Developing a Greek Biomedical Corpus towards Text Mining

Tsalidis Christos,[1] Orphanos Giorgos,[1] Mantzari Elena,[1]
Pantazara Mavina,[1] Diolis Christos[2] and Vagelatos Aristides[2]

Short Summary

Collection and annotation of specialised corpora, for less-spoken languages such as Greek, is a crucial endeavour for the development and growth of the language technology research for these languages. This paper presents the design and compilation of a biomedical corpus in the framework of the national R&D project "IATROLEXI". The aim of IATROLEXI is to create the critical infrastructure for the Greek language, i.e. linguistic resources and tools, to be used in advanced NLP applications in the domain of biomedicine.

Abstract

The project IATROLEXI (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, text mining, question answering systems, etc. To accomplish this, a number of essential tools and resources will be constructed for the Greek language, which will allow better management and processing of the information in the biomedical field. This will be made possible through the compilation of a representative corpus of biomedical texts and the construction of NLP tools for structural, lexical and semantic annotation of those texts. In this paper, we present the design and compilation of the Greek biomedical corpus.

The collection criteria of the texts were originally imposed by the project requirements: the corpus should comprise of written texts only. Due to time constraints, downloading texts from websites was proved to be the only viable and certainly less time consuming solution.

Overall, forty Greek websites were identified to contain appropriate medical documents for IATROLEXI. Most of the documents are paper abstracts, full papers, and conference proceedings. The majority of them, apart from the body text, contained additional information like images, tables, graphical representations, *etc*. The total number of documents that were collected up to now is approximately 6,250, from which the 69.8 percent is in hypertext markup language (.html) while the rest (30.2 percent) is in portable document format (.pdf).

[1] Neurosoft S.A.
e-mail: tsalidis@neurosoft.gr, orphan@neurosoft.gr, mantzari@neurosoft.gr, mavina @neurosoft.gr
[2] Research Academic Computer Technology Institute
e-mail: diolis@cti.gr, vagelat@cti.gr

1 Introduction

The amount of biomedical information contemporarily produced by the medical society has been enormously increased. With an overwhelming amount of textual information in the domain of biomedicine, there is a need for effective text processing that can help medical practitioners, researchers, patients, or other parts interested in the medical market to access the information encoded in text documents. Different text processing techniques have been developed in order to facilitate efficient extraction of information contained in large collections of scientific texts. The long-term goal is to support text mining, thus discovering the tacit knowledge "buried" in the texts and present it to users. Text mining's advantage, compared to knowledge discovery carried out manually or through unsophisticated search engines is based on the ability to process enormous amounts of texts efficiently and systematically. However, 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 (e.g. complex morphology, free word order).

In this paper, we discuss the design principles and the compilation of a Greek biomedical corpus, as well as the production of text annotations that will provide sufficient metadata input to advanced information retrieval, information extraction and text mining algorithms. We also give concrete examples of applications that are built on top of the annotated corpus: a) a biomedical vocabulary collector and b) a sophisticated concordancer.

The development of the biomedical corpus is carried out in the framework of IATROLEXI, a project that is partially funded by the General Secretariat of Research and Technology – Greek Ministry of Development, within Measure 3.3 of "Information Society" Operational Program.

The paper is structured as follows: section 2 gives some background information on design and encoding issues for developing text corpora and a brief overview of biomedical corpora developed so far for NLP applications; section 3 presents the design principles of the IATROLEXI corpus; section 4 discusses the process of document collection and classification; section 5 describes the production of document annotations; section 6 gives two application examples; and, finally, section 7 gives the directions of work in progress.

