Implementing the NLP infrastructure for Greek Biomedical Data Mining

Aristides Vagelatos
RACTI
10 Davaki str.
GR-11526 Athens,
Greece
vagelat@cti.gr

Elena Mantzari
Neurosoft S.A.
32 Kifisias Ave.
GR-15125 Athens
Greece
emantz@tee.gr

Giorgos Orphanos
Neurosoft S.A.
32 Kifisias Ave.
GR-15125 Athens
Greece
orphan@neurosoft.gr

Christos Tsalidis
Neurosoft S.A.
32 Kifisias Ave.
GR-15125 Athens
Greece
tsalidis@neurosoft.gr

Chryssoula Kalamara
Athens' Euroclinic
9 Athanasiadou str.
GR-15121 Athens
Greece
kalamach@otenet.gr

Christos Diolis
RACTI
10 Davaki str.
GR-11526 Athens,
Greece
diolis@cti.gr

Abstract

This paper presents the design and implementation of terminological and specialized textual resources that are produced in the framework of the Greek national R&D project "IATROLEXI". The aim of the project is to create the critical infrastructure for the Greek language, i.e. linguistic resources and tools, to be used in high level Natural Language Processing (NLP) applications in the domain of Biomedicine. The project builds upon existing resources that have been developed by the project partners, i.e. a Greek morphological lexicon of about 100.000 words, and language processing tools such as a lemmatiser and a morphosyntactic tagger, and it will further develop new resources such as a specialised corpus of biomedical texts and an ontology of medical terminology.

Keywords

Ontologies, data mining, biomedical terminology.

1. Introduction

The amount of biomedical information which is contemporarily produced by the medical society, i.e. health institutions, educational organisms and research institutes, has been enormously increased. This information which is mainly available in digital form and mostly accessible through Internet has been characterized by Eysenbach [4] as "information jungle" of narrative form, due to its enormous size and its unstructured form. However, information is only valuable to the extent that it is accessible, easily retrieved and relevant to the users' interests. The growing volume of data, the lack of structured information, and the information diversity have made information and knowledge management a real challenge towards the effort to support the medical society. It has been realised that added value is not gained merely through larger quantities of data, but through structuring of the data into knowledge for more sophisticated access to the required information.

In order to access information, medical practitioners, researchers, patients, or other interesting parts in the medical market are usually provided with unsophisticated tools, such as simple search engines which are seriously limited by their reliance on keyword-matching. These search mechanisms are unable to find information described by different terms and they often return information results that use the same words with a different meaning, while they are unable to combine information from diverse sources. These problems can be alleviated if search engines no longer search for matching keywords but for matching semantic concepts that underlie the information in web pages. The lack of high level language tools to facilitate accuracy and precision in accessing and retrieving the relevant information is harder in a less-used language like Greek, due to the limited research funding and the restricted interest by the medical industry, and also due to the intrinsic particularities of the Greek language morphology.

The project IATROLEXI[1] (http://www.iatrolexi.gr) aims at the creation of the critical infrastructure for the Greek language which will constitute the groundwork for advanced NLP applications in the domain of biomedicine: i.e. text indexing, information extraction and retrieval, data mining, question answering systems, etc. To accomplish this, a number of essential tools and resources for the Greek language are under construction, which will allow better management and processing of the digitally encoded information in the biomedical field.

More specifically, the expected output of the project are tools that will address directly the final user of the biomedical information, such as a spelling checker of Greek medical terms as well as a specialized search engine, and also tools that will mainly assist processing of the Greek biomedical texts and improve search and retrieval of biomedical data, such as a tagger for morphosyntactic annotation appropriately tuned to the particularities of the biomedical sublanguage and an ontology of the Greek biomedical terminology.

This paper is structured as follows: Section 2 presents some background information on Natural Language Processing in the biomedical domain towards data mining.

[1] "IATROLEXI" project is being partially funded by General Secretariat of Research & Development (project code: 9) within Measure 3.3 of "Information Society" Operational Program.

Section 3 gives a description of IATROLEXI project's goals and presents the environment and their main components. Finally in section 4 the conclusions are given.

