From medical language processing to BioNLP domain

Gabriella Pardelli*, Manuela Sassi*, Sara Goggi*, Stefania Biagioni**
* Istituto di Linguistica Computazionale "Antonio Zampolli"
** Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo"
Via Giuseppe Moruzzi 1 – 56124 Pisa, Italy
gabriella.pardelli@ilc.cnr.it, manuela.sassi@ilc.cnr.it
sara.goggi@ilc.cnr.it, stefania.biagioni@isti.cnr.it

Abstract

This paper presents the results of a terminological work on a reference corpus in the domain of Biomedicine. In particular, the research tends to analyse the use of certain terms in Biomedicine in order to verify their change over the time with the aim of retrieving from the net the very essence of documentation. The terminological sample contains words used in BioNLP and biomedicine and identifies which terms are passing from scientific publications to the daily press and which are rather reserved to scientific production.
The final scope of this work is to determine how scientific dissemination to an ever larger part of the society enables a public of common citizens to approach communication on biomedical research and development; and its main source is a reference corpus made up of three main repositories from which information related to BioNLP and Biomedicine is extracted.
This study is divided in three sections: 1) an introduction dedicated to data extracted from scientific documentation; 2) the second section devoted to methodology and data description; 3) the third part containing a statistical representation of terms extracted from the archive: indexes and concordances allow to reflect on the use of certain terms in this field and give possible keys for having access to the extraction of knowledge in the digital era.

Keywords: terminology; biomedicine; natural language processing

1. Introduction

Since the advent of the www technologies in the '90s, computing has had a strong impact on modern society offering new opportunities of expansion for future research. In these years the Internet has evolved to such an extent that the important changes in the field of acquisition, storage and transmission of data have provided new resources to the web society, satisfying its needs for constantly updated information. Such information can appear in journals published online, either started as electronic publications or made available in electronic format only afterwards. New forms of publication have emerged, providing the scientific community with free and easy access to research studies and establishing a direct relation between producers and consumers who share knowledge on the web.
Different terms used to communicate the same idea can generate linguistic ambiguity, since the same word or phrase can allow for more than one interpretation, thus affecting the information retrieval process. It follows that the queries which are made through these linguistic variations do not always obtain the response looked for and large amounts of information, although available, do not emerge from the web because the term is not present in the document requested. Access to the semantic contents of a document can become extremely difficult in the case of polysemy (when a word has two or more similar meanings) or of synonymy (when a word means the same as another word). Computer science is increasingly and progressively permeating every area of human life and human sciences: since some years the technological advancements in computer science find applications in many scientific fields thus creating new interdisciplinary domains.
At the intersection of information, computation and biological sciences, there exists an interdisciplinary research field that has been growing steadily over the last decade [Prosdocimi et al. 2009]. As a matter of fact, biology, medicine and computer science combine for shaping new research fields such as Biocomputing, Bioinformatics, Biotechnology, Biomedical Informatics, Medical Informatics. The need to name these new-born domains leads to the use of neologisms and compound terms.
With regard to this matter Dardano says: "Il progressivo diffondersi dei composti è uno dei percorsi più fruttuosi dell'evoluzione delle lingue romanze" [Dardano 2009].
In Italian the compound terms are commonly widespread, as for example the words with the prefix *bio-*: also the Treccani Encyclopaedia notes that "the invasion of compound terms with bio- ("bio-industry", "bio-recipe", etc.) is a phenomenon to be reported" [Treccani].
The Osservatorio Neologico della Lingua Italiana (Onli) is a research project by Giovanni Adamo e Valeria Della Valle with the aim of studying lexical innovation in the Italian language (with particular reference to neologisms taken from newspapers) for trying to define current trends in their creation. Here is what they report about their language in 2003: "One aspect worth mentioning of the semantic evolution of some Italian forms is the new value given to certain affixes. Our databank demonstrates, for example, that in the last few decades a new semantic value has developed for the prefix bio-, as a result of the development of biotechnology" [Adamo et al 2003].
From Greek *bios* (life), the prefix *bio-* is used for indicating the natural elements and generates several combining forms in many different languages: the most significant example is given by the English *BioNLP*, built by combining the adjective bio- with the acronym NLP (Natural Language Processing). The following are recent