2 Background

It is well known that "the beginning of any corpus study is the creation of the corpus itself. The decisions that are taken about what is to be in the corpus, and how the selection is to be organised, control almost everything that happens subsequently. The results are only as good as the corpus." (Sinclair, 1997: 13)

A specialised corpus is a text collection of a specific domain or sublanguage designed for specific research purposes. A well-designed specialised corpus should satisfy some general conditions such as: *representativeness* of the collected language samples, *coherence* of its internal structure, *homogeneity* according to the selection criteria, *variety* of language uses and text types, *balance* between text types and genres, and *coverage* resulting from the size of both the samples and the total corpus. Typically, design specifications for specialised corpora take account of the following

(Bowker, 1996; Friedbichler and Friedbichler, 1997): a) *the types and genres of texts* (i.e. specialised corpora must include scientific texts, educational texts as well as popularised articles), b) *the number of words per text* (i.e. it is highly recommended that specialised corpora include full texts and not samples), and c) *the size of the corpus* (i.e. 500.000 - 5 million word forms are sufficient enough for a specialised corpus).

Large electronic text corpora and machine-readable dictionaries (MRDs) belong to the so-called language resources (LRs), which are bodies of large electronic language data used as primary source to support research and applications in the field of natural language processing (NLP). Typically, such textual data are enhanced with extra information by a process called *encoding*, which makes explicit certain features and properties of texts in such a way as to aid their processing by distinct computer applications.

Since the early 1990s, several projects have worked on issues concerning standardisation of the representation and annotation (encoding) of language resources, with the basic aim to improve their interchangeability, reusability and processing efficiency by distinct language engineering applications. The guidelines or standards came up from those initiatives apply mainly to the *format* (i.e. SGML, XML, Lisp-like structures, annotation graphs, database format, *etc*.), the *annotation content* (i.e. categories for morphosyntactic, syntactic, or semantic annotation), and the *general architectural principles* of the LRs (i.e. pipeline architecture, stand-off annotation, *etc*.). Among the most remarkable projects are: TEI, CES/XCES, MATE and EAGLES/ISLE for *standardization of resource representation and annotation*; RDF/OWL and XTM for *knowledge representation*; Dublin Core, OAI-PMH and XMI for *metadata representation and interchange*; and TIPSTER, GATE, ATLAS, NITE and UIMA for *general text processing architecture*.

Several R&D projects for biomedical language processing have worked on the collection and annotation of biomedical corpora mostly for English (see among others, Zweigenbaum, 2001; Teufel and Elhadad, 2002; Smith *et al*., 2005; Kokkinakis, 2006). Moreover, Cohen *et al*., 2005, make an evaluation of six, publicly available, biomedical corpora for English (these are: PDG, Wisconsin, GENIA, MEDSTRACT, Yapex, GENETAG), according to various corpus design features, in order to set the bases for the design of the next generation biomedical corpora. Particularly, the GENIA corpus (Kim *et al*., 2003) is considered to be the most appropriately annotated corpus for use in biomedical NLP related activities. To the best of our knowledge there is no Greek biomedical corpus yet.

3 Corpus design principles

The design principles adopted for the IATROLEXI biomedical corpus deal with its balance and representativeness, as well as with its annotation.

Some of the document selection criteria were originally imposed by the specific requirements of IATROLEXI. According to these requirements, the scope was to develop a Greek corpus of *written* texts, coming from all *different* domains of biomedicine. Moreover, the corpus should contain representatives from as many biomedical text genres as possible: abstracts, articles, conference presentations, books, dictionaries, definitions, databases, clinical reports, patient records, *etc*. As for the size of documents to be collected, a growing body of recent research makes clear that full-text articles are different from abstracts, and full-text articles must be tapped if we

want to build high-recall text mining systems (Cohen *et al*., 2005). Experience has shown that significant amounts of data are not found in abstracts, but only in the full texts of the articles, or even in tables and figure captions (Shatkay and Feldman, 2003). 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.

Even a perfectly balanced corpus would have been of little utility –especially for text mining– if we did not anticipate a document annotation process. The annotation process of the IATROLEXI corpus involves almost all NLP components adopted, constructed or are 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. The document annotations fall into the following categories:

- **Global annotations.** They encode global document properties such as title, author(s), affiliation(s), source URL, genre, date of creation, *etc*.
- **Structural annotations.** They are used to define the physical structure of the document, e.g. its organization into sections, headings, paragraphs, sentences and tokens.
- **Lexical annotations.** They are associated to short text spans (smaller than a sentence), referring to one or more underlying tokens, and identify lexical units of some significance, e.g. single-word or multi-word biomedical terms, person names, company names, temporal expressions, *etc*.
- **Semantic annotations.** They extend lexical annotations with some type of semantic information, e.g. with concept identifiers and semantic categories acquired from an ontology.