2. Background

Natural Language Processing (NLP) has been applied to biomedical text for decades, in fact, soon after computerized clinical record systems were introduced in the mid 1960s [2]. The computerization of clinical records increased the tension in the field of medical reporting and recording. In [12] a broad overview of NLP in medicine can be found, with special attention to milestone projects and systems such as the Linguistic String Project, Specialist, Recit, MedLEE, and Menelas. The overview of [6] also concentrates on NLP with clinical narrative, giving a short summary of earlier projects and the state of the art at that point in time.

In recent years, research has continued to focus on text indexing and document coding to allow powerful, meaningful retrieval of documents. Document indexing uses terms from a glossary or ontology (MeSH, Gene Ontology, Galen4) or text features such as words or phrases. Most NLP systems in clinical medicine work with text from patient records such as discharge summaries and diagnosis reports. NLP systems in bioinformatics use mostly articles or abstracts from the scientific medical literature. Differences between these two types of text affect the choice of techniques for NLP. Biomedical literature is carefully constructed and meticulously proofread, so spelling errors and incomplete parses are less of a problem. On the other hand, new concepts may be introduced, such as a newly unraveled molecule.

Ontologies are considered to be a fundamental prerequisite for advanced language processing, knowledge management and the Semantic Web, since they offer the mechanisms for the formal representation and the description of the concepts in a given domain[1], [13]. Typically, an ontology identifies classes of objects that are important in a domain and organises these classes in a hierarchy. Each class is characterised by some properties and is related to other classes or to elements of other classes through a number of significant relations. The predominance of ontologies as knowledge sources in information processing lies on their power to represent knowledge in a model that is comprehensible equally by either humans or machines, thus assisting communication between human agents, achieving interoperability among computer systems, and advancing the systems' quality performance on indexing, processing, retrieval and extraction of required information.

A significant amount of work in developing an NLP system concerns extending lexical knowledge. Since there is a very large number of words and phrases associated with clinical concepts, the task of adding entries to the lexicon is considerable demanding [11]. The National Library of Medicine has undertaken a large-scale effort to facilitate access to biomedical information. The development of the UMLS (http://umlsinfo.nlm.nih.gov/) and the release of the SPECIALIST lexicon will substantially benefit NLP systems. A UMLS concept is given a unique identifier, and all synonymous concepts have the same identifier. This feature provides a substantial body of knowledge that NLP systems need: link words in text to a controlled vocabulary (the UMLS or to one of the other source vocabularies). The UMLS also has a semantic network and assigns semantic categories to all concepts. For example, "fever" is assigned the category SIGN/SYMPTOM. The categorization provides the semantic knowledge needed by NLP systems to identify relevant units of information. The SPECIALIST Lexicon, which has over 250,000 entries, assigns syntactic categories to words and phrases in biomedical text. The lexicon is not only useful for NLP extraction tasks, but also for indexing and vocabulary development.

Other nomenclatures are also important knowledge sources. Some work has been published investigating the use of SNOMED (http://www.snomed.com/) and ICD10 (http://www.who.int/classifications/icd/en/) as knowledge sources for lexical work. Like the UMLS, these nomenclatures are also effective for identifying relevant clinical terms and semantic categorization. Both SNOMED and ICD10 are particularly useful to groups involved in multilingual work because they are available in other languages and because the codes provide a way to link a concept to a similar concept in other languages.

Other types of knowledge sources needed by NLP systems, such as grammars, and domain models, are not available to NLP researchers. These are usually developed by each individual research group, and are more complex and interrelated than nomenclatures. They are also typically very difficult to adapt to different systems.

3. Project's main goals: Resources and tools

In order to apply data mining techniques Greek biomedical texts, it is inevitable that a number of text analysis tools and linguistic resources need to be developed. These tools constitute the basis for any application regarding data mining, NLP, indexing, etc. In this chapter the main goals of the project are discussed as well as the environment and its constituent parts are presented.

3.1 Corpus of biomedical texts

To the best of our knowledge there are no Greek electronic medical corpora exist, structurally or linguistically annotated. Thus within the projects' framework, a medical corpus is under construction, mainly from the literature that is already published on the web.