definitions of the compound form:
"BioNLP recently emerged from the combined expertise of molecular biology and computational linguistics" [Van Landenghem et al. 2009];
"BioNLP has developed its own characteristics to process the domain language of biology and medicine" [Prince et al. 2009];
"BioNLP is the branch of computational linguistics developing tools and algorithms tailored to the life sciences domain"[Engelken 2009].
The combination of terms such as computer, computational, natural language processing with terms of the biomedical domain might seem an improbable union but it is actually witnessed by the existence of an area of interest between the processing of medical information and the analysis and processing of language. This terminological combination arose only some decades ago while in the past expressions such as *Automated Processing of Medical Language* (1969), *Computational Linguistics in Medicine* (1977), *Computers and medical language* (1979), *Medical Language Processing: Computer Management of Narrative Data* (1987) were used when referring to developing techniques for analysing and processing the natural language form of medical information.
To this respect, here is what Pratt and Milos say in 1969 about the pertinence of linguistic tools to the medicine domain: "Several noted scientists, such as Bar-Hillel, have expressed a pessimistic view in regard to practical implementation of machine translation. Nevertheless, there is merit in continuing efforts for more fundamental research in the area of formal and applied linguistics and computer applications. Even if we are not able to resolve all the problems in language processing at once, limited goals can be attained and tested for validity by design of a model for language processing within a restricted language domain, such as medicine" [Pratt et al. 1969].
Over the last decades the development of NLP techniques allowed to retrieve specialistic knowledge from repositories like Medline and PubMed which are archives specifically dedicated to the fields of medicine, biomedicine and molecular biology. The scientific production on this topic progressively increased, the research ranging from information retrieval to knowledge extraction, knowledge classification, taxonomy, text mining, etc. [Prince et al. 2009].

2. Methods

The source of our analysis for this terminological study is constituted by two institutional repositories and one newspaper corpus. Subject-based information are related to BioNLP and Biomedicine:
1) first of all the corpus called ALPAC, created using the DBT software (Textual Data-Base, CNR-ILC patent) and containing 14,800 titles of articles presented at international conferences (mainly from ACL Anthology and LRECs, implemented by other titles from conferences of the NLP field) as well as data coming from the bibliographical analysis of the early '60s issues of the *Computers and the Humanities* Journal (now *Language Resources and Evaluation*) [Pardelli et al. 2004 and 2006]. From this corpus containing 175,050 words (titles and authors only), the terms with the prefix *bio/bio-* have been extracted and the results are shown in Figure 1: in the period 1969-2008 (2010 is only LREC), it can be observed, for instance, that till the first years of the XXIst century the term *medical* was the most employed while from 2003 onwards the term *biomedical* progressively became the most used.