4 Document collection and classification

Due to time limitations and after a few discouraging contacts with Greek publishers, we decided to consider only easily-accessible document sources, i.e. 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). Table 1 presents the websites that contributed the most to the IATROLEXI corpus:

Document source	**n. of docs**
Εγκέφαλος (Brain) http://www.encephalos.gr/index.html	152
Οφθαλμολογικά Χρονικά (Ophthalmology Annals) http://www.eyenet.gr/edition_gr/	135
Ελληνική Καρδιολογική Επιθεώρηση (Greek Cardiovascular Review) http://www.hcs.gr	380
Ελληνική Ακτινολογία (Greek Radiology) http://www.helrad.org/	320
Επιθεώρηση (Review) http://www.psnrenal.gr/periodiko/	103
Αρχεία Ελληνικής Ιατρικής (Greek Medical Archives) http://www.mednet.gr/archives/index.html	309
Ιατρικό Βήμα (Medical Tribune) http://www.iatrikionline.gr/	240
Παιδιατρική Βορείου Ελλάδος (Northern Greece Paediatrics) http://www.paediatriki.gr/	209
Δελτίο Α΄ Παιδιατρικής Κλινικής Πανεπιστημίου Αθηνών (Bulletin of 1st Paediatric Clinic of Athens University) http://www.iatrikionline.gr	115
Θέματα Μαιευτικής, Γυναικολογίας (Issues of Gynecology and Obstetrics) http://www.iatrikionline.gr/index1.htm	220
Ωτορινολαρυγγολογία (Otorinolaringology) http://www.iatrikionline.gr/index1.htm	222
Ελληνική Μαιευτική και Γυναικολογία (Greek Gynecology and Obstetrics) http://www.iatrikionline.gr/index1.htm	225
Info Gastroenterology http://www.iatrikionline.gr/index1.htm	130
Info Respiratory Medicine http://www.iatrikionline.gr/index1.htm	561
Info Urology http://www.iatrikionline.gr/index1.htm	151
Διαβητολογικά Νέα (Diabetes News) http://www.mednet.gr/greek/soc/ede/top.htm	196
Ελληνική Χειρουργική (Greek Surgery) http://www.mednet.gr/hss/	235
Πνεύμων (Lung) http://www.mednet.gr/pneumon/top.htm	238
22ο Ετήσιο Πανελλήνιο Ιατρικό Συνέδριο (22nd Annual Greek Medical Congress) http://www.mednet.gr/greek/epis/form5.htm	463

Table 1: Greek websites that were identified to contain useful biomedical documents

The documents were downloaded using a web crawler that was adapted for the needs of the project (e.g. Greek character handling). The storage media was the file system, thought special care was taken in order to preserve the document related information, e.g. the source URL, the document creation and download dates, *etc.*

The medical documents that were collected for IATROLEXI corpus were paper abstracts, full papers, conference proceedings, and documents with more than one article in the same file. Most of them, apart from the text body, contained additional information like images, tables, graphical representations, *etc*. Moreover, part of the corpus also contained some English text (mostly, the abstract in English) which may

help in a future construction of some kind of parallel corpus. Overall 6,276 documents were collected, from which the 69.8 percent was in hypertext mark-up language (.html) while the rest (30.2 percent) was in portable document format (.pdf).

The Greek health directories found on the Web included magazine titles without any distinction regarding: the type of the publication (e.g. electronic or printed), the type of the magazine (e.g. scientific or mainstream), the type of the content (e.g. full text, abstract, *etc*.), the format of the text (e.g. html, pdf, txt, doc, jpeg, *etc*.), the current status (e.g. magazines that are no longer on publication, *etc*.) and the accessibility (e.g. free or limited access, *etc*.).

