# Mining Biomedical Abstracts: What's in a Term?

Goran Nenadic[1,4], Irena Spasic[2,4], and Sophia Ananiadou[3,4,*]

[1] Department of Computation, UMIST, Manchester M60 1QD, UK
G.Nenadic@co.umist.ac.uk
[2] Department of Chemistry, UMIST, Manchester M60 1QD, UK
I.Spasic@umist.ac.uk
[3] Computer Science, University of Salford, Salford M5 4WT, UK
S.Ananiadou@salford.ac.uk
[4] National Centre for Text Mining, Manchester, UK

**Abstract.** In this paper we present a study of the usage of terminology in the biomedical literature, with the main aim to indicate phenomena that can be helpful for automatic term recognition in the domain. Our analysis is based on the terminology appearing in the Genia corpus. We analyse the usage of biomedical terms and their variants (namely inflectional and orthographic alternatives, terms with prepositions, coordinated terms, etc.), showing the variability and dynamic nature of terms used in biomedical abstracts. Term coordination and terms containing prepositions are analysed in detail. We also show that there is a discrepancy between terms used in the literature and terms listed in controlled dictionaries. In addition, we briefly evaluate the effectiveness of incorporating treatment of different types of term variation into an automatic term recognition system.

## 1 Introduction

Biomedical information is crucial in research: details of clinical and/or basic research and experiments produce priceless resources for further development and applications [16]. The problem is, however, the huge volume of the biomedical literature, which is constantly expanding both in size and thematic coverage. For example, a query "*breast cancer treatment*" submitted to PubMed[1] returned nearly 70,000 abstracts in 2003 compared to 20,000 abstracts back in 2001. It is clear that it is indeed impossible for any domain specialist to manually examine such huge amount of documents.

An additional challenge is rapid change of the biomedical terminology and the diversity of its usage [6]. It is quite common that almost every biomedical text introduces new names and terms. Also, the problem is the extensive terminology variation and use of synonyms [5, 6, 11]. The main source of this "terminological confusion" is that the naming conventions are not completely clear or standardised, although some attempts in this direction are being made. Naming guidelines do exist for some types of biomedical concepts (e.g. the Guidelines for Human Gene Nomenclature [7]). Still, domain experts frequently introduce specific notations, acronyms, ad-hoc and/or in-

* This research has been partially supported by the JISC-funded National Centre for Text Mining (NaCTeM), Manchester, UK.

[1] http://www.ncbi.nlm.nih.gov/PubMed/

K.-Y. Su et al. (Eds.): IJCNLP 2004, LNAI 3248, pp. 797–806, 2005.

novative names for new concepts, which they use either locally (within a document) or within a wider community. Even when an established term exists, authors may prefer – e.g. for traditional reasons – to use alternative names, variants or synonyms.

In this paper we present a detailed analysis of the terminology usage performed mainly on a manually terminologically tagged corpus. We analyse the terminology that is used in the literature, rather than the terminology presented in controlled resources. After presenting the resources that we have used in our work in Section 2, in Section 3 we analyse the usage of "ordinary" term occurrences (i.e. term occurrences involving no structural variation), while in Section 4 we discuss more complex terminological variation (namely coordination and conjunctions of terms, terms with prepositions, acronyms, etc.). We also briefly evaluate the effectiveness of accounting for specific types of term variation in an automatic term recognition (ATR) system, and we conclude by summarising our experiments.

## 2 Resources

New names and terms (e.g. names of genes, proteins, gene products, drugs, relations, reactions, etc.) are introduced in the biomedical scientific vocabulary on a daily basis, and – given the number of names introduced around the world – it is practically impossible to have up-to-date terminologies [6]. Still, there are numerous manually curated terminological resources in the domain: it is estimated that over 280 databases are in use, containing an abundance of nomenclatures and ontologies [4]. Although some cross-references do exist, many problems still remain related to the communication and integration between them.