Balance and representativeness are the main requirements for corpus design. According to these requirements, the scope was to develop a Greek corpus of written texts, coming from all different domains of biomedicine. The corpus should contain documents from as many biomedical text fields as possible. Recent research makes clear that full-text articles are preferable from abstracts, if we want to build high-recall text mining systems [3]. Therefore, it seems clear that a corpus that is to be used for biomedical text mining systems should include full text and not samples, which we seriously took under consideration in the development of the IATROLEXI corpus.

Corpus annotation is the distillation procedure adding (or extracting the) value to the texts. The annotation process of the IATROLEXI corpus involves almost all NLP components adopted, constructed or under construction in the framework of IATROLEXI: a tokeniser, a sentence splitter, a morphosyntactic tagger, a biomedical gazetteer, a multi-word term recogniser, and an ontology-based semantic tagger.

Due to time limitations we considered only documents from Internet sites, thus we recorded portals or other websites that included directories of health-related information. We started our investigation from websites of research and academic institutions, e.g.:

- MedNet Hellas – http://www.mednet.gr (a Greek Medical Network),
- Greek National Documentation Center – http://www.ekt.gr,
- Library of University of Macedonia – http://www.lib.uom.gr

The above sites proved to be very helpful, since they contained a rather exhaustive list of directories of Greek biomedical journals. Next, we utilised popular search engines in order to identify additional websites that might contain interesting texts, e.g.:

- Google – http://www.google.com
- Yahoo – http://www.yahoo.gr
- Live Search – http://search.live.com

Through these search engines, we mainly acquired the web addresses of Greek medical conferences that were not listed in the directories mentioned above. Overall, forty websites were identified to contain appropriate medical documents for IATROLEXI. So far, the total number of documents is touching 6,250 (about 11.5 million words).

3.2 Creation, enhancement and/or adaptation of existing resources and tools

A number of resources have been created, enhanced and/or adapted in order to constitute an environment supporting a) the discovery of syntactic patterns that can be candidate multiword terms, b) the construction of the ontology, c) the detection of medicine terms in the documents of the corpus d) semantic indexing of the documents. The core mechanism for the most of the software components working on the documents of the corpus is annotation.

The software implementation platform of all NLP components is Java v 1.5. The operational environment integrating and orchestrating the software components working with annotations is the Apache UIMA platform. UIMA stands for Unstructured Information Management Architecture; it was developed by teams from IBM Research and IBM Software Group and is now released to the open-source community as an Apache project.

The main components constructed or are under construction, participating in the analysis, annotation and indexing of the documents, along with the resources they use, are presented in the following sections.

3.2.1 Document conversion

The documents collected from the internet are either in html or in pdf format. On the other side all the tools process documents in a common format which is pure text decorated with annotations. The UIMA terminology for this common format is CAS (Common Annotation Structure). To satisfy the requirement of feeding the annotation process with documents of a common format, we decided this format to be *plain text*, for the reason that only the textual content of the documents is of interest; scripting, styling, formatting and page rendering information had to be filtered out. Therefore, we developed two document converters: an html-to-txt converter and a pdf-to-txt converter.

The html-to-txt converter incorporates the functionality of the CyberNeco HTML Parser along with the xpath facilities provided by Apache Xalan. To convert an html document to plain text, it is first parsed by the HTML parser and an HTML DOM (Document Object Model) is constructed into memory; noisy elements, such as <style>, <script> and <applet>, are filtered out during parsing. Then, the textual content is selected from the DOM with the help of xpath queries.

The pdf-to-txt converter is based on the PDFBox library. The main problems we faced during pdf-to-txt conversion were: a) the incorrect interpretation of Greek characters, especially for pdf documents produced on Mac systems and b) the injection of newline ('\n') characters in unwanted positions, even in the middle of words.

The output of document conversion is one CAS per input document, which contains the plain text extracted from the document along with global annotations

3.2.2 *Tokenisation and sentence splitting*

Content analysis starts with tokenization, i.e. conversion of the character stream to a token stream. Tokenisation is carried out in two steps. In the first step, a text stream is roughly converted into a token stream based on white space delimiters and some symbol characters. At the same time, the orthography of each token is recorded. By "token orthography" we mean the classes of the constituent characters, e.g. `νόσος` is a *Greek-letter-lower-case* token, `Disease` is an *English-letter-first-capital* token, `H.I.V.` is an *English-letter-all-capital* + *middle-dots* + *ending-dot* token. In the second step, the token stream passes through a refinement module. Tokens of a specific orthography may further split into two or three tokens. For example, a token that ends with a comma or question mark or exclamation mark or colon or semi-colon will split into two tokens; a token that starts with a quote and ends with a quote will split into three tokens.