The biomedical documents collected so far have been classified as regards to medium and topic. The "medium" classification was more or less straightforward since they were either periodical articles (abstracts or full papers) or conference papers. Regarding the "topic" classification, an appropriate scheme was developed manually by the medical experts, based on medical specialties. Documents coming from websites of specific medical societies were easily classified according to this scheme. The rest were classified through content examination. Table 2 illustrates the number of documents per topic:

Topic	**n. of docs**	**Topic**	**n. of docs**
Allergy	4	Neurology	78
Anaesthesiology	12	Neurosurgery	104
Cardiology	454	Ophthalmology	137
Cytology	4	Orthopaedics	162
Dermatology	1265	Otorinolaringology	231
Endocrinology	29	Pathologoanatomy	612
Forensic Medicine	2	Paediatrics	324
Gastroenterology	143	Pneumonology	525
General Medicine	4	Psychiatry	26
Genetics	4	Radiology	341
Gynaecology – Obstetrics	403	Rheumatology	15
Haematology	20	Social Medicine	14
Medical Issues (in general)	810	Surgery	283
Microbiology	19	Urology	163
Nephrology	14		

Table 2: Number of documents per medical topic

5 Document annotation

From the entire set of Greek NLP components specified in the plan of IATROLEXI, some components were available at the commencement of the project (as partners' contributions), some were developed at the early project stages and others are under development; the latest will become available progressively till the end of the project.

This led us to divide the document annotation process into two phases: a) production of basic annotations (through utilisation of available components) and b) production of advanced annotations (through utilisation of forthcoming components). We call "basic annotations" the annotations that represent the outcome of tokenisation, sentence splitting, morphosyntactic tagging and biomedical word identification. Annotations that characterise a document globally (e.g. title, author(s), affiliation(s), source) are also basic annotations. We call "advanced annotations" the annotations that represent the outcome of multi-word term recognition and semantic tagging.

The software implementation platform of all NLP components is Java v 1.5. To integrate these components into the annotation process, we adopted 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. Among the many useful (and sophisticated) features of UIMA, we were mainly attracted by a) its pretty straightforward mechanism of composing document analysis engines from primitive NLP components that cooperate via well-defined interfaces, and b) its powerful annotation representation model, called Common Annotation Structure (CAS), which borrows many ideas from the object-oriented world: annotations are objects; object types may be related to each other in a single-inheritance hierarchy; a sufficient set of basic types is already defined (in accordance with the primitive data types and data structures of programming languages, i.e. integer, real, boolean, string, array, list, structure); the developer can extend these types and define an arbitrarily rich type system.

In UIMA parlance, NLP components are called *annotators*. One annotator is combined with other annotators in a document processing *flow*. During runtime, each annotator processes one CAS at a time; the CAS contains the document text along with annotations produced by preceding annotators (if any). The annotator examines the document text and/or the available annotations and produces new annotations; it then adds the new annotations to the CAS and returns it to the UIMA runtime environment so as to be delivered to the next annotator in the flow.

In the following subsections we present the flow of processing towards the acquisition of fully annotated documents. Subsections 5.1 through 5.4 describe the production of "basic annotations". Subsections 5.5 and 5.6 describe the production of "advanced annotations", from components that are currently under development.

5.1 Document conversion

As already mentioned in Section 4, the collected documents were either in html or in pdf format. 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. For example, the xpath expression `HTML/BODY//P//text()` selects all the text nodes of all paragraphs found in the body of an html document; the xpath expression `HTML/BODY//TD[1][@width='620px']//text()` selects all the text nodes of the first table cell that has a width of 620 pixels and is found in the body of an html document; the xpath expression `HTML/HEAD/TITLE/text()` selects the text found in the title of an html document. The xpath expressions are closely related to the internal structure of the html documents. As documents downloaded from the same site have (more or less) the same internal structure, we grouped the html documents by origin and defined xpath expressions for each group: a) expressions for textual content extraction and b) expressions for global metadata extraction (e.g. title, author, creation date).

