processes, which occur within technical languages, we studied some specific domains of knowledge, as e-government one.

The building process of our LR follows three steps. The first step is the extraction of compounds from corpora concerning a specific domain. The second step concerns the analysis of their lexical structures, which is achieved in order to describe the distributional behaviors and co-occurrences of all compound words that are registered as fixed, frozen or non-compositional inside glossaries and corpora. Then we build electronic dictionaries in each specific domain that are used in NLP software. The third step is the building of FSA/FST, which are used to recognize given sequences of words, and which successively will give specific output results matching with what requested by the set of input data.

In order to develop an e-government device, we selected EU Information domain whose lexicon is miscellaneous as it particularly includes languages of Law, Economics and Informatics. Analyzing EU Information domain is crucial because of its socio-economic and political relevance for the Member States citizens. Also, the terminological compound words analysis is relevant especially in the translations activities of European Community languages (e.g. the transposition from Community law to National law) [6].

Moreover, the massive presence of compound lexical units that mostly characterize Community lexicon entries [7] may be proved by a simple consultation of:

- all different digitalized documents, produced in the form of specialized terminology glossaries;
- several other resources, such as terminological and multilingual databases, or thesauri, relating to politics and to sectors of EU legal-institutional activity.

As above mentioned the description of terminological compound words is key step for the building and implementing of electronic dictionary in any domain: as for EU Information domain we developed a new dictionary named DIE – Dizionario Informazione Europea (European Information Dictionary). The dictionary was built retrieving data from corpora and glossaries; for each entry, we registered the lexical variants used (in terms of synonyms, hyponyms, hypernyms, and spelling variants) in order to analyze the processes of compound word formation. This led us to identify different types of lexical unit structures: bi-gram, tri-gram, fourth-gram, fifth-gram, and so on, up to longest structures consisting of more than 10 elements. The description of compound words formalized in DIE, as well as in the rest of our electronic dictionaries, is useful not only for those NLP systems achieving automatic recognition plus extraction of morpho-grammatical and lexical structures, but also for those automated processes concerning plain and clear writing of technical and specialized language.

The DIE, together with developed automata to automatically recognize a given set of FAQs about European Community Information, allows us the implementation of an Information Retrieval System for e-Government.

2. FRAMEWORK

2.1 RELATED WORKS ON INFORMATION RETRIEVAL

This work focuses on applying linguistically enhanced data mining for LT development. The main aim is to retrieve terminological multiword-expressions in order to process, in a semi-automatic way, technical and specialized languages. Most methods rely on statistical approaches to retrieve the same linguistic facts. We refer to [8] that developed a language model in a probabilistic framework and to their *vector space model* where documents are represented as vectors in a common vector space [9]. We also refer to *Topic model* approach of [10], [11] and [12] that is based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. A topic model is a generative model for documents: it specifies a simple probabilistic procedure by which documents can be generated. Another approach of [13] is the *Latent Semantic Analysis* one (LSA) that makes three claims: that semantic information can be derived from a word-document co-occurrence matrix; that dimensionality reduction is an essential part of this derivation; and that words and documents can be represented as points in Euclidean space.

On the contrary, our approach gets linguistically motivated text mining based on a systematic and exhaustive formalization of natural language. Unlike *Generative Grammars* [14] and [15], we assume that syntactic rules must consider the lexical phenomena. Of course, while automatic statistical methods get almost good results in a faster way, our manual method, even if time consuming and much more expensive, gives us more accurate and quite complete results useful to accomplish SW project [16].

2.2 LEXICON-GRAMMAR FOR INFORMATION RETRIEVAL

The linguistic methodology adopted for our research is Lexicon-Grammar (LG) theoretical and practical analytical framework, which gives an exhaustive description of lexical and syntactic structures of languages². Lexicon-Grammar is actually one of the most profitable and consistent methods for natural language formalization. The main goal of LG is to describe all mechanisms of word combinations closely related to concrete lexical units and sentence creation. It was set up by French linguist Maurice Gross during the '60s, and subsequently developed for and applied to Italian by Annibale Elia, Emilio D'Agostino and Maurizio Martinelli. In the last twenty years, LG has also reached important results in the domain of automatic textual analysis and parsing, with the creation of software and lingware fully oriented toward NLP,

² [17] is an updated presentation of the Italian Lexicon-grammar and of the linguistics and computational tools developed at the Department of Political, Social and Communication Sciences of the University of Salerno (Italy). For other books and publications about multilingual Lexicon-Grammars, see <http://infolingua.univ-milv.fr/>.

such as INTEX and UNITEX³, and more recently NOOJ⁴ and CATALOGA⁵.