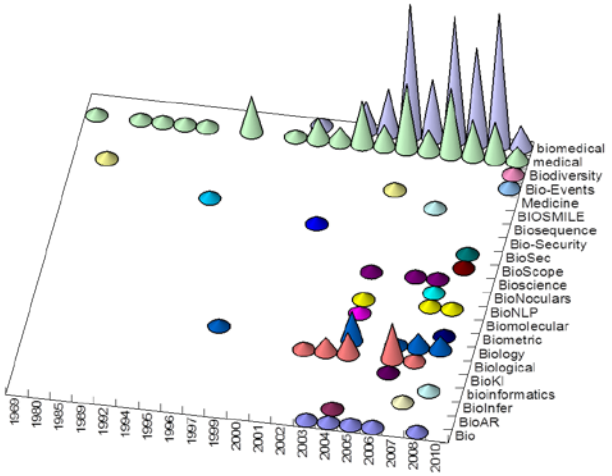

Figure 1

2) The second set of sample terms is extracted from the scientific publications of the biomedical domain which can be found in the Publication Management system (PuMa) of the Institute of Information Science and Technology (ISTI) of the National Research Council of Italy (CNR). PuMa is then a software infrastructure, user-focused and service-oriented, developed by the ISTI Institute: the system functionalities manage, for different collections, the whole life-cycle of different types of documents, from their submission by authors to their dissemination through web access. The most important PUMA feature is its capability to allow stored content to be reusable for different purposes, so that researchers and librarians can manipulate this content to fulfill scientific and administrative issues.
PUMA also constitutes the first step towards creating the Italian network of CNR institutional repositories, looking at the DRIVER vision, i.e., building an infrastructure that allows European research institutions to share content and functionality. It presently manages CNR institutional repositories containing globally about 21000 documents covering different disciplines. Biomedicine repository contains about 3500 items describing documents of different types, i.e. published documents – journal papers, books or book chapters, conference papers etc. – and Grey Literature ones – project deliverables, technical reports, theses and so on. A great part of these documents are Open Access. The textual database of this bibliographical cards (title, abstract and keyword only) contains 732,412 words in total and covers the same period (1970-2010).

This digital archive is a repository indeed rich of compound terms with the prefix *bio-*, among which the most frequent are listed here in Figure 2 [Sassi et al. 2010].

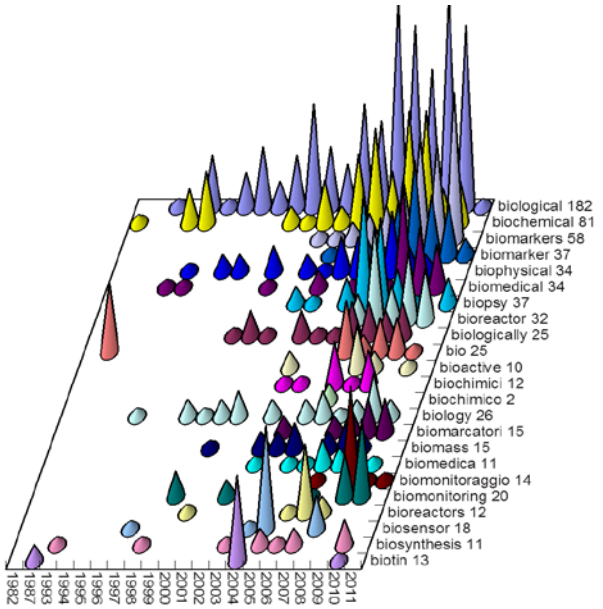

Figure 2

3) Finally, a third sample is constituted by the results of the analysis of three Italian newspapers *(Il Corriere della Sera, la Repubblica, La Stampa*) in the years 1999-2010, from which the articles pertaining to the health domain have been retrieved; then, the analysis concentrated only on words with the prefix *bio-* and, more in depth, on those common words used in scientific communication as well as in articles of Italian newspapers in the given period. This corpus contains 33,683 full-text articles for a total of 14,552,196 words (see Figure 3).

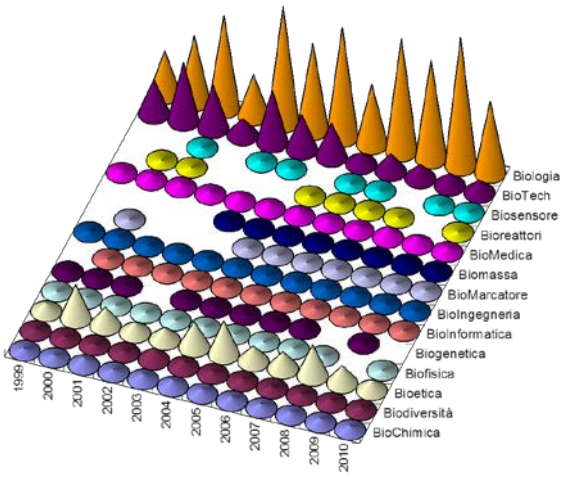

Figure 3

Given the high number of lexical forms, the graph of Figure 3 has been built by grouping the forms by entry/hyperonym (for example, the following forms have been grouped under the entry *biologia*: *biologa, biologhe, biologi, biologia, biologica, biological, biologically, biologicals, biologiche, biologici, biologico, biologie, biologique, biologismo, biologista, biologistica, biologo, biologos, biology*).

The first step of the research focused on the process of indexing the three archives by using the DBT[1] software; in the second phase concordances have been created for the three repositories while the third step consisted in the extraction of words with the prefix *bio* (with the respective date). Those indexes have then been manually checked and words from the second corpus (PuMa) have been chosen and collected for allowing a graphical elaboration; in many cases, English terms and the corresponding Italian translations have been grouped under the same concept.

