Development of a Greek biomedical corpus

Mavina Pantazara[2], Elena Mantzari[2], Aristides Vagelatos[1], Chryssoula Kalamara[3], Christos Tsalidis[2], Anna Iordanidou[4]

RACTI[1], Neurosoft S.A.[2], Athens' Euroclinic[3], University of Patras[4]
vagelat@cti.gr
{mavina, mantzari, tsalidis}@neurosoft.gr
anna_ior@hol.gr

Abstract

Collection and annotation of specialized corpora, for less-spoken languages such as Greek, is 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 that took place in the framework of the national R&D project "IATROLEXI" (*http://www.iatrolexi.gr*). 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 natural language processing (NLP) applications, i.e. information extraction, data mining, etc., in the domain of biomedicine. The project will build 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 lemmatizer and a morphosyntactic tagger, and it will further develop new resources such as a specialised corpus of biomedical texts that is presented in this paper and an ontology of medical terminology.

Keywords: Corpus Linguistics, NLP, biomedical corpus, 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 [Eysenbach (2003)] 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. 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. 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 (*has been approved for partially funding by General Secretariat of Research & Development, within Measure 3.3 of "Information Society" Operational Program*) 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 will be constructed for the Greek language, which will allow better management and processing of the digitally encoded information in the biomedical field. This will be made possible through the compilation of a representative specialised corpus of biomedical texts and the construction of NLP tools for morphosyntactic and semantic annotation of those texts. In this paper, the design and compilation of the Greek biomedical corpus is presented.

The paper is structured as follows: section 2 presents some background information on design and encoding issues for developing text corpora; section 3 gives a brief overview of the biomedical corpora developed so far for NLP applications. In section 4 the present state of the Greek biomedical corpus compiled for the project IATROLEXI is presented. Finally, in section 5 the conclusions are given.

2. Design and encoding issues for text corpora

Over the last years, terminology studies are mostly based on text corpora for term recognition and extraction, as well as for the development of computational tools for terminology processing, documentation and evaluation. A modern corpus is "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" [Sinclair (2005)]. The four main characteristics of the modern corpus are: a) sampling and representativeness, b) finite size, c) machine-readable form, and d) a standard reference [McEnery & Wilson (1996)]. Among the different types of text corpora (general vs. specialized, synchronic vs. diachronic, annotated vs. unannotated, monolingual vs. multilingual, etc.), IATROLEXI aims at building a *monolingual* (Greek) *specialized* (biomedical) *annotated* corpus.

2.1 Design principles for specialized corpora

The guiding principles for transforming a collection of texts into a corpus are not strictly definable, but rely heavily on the good sense of the designers and the feedback from a consensus of users. A well-designed corpus, however, should satisfy a number of conditions: a) conditions of acceptability: *representativeness* of the research purpose, *coherence* of its internal structure, *homogeneity* according to the selection criteria, and b) conditions of significance: *variety* of language uses and text types,

balance between text types and genres, *coverage* resulting from the size of both the samples and the total corpus.

Typically, design specifications for specialized corpora take account of the following ([Bowker (1996)], [Friedbicher & Friedbicher (1997)]): a) *the types and genres of texts* (i.e. specialized corpora must include scientific texts, educational texts as well as popularized articles), b) *the number of words per text* (i.e. it is highly recommended that specialized 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 specialized corpus).

2.2 Text encoding and annotation

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. Information which is of direct relevance to the encoding of text corpora may be of the following categories:
a) *Documentation*: global information about the text, its content, and its encoding (i.e. bibliographic description, documentation of character sets and entities, description of encoding conventions, etc.).
b) Structures of primary data:

- text elements appearing down to the level of paragraph, which is the smallest unit that can be identified language-independently (i.e. volume, chapter, footnotes, titles, headings, tables, figures, etc.),
- text elements appearing at the sub-paragraph level which are usually signalled by typography in the text and which are language dependent (i.e. quotations, orthographic words, abbreviations, names, dates, highlighted words, etc.),

c) Linguistic annotation: information enriching the text with the results of linguistic analyses (i.e. morphological information, part of speech information, parser output, alignment of parallel texts, prosody markup, etc.).

Since the early 1990s, several projects have worked on issues concerning standardization of the representation and annotation (endoding) of LRs, 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: the Text Encoding Initiative (TEI - http://www.tei-c.org), the Corpus Encoding Standard (CES and XCES - http://www.xml-ces.org), and MATE/NITE for the *representation of primary data and annotations*; EAGLES (http://www.ilc.cnr.it/EAGLES96/home.html) for *annotation content*; RDF/OWL (http://www.w3.org/2004/OWL/) and Topic Maps for *knowledge structures*; Dublin Core and the Open Archives Initiative (OAI) for *general metadata*; MPEG7, IMDI, and OLAC for *domain-specific metadata*; and MULTEXT (http://www.lpl.univ-aix.fr/projects/multext), Edinburgh's LT framework, TIPSTER, GATE, and ATLAS (http://www.nist.gov/speech/atlas/) for general architecture.

In late 2002, within ISO/TC 37/SC4, the Working Group 1 has been established to develop a Linguistic Annotation Framework (LAF) that can serve as a basis for harmonising existing LRs or creating new ones. Its design is intended to allow for, maximum flexibility and maximum processing efficiency and reusability, by separating annotation formats from the exchange/processing format [Ide & Romary (2003)].

3. Background - Biomedical corpora

NLP applications are becoming increasingly recognized as central to medical informatics and even more so to bioinformatics. Several R&D projects for biomedical language processing have worked on the collection and linguistic annotation of biomedical corpora (see among others, [Zweigenbaum (2001)], [Teufel & Elhadad (2002)], [Kokkinakis (2006)], [Smith et al. (2005)]). Moreover, Cohen [Cohen et al. (2005)] make an evaluation of six, publicly available, biomedical corpora for English (i.e. 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.