Special care is taken for tokens that end with a dot, so as to decide whether this dot is part of the token (e.g. the token is an abbreviation) or the dot is a punctuation mark (i.e. a full stop). Among the various tests performed towards the disambiguation of the ending dot, the one worth-mentioning (because it covers the ninety percent of the cases) refers to tokens where all the characters before the dot are Greek letters. If these letters are more than two and constitute a valid Greek word, then the token splits into two tokens: a Greek-word token and a full-stop token. The validity of a Greek word is examined through lookup in Neurosoft's Morphological Lexicon, a broad-coverage lexicon of Modern Greek (~90.000 words, ~1.200.000 word-forms).

Sentence splitting examines the token stream produced from the second step of tokenization and locates tokens that traditionally play the role of sentence delimiters, i.e. full stops, question marks, exclamation marks and dot-ending tokens. It then examines the local context of the candidate sentence delimiters and sets the sentence boundaries on tokens that are proved to be real sentence delimiters.

3.2.3 *Morphosyntactic tagging*

Morphosyntactic tagging is based on the Morphological Lexicon. The contents of the lexicon are organised into morphological lemmas. Each lemma contains all the word-forms of a Greek word accompanied by the values of their morphosyntactic attributes. The basic morphosyntactic attribute of a word-form is its part-of-speech. The value of part-of-speech determines what other morphosyntactic attributes characterise a word-form: gender, number and case for nouns, adjectives, articles, pronouns and present perfect participles; voice, tense, mood, number and person for verbs. The first word-form of a morphological lemma, the headword, plays the role of lemma representative; referring to the headword is the same as referring to the lemma. As the morphological lexicon is monolingual, morphosyntactic annotations are assigned only to Greek words.

Each Greek-letter token identified during tokenization is assumed to be a Greek word-form. Every word-form is looked-up in the morphological lexicon. The possible outcomes are three: a) the word-form is found in one morphological lemma, b) the word-form is found in two or more morphological lemmas and c) the word-form is not found. Since the goal of morphosyntactic analysis is to assign unambiguous morphosyntactic annotations to word-forms, outcomes (b) and (c) are problematic; outcome (b) introduces ambiguity while outcome (c) introduces failure. If the morphological lemmas of outcome (b) have different part-of-speech values (which is the most frequent), the selection of the appropriate lemma can be interpreted as the selection of the appropriate part-of-speech value. Also, to overpass the failure of outcome (c), the only way is to guess the values of as many morphosyntactic attributes as possible – at least the part-of-speech. Part-of-speech disambiguation and guessing is carried out with the help of decision trees through examination of the local context (see [10]), achieving an accuracy of ninety-seven percent in part-of-speech disambiguation and eighty-nine percent in part-of-speech guessing.

3.2.4 *Biomedical word identification*

The next step was to mark words that belong to the biomedical domain. This marking was crucial for the next processing steps. Every single biomedical word may be a biomedical term by itself (which can be certified through look-up in a biomedical dictionary or ontology) or may be part of a multi-word biomedical term.

Biomedical words are identified with the help of a gazetteer that currently contains ~52,000 biomedical word-forms (that correspond to ~9,000 biomedical words). The contents of the gazetteer partly come from the Morphological Lexicon and partly were collected through a process described in section 3.3.

3.2.5 *Multi-word term recognition*

The multiword recognition mechanism is one of the advanced outcomes of the project. It is based on a rule description system where every rule recognizes a syntactic pattern in the input text. Rules can be applied in a consecutive and aggregative manner. Consecutive means that rules are applied in the same sequence of annotated text spans repeatedly i.e. as far as we can apply rules and the size of the text span's sequence is decreased, the processing continues. Aggregative means that a set of rules can be applied after another set of rules.