The overall set of sample words with the prefix *bio* contains names of domain areas, acronyms of application systems, names of research projects and initiatives in the BioNLP field, in the medicine field and also words extracted from the Italian press related to the biomedical domain.

Figure 4 reproduces the result of processing the PuMa corpus: the chosen keyword for searching the scientific documentation of the archive is *BioMarker/BioMarcatore*.

Figure 4

Figure 5 is the graphical visualization of the relations between terms and clusters of meaning, as a result of the above mentioned search within the PuMa corpus.

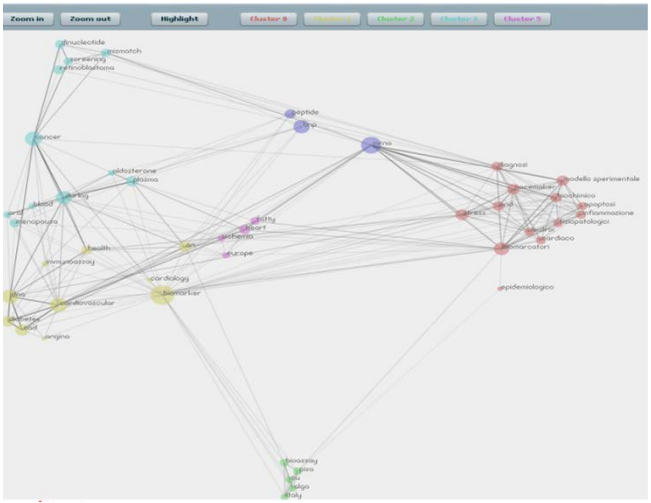

Figure 5

Instead in Figure 6 are shown the results of the same search on the corpus of Italian newspapers.

[1] Data Base Testuale software by Eugenio Picchi, CNR patent

Figure 6

3. Results

Nowadays research in this field make use of a terminology characterized by words formed by prefixes or acronyms.

The graph of Figure 1 compares the overall use of single words and acronyms employed and coined in the period 1969-2008 in the biomedical domain by means of the analysis of titles of the ALPAC corpus which contains ACL Anthology and LREC articles (the latter pertaining only to the years 1998-2010). In this time span the adjective *medical* remains the leading one while since 1999 the adjective *biomedical* begins to be spread in the lexicon of the field as well, with high peaks in the years 2008-1010; while the substantive *biology* is used only from 1996 onwards.

The origin of acronyms in the ALPAC repository is witnessed since 1995 with the use of *BioNoculars* and from 1997 onwards there is a steady creation of words with the prefix *bio*. Some examples follow:

BioSMILE / Adapting Semantic Role Labeling for Biomedical Verbs; using a maximum entropy model with automatically generated template features;

BioSequence / *Bio::Sequence* objects represent annotated sequences in bioruby. A Bio::Sequence object is a wrapper around the actual sequence, represented as either a Bio::Sequence::NA or AA object;

Bio-Security / Biological surveillance systems in the United States rely most heavily on human medical data for signs of epidemic activity;

BioSec / Biometry and Security/ a name for multimodal database for biometric recognition;

BioScope / annotation for negation, uncertainty and their scope in biomedical texts the name of the BioScope corpus. The corpus consists of three parts, namely medical free texts, biological full papers and biological scientific abstracts;

Bioscience / The late 1990s saw the beginning of a trend towards significant growth in the area of biomedical language processing, and in particular in the use of natural language processing techniques in the molecular biology and related computational bioscience domains;

BioNoculars / a statistical unsupervised system, called BioNoculars, for extracting protein-protein interactions from biomedical text;

Biomolecular /events are a major category of relationships in the biomedical domain in which, among others, relationships involving non molecular entities such as diseases and static relations such as protein family memberships are also of interest;

Biometric / The term biometrics, in Information Technology, refers to an array of techniques to identify people based on one or more unique behavioural or physiological characteristics;

Biomedical / this adjective has 1470 occurrences in the repository;

Biology / this noun has 810 results in the repository;

Biological / this adjective has 959 occurrences in the repository;

BioKI / *BioKI* Enzymes is a literature navigation system that uses a two-step process;

Bioinformatics / this noun has 543 occurrences in the repository;

BioInfer /Bio Information Extraction Resource /;

BioAR / biomedical anaphora resolution;

The analysis of information extracted from titles of scientific publications stored in the PuMa System and visible in the graph of Figure 2 provide words with the prefix *bio* in the time spam 1982-201. The fact that within PuMa compound words are predominant outlines how the need to name leads to the massive use of the prefix *bio/bio-*.