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. Problem (a), caused by 8-bit character sets and proprietary typographic fonts, was solved with the application of ad hoc character conversion filters. Problem (b) is inherent in pdf documents: each line of text constitutes a separate text chunk; the continuity of text chunks is apprehensible to the human eye when seen on the page, but for the computer these are just separate text chunks (accompanied by page positioning coordinates and formatting instructions). Problem (b) was solved (not in its entirety) with the application of heuristics that aim to remove the unwanted newline characters, e.g. replace the newline character with the space character when found between two lines where the first line ends with a lowercase word and the second line also starts with a lowercase word.

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

5.2 Tokenisation and sentence splitting

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 morphology of each token is recorded. By "token morphology" 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 morphology 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.

Upon receipt of a CAS, the tokeniser accesses the document text stored in the CAS and performs the processing of steps one and two described above. It then augments the CAS with annotations that encode the begin-offsets, the end-offsets and the morphology of the tokens. Next, the CAS passes to the sentence splitter, which examines the token annotations and augments the CAS with annotations that encode the begin- and end-offsets of sentences.

5.3 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 Orphanos and Christodoulakis, 1999), achieving an accuracy of ninety-seven percent in part-of-speech disambiguation and eighty-nine percent in part-of-speech guessing.

Upon receipt of a CAS, the morphosyntactic annotator iterates through the token annotations produced by the tokeniser and focuses on Greek-letter tokens. For each such token, it produces annotations that encode the morphological lemma (if found), the part-of-speech and the rest morphosyntactic attributes (if any).

5.4 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 subsection 6.1.

The functionality of the biomedical word identification module is rather simple. It iterates through the token annotations of a CAS and, for each Greek-word token found in the gazetteer, produces an extra annotation denoting that this token is a biomedical word.

5.5 Multi-word term recognition

Towards the recognition of multi-word biomedical terms, the following tasks were carried out:

- We collected a sufficient body of multi-word terms from MeSH-Hellas (IATROTEK, 1997), a Greek-English and English-Greek dictionary of biomedical terminology.
- We classified the collected multi-word terms according to their structure and defined draft phrase-structure rules.
- We specified a unification grammar formalism for the formal expression of phrase-structure rules (for more on unification grammars, see Shieber, 1986). The morphosyntactic agreement constraints were expressed with the help of typed feature structures. For example, the rule

  ```
  Candidate_Biomedical_Term(Gender1, Number1, Case1) =>
     Biomedical_Noun(Gender1, Number1, Case1),
     Biomedical_Noun(Gender2, Number2 = singular, Case2 = genitive)
  ```

  describes the formation of biomedical terms like όζος θυρεοειδούς (thyroid nodule), νεοπλάσματα δέρματος (skin neoplasms), *etc*.
- We constructed a parser generator based on ANTLR. The parser generator takes as input a unification grammar (written according to the already specified formalism) and produces a parsing model that encodes the grammar rules and the actions to be taken upon rule application. The actual parsing is performed by an execution engine, which loads the parsing model 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 Knight, 1989).

Given a CAS, the parser iterates through morphosyntactic annotations in the neighbourhood of biomedical words and tests the applicability of phrase-structure rules; for every successful rule application, it produces an annotation that encodes the

constituent words (by reference) and the morphosyntactic attributes of the identified phrase. All phrases identified with this process are candidate multi-word terms.

5.6 Ontology-based semantic tagging

According to Kiryakov *et al*., 2003, 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.

6 Application examples

6.1 Biomedical vocabulary collector

When developing tools that aim to process domain-specific texts, the bottom-line requirement is to possess the vocabulary of the domain. When IATROLEXI started, we had ~5,000 words in the Morphological Lexicon that were marked to belong to the biomedical domain. Of course, a biomedical corpus of 11.5 million words is a very good source of biomedical words. To be exact, the corpus contains 11.5 million word-forms (i.e. morphological variations of words), thus it is a very good source of biomedical word-forms.