For LG, “simple sentence”⁶ is the minimal linguistic meaning context that can be analyzed; on the basis of this “simple sentence” it is possible to achieve concrete studies on natural languages. LG theory is prevalently based on Zellig Harris’ Operator-Argument Grammar⁷. LG range of analysis also invests lexicon, and especially the concept of “meaning unit” and “lexical unit” and of word group, for which LG identifies four different combinatorial behaviors (see [23]):

- Combinations with a high rank of variability of co-occurrence between words. In this case we have combinations based on opened distribution with a compositional and signified meaning;
- Combinations with a low rank of variability of co-occurrence between words. In this case we have combinations based on constrained distribution;
- Combinations with zero or almost zero rank of variability of co-occurrence between words. In this case we have combinations based on fixed distribution;
- Combinations without variability of co-occurrence between words. In this case we have proverbs.

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

a) (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) (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) (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) (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 each lexical element meaning). 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, as for instance information retrieval and parsing.

Today LG methodology is used by several groups of researchers working in different universities, laboratories and international research structures, which form the RELEX network. The Department of Communication Sciences of Salerno University has had a long scientific collaboration with the LADL (Laboratoire d’Automatique Documentaire et Linguistique) of Paris 7 University, directed for over twenty years by Maurice Gross. This closed scientific collaboration had the aims to refine lingware in order to adapt it to NLP software.

Finally, our research is part of a complex LG study on specialty languages (see also [24], [25], [26], [27], [28] and [29]); it is based on the use of multiword expressions and compound words classifiable as (...) “gruppo di parole che ha un significato unitario, non desumibile da quello delle parole che lo compongono, sia nell’uso corrente sia nei linguaggi tecnico-specialistici”⁸.

³ More information on the website <http://www-igm.univ-mlv.fr/~unitex/>.

⁴ See <http://www.nooj4nlp.net/pages/nooj.html>.

⁵ CATALOGA, is a software developed by Alberto Postiglione, Mario Monteleone and Annibale Elia at Department of Political, Social and Communication Sciences at University of Salerno.

⁶ In LG, 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 the predicate in order to obtain an acceptable and grammatical sentence. The study of simple sentences is achieved analyzing the so-called rules of co-occurrence and selection restriction, i.e. distributional and transformational rules based on predicate syntactic-semantic properties. For more specification on simple sentence definition, see [18].

⁷ Operator-Argument Grammar [19] is based on the assumption that each human language is a self-organizing system, in which 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. See also [20], [21] and [22].

⁸ “(...) group of words having an overall meaning, not inferable from that of the single words by which it is formed, both in current use and in technical and specialized languages”. [30], page VIII; translation by Mario Monteleone.

2.3 Electronic dictionaries vs computerized dictionaries

As for the differences existing between electronic and computerized dictionaries, it has been evidenced in [31] that the term “computerization” has somehow confused the two categories. Print modernization process requires paper dictionaries texts to be typographically composed on computerized supports, but this computerization process doesn’t affect the content of these dictionaries, which remains unchanged. So, both paper and computerized dictionaries are only used by humans having a solid and already existing expertise. On the contrary, electronic dictionaries are only used by computers within specific software routines, and are managed by specialized human users. Therefore, the data included inside electronic dictionaries are formalized by means of codes which are not intelligible for common readers. On such basis, it is also possible to state that a classic paper dictionary is not fully (re)usable within automatic textual analysis.

2.4 The DELA system of electronic dictionaries

Due to its formal characteristics, the DIE is classified as an electronic dictionary of the DELA system, i.e. as a formalized lexical databases⁹, homogeneously structured and in which the morpho-grammatical characteristics of lexical entries (gender, number and inflection) are formalized by means of distinctive and non-ambiguous alphanumeric tags.

DELA electronics dictionaries work as a linguistic engine embedded in automatic textual analysis software and parsers. DELA electronic dictionaries are of two types:

- simple word electronic dictionaries (DELAS-DELA), which include lexical units semantically autonomous and formed by sequences of characters delimited by blanks. This is the case of words as *casa* (house) and *battello* (boat);
- compound word electronic dictionaries (DELAC-DELA), which include lexical units composed by of two or more simple words having an overall meaning. This is the case of sequences as *casa di cura* (nursing home), and *battello a vapore* (steamboat).