The following are some relevant examples:

Bioactive / Relating to a substance that has an effect on living tissue;

Biomarker / A specific physical trait used to measure or indicate the effects or progress of a disease or condition;

Biomass / The total mass of living matter within a given unit of environmental area;

Biomonitoring / is the measurement of the body burden of toxic chemical compounds, elements, or their metabolites, in biological substances;

Biopsy / The removal of a sample of tissue for purposes of diagnosis. (Many definitions of "biopsy" stipulate that the sample of tissue is removed for examination under a microscope.

Bioreactor / A bioreactor is a vessel in which is carried out a chemical process which involves organisms or biochemically active substances derived from such organisms.

Biosensor / is an analytical device for the detection of an analyte that combines a biological component with a physicochemical detector component;

Biosyntesis / the production of a chemical compound by a

living organism;
Biotin / Biotin is a vitamin that is found in small amounts in numerous foods.

4 Concluding remarks

The analysis of documentation related to scientific terminology in the field of Biomedicine is especially relevant because its continuous evolution in the last decades is very remarkable from a terminological point of view, especially from the 1960s when the use of the computer spread over all human activities thanks to the possibility of using the natural language. In these years, the need of naming rapidly grew within the computer science community and the coining of new terms and compound names satisfied this initial need by linking terms to their respective semantic and cognitive value.
Nowadays the need of retrieving the great amount of digital knowledge available on the web is ever more important: the vast majority of this knowledge is conveyed by means of textual material stored in scientific documentary repositories and digital archives. This documentary world preserves inside the wealth of far-off terms belonging to the past which have often fallen out of use as well as a more recent terminological production derived from the feverish need of coining new terms, a very common trend in certain branches.
The technical-scientific vocabulary represents a significant part of the lexicon of any language and plays an important role in the information exchange between the experts of the field and the users of the Internet community, the new market of knowledge where people producing informative contents can meet (thanks to a common lexicon) those who are actually looking for that specific information. Information dissemination on the World Wide Web contributes to make terminology ever more important: in particular, the use of certain specific terms which become the keys for retrieving information.
From this terminological analysis it can be inferred that a good amount of the set of newly-created words in certain research fields are not passing from the scientific production to the daily press: this is the case of the topics treated by the ACL Anthology as well as by the PuMa repository. It is rather a terminology for a niche - created by and destined to experts only - while the press treats just the more general and comprehensible sense of these terms, especially when they are connected with certain news which emphasize their echo on the media.
Moreover, the terminology used in press is often connected with important ethical themes commonly debated such as *bioethics*, which in newspapers and journals is paired to notions like *assisted reproduction* and *living will*. These themes, although largely debated by the media, do not appear as such in scientific literature but are mentioned just in connection with research, for instance, on *stem cells*.

The words shared by the three repositories analyzed in this study (ALPAC corpus, PuMa system and Italian newspapers corpus) are few: as for the adjectives, only *biomedical* is found in all of them.
This example shows an increasing interdisciplinarity among the fields of biology and medicine, also witnessed by the unprecedented growth of the volume of biomedical literature available on the Web which is going alongside with a growing demand for NLP robust and efficient methods and techniques for processing biomedical language.
The union between the medicine domain and Natural Language Processing dates back to the ‘60s. Since then NLP - on the one side – has progressively become a tool for querying knowledge from a major interdisciplinary domain as biomedicine and, on the other side, it has lately proved to have the capacity to revamp itself with the acquisition of new paradigms and evolve to the extent of creating its own domain: *BioNLP*.

5 References

Heimonen J., Björne J., Salakoski T. (2010). *Reconstruction of semantic relationships from their projections in biomolecular domain*. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010. Uppsala, Sweden. The Association for Computational Linguistics, pages 108-116.

