Journal of Biomedical Informatics 46 (2013) 228–237

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Acquisition and evaluation of verb subcategorization resources for biomedicine

Laura Rimell [a], Thomas Lippincott [a,*], Karin Verspoor [b], Helen L. Johnson [c], Anna Korhonen [a]

[a] Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK
[b] National ICT Australia, Victoria Research Lab, Melbourne, VIC 3010, Australia
[c] Department of Pharmacology, Center for Computational Pharmacology, University of Colorado School of Medicine, Aurora, Colorado, Denver, CO, USA

ARTICLE INFO

Article history:
Received 24 July 2012
Accepted 5 January 2013
Available online 22 January 2013

Keywords:
Verb subcategorization
Lexical resources
Natural language processing
Biomedical text processing

ABSTRACT

Background: Biomedical natural language processing (NLP) applications that have access to detailed resources about the linguistic characteristics of biomedical language demonstrate improved performance on tasks such as relation extraction and syntactic or semantic parsing. Such applications are important for transforming the growing unstructured information buried in the biomedical literature into structured, actionable information. In this paper, we address the creation of linguistic resources that capture how individual biomedical verbs behave. We specifically consider verb subcategorization, or the tendency of verbs to "select" co-occurrence with particular phrase types, which influences the interpretation of verbs and identification of verbal arguments in context. There are currently a limited number of biomedical resources containing information about subcategorization frames (SCFs), and these are the result of either labor-intensive manual collation, or automatic methods that use tools adapted to a single biomedical subdomain. Either method may result in resources that lack coverage. Moreover, the quality of existing verb SCF resources for biomedicine is unknown, due to a lack of available gold standards for evaluation.

Results: This paper presents three new resources related to verb subcategorization frames in biomedicine, and four experiments making use of the new resources. We present the first biomedical SCF gold standards, capturing two different but widely-used definitions of subcategorization, and a new SCF lexicon, BioCat, covering a large number of biomedical sub-domains. We evaluate the SCF acquisition methodologies for BioCat with respect to the gold standards, and compare the results with the accuracy of the only previously existing automatically-acquired SCF lexicon for biomedicine, the BioLexicon. Our results show that the BioLexicon has greater precision while BioCat has better coverage of SCFs. Finally, we explore the definition of subcategorization using these resources and its implications for biomedical NLP. All resources are made publicly available.

Conclusion: The SCF resources we have evaluated still show considerably lower accuracy than that reported with general English lexicons, demonstrating the need for domain- and subdomain-specific SCF acquisition tools for biomedicine. Our new gold standards reveal major differences when annotators use the different definitions. Moreover, evaluation of BioCat yields major differences in accuracy depending on the gold standard, demonstrating that the definition of subcategorization adopted will have a direct impact on perceived system accuracy for specific tasks.

1. Introduction

Biomedical natural language processing (NLP) applications can benefit from detailed resources describing the linguistic characteristics of biomedical language [1]. In particular, applications having information about the behavior of verbs in the biomedical literature have demonstrated improved performance on tasks such as relation extraction and syntactic or semantic parsing. Such applications are important for transforming the growing unstructured information buried in the biomedical literature into structured, actionable information.

In this paper, we address the creation of linguistic resources that capture how individual biomedical verbs behave. We specifically consider verb subcategorization, or the tendency of verbs to "select" co-occurrence with particular phrase types, which influences the interpretation of verbs and identification of verbal arguments in context. For example, the verb *detect* can be transitive (taking a single direct object) or it can take a clausal complement: *A routine X-ray of the thorax detected* [$_{NP}$ *pneumonia*] *and Researchers have detected* [$_{S}$ *that the tissues have high levels of Wnt signaling components*] are both fully grammatical sentences. In contrast, the verb

* Corresponding author.
E-mail addresses: Laura.Rimell@cl.cam.ac.uk (L. Rimell), Thomas.Lippincott@cl.cam.ac.uk (T. Lippincott), karin.verspoor@nicta.com.au (K. Verspoor), helen.linguist@gmail.com (H.L. Johnson), Anna.Korhonen@cl.cam.ac.uk (A. Korhonen).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.01.001

examine can be transitive, but cannot take clausal complements: *The study examined* [$_{NP}$ *the relationship between ankle-brachial index and stroke*] is fully grammatical, but not *The study examined* [$_{S}$ *that there is a relationship between ankle-brachial index and stroke*].