The characteristics of specific biomedical terminologies have been investigated by many researchers. For example, Ananiadou [1] analysed term formation patterns in immunology, while Maynard and Ananiadou [10] analysed the internal morpho-syntactic properties of multi-word terms in ophthalmology. Ogren and colleagues [13] further considered compositional characteristics of the GO ontology[2] terms.

Previous studies are mainly focused on controlled vocabularies. However, controlled terms can be rarely found as on-the-fly (or "running") terms in domain literature. For example, we analysed a collection of 52,845 Medline abstracts (containing around 8 million words) related to baker's yeast (*S. cerevisiae*) and experimented with locating terms from the GO ontology (around 16,000 entries). Only around 8,000 occurrences corresponding to 739 different GO terms were spotted, with only 392 terms appearing in two or more abstracts[3]. Occurrences of controlled terms are more frequent in full text articles: for example, in a set of 621 articles (around 2 million words) from the Journal of Biomedical Chemistry[4] we have located around 70,000 occurrences with almost 2,500 different GO terms. This discrepancy is mainly due to the fact that abstracts tend to represent a summary using typically new and specific

[2] http://www.geneontology.org/

[3] Many GO ontology terms (i.e. entries) are rather "descriptions" than real terms (e.g. *ligase, forming phosporic ester bonds or oxidoreductase*), and therefore it is unlikely that they would appear in text frequently.

[4] http://www.jbc.org

terms, while full texts additionally relate presented work to existing knowledge using (widely known) controlled terms.

In this paper we focus on the terminology that is used in biomedical abstracts. To conduct the experiments, we have used the Genia resources [14] developed and maintained at the University of Tokyo, which include publicly available[5] manually tagged terminological resources in the domain of biomedicine. The resources consist of an ontology and an annotated corpus, which contains 2,000 abstracts obtained from PubMed by querying the database with the MeSH terms *human*, *blood cells* and *transcription factor*. All term occurrences in the corpus are manually tagged by domain experts, disambiguated and linked to the corresponding nodes of the Genia ontology. Also, "normalised" term forms (typically singular forms) are supplied, but apart from inflectional and some orthographic variations, the "normalisation" does not include other types of variation (e.g. acronyms). However, more complex phenomena (such as term coordinations) are annotated.

A total of 76,592 term occurrences with 29,781 distinct terms have been annotated by the Genia annotators in the version we have analysed. Three quarters of marked terms occur only once and they cover one third of term occurrences, while terms with frequencies of 5 or more cover almost half of all occurrences (see Figure 1 for the distribution).

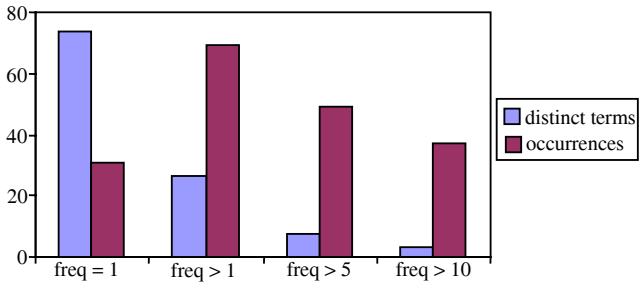


**Fig. 1.** Distributions (in %) of the Genia terms and their occurrences (coverage in the corpus)

## 3 Ordinary Term Occurrences

The vast majority of term occurrences (almost 98%) in the Genia corpus are "ordinary" term occurrences. An *ordinary occurrence* is a term occurrence associated with one term and is represented by a non-interrupted sequence of words (constituents), i.e. an occurrence that does not involve structural variation. Apart from ordinary occurrences, term constituents can be, for example, distributed within term coordination (e.g. *virus or tumor cells* encodes two terms, namely *virus cell* and *tumor cell*) and/or interrupted by acronym definitions (e.g. *progesterone (PR) and estrogen (ER) receptors*). However, only around 2% of Genia term occurrences are non-ordinary occurrences.