As already stated, compound word electronic dictionaries mostly lemmatize terminological entries¹⁰.

⁹ The term *database* is intended according to the most common meaning of Informatics, from both the theoretical and the practical point of view.

¹⁰ Unlike simple words, which are often polysemic and ambiguous, compounds have a polysemic rank almost always near to zero, which is an important characteristic as far as terminological and specialized languages are concerned. Besides, as already stated, from a formal and morphological point of view there are concrete differences between simple and compound words, which must be necessarily accounted both with reference to NLP routines and when building linguistic databases. For instance, automatic linguistic queries, information retrieval and automatic textual analysis procedures necessarily require different reading procedures these two types of meaning units, in order to grasp and reuse their main linguistic functions. In fact, the absence of separators within simple words and their presence inside compounds will necessitate

2.5 Example of Electronic dictionaries

As a sample, we give here an extract of the Italian-English Electronic Dictionary of Medicine and Biomedicine that is one of those specialized dictionaries useful in e-government services for Public Utility Information in Health domain:

facce/a/maschera,faccia/a/maschera.NPN:fp-+;MED/
=mask/face,.NN:s+/N
facce/a/maschera,faccia/a/maschera.NPN:fp-+;MED\$GCH/
=masked/facies,.AN:s+/N
facce/a/maschera,faccia/a/maschera.NPN:fp-+;MED/
=masklike/face,.NN:s+/N
facce/amimiche,faccia/amimica.NA:fp-+;MED\$MGH/
=mask/face,.NN:s+/N
facce/anteriori/del/braccio,faccia/anteriore/del/braccio.NAPN:
fp-+;MED/ =facies/anterior/brachii\NAN/N
facce/anteriori/del/cristallino,faccia/anteriore/del/cristallino.
NAPN:fp-+;MED/ =facies/anterior/lentis\NAN/N
facce/anteriori/del/femore,faccia/anteriore/del/femore.NAPN:
fp-+;MED/ =facies/anterior/femoris\NAN/N
facce/anteriori/del/pancreas,faccia/anteriore/del/pancreas.NA
PN:fp-+;MED/ =facies/anterior/pancreatis\NAN/N
facce/anteriori/del/radio,faccia/anteriore/del/radio.NAPN:fp-
+;MED/ =facies/anterior/radii\NAN/N
facce/anteriori/del/radio,faccia/anteriore/del/radio.NAPN:fp-
+;MED/ =facies/volaris/radii\NAN/N
facce/anteriori/del/rene,faccia/anteriore/del/rene.NAPN:fp-
+;MED/ =facies/anterior/renis\NAN/N
facce/anteriori/del/surrene,faccia/anteriore/del/surrene.NAP
N:fp-+;MED/
=facies/anterior/glandulae/suprarenalis\NANA/N
facce/anteriori/dell'avambraccio,faccia/anteriore/dell'avambr
accio.NAPN:fp-+;MED/ =facies/anterior/antebrachii\NAN/N
facce/anteriori/dell'avambraccio,faccia/anteriore/dell'avambr
accio.NAPN:fp-+;MED/ =facies/volaris/antebrachii\NAN/N
facce/anteriori/dell'iride,faccia/anteriore/dell'iride.NAPN:fp-
+;MED/ =facies/anterior/iridis\NAN/N
facce/anteriori/dell'ulna,faccia/anteriore/dell'ulna.NAPN:fp-
+;MED/ =facies/anterior/ulnae\NAN/N
facce/anteriori/dell'ulna,faccia/anteriore/dell'ulna.NAPN:fp-
+;MED/ =facies/volaris/ulnae\NAN/N
facce/anteriori/della/cornea,faccia/anteriore/della/cornea.NA
PN:fp-+;MED/ =facies/anterior/corneae\NAN/N
facce/anteriori/della/gamba,faccia/anteriore/della/gamba.NA
PN:fp-+;MED/ =facies/anterior/cruris\NAN/N
facce/anteriori/della/gamba,faccia/anteriore/della/gamba.NA
PN:fp-+;MED/ =shin\N/N
facce/anteriori/della/mascella,faccia/anteriore/della/mascella.
NAPN:fp-+;MED/ =facies/anterior/maxillae\NAN/N
facce/anteriori/della/prostata,faccia/anteriore/della/prostata.
NAPN:fp-+;MED/ =facies/anterior/prostatae\NAN/N
facce/anteriori/della/rotula,faccia/anteriore/della/rotula.NAP
N:fp-+;MED/ =facies/anterior/patellae\NAN/N

different settings and choices with regard to automatic data processing: simple words will only need alphabetic and eventually numeric characters analysis; on the contrary, compounds and more generally multi-word units, which actually do present separators, will need an additional level of data analysis, i.e. the one concerning not only single words, but also the overall sequences of words which compose meaning units and which unique and unambiguous values must be assigned to. These values will also and necessarily be different from those assigned to alphabetic and numeric characters.