Any natural language processing (NLP) application that makes use of predicate-argument structure can make use of subcategorization information. *Subcategorization frames* (SCFs) were recently used by [2,3] to improve event extraction from UKPubMedCentral documents. SCFs have the potential for wide application in many other tasks such as entailment detection, relation extraction, syntactic and semantic parsing, all of which are important in biomedical NLP. Moreover, SCF information is more easily acquired from text corpora than similar linguistic structures used in biomedicine, such as Predicate Argument Structures [1].

In [1] we reviewed the state of the art with regard to verb subcategorization for biomedicine. We observed that there are a limited number of existing biomedical verb SCF resources, and to date their development has relied on either introspective, manual collation of SCFs, which results in resources that lack coverage, or automatic identification of SCFs using tools adapted to a single biomedical subdomain. Adaptation of such tools is labor-intensive, and the resulting resources may still lack coverage because the tools are not adapted to the broader biomedical literature. We showed that biomedical subdomains show notable and complex variation in verb subcategorization behavior, highlighting the need for minimally-supervised tools to automatically acquire SCF information, since such tools can be applied to different subdomains with minimal manual intervention.

Moreover, we observed in [1] that the quality of existing verb SCF resources for biomedicine is unknown, due to a lack of available gold standards which can be used for evaluation. The effect of adopting a more lenient definition of subcategorization compared to the traditional linguistic definition, as is typically done in biomedical NLP, has also not been ascertained due to the lack of gold standards.

In this paper we present three new resources related to verb SCFs in biomedicine, and four experiments making use of the new resources. We present the first biomedical SCF gold standards, capturing two different but widely-used definitions of subcategorization. These novel gold standards make it possible for the first time to perform quantitative and qualitative evaluation of automatically acquired SCF resources in biomedicine. We also present a new SCF lexicon, BioCat, acquired automatically from the PubMed Open Access collection (PMC OA), covering a large number of biomedical subdomains, and using state of the art tools developed for general language SCF acquisition.

Our first experiment evaluates BioCat using two different filtering methods to remove noise from the SCF lexicon, one of which is a new method for filtering hypothesized SCFs that improves accuracy by drawing on knowledge of subcategorization tendencies in general language. Our results show a respectable level of accuracy considering that no adaptations were made to the SCF acquisition system besides using a large biomedical corpus as input. We then compare the accuracy of BioCat to that of the BioLexicon [4–7], the only previously existing automatically-acquired SCF lexicon for biomedicine, which was extracted from corpus data in the E. Coli subdomain using NLP technology adapted to the subdomain of molecular biology, but which has not previously been evaluated. Our results show that the BioLexicon has greater precision while BioCat has better coverage of SCFs. Although the BioLexicon shows better overall accuracy, it is still considerably lower than that reported with general English lexicons, demonstrating the need for domain- and subdomain-specific SCF acquisition tools for biomedicine.

Our final two experiments explore the definition of subcategorization and its implications for biomedical NLP. It is well known that the standard definition of subcategorization in biomedicine is different than the traditional linguistic definition, since it includes more adjuncts (modifiers) as part of the SCF, because they are important for relation extraction. However, the effect of adopting this more lenient definition of subcategorization has not previously been ascertained due to the lack of gold standards. Our new gold standards reveal major differences when annotators use the different definitions. Moreover, evaluation of BioCat yields major differences in accuracy depending on the gold standard, demonstrating that the definition of subcategorization adopted will have a direct impact on perceived system accuracy. All resources are made publicly available.[1]

The rest of the paper is organized as follows. Section 2 provides a brief overview of subcategorization in biomedicine, and Section 3 summarizes our research questions. Section 4 introduces the new gold standards, and Section 5 introduces the new SCF lexicon, BioCat. The evaluation methodology is described in Section 6 and the results in Section 7.