4. IATROLEXI biomedical corpus: collecting the data

The collection criteria of the texts were originally imposed by the specific requirements of the project. According to those, the corpus should comprise only by written texts, which they should be collected in a short time period (less than six months). Due to time constraints, sourcing texts from Internet was proved to be less time consuming.

The development of the IATROLEXI biomedical corpus was designed to be accomplished through the following steps:
a) Identification of the resources that fall into the project’s scope, evaluation of their content and downloading of the appropriate ones.

b) Classification of the collected text.
c) Development of the appropriate encoding scheme for corpus annotation.
d) Implementation of the appropriate software (chunker, tokenizer, etc.)
e) Corpus Annotation.

Up to now the first two steps have already been completed, whilst the encoding scheme is under development.

4.1 Identification of text sources

The Greek Internet health directories 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 full access, etc.).

Overall, forty internet sites were identified to contain appropriate medical documents for IATROLEXI. So far, the total number of documents is touching 6,250. In the following table information is illustrated on the websites from which most of the text documents were fetched, as well as the number of documents per each source:

Table 1: *Greek Internet sites that identified to contain useful biomedical texts*

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 Pediatrics) (http://www.paediatriki.gr/)	209
Δελτίον Α' Παιδιατρικής Κλινικής Πανεπιστημίου Αθηνών (A' Pediatric	115

Clinic of Athens' Un. Bulletin) (http://www.iatrikionline.gr)	
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

4.2 Document classification

The biomedical texts 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 by the medical doctors manually, based on medical specialties. Most of the documents were already classified according to this scheme, due to their origin: they were collected from a periodical or a web site of a certain medical society. The rest were classified by the medical experts of the team. The result is illustrated in the following table:

Table 2: *Topic classification and number of documents per category*

Topic	**n.**	**Topic**	**n.**
Allergy	4	Neurology	78
Anesthesiology	12	Neurosurgery	104
Cardiology	454	Ophthalmology	137
Cytology	4	Orthopedics	162

Dermatology	1265	Otorinolaringology	231
Endocrinology	29	Pathologoanatomy	612
Forensic Medicine	2	Pediatrics	324
Gastroenterology	143	Pneumonology	525
General Medicine	4	Psychiatry	26
Genetics	4	Radiology	341
Gynecology - Obstetrics	403	Rheumatology	15
Haematology	20	Social Medicine	14
Medical Issues (in general)	810	Surgery	283
Microbiology	19	Urology	163
Nephrology	14		

4.3 Type and format of the documents

The medical documents that were collected for IATROLEXI corpus were paper abstracts, full papers, conference proceedings, and documents with more than one paper in the same file. Most of them, apart of the body text, contained additional information like images, tables, graphical representations, etc. Moreover, part of the corpus contained also some English text (mostly, the abstract in English) which may help in a future construction of some kind of parallel text. Overall 6,276 documents were collected, from which the 69,8% was in hypertext markup language (.htm) while the rest (30,2%) was in Portable Document Format (.pdf). At next stage the original format will be converted into XML.

5. Next steps – Conclusions

Being at the stage of selecting the annotation scheme, an extensive bibliographic research was made in order to figure out the most appropriate scheme for IATROLEXI corpus. The formalism that seems to be the most appropriate is TEI lite [Burnard et al (2002)]. Based on that formalism, a proper scheme will be developed for IATROLEXI corpus.

Next, in order to fully annotate the collected medical corpus with the selected annotation scheme in an automatic way, certain templates should be designed and a number of utilities will be developed to analyze and appropriately annotate the existing documents. Additionally a suitable database must be selected to store the annotation as well as the original documents, thus allowing for further processing of the corpus.

References

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

Burnard L., Sperberg-McQueen C. (2002), *TEI Lite: An Introduction to Text Encoding for Interchange*, http://www.tei-c.org/Lite/teiu5_en.pdf (accessed J. '07).

Cohen B., Fox L., Ogren P., Hunter L. (2005), *Corpus Design for biomedical natural language processing*, ACL-ISMB workshop on Linking Biological Literature Ontologies and Databases.

Eysenbach G. (2003): *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, pp. 194 -212.

Friedbicher I., Friedbicher M. (1997), *The potential of domain-specific target language corpora for the translator's workbench*, available from: http://www.sslmit.unibo.it/cultpaps/fried.htm (accessed J. '07).

Ide N., Romary L. (2003), *Outline of the International Standard Linguistic Annotation Framework,* in Proceedings of the ACL 2003 - Workshop on Linguistic Annotation: Getting the Model Right, pp. 1-5.

Kim J.-D., Ohta T., Tateisi Y. and Tsujii J. (2003), *GENIA Corpus - a Semantically Annotated Corpus for Bio-textmining,* Bionformatics, vol. 19, Suppl. 1, pp. 1180-1182.

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

McEnery, T., Wilson, A. (1996), *Corpus Linguistics*, Edinburgh, Edinburgh University Press.

Sinclair J., Ball J. (1996), *EAGLES Text typology*.

Sinclair J. (2005), *Corpus and Text - Basic Principles*, in Developing Linguistic Corpora: a Guide to Good Practice, Oxford, Oxbow Books, pp. 1-16.

Smith L., Rindflesch T., Wilbur W. (2005), *The importance of the lexicon in tagging biological text*, Natural Language Engineering, vol. 10.

Teufel S., Elhadad N. (2002), *Collection and Linguistic processing of large scale corpus of medical articles*, in Proceedings of the Language Resources and Evaluation Conference.

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