

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

**Goran Nenadić**

Dept. of Computation  
UMIST, Manchester

G.Nenadic@umist.ac.uk

**Irena Spasić**

Dept. of Chemistry  
UMIST, Manchester

I.Spasic@umist.ac.uk

**Sophia Ananiadou**

Computer Science  
University of Salford

S.Ananiadou@salford.ac.uk

## Abstract

In this paper we present a study of the usage of terminology in biomedical literature, with the main aim to indicate phenomena that can be helpful for automatic term recognition in the domain. Our comparative analysis is based on the terminology used in the Genia corpus. We analyse the usage of ordinary biomedical terms as well as 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 show that there is a discrepancy between terms used in literature and terms listed in controlled dictionaries. We also evaluate the effectiveness of incorporating 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 (Pustejovsky *et al.*, 2002). The problem is, however, the huge volume of the literature, which is constantly expanding both in size and thematic coverage. For example, a query “*breast cancer treatment*” submitted to PubMed (NLM, 2003a) returned nearly 70,000 abstracts in 2003 and 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 abstracts.

An additional challenge is the rapid change of the biomedical terminology and the diversity of its usage. It is quite common that almost every new biomedical text introduces new names and terms. Also, the problem is the extensive terminology variation and use of synonyms. For example, a study reported by Ding *et al.* (2002) found that – when querying Medline – target interactions contained in an abstract were “often described using a synonym of the query term”.

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 (Lander *et al.*, 2001)). However, domain experts also frequently introduce specific notations, acronyms, ad-hoc and/or innovative names for new concepts, which they use either locally (within a document) or within the wider community. Even when an establish name exists, authors may prefer – e.g. from 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 literature, rather than the terminology presented in controlled resources. After presenting the resources that we have used for 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 evaluate the effectiveness of incorporating specific types of term variation into 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. 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 (cf. (Hirschman *et al.*, 2003)). 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 (1994) analysed term formation patterns in immunology, while Maynard (2000) analysed the internal morpho-syntactic properties of multi-word terms in ophthalmology. Estopa *et al.* (2000) considered only single word terms in the biomedical domain, and classified them into 3 classes: basic, inflected and terms containing neo-classical compounds.

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 the baker’s yeast (*S. cerevisiae*) and experimented with locating terms from the GO ontology<sup>1</sup> (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.<sup>2</sup> 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<sup>3</sup> 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

<sup>1</sup> <http://www.geneontology.org/>

<sup>2</sup> Many GO ontology terms (i.e. entries) are more “descriptions” than real terms (e.g. *ligase*, *forming phosphoric ester bonds* or *oxidoreductase*), and therefore it is unlikely that they would appear in text frequently.

<sup>3</sup> <http://www.jbc.org>

represent a summary using typically new and specific terms, while full texts usually 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 developed at the University of Tokyo (cf. (Ohta *et al.*, 2002)), which provide publicly available<sup>4</sup> manually tagged terminological resources in the domain of biomedicine. The resources include the Genia ontology and the Genia corpus, which contains 2,000 abstracts from Medline. These abstracts have been obtained from PubMed by querying the database with the MeSH (NLM, 2003b) 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 (e.g. singular forms) are supplied, and the results are encoded using an XML-based annotation scheme. 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 manually tagged.

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

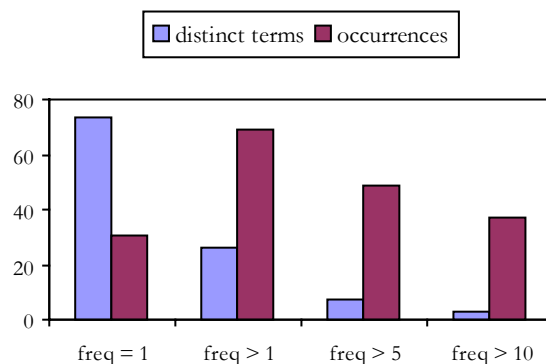


Figure 1: Distributions (in %) of the Genia terms and their occurrences (coverage in the corpus)

<sup>4</sup> <http://www-tsuji.is.s.u-tokyo.ac.jp/~genia/>

### 3 Ordinary terms

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 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) 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 term 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 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*, *nuclear factor of activated T cells family protein*). Such terms are typically hapax legomena in the Genia corpus.