Ordinary terms are mostly multi-word units (terms containing at least one "white space"): 85.07% of all Genia terms are compounds, or almost 90% if we consider

[5] http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/

terms with hyphens as multi-words (e.g. *BCR-cross-linking*, *DNA-binding*). The multi-word Genia terms typically contain two or three words (see Figure 2 for the distribution of the term lengths). Terms with more than six words are rare, although they do exist (e.g. *tumor necrosis factor alpha induced NF kappa B transcription*). Such terms are typically hapax legomena in the Genia corpus.

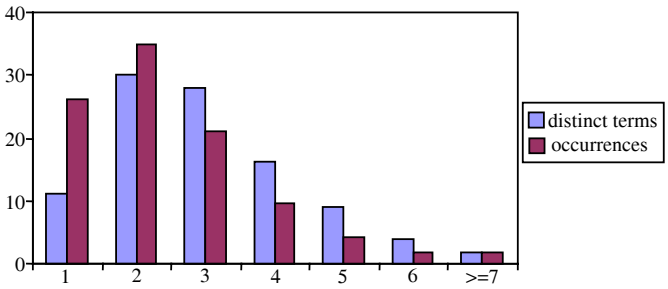


**Fig. 2.** Distributions (in %) of the Genia terms and their occurrences with respect to the length

Apart from using different orthographic styles, a range of specific lexical expressions characterise the common biomedical terminology. For example, neoclassical combining forms (e.g. *NF-kappa B*), adjectival and gerund expressions (e.g. *GTPase-activating protein*), as well as nominalizations and prepositional phrases (e.g. *activation of NF-kappaB by SRC-1*) are frequently used. Many terms in the domain incorporate complex relationships that are represented via nested terms. A nested term is an individual term that may occur within longer terms as well as independently [3, 13]. For example, the term *T cell* is nested within *nuclear factor of activated T cells family protein*. In the Genia corpus, nested terms appear in 18.55% of all term occurrences, with only 8.42% of all distinct Genia terms occurring as nested[6]. Almost a third of all nested terms appear more than once as nested, while more than a half of nested terms do not appear on their own elsewhere in the corpus. These facts suggest the recognition of inner structures of terms cannot rely only on spotting the occurrences of the corresponding sub-terms elsewhere in a corpus.

## 4 Terminological Variation

Terminological variation and usage of synonyms are extremely prolific in biomedicine. Here we discuss two types of term variation: one affecting only term candidate constituents (e.g. different orthographic and inflectional forms) and the other dealing with term structure (prepositional and coordinated terms). We also briefly examine how the integration of term variation into ATR influences the precision and recall performance (Subsection 4.4).

Variations affecting only term constituents are the simplest but the most prolific. For example, in Genia, a third of term occurrences are affected by inflectional variations, and – considering only distinct terms – almost half of the Genia terms had inflectional variants occurring in the corpus (i.e. almost half of occurrences are "nor-

[6] However, 2/3 of GO-ontology terms contain another GO-term as a proper substring [13].

malised" by the experts with respect to inflectional variation). Variations affecting term structure are less frequent, but more complex and ambiguous. Only around 7% of distinct Genia terms are affected exclusively by structural variation. We will examine in turn the most productive of these variations.

### 4.1 Terms Containing Prepositions

Terms containing prepositions are scarce: in the Genia corpus only 0.45% of all terms (or 0.5% of all multi-word terms) is constructed using a preposition[7]. Such terms are also extremely infrequent: 90% of the prepositional Genia terms appear only once in the corpus. The most frequent preposition is *of* (85% of prepositional terms) followed by only three other prepositions (*in*, *for* and *by*, see Table 1). In some cases terms can be "varied" by different prepositions (e.g. *nuclear factor* ***of*** *activated T-cells* and *nuclear factor* ***for*** *activated T cell*), and they can contain several prepositions (e.g. *linker* ***of*** *activation* ***of*** *T-cells*).