2. Subcategorization frames in biomedicine

Verb subcategorization information is typically captured in "frames" that indicate which syntactic phrase types the verb co-occurs with, including noun phrases (NPs), prepositional phrases (PPs), subordinate clauses, and adjectives. Some examples of subcategorization frames (SCFs) can be seen in Table 1.[2] Most verbs take several SCFs, with varying frequencies.

SCFs are similar to Predicate Argument Structures (PASs), another argument structure representation used in biomedical NLP. PASs have been used in Semantic Role Labeling for biomedicine [9–11]. However, PASs are built up from very specific verb-specific roles such as, for the verb *delete*, "entity doing the removing", "thing being removed", and "removed from". Because SCFs are built from phrase types which are common across verbal argument structure, they are general enough to be automatically acquired for a large number of verbs, while still providing a basic level of argument structure information which can aid in event identification. As yet there are only a small number of studies using SCFs in biomedical NLP, perhaps because resources have previously been limited, but initial results are promising: [2,3] used SCFs to improve event extraction from UKPubMedCentral documents.

In the traditional linguistic view of subcategorization, a distinction is made between *arguments* and *adjuncts*. Arguments are phrases that obligatorily co-occur with the verb, while less closely associated modifiers such as location, manner, or temporal phrases are adjuncts and not part of the SCF. For example, in Fig. 1, the PP *on a pre-warmed operation table* is optional, elaborating on the event description by describing the location at which it took place. The PP *on the patient* is obligatory and exhibits a special, idiomatic meaning in the context of the verb *operate*. The biomedical NLP field, however, has adopted a more inclusive view of subcategorization, including adjuncts in the SCFs, because the information they contain is considered important for information extraction in biomedicine [12,9]. We will refer to the traditional linguistic view as "syntactic" (i.e. grammar-level), since it emphasizes syntactic obligatoriness, and the biomedical view as "semantic" (i.e. meaning-level), since it emphasizes semantic importance. We observed in [1] that the effects of these two views on resource creation and automatic acquisition have not been quantitatively

[1] http://www.cl.cam.ac.uk/~tl318/BioCat.tgz.

[2] For the SCF names we use COMLEX Syntax notation [8], which includes an abbreviation for each phrase type in the SCF. Thus the SCF for a transitive verb (taking one direct object noun phrase) is NP, and for a verb taking a direct object and a prepositional phrase NP-PP. We do not specify the subject NP as part of the SCF, since subjects are obligatory in English.

Table 1
Sample SCFs for *decrease*, *compare*, and *reveal*. All examples adapted from the PubMed Open Access (PMC OA) corpus.

SCF	Example
NP	This physical, mental, or emotional tension of an individual decreases [$_{NP}$ the feeling of being in control]
NP-PP	Heterozygosity for twine also decreases [$_{NP}$ the frequency of precocious NEB] [$_{PP}$ to less than 10%]
PP-PP	The proportion of subjects with moderate-to-severe symptoms decreased [$_{PP}$ from 29.6%] [$_{PP}$ to 2.3%]
Intransitive	In the control group SV decreased
NP	We compared [$_{NP}$ the performance of the Charlson and the Elixhauser comorbidity measures]
NP-PP	We compared [$_{NP}$ mandatory celiotomy] [$_{PP}$ to laparoscopy]
NP	A post hoc analysis revealed [$_{NP}$ a statistically significant relationship between timing of fondaparinux dose and bleeding]
THAT-S	This observation revealed [$_{S}$ that systolic and diastolic BP increased during the interdialytic period]
NP-TOBE	[$_{NP}$ The incidence for cardiovascular events] was revealed [$_{S}$ to be 1.13%]

ADJUNCT:
. . . and operated [$_{PP}$ on a pre-warmed operation table] . . .
ARGUMENT:
HW provided clinical care for this patient and operated [$_{PP}$ on the patient].