The format of the rules resembles the context free BNF rules where every symbol is presented as a set of feature value pairs. The grammar is strongly typed in the sense that every feature has a type which specifies the values of its instances in the rules. The syntax of the rules is depicted in the following sample grammar consisting of two rules:

```
options:
  grammar  = "Article";
  maxdepth = "8";
types:
  ATTRS is set of external
     "com.neurolingo.NLP3.morphology.IMorphology";
features:
  MORPHO is object;
  ONTO   is object;
functions:
  Contains in module Morpho of file internal
    is object of (ATTRS) as object;
  GNC_Agreement in module Agreement of file
    internal is predicate of (number, ATTRS)
    as object;
  GNC_Reduction in module Reduction of file
    internal is rule of (number, ATTRS,
                         ATTRS, text) as object;
//GNC_Reduction is called in order to create the
// reduced predicate. The arguments are:
//  pivot:   number is the pivot predicate
//           (in our case the second).
//  select_attrs: The attributes of the pivot
//               element (in case that the pivot
//               predicate has more than one alts)
//  common_attrs: The result attributes (or the
//           common one) The Gender, Case & Number
//           attributes are taken from pivot
//           predicate. The remaining attributes
//    are these attributes
//  lemma_frmt: Is an format string describing
//      how the headword (lemma) of the
//      multiword text span will be computed
rules:
  /* A_R1 */
  [MORPHO=GNC_Reduction(2,[N],[N],"%2")] =>
       \
       [MORPHO=Contains([ART]),
        ONTO=$x:GNC_Agreement(1,[ART])],

       [MORPHO=Contains([N]),
        ONTO=$x:GNC_Agreement(1,[N])]
       /
       ;

/* A_R2 */
[MORPHO=GNC_Reduction(2,[ADJ], [ADJ],"%2")] =>
       \
       [MORPHO=Contains([ART]),
        ONTO=$x:GNC_Agreement(1,[ART])],
       [MORPHO=Contains([ADJ]),
        ONTO=$x:GNC_Agreement(1,[ADJ])]
       /
       ;
```

Figure 1 Sample grammar

The parts presented in a rules file are:

- **options** affects the way the rules will be processed and used. In the sample grammar of Figure 1 we set two options. The name of the grammar rules "*Article*" which specifies the name of the annotator that will apply these rules to text spans. The other option with name "*maxdepth*" specifies the number of levels that operators like * (Kleene star) and + will be expanded.

- **types** presents new derived types that features can use. Our formalism uses the primitive types **number** and **text** and the derived types of **set** and **value**. In our sample we can see another important characteristic of the formalism, the ability to communicate with the implementation Java environment. The members of set ATTRS is defined in the interface class identified with the full path name `"com.neurolingo.NLP3.morphology.IMorphology"`. This way we can use the morphological attributes of our lexical resources in grammar rules and in the software components we develop without the need to have duplicate definitions.

- **features** defines the names and types of the features we are going to use in the grammar rules. All features that are going to define grammar symbols in the following rules must be defined in this section. There are no untyped features, as we already mentioned, and the system accomplishes a strong type checking of how values and types are used in the rules. In our sample we define two features with names MORPHO and ONTO which both are of type **object**. This is another extension characteristic of our formalism permitting incomplete or generic types that are defined in the Java environment. The way these types are instantiated and used in the rules will be shown in the following paragraphs.

- **functions** section defines functions that can appear in expressions specifying the values of features in the rules. There are four types of functions

  1. **Object** functions can appear in the body symbols (predicates) of a rule. There are object (instance) methods (in Java parlance) that can take a list of parameters and return a value assigned in a feature. Function *Contains* in the sample grammar takes as input parameter a set of attributes and its return type is the superclass type **object**. The module *Morpho* must be known to the environment executing these rules. This module contains the definition of the object's actual type where this function is encapsulated. The system can accept external modules placed in jar files and loaded dynamically where needed permitting the extension or incorporation of the rules component with external systems.