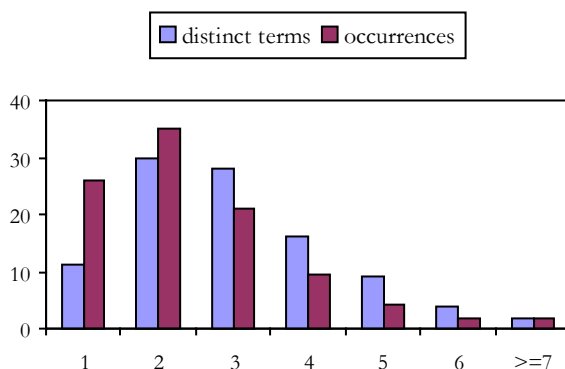


Figure 2: Distributions (in%) of the Genia terms and their occurrences with respect to the length

Apart from using different orthographic styles, a range of specific lexical expressions characterise the common biomedical terminology. For example, neo-classical Greek/Latin 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 (Frantzi *et al.*, 2000). 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.<sup>5</sup> 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 corpora.

### 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 the term candidate constituents (e.g. orthographic and inflectional forms) and one dealing with term structure (prepositional and coordinated terms). We also examine how the integration of term variation into ATR influences the precision and recall performance (Subsection 4.3).

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 appearing in the corpus (i.e. almost half of occurrences are “normalised” by the experts with respect to inflectional variation).

On the other hand, 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.

<sup>5</sup> On the other hand, Ogren *et al.* (2004) reported that two thirds of GO-ontology terms contained another GO-term as a proper substring.

#### 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.<sup>6</sup> 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*).

prep.	number of terms	examples
<i>of</i>	113	<i>promoter of gene</i>
<i>for</i>	9	<i>binding site for API</i>
<i>in</i>	8	<i>increase in proliferation</i>
<i>by</i>	2	<i>latency by expression</i>

Table 1: Distribution and examples of the Genia terms containing prepositions

Many potential term occurrences containing prepositions are not marked as terms by the experts, although equivalent occurrences without prepositions are marked as terms. 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 in literature, 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 –

<sup>6</sup> 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*).

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, and the recognition of such expressions may be beneficial for many text-mining tasks (such as document indexing, term or document clustering and classification, etc.).

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

#### 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 (appears only once), while their distinct lexical parts are enumerated and coordinated with a coordination conjunction (CC). Therefore, 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 (Jacquemin, 2001).

In the Genia corpus, term coordinations are 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.<sup>7</sup> 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).

<sup>7</sup> Only 1.4% of the GO-ontology terms contain CCs. However, these nodes mainly represent single concepts, and not coordinations of different terms.

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).

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

Table 2: Distribution of term coordinations in the Genia corpus

number of terms	number of coordination	
	<i>and</i>	<i>or</i>
2	1230 89.08%	141 85.97%
3	101 7.31%	19 11.59%
4	31 2.24%	1 0.61%
5	14 1.01%	2 1.22%
6	4 0.29%	0 0.00%
7	1 0.07%	0 0.00%

Table 3: Number of terms in term coordinations in the Genia corpus

Coordinations 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.

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 patterns used in expressing term coordinations. Although patterns were highly variable, the simplest patterns<sup>8</sup> (such as  $(N|A)^+ CC (N|A)^* N^+$ ) covered more than two thirds of term coordination occurrences.

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.

example	<i>adrenal glands and gonads</i>
head coordination	[ <i>adrenal [glands and gonads]</i> ]
term conjunction	[ <i>adrenal glands</i> ] <i>and</i> [ <i>gonads</i> ]

Table 4: Ambiguities within coordinated structures

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

<sup>8</sup> A and N denote an adjective and a noun respectively.

problem is that coordination patterns have to be more specific, but there are no reliable morpho-terminological clues indicating genuine term coordinations and their types. 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 for head coordinations.

Not only the recognition of term coordinations and their subtypes is ambiguous, but also internal boundaries of coordinated terms are blurred. For example, in the coordination *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 the pattern  $N_1$  and  $N_2$  PCP  $N_3$ )<sup>9</sup> 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. Manually collected acronym dictionaries are widely available (e.g. BioABACUS (Rimer and O'Connell, 1998) or

<sup>9</sup> PCP denotes an ing-form of a verb.

acronyms within the UMLS thesaurus (NLM, 2003b), etc.). However, when acronym dictionaries are static, both their coverage and accuracy might be doubtful: some repositories cover only up to one third of acronyms appearing in documents (Larkey *et al.*, 2000).