Fig. 1. Example adjunct and argument PPs from the PMC OA corpus for the verb *operate*.

evaluated, and in this paper we perform the first investigation of SCF acquisition that explicitly compares the two definitions of subcategorization.

3. Research questions

The motivation for this paper is to address the following research questions. First, we want to know whether general purpose NLP tools can be used to extract SCFs from biomedical documents from various biomedical subdomains, despite the differences between biomedical and general language. Specifically, since it is not possible to develop tools for each subdomain, we would like to compare general purpose tools against tools adapted for a single subdomain, when applied to a broad corpus and a gold standard drawn from across subdomains.

Second, we aim to understand the implications of the definition of subcategorization used to create the gold standards. We investigate whether the gold standards produced from the same data but using two different definitions, one syntactic and the other semantic, are substantially different, and whether general purpose tools perform better on one gold standard than another.

4. Biomedical SCF gold standards

We have produced the first set of biomedical SCF gold standards, which we make available as a resource to the biomedical NLP community. Each resource has been produced by selecting a set of representative verbs, then manually annotating 150–200 sentences randomly chosen from across the PMC OA for each verb, in order to provide broad coverage of multiple subdomains. Annotation of corpus data is crucial to avoid missing SCF types [1], as well as to gather statistical information about SCF frequency, which can be important for resource evaluation.

4.1. SCF inventory

For annotation of corpus data we chose to use the SCF inventory of [13], a rich, manually-developed inventory previously used for general language. It consists of 163 SCFs obtained by manually merging the SCFs exemplified in the COMLEX Syntax [8] and ANLT [14] dictionaries, along with some additional frames identified by inspection of general language data. We refer to this inventory as the "Cambridge inventory" because it was developed at the University of Cambridge.

Many of the 163 SCF types in the Cambridge inventory are rare, and in any given dataset only a subset of these SCFs will actually be found. The annotators, who were familiar with the entire inventory, ended up using a total of 27 SCF types from the inventory while annotating the corpus data used for our gold standards.

4.2. Annotation tool

A custom tool was developed and used for annotation. The tool highlighted the target verb in each sentence and allowed the annotator to select an SCF from a drop-down menu. The annotator could also customize responses to particular sentences, for example by flagging problematic examples, or adding comments for later reconciliation. Problematic examples included sentences needing new frames not in the Cambridge inventory, of which there were only a few (see Section 4.3), or cases where the annotators wished to seek guidance from the authors before deciding on a final frame. A screen shot of the annotation tool is shown in Fig. 2.

4.3. Semantic gold standard

Our main gold standard contains 30 verbs annotated using the "semantic" definition of subcategorization favored for biomedicine (see Section 2). We refer to this gold standard as SEM-30. The verbs were chosen based on frequency,[3] occurrence across both biomedical and general language text, and the fact that they are known to take multiple SCFs in biomedical text. We also preferred verbs that we believed may have developed specialized senses in biomedicine – e.g. *activate* – since specialized senses often correspond to specialized SCFs. The first column of Table 2 shows the verbs in SEM-30.

The annotator, a biomedical NLP expert, was instructed to include in the SCF all phrases attached to the verbal head which were important for biomedicine, and also aimed for similarity with the semantic role types in PropBank [15], a corpus of verbal propositions and their arguments.

Prior to beginning the annotation we did not know whether the Cambridge inventory, developed for general language, would be appropriate for biomedical text. We found that it was; during annotation of SEM-30, only 20 sentences, or about 0.3% of the 6,473 total annotated sentences, were identified by the annotator as involving an SCF not included the Cambridge inventory. Since the number of examples was so small, we chose to discard these sentences rather than modify the inventory.

[3] Verbs needed to be frequent enough to ensure enough data to annotate for the gold standard, as well as enough raw corpus data for the SCF acquisition system to produce a comprehensive lexicon.