  2. **Predicate** functions are static methods of a Java class. They can be presented only in the body predicates. Except from the defined parameter list, these functions enriched with an extra parameter. This parameter is the table of all feature value pairs assigned to the predicate they appear in. The

function *GNC_Agreement* checks the agreement of Gender, Number and Case of the neighbor symbols found in input. The first parameter appearing in its definition is a number denoting the way this agreement must be checked. We can specify if we want full agreement in Gender, Number and Case or partial agreement in Gender and Number, in Number and Case, only in Case, etc. The second parameter specifies a set of attributes that the symbol must possess as an extra matching condition.

3. **Rule** functions are used in the head symbol (predicate) of a rule and are mapped to static methods of a class. They take as extra input parameter, a representation of the reduction i.e. the predicates recognized accompanied with the values of the features they contain. Function *GNC*_Reduction is used in order to compute the morphological attributes and the headword of the multiword reduced text span of the rule. The interpretation of the parameters, appearing in line comments, follows the definition.

4. **Feature** functions are the fourth type of functions. They appear in the head predicates and mapped also to static functions of a class. They take as input parameters a list of feature names. These feature names must appear in the body predicates and when called by the system all values of these body features have been evaluated.

- **rules** section contains the actual grammar rules. Every rule contains a head predicate and one or more body predicates. Head is defined in terms of the body predicates and this means that if a sequence of symbols (text spans) matches the body predicates then we can reduce these predicates to the one of the body. Rules are independent of each other. Their order does not matter the way they are evaluated. The system can use different heuristics about which rule to choose for reduction in case that multiple rules match an input sequence of symbols. The current applied technique chooses the longest (in terms of size of predicates in the body of a rule) rule. The symbols '\' and '/' specify the left and right context of a reduction. We can have a list of predicates at the left of the '\' symbol denoting the left context of the reduction. The meaning of the left context is that we expect to match all the predicates presented in the left context but we will not use them in the reduction. The same holds for the right context. Only the predicates presented between the '\' and '/' symbols will be reduced. Parentheses can also be used to group sequence of predicates. A body predicate or group can be right followed by a repeating operator of the '*', '+', {m,n}. The meaning of '*' is zero or more instances of the predicate or group existing in the left of the operator must be matched. The '+' operator is interpreted as one or more instances while the expression {m,n} means that we expect to match at least m and an most n instances.

We constructed a parser based on ANTLR. The parser takes as input a unification grammar (written according to the already specified formalism) and produces a compiled representation of the rules. The actual application of the rules is performed by an execution engine, which loads the compiled rules at start-up (i.e. the parser is the execution engine plus the parsing model). The execution engine incorporates a prototype unification algorithm for the efficient handling of multi-valued features, which facilitates the treatment of the inherent morphosyntactic ambiguity (for more on unification, see [8]).

3.2.6 Ontology-based semantic tagging

According to Kiryakov *et al.* [7], there are a number of basic prerequisites for the representation of semantic annotations:

- an ontology (or taxonomy, at the least), defining the entity classes;
- entity identifiers, which allow those to be distinguished and linked to their semantic descriptions;
- a knowledge base with entity descriptions.

As the aim of IATROLEXI is to build a generic and application independent infrastructure for the language processing of the Greek biomedical data, the project team opted for the adoption of the UMLS knowledge resources, namely UMLS Metathesaurus (MT) and UMLS Semantic Network (SN). Adopting UMLS semantic network as an initial top-level ontology, and mapping it into Greek, we gain access to the conceptual information for some thousands of biomedical terms. Up to now, the whole number of the SN semantic types and semantic relations have been translated into Greek, while both English and Greek versions of the SN have been fed into Protégé for further processing and evaluation.

By semantic tagging in the context of IATROLEXI we mean providing automatic annotations with references to the semantic types of the Greek version of the UMLS Semantic Network.

3.3 A methodology for the development of a biomedical ontology

The methodology will combine bottom-up and top-down approaches for the determination of the semantic/conceptual framework to be used for the knowledge representation of the biomedical domain (i.e. selection of a conceptual hierarchy, semantic classes, relations between concepts, etc.) and the selection of the relevant biomedical terms that designate and instantiate the concepts of those hierarchy nodes. The UMLS semantic network will be used as a frame basis for expressing the IATROLEXI's ontology. The construction and the gradual enrichment of the ontology will be accomplished through the following steps:

1. determination of an initial up-level taxonomy which will be gradually enriched with lower level information on concepts and terms,
2. collection of specialized texts in the biomedical domain,
3. semi-automatic excerption of the texts' terminology,
4. determination of the morpho-syntactic rules that describe the structures in which the relevant terms are realised,
5. extraction of candidate terms,
6. enrichment of ontology with selected terms and relations, and
7. a loop of steps 4, 5 and 6, for as many times as needed.