The corpus-based biomedical vocabulary collection is based on a simple hypothesis: if a word-form found in a biomedical text is unknown to a broad coverage lexicon (such as our Morphological Lexicon), then there are good chances to be a biomedical word-form. Other chances are to be a misspelled word-form or a word-form that does not belong to the biomedical domain, e.g. a person or company name. Filtering out word-forms of low frequency (measured in the entire corpus) is a good heuristic to get rid of misspelled word-forms. Filtering out first-capital or all-capital word-forms is a good heuristic to get rid of entity names.

Implementing the above ideas in a software module that works on the output of the tokeniser, we collected more than 25,000 unknown word-forms. After having examined a sample of 6,000 word-forms, we found out that 5,220 word-forms (eighty-seven percent) belong to the biomedical domain; these word-forms correspond to 3,551 biomedical words, of which 64 percent are nouns, 35 percent are adjectives, 0.6

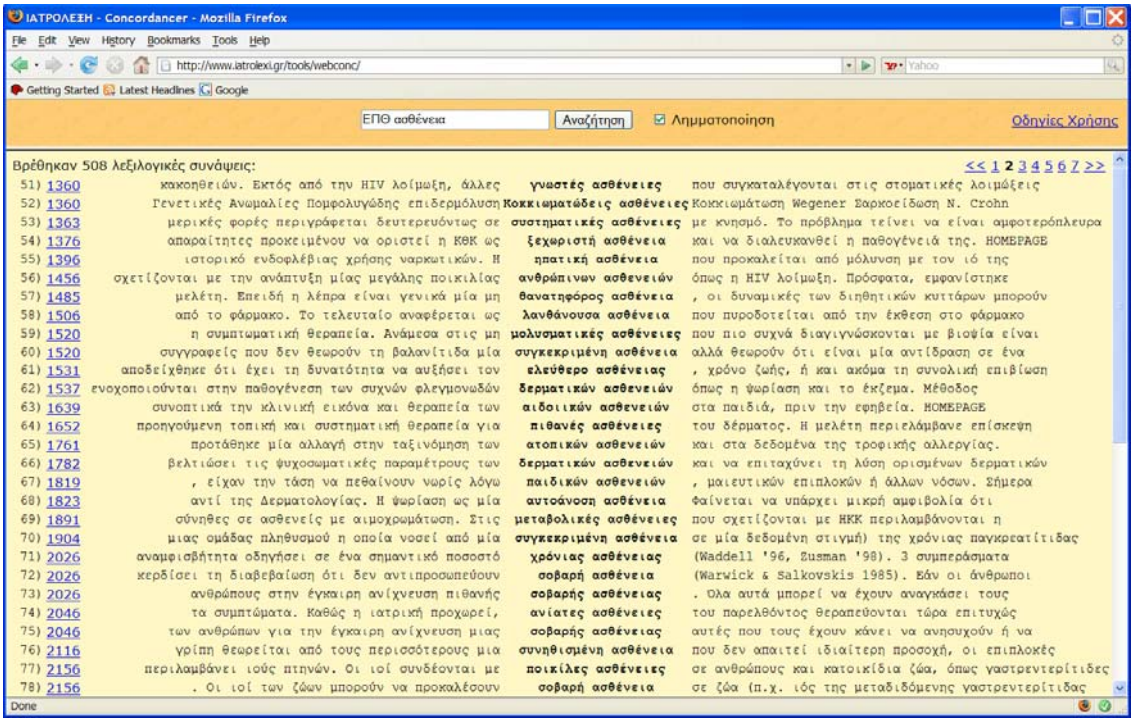
percent are participles and 0.4 percent are verbs. The non-biomedical word-forms of the sample (the rest thirteen percent) are:

a) Word-forms that follow the inflectional system and the orthography of Katharevousa, a fabricated form of the Greek language that was to be the midpoint between Ancient and Modern Greek. As Katharevousa was the official Greek language up to 1976, it is not surprising to be still in use, especially in scientific texts.
b) Misspelled word-forms. These are due to the fact that we selected a frequency cut-off threshold of tree whereas the average frequency of misspelled word-forms (measured in the sample) is six.
c) Nonsense words, e.g. words that miss their initial or final part. These are mainly due to document conversion errors (mostly from pdf-to-txt conversions).