**Table 1.** Distribution and examples of the Genia terms containing prepositions

| Preposition | Number of terms | Examples |
|---|---|---|
| *of* | 113 | *promoter* ***of*** *gene* |
| *for* | 9 | *binding site* ***for*** *AP1* |
| *in* | 8 | *increase* ***in*** *proliferation* |
| *by* | 2 | *latency* ***by*** *expression* |

Interestingly, many potential term occurrences containing prepositions have not been marked as terms by the experts, unlike their semantically equivalent occurrences without prepositions. For example, in Genia, *HIV-1 replication* is marked as a term, while *replication of HIV-1* is not; similarly, *level of expression* is never marked as a term as opposed to *expression level*. Only in one case a prepositional term has been marked in an equivalent form without preposition elsewhere in the Genia corpus (*nuclear factor for activated T cell* appeared also as *activated T cell nuclear factor*). This analysis shows that biomedical experts seem to "prefer" nominal term forms, rather than prepositional expressions. Still, a number of terminologically significant expressions contain prepositions (e.g. *activation of PKC*, *NF kappa B activation in T-cells*, *expression of genes*, *production of cytokines*, *binding of NF kappa B*, *activation by NF kappa B*). These expressions – when individually presented to experts – are typically considered as terms. Therefore, the number of terminologically relevant prepositional expressions is much higher than the number of terms marked in the Genia corpus.

Still, the recognition of prepositional term expressions is difficult. Firstly, such expressions are extremely infrequent (for example, in the Genia corpus, only around 200 out of 60,000 preposition occurrences (i.e. 0.33%) have been marked as part of terms). Secondly, there are no clear morpho-syntactic clues that can help differentiate between terminologically relevant and irrelevant prepositional phrases.

---

[7] On the other hand, almost 12% of the GO-ontology terms contain prepositions (e.g. *regulation* ***of*** *R8 fate*), with prepositions frequently appearing in "description" parts (e.g. *oxidoreductase activity, acting* ***on*** *sulfur group* ***of*** *donors*).

### 4.2 Terms Encoded in Coordinations

Term coordination is a multi-word variation phenomenon where a lexical constituent(s) common for two or more terms is shared (appearing only once), while their distinct lexical parts are enumerated and coordinated with a coordination conjunction (CC). Consequently, term coordination encodes at least two terms. Apart from the pragmatic reasons of the language economy, stylistic motivations are also very important for the introduction of coordinations, as authors try to avoid recurrence of shared lexical units [5].

In the Genia corpus, term coordinations have been manually marked and they appear 1,585 times (1,423 distinct coordinations), out of 76,592 term occurrences, which is only 2.07% of all term occurrences. Still, a total of 2,791 terms are involved in coordinations, which makes 9.38% of all distinct Genia terms[8]. However, only one third of coordinated terms appear also as ordinary terms elsewhere in the corpus, which means that even 6.37% of all Genia terms appear exclusively as coordinated (i.e. they do not have any ordinary occurrence in the corpus, and can be extracted only from coordinations).

Coordinations containing conjunction *and* are by far the most frequent (87% of all term coordination occurrences), with *or*-coordinations contributing with more than 10% (see Table 2). Coordinated expressions encode different numbers of terms, but in the majority of cases (85-90%) only two terms are coordinated (see Table 3 for the detailed distributions for *and*- and *or*-coordinations).

In our analysis we distinguish between head coordinations of terms (where term heads are coordinated, e.g. *adrenal glands and gonads*) and argument coordinations (where term arguments (i.e. modifiers) are coordinated, e.g. *B and T cells*). In almost 90% of cases term arguments are coordinated, and as much as 94% of *or*-coordinations are argument coordinations.