Picchi E., Sassi M., Biagioni S., Giannini S. (2010) *Extending the "Facets" concept by applying NLP tools to catalog records of scientific literature*. In: GL 12 - Twelfth International Conference on Grey Literature : Transparency in Grey Literature, Grey Tech Approaches to High Tech Issues (Praga, 6-7 December 2010). Abstract, pp. 82-87. D.J. Farace, J. Frantzen, GreyNet (eds.). TextRelease.

Sassi M., Pardelli G., Biagioni S., Carlesi C., Goggi S. (2010). *A Digital Archive of Research Papers in Computer Science*. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, Daniel Tapias (eds.), LREC'10 - Seventh International Conference on Language Resources and Evaluation, Proceedings. European Language Resources Association (ELRA), Paris. pp. 1245 – 1248.

Täckström O., Eriksson G., Velupillai S., Dalianis H., Hassel M., Karlgren J. (2010) . *Uncertainty Detection as Approximate Max-Margin Sequence Labelling*. Proceedings of the Fourteenth Conference on Computational Natural Language Learning: Shared Task, Uppsala, Sweden. The Association for Computational Linguistics, pages 84–91.

Dardano M. (2009). *Costruire parole*, Bologna, Il Mulino.

Engelken E. (2009). *A System for Semantic Analysis of Chemical Compound Names.* Proceedings of the ACL-IJCNLP 2009 Student Research Workshop, Suntec, Singapore. ACL and AFNLP. p.36.

Prince V., Roche M. (eds.). (2009). *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration*. Book News Inc, Portland.

Prosdocimi F., Chisham B., Pontelli E., Stoltzfus S. and Julie D. Thompson (2009). *Knowledge Standardization in Evolutionary Biology: The Comparative Data Analysis Ontology*. Evolutionary Biology, Springer Verlag, Berlin , Part 1, Pages 195-214.

Van Landeghem S., Saeys Y., De Baets B., Van de Peer Y. (2009). *Analyzing text in search of bio-molecular events: a high-precision machine learning framework*. Proceedings of the Workshop on BioNLP: Shared Task, pages 128–136,Boulder, Colorado, June 2009. Association for Computational Linguistics. p. 128.

Täckström O., Velupillai S., Hassel M., Eriksson G., Dalianis H., Karlgren J. (2009). *Uncertainty Detection as Approximate Max-Margin Sequence Labelling*. Proceedings of the Workshop on BioNLP: Shared Task, Boulder, Colorado, 2009. The Association for Computational Linguistics, pages 107–110.

György Szarvas G., Veronika Vincze V., Farkas R., Csirik J. (2008). *The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts*. *BioNLP 2008*: Current Trends in Biomedical Natural Language Processing, Columbus, Ohio, USA. The Association for Computational Linguistics, pages 38-45.

Demner-Fushman D., Ananiadou, Sophia K., Cohen B., Pestian J., Tsujii J., Webber B. (2008). *Current trends in biomedical natural language processing: the view from computational linguistics*. BioNLP 2008: Current Trends in Biomedical Natural Language Processing, Columbus, Ohio, USA. The Association for Computational Linguistics, pages iii-xi.

Cohen K.Bretonnel, Demner-Fushman D., Friedman C., Hirschman L., Pestian John P. (2007). *Biological, translational, and clinical language processing*. Proceedings of the Workshop on BioNLP 2007, Biological, Translational, and Clinical Language, Prague, Czech Republic. The Association for Computational Linguistics, pages i-XVI.

Hernández Á., López B., Díaz D., Fernández R., Hernández L. , Caminero J. (2007*). A "person" in the interface: effects on user perceptions of multibiometrics*. Proceedings of the Workshop on Embodied Language Processing, Prague, Czech Republic. The Association for Computational Linguistics, pages 33-40.

Madkour A., Darwish K., Hassan H., Hassan A., Emam O. (2007). *BioNoculars: Extracting Protein-Protein Interactions from Biomedical Text*. Proceedings of the Workshop on BioNLP 2007, Biological, Translational, and Clinical Language, Prague, Czech Republic. The Association for Computational Linguistics, pages 89-96.