6.2 Concordancer

The task of developing a concordancer was very high in the agenda of IATROLEXI, due to its importance in the advancement of research within the project but also because we believe that this tool is perhaps the most comprehensible demonstrator of the corpus usefulness.

The concordancer we developed uses the Apache Lucene engine to index and search annotated documents produced by the morphosyntactic annotator. At indexing time, for each Greek-word token found in a document we also index its morphological lemma and its morphosyntactic attributes. This way, we can search by lemma and thus retrieve concordance lines with all morphological variations of a word. We can also search by part-of-speech/gender/number/*etc*. and thus retrieve concordance lines of word classes, e.g. of nouns that are in genitive.



Picture 1: The IATROLEXI Concordancer

The concordancer is a web application accessible at http://www.iatrolexi.gr/tools/webconc/. Picture 1 shows a screen-dump of the concordancer. The query we posed is ΕΠΘ ασθένεια, which means that we wanted to find concordance lines where the word ασθένεια (sickness) is preceded by an ΕΠΘ (ΕΠίΘετο – adjective). The checkbox Λημματοποίηση (lemmatisation) instructs the concordancer to search for all word-forms.

7 Work in progress

We presented various aspects of the work done up to now in the context of IATROLEXI for the development of an annotated biomedical corpus.

Currently, a part of our efforts focuses on the completion of the multi-word term recogniser. In subsection 5.5 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 Fratzi and Ananiadou, 1999).

In parallel, we are developing a bilingual biomedical dictionary, by aligning Greek biomedical terms (collected from biomedical dictionaries and from the corpus) with American biomedical terms found in the UMLS. So far, we have coded a critical mass of 17,000 terms. Mapping a Greek term to a UMLS term is very important, as through UMLS we gain access to other very significant pieces of information about the term (of course in English): its classification (semantic type) in the Semantic Network, its definition, its synonyms and its relations with other terms from over 100 vocabularies contained in the Metathesaurus. We envisage (at least) three applications of the bilingual biomedical dictionary:

a) 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.
b) Bilingual term searching. A Greek term can be translated to its American equivalent(s) and then searched in American texts, and vice-versa.
c) 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

Bowker, L. (1996) ‘Towards a corpus-based approach to terminography’. *Terminology*, vol. 3, pp. 27–52.

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.*

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, 145–79.

Friedbichler I. and M. Friedbichler (1997) The potential of domain-specific target language corpora for the translator's workbench. *1st International Conference on Corpus Use and Learning to Translate*, *Bertinoro, Italy*.

IATROTEK (1997) *Ελληνοαγγλικό και Αγγλοελληνικό Λεξικό Βιοϊατρικών Όρων (MeSH Hellas)*. Εταιρία Ιατρικών Σπουδών.

Kim J.-D., T. Ohta, Y. Tateisi and J. Tsujii (2003) 'GENIA Corpus - a Semantically Annotated Corpus for Bio-textmining'. *Bionformatics*, vol. 19, Suppl. 1, pp. 1180–82.

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

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

Kokkinakis D. (2006) Developing resources for Swedish Bio-Medical text mining. *2nd International Symposium on Semantic Mining in Biomedicine, Jena, Germany*.

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

Shatkay, H. and R. Feldman. (2003) 'Mining the biomedical literature in the genomic era: an overview'. *Journal of Computational Biology*, 10(6), pp. 821–55.

Shieber, S. M. (1986) *An Introduction to Unification-Based Approaches to Grammar*. University of Chicago Press, Chicago.

Sinclair J.M. (1997) *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Smith L., T. Rindflesch and W. Wilbur (2005) 'The importance of the lexicon in tagging biological text'. *Natural Language Engineering*, vol. 10.

Teufel S. and N. Elhadad (2002) Collection and Linguistic processing of large scale corpus of medical articles. *LREC 2002, Las Palmas, Canary Islands, Spain*.

Zweigenbaum P., P. Jacquemart, N. Grabar and B. Haber (2001) Building a Text Corpus for Representing the Variety of Medical Language. *10th World Congress on Medical Informatics, Medinfo 2001, London, UK*.