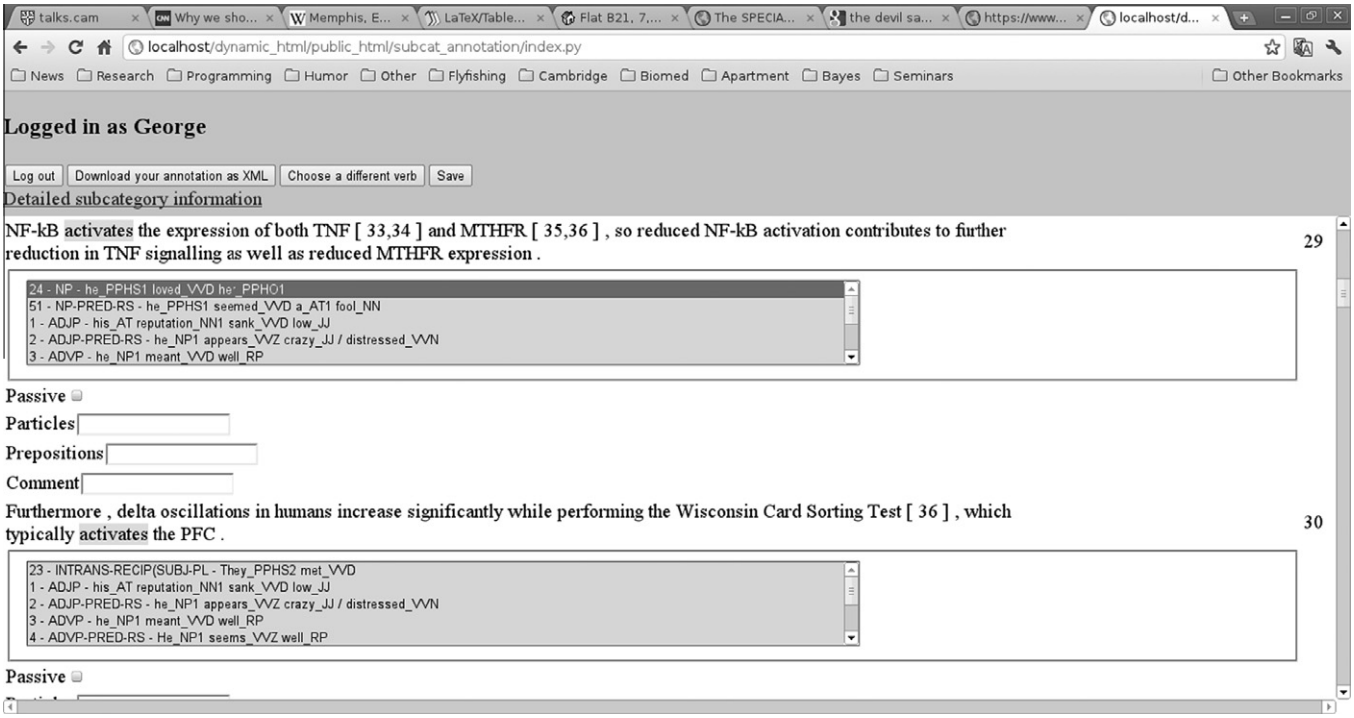

Fig. 2. Annotation interface.

Table 2
Verbs in the gold standards SEM-30 (full gold standard, Section 4.3), SEM-26 (overlap of SEM-30 with verbs in the BioLexicon, Section 4.3), SYN-10 and SEM-10 (comparative syntactic and semantic gold standards, Section 4.4), and those that overlap with the general language gold standard of [13] (Section 6.3).

Verb	SEM-30	SEM-26	SYN-10 and SEM-10	Overlap with [13]
activate	•	•	•	
analy (z/s) e	•	•	•	•
associate	•	•	•	
cause	•	•		•
compare	•	•	•	•
contain	•	•		
decrease	•	•	•	
detect	•	•		
develop	•	•		
enhance	•	•		
examine	•			
express	•	•	•	
fail	•			
follow	•	•		
generate	•	•		
improve	•	•	•	
increase	•	•		
induce	•	•		•
inhibit	•	•		
modify	•	•		
mutate	•	•	•	
occur	•	•	•	
perform	•			
predict	•		•	
produce	•	•		•
recogni (z/s) e	•	•		
reduce	•	•		
regulate	•	•		
transcribe	•	•		
treat	•	•		