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.<sup>10</sup>

### 4.4 Term variation and ATR

Although terminology is highly variable, only few methods for the incorporation of term variants into the ATR process have been suggested (e.g. (Jacquemin, 2001)). The integration of term variation into an ATR system is not only important for boosting precision and recall values, but also crucial for terminology management and linking synonymous term occurrences across documents.

In our experiments we evaluated the effectiveness of incorporating specific types of term variation (mentioned in 4.1- 4.3) into an ATR system. We compared a baseline method (namely the C/NC-value method (Frantzi *et al.*, 2000)), which considered term variants as separate terms, with the same method enhanced by the incorporation and conflation of term variants (Nenadic *et al.*, 2002). The base-line method suggests term candidates according to “termhoods” based on a corpus-based 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 base-line C/NC-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, instead of individual term candidates we use the notion of *synterms*, i.e. sets

<sup>10</sup> For more information on acronyms in the biomedical domain see (Pustejovsky *et al.*, 2001), (Chang *et al.*, 2002), (Nenadic *et al.*, 2002), (Liu *et al.*, 2002), (Yu *et al.*, 2002), etc.

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 mapped to equivalent forms without prepositions. Similarly, 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. We hypothesised that – in this case – precision would be enhanced by considering joint frequencies for all candidate terms from synterms, while recall would benefit by the introduction of new candidates through support of different variations.

In the experiments with the Genia corpus, the incorporation of the simplest variations (such as inflectional) into the base-line ATR method resulted in a significant improvement of performance: the precision<sup>11</sup> improved by 30-40% (see Figure 3). Furthermore, the integration of the acronym recognition resulted in considerable improvements in both precision and recall, in particular for more frequent terms. When individually integrated into the base-line ATR, acronyms increased precision by 20-70% for top ranked synterms’ intervals. Recall was generally improved by 10-50%, as the acronym acquisition extracted variants that had more complex internal structures not targeted by the base-line method (such as acronyms containing prepositions (e.g. *repressor of estrogen activity* ↔ *REA*) or prepositions and coordinations (e.g. *silencing mediator of retinoic and thyroid receptor* ↔ *SMRT*)). Further integration of acronyms and inflectional variants proved to be the most effective in improving both precision and recall.

On the other hand, more complex structural phenomena had moderate positive influence on recall (5-12%), but, in general, the negative effect on precision. Prepositional term candidates, in particular, significantly reduced the precision

(mainly by introducing “false positives”)<sup>12</sup> without considerably improving recall. The main reason for such performance was structural and terminological ambiguity of these expressions, in addition to their low frequency (compared to the total number of term occurrences).

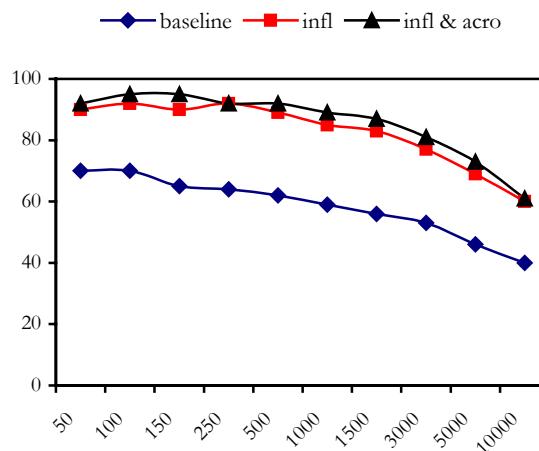


Figure 3: Comparison of ATR interval precisions on the Genia terms (intervals are on the X-axis)

As indicated earlier, coordinated term candidates or candidates with prepositions need to be additionally semantically analysed in order to suggest more reliable extractions, and to reduce the number of false candidates. Therefore, for handling structural variants a knowledge-intensive and domain-specific approach is needed (e.g. ontological information on adjectives and nouns that can be combined within coordination or with a given preposition). In addition, as the frequencies of such term phenomena are very low, statistically-based morpho-syntactic features cannot be used as reliable criteria for the recognition of affected terms.

## 5 Conclusion

In this paper we analysed the terminology that is used in biomedical abstracts and the influence of term variability on the performance of an ATR

<sup>11</sup> Since the ATR method ranks extracted terms according to their termhoods, we used interval precisions to estimate the accuracy. Interval precision measures precision in fixed rank cut-offs (i.e. “intervals”). For example, we considered precisions in intervals containing first 50, first 100, first 150 etc. suggested terms.