Kim Jung-jae , Park Jong C. (2004). *BioAR: Anaphora Resolution for Relating Protein Names to Proteome Database Entries.* ACL 2004.Workshop on Reference Resolution and its Applications. The Association for Computational Linguistics, Page 79.

Madkour A., Darwish K., Hassan H., Hassan A., Emam O. (2007). BioNoculars: *Extracting Protein-Protein Interactions from Biomedical Text*. BioNLP 2007: Biological, translational, and clinical language processing, Prague, Czech Republic. The Association for Computational Linguistics, page 89.

Bergler S., Schuman J., Dubuc J., Lebedev A. (2006). *BioKI: Enzymes—an adaptable system to locate low-frequency information in full-text proteomics articles*. BioNLP'06 Linking Natural Language Processing and Biology: towards Deeper Biological Literature Analysis at HLT-NAACL 06, Proceedings of the Workshop, New York . The Association for Computational Linguistics, pages 91-92.

BioNLP'06 *Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*.(2006). Proceedings of the Workshop. The Association for Computational Linguistics, page iii.

Richard Tzong-Han Tsai, Wen-Chi Chou, Yu-Chun Lin, Cheng-Lung Sung, Wei Ku, Ying-shan Su, Ting-Yi Sung and Wen-Lian Hsu. (2006). *BIOSMILE: Adapting Semantic Role Labeling for Biomedical Verbs: An Exponential Model Coupled with Automatically Generated Template Features*. Proceedings of BioNLP-2006, pages 57-64.

BioNLP'06 *Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis.*(2006). Proceedings of the Workshop. The Association for Computational Linguistics, pages iii-x.

Pardelli G., Sassi M., Goggi S., Orsolini P. (2006). *Natural Language Processing A Terminological and Statistical Approach*. LREC 2006 - 5th International Conference on Language Resources and Evaluation, Proceedings. European Language Resources Association (ELRA), Paris. pp. 2395 - 2398.

Kim Jung-jae , Park Jong C. (2004). *BioAR: Anaphora Resolution for Relating Protein Names to Proteome Database Entries*. ACL 2004.Workshop on Reference Resolution and its Applications. The Association for Computational Linguistics, pages 79-86.

Damianos Laurie E., Bayer S., Chisholm Michael A., Henderson J., Hirschman L., Morgan W., Ubaldino M., Zarrella G., Wilson James M., V, MD, Polyak Marat G. (2004). *MiTAP for SARS Detection.* HLT-NAACL , Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, pages 13-16.

Pardelli G., Sassi M., Goggi S. (2004). *From Weaver to the ALPAC Report*. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, Raquel Silvia (eds.), LREC 2004, Fourth International Conference on Language Resources and Evaluation, Proceedings. European Language Resources Association (ELRA), Paris. Volume VI pp. 2005 - 2004.

Biosec: *Biometry and Security* (2004). Deliverable D6.3: Report on results of first phase usability testing and guidelines for developers.

Adamo G., Della Valle V. (2003). *Neologismi quotidiani. Un dizionario a cavallo del millennio*, (Lessico Intellettuale Europeo, 95), Firenze, Leo S. Olschki Editore. http://www.cnr.it/istituti/FocusByN_eng.html?cds=047&nfocus=1

Pacak M.G. , Dunham G.S. (1979) *Computers and medical language*, Med. Inform. 4, I, 13-27.

Schneider W., Hein A-L S. (eds.) (1977). *Computational Linguistics in Medicine*, North-Holland Publishing Co., New York.

Sager N., Friedman C., Lyman Margaret S. (1987). *Medical Language Processing: Computer Management of Narrative Data*. Reading, MA: Addison-Wesley.

Pratt, W., Pacak Milos G. (1969) *Automated Processing of Medical English* . International Conference on Computational Linguistics COLING 1969: PREPRINT NO. 10. (8).

Bar-Hillel Y. (1964). *Language and Information*. Addison-Wesley, Reading~Massachusetts.

Treccani.it. Enciclopedia italiana. *Neologismi.*

http://bioruby.rubyforge.org/classes/Bio/Sequence.html

http://www.it.utu.fi/BioInfer

http://www.adnkronos.com/IGN/Regioni/Lazio/?id=3.1.2988428030

http://www.treccani.it/magazine/lingua_italiana/parole/delleconomia/biocapitalismo.html