**Table 2.** Distribution of term coordinations in the Genia corpus

| CC | Number of occurrences | Examples |
|---|---|---|
| *and* | 1381 (87.07%) | *B-cell expansion* ***and*** *mutation* |
| *or* | 164 (10.34%) | *natural* ***or*** *synthetic ligands* |
| *but not* | 20 (1.26%) | *B-* ***but not*** *T-cell lines* |
| *and/or* | 8 (0.50%) | *cytoplasmic* ***and/or*** *nuclear receptors* |
| *as well as* | 3 (0.19%) | *PMA-* ***as well as*** *calcium-mediated activation* |
| *from to* | 3 (0.19%) | ***from*** *memory* ***to*** *naive T cells* |
| *and not* | 2 (0.12%) | *B* ***and not*** *T cells* |
| *than* | 2 (0.12%) | *neonatal* ***than*** *adult T lymphocytes* |
| *not only but also* | 1 (0.07%) | ***not only*** *PMA-* ***but also*** *TNF-induced HIV enhancer activity* |
| *versus* | 1 (0.07%) | *beta* ***versus*** *alpha globin chain* |

[8] Only 1.4% of the GO-ontology terms contain CCs. However, these nodes mainly represent single concepts, and not coordinations of different terms.

**Table 3.** Number of terms in term coordinations in the Genia corpus

| CC | Number of terms | | | | | |
|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 |
| *and* | 1230<br>89.08% | 101<br>7.31% | 31<br>2.24% | 14<br>1.01% | 4<br>0.29% | 1<br>0.07% |
| *or* | 141<br>85.97% | 19<br>11.59% | 1<br>0.61% | 2<br>1.22% | 0<br>0.00% | 0<br>0.00% |

In order to further analyse the inner structure of coordinations occurring in the Genia corpus, we automatically extracted a set of regular expressions that described the morpho-syntactic patterns used for expressing term coordinations. Although the patterns were highly variable, the simplest ones[9] (such as $(N|A)^+$ CC $(N|A)^*$ $N^+$) covered more than two thirds of term coordination occurrences.

**Table 4.** Ambiguities within coordinated structures

| Example | *adrenal glands and gonads* |
|---|---|
| head coordination | [*adrenal* [*glands* *and* *gonads*]] |
| term conjunction | [*adrenal glands*] *and* [*gonads*] |

Still, the structure of term coordinations is highly ambiguous in many aspects. Firstly, the majority of patterns cover both term coordinations and term conjunctions (where no term constituents are shared, see Table 4), and it is difficult (in particular in the case of head coordinations) to differentiate between the two. Furthermore, term conjunctions are more frequent: in the Genia corpus, term conjunctions appear 3.4 times more frequently than term coordinations.

In addition, some patterns cover both argument and head coordinations, which makes it difficult to extract coordinated constituents (i.e. terms). For example, the above-mentioned pattern describes both *chicken and mouse receptors* (an argument coordination) and *cell differentiation and proliferation* (a head coordination). Of course, this pattern also covers conjunction of terms (e.g. *ligands and target genes*). Therefore, the main problem is that coordination patterns have to be more specific, but there are no reliable morpho-terminological clues indicating genuine term coordinations and their subtypes. In some cases simple inflectional information can be used to identify an argument coordination expression more accurately. For example, head nouns are typically in plural (like in *Jun and Fos families*, or *mRNA and protein levels*), but this is by no means consistent: singular variants can also be found, even within the same abstract (e.g. *Jun and Fos family*, or *mRNA and protein level*, or *RA receptor alpha, beta and gamma*). Also, optional hyphens can be used as additional clues for argument coordinations (e.g. *alpha- and beta-isomorphs*). However, these clues are typically not applicable to head coordinations.