4. Conclusions

NLP infrastructure is a key element in the further development of informatics applications in several areas, such as data mining, knowledge-based decision support, terminology management, and systems interoperability and integration. A significant body of work now exists that reports on experiences with various approaches in important problem areas of research. On the contrary in the biomedical field and especially for the Greek language, there is not much work implemented.

Currently, a part of our efforts focuses on the completion of the multi-word term recogniser. In subsection 3.2.5 we presented the extraction of candidate multi-word terms from the corpus, based on linguistic knowledge. To automatically decide upon real multi-word terms, we have to exploit some type of statistical evidence which will help us to compute a term-validity metric (e.g. the C/NC-value metric, see [5]).

Project IATROLEXI aims to cover this certain gap by developing a number of NLP resources as well as application for the scientific community. On the one hand the scientist may use the outcomes of the project in his/her own way towards his/her special research needs. On the other hand, the user may look for information in texts or make searches with specific terms or combination of terms or relations that relate terms to each other.

We envisage (at least) three applications of the bilingual biomedical dictionary:

1. Semantic tagging. Any term found in the dictionary can receive an annotation that encodes its semantic type and thus links the term with the UMLS Semantic Network.
2. Bilingual term searching. A Greek term can be translated to its American equivalent(s) and then searched in American texts, and vice-versa.
3. Ontology-based query expansion. A query that contains a term of a specific semantic type can be enriched with other terms of the same semantic type or with terms of narrower semantic types.

References

[1] Alexander, U. (2006). Methods in Biomedical Ontology. Journal of Biomedical Informatics, Vol. 39 (2006) 252--266

[2] Bruijn, B., Martin J: Getting to the (c)ore of knowledge: mining biomedical literature. Int. Journal of Medical Informatics, Vol. 67. (2002) 7—18

[3] Cohen B., L. Fox, P. Ogren and L. Hunter (2005) Corpus Design for biomedical natural language processing. *ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, Detroit.*

[4] Eysenbach, G.: The Semantic Web and healthcare consumers: a new challenge and opportunity on the horizon?. International Journal of Healthcare Technology and Management, Vol. 5, No.3/4/5 (2003) 194 – 212

[5] Frantzi, K. T. and S. Ananiadou (1999) 'The C-value/NC-value domain-independent method for multi-word term extraction'. *Journal of Natural Language Processing*, Vol. 6, No. 3, 145-179.

[6] Friedman C., Gripcsak G.: Natural language Processing and its future in medicine. Acad. Med., Vol. 74, No.8 (1999) 890 – 895

[7] Kiryakov, A., B. Popov, I. Terziev, D. Manov and D. Ognyanoff (2003) Semantic Annotation, Indexing and Retrieval. *ISWC' 2003*, *Florida*.

[8] Knight, K. (1989) 'Unification: A multidisciplinary survey'. *ACM Computing Surveys*, 21(1), pp. 93-124.

[9] Kokkinakis, D.: Developing resources for Swedish Bio-Medical text mining. Proceedings of the 2nd Int. Symposium on Semantic Mining in Biomedicine (2006), Jena, Germany.

[10] Orphanos G. and D. Christodoulakis (1999) Part-of-speech Disambiguation and Unknown Word Guessing with Decision Trees. *9th EACL Conference, Bergen, Norway.*

[11] Rosse, M.: A Reference Ontology for Biomedical Informatics: The Foundational Model of Anatomy. Journal of Biomedical Informatics, Vol. 36 (2003) 478--500

[12] Spyns, P.: Natural Language Processing in medicine: an overview. Methods Inf. Med. Vol. 35 (1996) 285--301

[13] Vickery, C.: Ontologies. Journal of Information Science. Vol. 23 (1997) 277--28