*facce/anteriori/delle/palpebre,faccia/anteriore/delle/palpebre.
 NAPN:fp-+;MED/ =facies/anterior/palpebrarum\NAN/N
 facce/anterolaterali/dell'omero,faccia/anterolaterale/dell'ome
 ro.NAPN:fp-+;MED/
 =facies/anterior/lateralis/humeri\NAAN/N*

2.6 Automatic Textual Analysis Software

To build and test electronic dictionaries and local grammars (as finite state automata) we use two software packages: NOOJ and CATALOGA.

As for NOOJ¹¹, it is a complex NLP environment in which it is possible to automatically read digitalized texts, retrieving from them specific linguistic patterns in the form of concordances. NOOJ engine is based on the DELA system of electronic dictionaries, on LG syntactic tables and on FSA/FST, built in the form of graphs and used in LG to parse texts.

CATALOGA is data mining software that can read text files and establish, without any human intervention:

- if a given text deals with a generic or a terminological topic;
- which is the eventual main specific knowledge domain dealt with in that text;
- as for the same text, if other terminological knowledge domains are dealt with, and which statistical relevance they have with reference to the main one.

This analytical inferential procedure is achieved by means of two specific elements:

- a lingware composed by terminological electronic dictionaries, both monolingual (Italian) and bilingual bidirectional (Italian-English); the entries of such dictionaries are all marked with terminological tags;
- two algorithms which achieves the matching between the text to analyze and the already mentioned electronic dictionaries.

The procedure achieved by CATALOGA is based on the following steps:

- inside a given text, the localization of all occurrences of any of a finite number of terminological words;
- the achievement of a statistical analysis on the two word classes (i.e. non terminological vs. terminological);
- the statistical weight comparison of both words classes;
- the creation and display of a list in which the knowledge domains dealt with are ranked in a statistically decreasing order;
- also, the creation and display of a list of all the terminological words retrieved, alphabetically ordered and tagged according to the knowledge domain (or domains) they belong to.

All terminological words can be simultaneously recognized in one pass. During the analysis procedure, the terminological

electronic dictionaries are completely allocated in RAM, so effective software execution time is very short and does not depend on dictionary or text size.

Up to today, CATALOGA has been used to analyze large and heterogeneous text corpora. The results achieved reach a 71% of correct analyses, a 29% of partially-correct analyses, and a 0% of incorrect analyses. Partially-correct analyses depend on electronic dictionaries that need of continuous update. It is important to stress the CATALOGA achieves detailed and successful analyses also with very short text files.

3. LEMMATIZATION METHOD

As already stated, the aim of study is to build linguistic resources in European Community knowledge domain, by means of lexicographic and computational methodologies. To accomplish such an aim, we used corpora which include 3.071.610 tokenization; also, 16 institutional glossaries were collected and tested. The matching between corpora and glossaries was achieved to validate all the entries and also to record all lexical variants used for each terminological word. In order to build a corpus useful to detect multiword-expressions (i.e. compound words), texts were classified on the basis of their different communication and/or pragmatic aims. A great importance had the different processes of information and public communication [32]. Therefore, the texts of the corpora were classified on the basis of the communication forms used and of the related textual typologies, as shown in the following scheme:

- Juridical Communication (674.746 tokenizations) → type of text: juridical → example of text: Lisbon Treaty;
- Communication for Institutional Administration (2.390.093 tokenizations) → type of text: document of economic planning → example of text: Programming Complement;
- Communication for Legal Publicity (12.490 tokenizations) → type of text: Gazzetta → example of text: Notice;
- Communication for Accountability (133.671 tokenizations) → type of text: document of economic policy → example of text: General Report on the European Union; example of text: European Union general budget for the financial year 2011;
- Public and Institutional Communication for Information of Public Utility (6.771 tokenizations) → type of text: FAQ → example of text: Requests for information about funding from the Structural Funds.