Not only recognition of term coordinations and their subtypes is ambiguous, but also internal boundaries of coordinated terms are blurred. For example, in the coordi-

[9] In these patterns, A and N denote an adjective and a noun respectively, while PCP denotes an *ing*-form of a verb.

nation *glucocorticoid and beta adrenergic receptors* it is not "clear" whether receptors involved are *glucocorticoid receptor* and *beta adrenergic receptor*, or *glucocorticoid adrenergic receptor* and *beta adrenergic receptor*. Furthermore, from *chicken and mouse stimulating factors* (a coordination following pattern $N_1$ *and* $N_2$ PCP $N_3$) one has to "generate" *chicken stimulating factor* (generated pattern $N_1$ PCP $N_3$) and *mouse stimulating factor* (pattern $N_2$ PCP $N_3$), while from *dimerization and DNA binding domains* (the same coordination pattern, $N_1$ *and* $N_2$ PCP $N_3$) terms *dimerization domain* ($N_1$ $N_3$) and *DNA binding domain* ($N_2$ PCP $N_3$) have to be extracted.

Therefore, we can conclude that significant background knowledge needs to be used to correctly interpret and decode term coordinations, and that morpho-syntactic features are not sufficient neither for the successful recognition of coordinations nor for the extraction of coordinated terms.

### 4.3 Terms and Acronyms

Acronyms are a very common term variation phenomenon as biomedical terms often appear in shortened or abbreviated forms [6]. Manually collected acronym dictionaries are widely available (e.g. BioABACUS [17] or acronyms within the UMLS thesaurus, etc.). However, many studies suggested that static acronym repositories cover only up to one third of acronyms appearing in documents [8].

In our experiments with acronyms we have found that each abstract introduces 1.7 acronyms on average: in a random subset of the Genia corpus (containing 50 abstracts) 85 acronyms have been defined. However, coining and introducing new acronyms is a huge topic on its own, and we will not discuss it here[10].

### 4.4 Term Variation and ATR

Although biomedical terminology is highly variable, only few methods for the incorporation of term variants into the ATR process have been suggested (e.g. [5, 11, 12]). In our experiments we evaluated the effectiveness of incorporating specific types of term variation (presented in 4.1– 4.3) into an ATR system (see [12] for details). We compared a baseline method (namely the C-value method [3]), which considered term variants as separate terms, with the same method enhanced by the incorporation and conflation of term variants [11, 12]. The baseline method suggests term candidates according to "termhoods" based on a corpus-dependent statistical measure, which mainly relies on the frequency of occurrence and the frequency of occurrence as a substring of other candidate terms (in order to tackle nested terms). When the baseline C-value method is applied without conflating variants, frequencies are distributed across different variants (of the same term) providing separate values for individual variants instead of a single frequency calculated for a term candidate unifying all of its variants. In the enhanced version [12], instead of individual term candidates we use the notion of *synterms*, i.e. sets of synonymous term candidate variants that share the same normalised, canonical form. For example, plural term occurrences are conflated with the corresponding singular forms, while prepositional term candidates are

[10] For more information on acronyms in the biomedical domain see [2, 6, 9, 11, 15].

mapped to equivalent forms without prepositions. Further, acronym occurrences are linked and "counted" along with the corresponding expanded forms. Then, statistical features of occurrences of normalised candidates from synterms are used for the calculation and estimation of termhoods.

The experiments with the Genia corpus have shown that the incorporation of the simplest variations (such as inflectional variants and acronyms) resulted in a significant improvement of performance: precision improved by 20-70%, while recall was generally improved by 2-25% (see [12] for further details). However, more complex structural phenomena had moderate positive influence on recall (5-12%), but, in general, the negative effect on precision. The main reason for such performance was structural and terminological ambiguity of these expressions, in addition to their extremely low frequency (compared to the total number of term occurrences).

## 5 Conclusion

In this paper we have analysed the terminology that is used in biomedical abstracts. The analysis has shown that the vast majority of terms are multi-words and they typically appear as ordinary terms, spanning from two to four words. Terms also frequently appear as nested in longer terminological expressions in text, while controlled dictionaries – having a more "complete world" of terms – have even higher proportion of nested terms than the literature. We also show other discrepancies (such as in prepositional and coordinated expressions) between variations occurring in literature and those found in dictionaries.