<sup>12</sup> Precision was further reduced by the fact that some of the “true positive” prepositional term candidates were considered as “negatives” because they have not been marked by the Genia annotators as terms.

system. 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 larger terminological expressions. Controlled dictionaries – having a more “complete world” of terms – have even higher proportion of nested terms than literature.

Regarding the term variation, the biomedical terminology is mainly affected by simple term variations (such as inflectional variation) and acronyms, which also have the most significant impact on ATR. On the other hand, only around 7% of terms involve more complex structural phenomena (such as term coordination or the usage of prepositional term forms). We also show that there is a discrepancy between variations used in literature and found in dictionaries. 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. On the other hand, the recognition of such expressions will be beneficial for many text-mining tasks (such as information retrieval, information extraction, term or document clustering and classification, etc.).

## References

- Ananiadou S. 1994. A Methodology for Automatic Term Recognition. Proc. of COLING-94, 1034-1038
- Chang J., H. Schutze, and R. Altman. 2002. Creating an Online Dictionary of Abbreviations from Medline. Journal of the American Medical Informatics Association. 9(6): 612-620
- Ding J., D. Berleant, D. Nettleton, and E. Wurtele. 2002. Mining Medline: Abstracts, Sentences, or Phrases? Proc. of PSB 2002
- Frantzi K., S. Ananiadou and H. Mima. 2000. Automatic Recognition of Multi-word Terms: the C-value/NC-value Method. Int. J. on Digital Libraries, 3/2, pp. 115-130.
- Hirschman L., C. Friedman, R. McEntire, and C. Wu. 2003. Linking Biological Language Information and Knowledge. Proc. of PSB 2003 (introduction)
- Estopa R., J. Vivaldi, and T. Cabre. 2000. Use of Greek and Latin Forms for Term Detection. Proc. of LREC-2000
- Jacquemin C. 2001. Spotting and Discovering Terms through NLP, MIT Press, Cambridge MA
- Lander ES, et al. (International Human Genome Sequencing Consortium). 2001. Initial sequencing and analysis of the human genome. Nature 409(6822), pp. 860-921.
- Larkey L., P. Ogilvie, A. Price, and B. Tamilio. 2000. Acrophile: An Automated Acronym Extractor and Server. Proc. of ACM Digital Libraries 2000
- Liu H., A.R. Aronson, and C. Friedman. 2002. A study of abbreviations in Medline abstracts. Proc. of AMIA Symposium 2002, pp. 464-468
- Maynard D. 2000. Term Recognition using Combined Knowledge Sources. PhD thesis, Manchester Metropolitan University, UK, 2000
- Nenadic G., I. Spasic, and S. Ananiadou. 2002. Automatic Acronym Acquisition and Term Variation Management within Domain-Specific Texts. Proc. of LREC-3, 2155-2162
- NLM (National Library of Medicine). 2003a. Medline, available at: <http://www.ncbi.nlm.nih.gov/pubmed/>
- NLM (National Library of Medicine). 2003b. UMLS - Unified Medical Language System, 2003
- Ogren P., K. Cohen, G. Acquaaah-Mensah, J. Eberlein, and L. Hunter. The Compositional Structure of Gene Ontology Terms. in Proc. of PSB 2004.
- Ohta T., Y. Tateisi, J. Kim, H. Mima, and J. Tsujii. 2002. Genia Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. Proc. of HLT-2002
- Pustejovsky J., J. Castaño, B. Cochran, M. Kotecki, M. Morrell, and A. Rumshisky. 2001. Extraction and Disambiguation of Acronym-Meaning Pairs in Medline. Proc. of Medinfo, 2001
- Pustejovsky J., J. Castaño, J. Zhang, M. Kotecki, and B. Cochran. 2002. Robust Relational Parsing Over Biomedical Literature: Extracting Inhibit Relations. Proc. of PSB 2002, pp. 362-373
- Rimer M. and M. O'Connell. 1999. BioABACUS: a database of abbreviations and acronyms in biotechnology and computer science. Bioinformatics, 1998. 14(10): p. 888-889.
- Yu H., G. Hripesak, and C. Friedman. 2002. Mapping abbreviations to full forms in biomedical articles. Journal of Am Med Inform Assoc, 2002. 9(3): 262-272