Both terminological and lexicographical linguistic information were recorded in a specific database. Each entry is described according to:

- a terminological definition with a related source statement;
- all lexical variants, i.e. synonyms, hyponyms and hypernyms;
- spellings (lowercase, uppercase, capitalized);
- acronyms (such as *AdG* that is a variant of the compound *Autorità di Gestione*, i.e. *Managing Authority*);
- other variants (such as the compound *Programma Operativo Regionale*, i.e. *Regional Operational Programme*, which may have variants as *POR* e *P.O.R.*);

¹¹ NOOJ Italian Module has been developed by Simonetta Vietri of the Department di Scienze della Comunicazione, in Salerno University (Italy). For more information on NooJ, see www.nooj4nlp.net.

- sociolinguistic variants; this is mainly the case of foreign compound words used in Italian but having also corresponding Italian entries (such as *gender mainstreaming*, which becomes *prospettiva di genere* and *ottica di genere*).

After this terminological and lexical indexing, morpho-grammatical tagging, PoS and flectional tagging were achieved. By means of a set of alphanumeric and not ambiguous marking, notations about gender, number, grammatical category and compound words internal structure were added to all entries.

3.1 EU Information Lexicon

As above mentioned, from a semantic and terminological point of view, EU Information lexicon is miscellaneous, as it may be considered as [...] *un linguaggio triplo-specialistico in quanto abbraccia sia il settore comunitario, sia quello normativo, sia quello specifico che i vari documenti via via trattano*¹².

A part from denominating juridical principles of European law, EU terms often denote:

- economic concepts: *Fondi Strutturali* (Structural Funds), *Accordo di Programma Quadro* (Framework Agreement Programme), *Cooperazione Territoriale Europea* (European Territorial Cooperation); institutions, *Autorità di Gestione* (Management Authority), *Comitato di Sorveglianza* (Monitoring Committee), *Organismo Intermedio* (Onlending Body), *Stato membro* (Member State), *Unione Europea* (European Union);
- titles of intervention programmes in economic policy: *Programma Operativo* (Operational Programme), with the registered variant *PO* (OP), an acronym also used to abbreviate the juridical concept of the *Pari Opportunità* (Equal Opportunities); *Settimo Programma Quadro* (Seventh Framework Programme), with the two registered variants *7° Programma Quadro* and *7PQ*;
- titles of official documents: *Trattato di Lisbona* (Treaty of Lysbon), *Libro Verde* (Green Paper), *Libro Bianco* (White Paper), *Curriculum Vitae Europeo* (European Curriculum Vitae), *Rapporto Annuale di Esecuzione* (Annual Report of Implementation), also registered with the acronym variant *RAE*.

The creation of such compound words has to be interpreted as the specific decision of an Institution, who aimed at acting as a juridical and administrative organization. Compound words creation process within EU lexicon is characterized by an increasing of the semantic specificity produced through the addition of constituents inside each syntagmatic structure. This creation process can be observed in the lexical composition determined by the simple nominal structure noun+adjective (N+A); this is the case of *sviluppo sostenibile* (sustainable development), *deficit democratico* (democratic deficit), *mercato comune* (common market). The same fact may be observed for those lexical structures in which this specification takes places thanks to the addition of the adjective *europeo/europea* (European); this is the case of *Agenzia Europea* (European agency), *Parlamento Europeo*

(European Parliament), *Integrazione europea* (European integration), *Commissione Europea* (European commission).

Passing from specialized communication – which is typical of EU experts – to mass-media communication, it is possible to note that this type of specification is consolidating the process of derivational lexical formation, especially by means of the apposition of the prefix *euro-*; this is the case of words as *Europarlamento* (Europarlament), *Europartenariato* (Europartenariat), *Eurobarometro* (Eurobarometer), *Eurozona* (Eurozone).

Another widely used structure which confirms the existence of a lexical production mechanism through semantic increasing is the sequence noun + prepositional phrase (N+PP), which is observable in *Piano d'Azione* (Action Plan), *Prospettiva di Genere* (Gender Mainstreaming), *Centro di Accoglienza* (Reception Centre).

Moreover, lexical combination of compound words is based on the gradually addition of lexical elements that increase the specificity of the institutional domain(s) concerned; this is the case of *Agenzia Europea dell'Ambiente* (European Environment Agency), *Agenzia Europea per i Medicinali* (European Medicines Agency), *Agenzia Europea per la Gestione della Cooperazione Operativa alle Frontiere Esterne* (European Agency for the Management of Operational Cooperation at the External Borders), *Personalità Giuridica dell'Unione* (Legal Entity of EU).