Regarding term variation, the biomedical terminology is mainly affected by simple term variations (such as orthographic and inflectional variation) and acronyms, which also have the most significant impact on ATR [12]. Only around 7% of terms involve more complex structural phenomena (such as term coordination or the usage of prepositional term forms). Although undoubtedly useful, attempts to recognise such variation in text may result in a number of false term candidates, as there are no reliable morpho-syntactic criteria that can guide the recognition process, and a knowledge-intensive and domain-specific tuning is needed (e.g. ontological information on adjectives and nouns that can be combined within coordination or with a given preposition). Still, the integration of term variation into an ATR system is not only important for boosting precision and recall, but also crucial for terminology management and linking synonymous term occurrences across documents, as well as for many text-mining tasks (such as information retrieval, information extraction, term or document clustering and classification, etc.).

## References

1. Ananiadou, S.: A Methodology for Automatic Term Recognition. Proc. of COLING-94 1034-1038
2. Chang, J., Schutze, H., Altman, R.: Creating an Online Dictionary of Abbreviations from Medline. Journal of the American Medical Informatics Association. 9(6): 612-620, 2002
3. Frantzi, K., Ananiadou, S., Mima, H.: Automatic Recognition of Multi-word Terms: the C-value/NC-value Method. Int. J. on Digital Libraries. 3(2), 115-130, 2000

4. Hirschman, L., Friedman, C., McEntire, R., Wu, C.: Linking Biological Language Information and Knowledge. Proc. of PSB 2003 (the introduction to the BioNLP track)
5. Jacquemin, C.: Spotting and Discovering Terms through NLP. MIT Press, Cambridge MA (2001)
6. Krauthammer, M., Nenadic, G.: Term Identification in the Biomedical Literature. Journal of Biomedical Informatics. 2004 (*in press*)
7. Lander, ES, et al. (International Human Genome Sequencing Consortium): Initial sequencing and analysis of the human genome. Nature 409(6822), 860-921
8. Larkey, L., Ogilvie, P., Price, A., Tamilio, B.: Acrophile: An Automated Acronym Extractor and Server. Proc. of ACM Digital Libraries 2000, 205-214
9. Liu, H., Aronson, AR, Friedman, C.: A study of abbreviations in Medline abstracts. Proc. of AMIA Symposium 2002, 464-468
10. Maynard, D., Ananiadou, S.: TRUCKS: A Model for Automatic Multi-Word Term Recognition. Journal of Natural Language Processing. 8(1): 101-125, 2000
11. Nenadic, G., Spasic, I., Ananiadou, S.: Automatic Acronym Acquisition and Term Variation Management within Domain-Specific Texts. Proc. of LREC-3 (2002), 2155-2162
12. Nenadic, G., Ananiadou, S., McNaught, J.: Enhancing automatic term recognition through recognition of variation. Proc. of COLING-2004 (*in press*)
13. Ogren P., Cohen, K., Acquaah-Mensah, G., Eberlein, J., Hunter, L.: The Compositional Structure of Gene Ontology Terms. In: Proc. of PSB 2004, 214-225
14. Ohta T., Tateisi, Y., Kim, J., Mima, H., Tsujii, J.: Genia Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. Proc. of HLT-2002, 73-77
15. Pustejovsky J., Castaño, J., Cochran, B., Kotecki, M., Morrell, M., Rumshisky, A.: Extraction and Disambiguation of Acronym-Meaning Pairs in Medline. Proc. of Medinfo, 2001
16. Pustejovsky J., Castaño, J., Zhang, J., Kotecki, M., Cochran, B.: Robust Relational Parsing Over Biomedical Literature: Extracting Inhibit Relations. Proc. of PSB 2002, 362-373
17. Rimer M., O'Connell, M.: BioABACUS: a database of abbreviations and acronyms in biotechnology and computer science. Bioinformatics. 14(10): 888-889, 1998