Also, there are frequent cases in which compound words become head nouns of other compounds, as it happens with denominations of actions and titles of official documents; for instance *Piano d'Azione* (Action Plan) from which derives *Piano d'azione per la Società dell'Informazione* (Action Plan for Information Society), *Piano d'Azione per l'Efficienza Energetica* (Action Plan for Energy Efficiency), *Piano d'Azione Nazionale per l'Inclusione* (National Action Plan for Inclusion).

3.2 Electronic Dictionaries: building and application

The building and managing of an electronic dictionary consists of three main steps:

- *Lexical acquisition*. During this phase, multiword-expressions are extracted, the terminological properties of which are established on the basis of the specialized domain coped with by the analyzed documents. The linguistic function of each lemma is analyzed in order to record lexical variants that are formally described according to their combination structure. The matching between corpus and glossaries is achieved to verify all the entries and also to record any linguistic variants for each terminological word.

- *Morpho-grammatical and syntactic tagging*. It is achieved by means of inflectional paradigms, and in order to inflect all entries. The following string gives a sample of this morpho-grammatical formalization procedure:

carte d'identità elettroniche, carta d'identità elettronica, N+NPNA+f+p+DIE

The elements that form the pattern of morpho-syntactic structure, “NPNA” (name + preposition + name + adjective), are preceded by the tag “N” (noun), which indicates the grammatical function of the whole compound. Other indications are “f” and “p” (feminine plural) and “DIE”, the terminological tag, which refers to our electronic dictionary of

¹² “[...] is a language formed by three specific domains, as it includes the EU area, the juridical area and a specific area that is gradually treated by documents.” [33], p. 205. Translation by Federica Marano.

In order to acquire information on compound words formation processes, we identify the typologies of multiword-expressions in the dictionary, as shown in the following table:

N° of constituents in lexical unit	Syntagmatic structure	Example
4 <i>bi-gram</i>	AN NA NN NV	pubblica amministrazione documento digitale interfaccia utente soggetto abilitato
11 <i>tri-gram</i>	NAA NPN ...	posta elettronica certificata servizio di autenticazione ...
20 <i>fourth-gram</i>	NAPN ...	Carta Nazionale dei Servizi ...
More than 10 constituents	NCNAPNP- NPNACA	detenzione e diffusione abusiva di codici di accesso a sistemi informatici o telematici

Currently, the most important dictionaries are those of Informatics (54000 entries ca.), Medicine (46000 entries ca.), Law (21000 entries) and Engineering¹³ (19000 entries ca.). Each dictionary has been created and verified under the supervision of domain experts.

¹³ Subset tags are also previewed for those domains that include specific subsectors. This is the case of 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).

- *Construction of FST.* An automatic analysis of each FAQ text is achieved by means of NooJ. Successively, each FAQ is transformed into an FSA/FST by means of the function “construct FST”. Finally, only automata nodes relevant to linguistic description is manually modified, in order to build correct and non-ambiguous local grammars for each FAQ (see examples in figures 1-2).

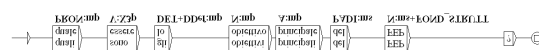


Figure 1. FST for recognition of FAQ: Quali sono gli obiettivi principali del FES? (Which are the main objectives of FES?)

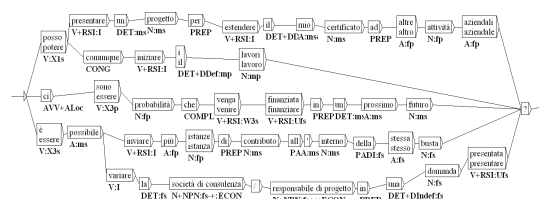


Figure 2. FST for recognition of more FAQs in the same graph.

- *Validation and generalization.* In this phase, the identification and classification of more general linguistic elements is achieved in order to construct more general automata, applicable also to other corpora. We replace each word of the nodes with morpho-grammatical tags. In this way, verbs such as *presentare*, *inviare* and *variare* are replaced with grammatical category <V>, which allows us to recognize all verbal entries. This method is applied also for other grammatical categories such as preposition <PREP>, adjectives <A>, adverbs <AVV> and so on (see figure 3).



Figure 3. Generalized FST based on grammatical categories.

- *Embedding*. By means of meta-nodes, in this phase we create a unique automaton made of all the single local grammars previously constructed for each corpus topic. Local grammars are subdivided according to the specific types of text they had been created for. The meta-nodes of the sing graph realized allow us to easily access to all sub-graphs (see picture 4). The main aim is to obtain a complete system of automatic recognition of “polite expressions” useful to develop a first step for a semi-automatic responder in which it will be possible to associate pertinent answers to each specific recognized FAQ.

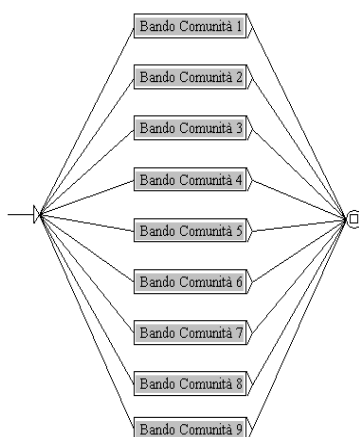


Figure 4. Embedded graphs.

4. QUALITATIVE EVALUATION

Our approach doesn't require a statistical evaluation of results but allows a qualitative one. As it is a supervised approach based on a manually-developed lingware, our linguistic resources are exhaustive and also enriched by assistance of domain experts and by implementation of institutional and acknowledged lexica. As a consequence our resources could be appropriately considered a *gold standard* useful, for example, in machine learning systems and/or machine translation. In other words, the real evaluation process is then to be carried out not on the procedure itself, but on the performativity of electronic dictionaries, which must always be updated to ensure consistent and reliable results.

5. CONCLUSIONS AND FUTURE PERSPECTIVES

This research concerned the building of language resources (electronic dictionaries and local grammars – i.e. FST/FSA) for e-Government useful to carry out an Information Retrieval System based on automatic recognition of sentences and syntagmatic structures. This procedure is also useful to verify the existence of recursive linguistic creation models for terminological compound words. Besides, the building of these automata is the first step towards the creation of a semi-automatic responder or a dedicated search engine.

Moreover, we have shown that in order to achieve efficient semantic data mining, the crucial step to take is the adoption of a well-founded, empirically coherent and solid linguistic formalization method. Also, great importance must be given to the study of language changes, which essentially occur due to new signification necessities.

Hence, as for the topics dealt with in this paper, our future research perspectives will be focused on the three main themes:

1. validation and updating of the lingware structure. This step will include:
 - i. creation and implementation of terminological electronic dictionaries for new-born semantic domains, such as E-government, Bioethics, Biomedicine, Renewable Energy and so on;
 - ii. implementation of multilingual semantic-based terminological analysis;
 - iii. design of an ontology-based query interface for Information Retrieval;
2. testing of the lingware described on more large corpora;
3. integration of lingware and software tools inside Web Sites and Portals, in order to let web-surfers test the whole system, and to let us receive quality and usability feedback.

These three steps will surely serve as a launching pad towards new forms of experimentation and, conclusively, towards a concrete formulation of Semantic Web platforms, that today do not seem to us so far to achieve.

Author contibutions. Annibale Elia is author of paragraphs 1, 2.2 and 4; Daniela Vellutino is author of paragraphs 3, 3.1 and 3.2 together with Federica Marano, and also paragraph 5; Federica Marano is author of paragraphs 2.3, 2.4, 2.6; Alberto Maria Langella is author of paragraph 3.3; Antonella Napoli is author of paragraph 2.1 and 2.5.

REFERENCES

1. Gross M. (1989). La construction de dictionnaires électroniques. Annales des Télécommunications, vol. 44, n° 1-2: 4-19, CENT, Issy-les-Moulineaux/Lannion.
2. Silberztein M. (1993). Dictionnaires électroniques et analyse automatique de textes, Masson, Paris.
3. Monteleone M. (2002). Lessicografia e dizionari elettronici. Dagli usi linguistici alle basi di dati lessicali. Fiorentino & New Technology, Napoli.
4. Vietri S. (2008). Dizionari elettronici e grammatiche a stati finiti, Plectica, Salerno.
5. De Bueriis G. et al. (2008). Le polirematiche dell'italiano. In De Bueriis G. e Elia A. (a cura di), Lessici elettronici e descrizioni lessicali, sintattiche, morfologiche ed ortografiche. Plectica, Salerno.
6. Cosmai D. (2007). Tradurre per l'Unione Europea, Hoepli, Milano.
7. Nystedt J. (2000). L'italiano dei documenti CEE: le sequenze di parole, in Atti del Convegno di studi "Linguistica giuridica italiana e tedesca: obiettivi, approcci, risultati" (Bolzano 1-2 ottobre 1998), Unipress, Padova.

8. Manning C.D. and Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press Cambridge, Massachusetts, London, England.
9. Manning C.D., Raghavan P. and Schütze H. (2008). *Introduction to Information Retrieval*, Cambridge University Press, New York, USA.
10. Blei D.M., Ng A.Y. & Jordan M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.
11. Griffiths T.L. & Steyvers M. (2002). A probabilistic approach to semantic representation. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*.
12. Hofmann T. (1999). Probabilistic Latent Semantic Analysis. *Proceedings of Fifteenth Conference on Uncertainty in Artificial Intelligence*.
13. Landauer T.K. & Dumais S.T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
14. Chomsky N.A. (1957). *Syntactic Structures*. Mouton, The Hague/Paris
15. Chomsky N.A. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge, Massachusetts.
16. Berners-Lee T., Hendler J. & Lassila O. (2001). The semantic web. *Scientific American*, May.
17. Vietri S., Elia A., D'Agostino E. (2004). "Lexicon-grammar, Electronic Dictionaries and Local Grammars in Italian", in Laporte, Leclère, C., Piot, M., Silberstein M. (eds.), *Syntaxe, Lexique et Lexique-Grammaire*. Volume dédié à Maurice Gross, *Lingvisticae Investigationes Supplementa* 24, John Benjamins, Amsterdam/Philadelphia.
18. Gross M. (1968). *Grammaire transformationnelle du français*. – I – *Syntaxe du verbe*, Paris, Larousse.
19. Harris Z. (1982). *A Grammar of English on Mathematical Principles*. John Wiley and Sons, New York, USA
20. Harris Z. (1964). Transformations in Linguistic Structure. *Proceedings of the American Philosophical Society* 108:5, pp. 418-122.
21. Bloomfield L. (1933). *Language*. Henry Holt, New York, USA.
22. Harris Z. (1976) *Notes du cours de syntaxe*, Paris, Seuil, 1976
23. D'Agostino, E., Elia. A. (1998). "Il significato delle frasi: un continuum dalle frasi semplici alle forme polirematiche", AA. VV., *Ai limiti del linguaggio*, Laterza, Bari, pp. 287-310.
24. Gross M. (1975). *Méthodes en syntaxe, régime des constructions complétives*, Hermann, Paris.
25. EMDA – Elia A., Martinelli M., D'Agostino E. (1981). *Lessico e strutture sintattiche*, Liguori, Napoli.
26. Elia A. (1984). *Le verbe italien. Les complétives dans les phrases à un complément*, Schena-Nizert, Fasano di Puglia – Parigi.
27. D'Agostino E. (1984). *Lessico e sintassi delle costruzioni locative*, Liguori, Napoli.
28. Vietri S., Elia A. (2000), "Electronic Dictionaries and Linguistic Analysis of Italian Large Corpora", in Rajman M. & J. C. Chappelier (eds.) *JADT 2000 – Actes des 5es Journées internationales d'Analyse statistique des Données Textuelles*, 9-11 Mars 2000, Ecole Polytechnique fédérale de Lausanne, Suisse, pp. 181-196.
29. Elia A. (2002). "Discorso scientifico e linguaggio settoriale. Un esempio di analisi lessico-grammaticale di un testo neuro-biologico", in Cicalese A., Landi A. (eds.), *Simboli, linguaggi e contesti*, Quaderni del Dipartimento di Scienze della Comunicazione – Università di Salerno, n. 2, Carocci, Roma.
30. De Mauro T. (2000), *Il dizionario della lingua italiana*. Paravia-Bruno Mondadori, Milano.
31. Vietri S. (2001). *Navigare nei testi. Teorie e applicazioni informatiche per la linguistica testuale*. Editoriale Scientifica Italiana, Napoli.
32. Vellutino D. (2009), "La comunicazione pubblica per la promozione delle pari opportunità", in V. D'Antonio e S. Vigliar (a cura di), *Studi di Diritto della Comunicazione. Persone, Società e Tecnologie dell'Informazione*, CEDAM, Padova.
33. Nystedt J. (1999). "L'italiano che si scrive a Bruxelles", *Italiano e oltre*, XIV, La Nuova Italia, Firenze
34. Silberstein M. 2002. *NooJ Manual*. Available for download at: www.nooj4nlp.net.