The SEM-30 gold standard, which includes SCFs and their relative frequencies for each verb, was derived directly from the annotations. A sample entry from SEM-30 for the verb *transcribe* is shown in Fig. 3, along with the automatically acquired lexicon entry for *transcribe* from our most accurate system. We will also refer to SEM-26, which is simply a subset of SEM-30 used for comparative evaluation. The second column of Table 2 shows the verbs in SEM-26. We used SEM-30 and SEM-26 to evaluate SCF acquisition systems (see Sections 6.1, 6.2, 7.1, 7.2).

transcribe			
Gold		Automatic	
NP	0.719424	NP	0.782702
NP-PP	0.215827	PP	0.091381
NP-as-NP	0.021583	INTRANS	0.075795
NP-PP-PP	0.014388	P-NP-ING	0.024450
PP	0.014388	NP-as-NP	0.017115
INTRANS	0.007194	ING-PP	0.008557
ADVP	0.007194		

Fig. 3. Sample gold standard entry for *transcribe* from SEM-30. Column 1 shows the SCF and column 2 shows the relative frequency across sentences annotated for the gold standard. Automatically acquired lexicon entry for *transcribe* in columns 3 and 4. Per-verb accuracy was Precision: 66.7, Recall: 57.1, F-Score: 61.5.

4.4. Syntactic gold standard

In order to investigate the difference between the semantic and syntactic annotation styles, we chose ten verbs from SEM-30 to annotate according to the syntactic definition of subcategorization. We chose verbs that appeared to occur in the corpus with a relatively large number of highly selected adjuncts, making them more likely to exhibit variation across the two definitions of subcategorization. We refer to this gold standard as SYN-10, and the same verbs annotated with the semantic definition as SEM-10 (i.e. SEM-10 is a subset of SEM-30). The third colum of Table 2 shows the verbs in SYN-10 and SEM-10.

A second annotator, a linguistics expert, performed the syntactic annotation for SYN-10, and was given different guidelines from

Sentence	The p53 mutant, which contains a disabled DNA-binding domain, does not activate [$_{NP}$transcription] [$_{ADVP}$significantly].
Semantic annotator	NP-ADVP (*transcription, significantly*)
Syntactic annotator	NP (*transcription*)
Sentence	Both receptors activate [$_{NP}$NF-kB] [$_{PP}$through the canonical and noncanonical pathways], with RANK specifically requiring TRAF6.
Semantic annotator	NP-PP (*NF-kB, through the ... pathways*)
Syntactic annotator	NP (*NF-kB*)

Fig. 4. Examples of sentences that the semantic and syntactic annotators treated differently. The syntactic annotator was judging by what was syntactically obligatory, while the semantic annotator by what was important to an understanding of the event.

those given to the semantic annotator. The syntactic annotator was instructed to use the traditional criterion of optionality to distinguish arguments from adjuncts. We again used our annotation tool (Fig. 2). The time and effort involved prevented us from undertaking double annotation by both annotators. However, the data in Table 7 indicates that the two annotators were choosing approximately the same number of SCF types per verb, which suggests that any disagreements between the two annotators on particular sentences were due to the targeted difference in annotation guidelines.

We used the Cambridge inventory of SCFs for syntactic as well as semantic annotation. However, the syntactic and semantic interpretations differ for the same SCFs. For example, according to the syntactic definition, the frame NP-ADVP would only be used for certain obligatory adverbs such as *there* as in the sentence *She put* [$_{NP}$*it*] [$_{ADVP}$*there*]. Under the semantic annotation, the use of this frame would be extended, e.g. for adverbs such as *significantly* or *normally*. Fig. 4 gives examples of sentences that the annotators treated differently.

As with SEM-30, the SYN-10 gold standard was derived directly from the annotations. A small number of sentences were discarded which had not been annotated by both annotators, due to differences in opinion about whether the sentence was a valid instance, e.g. if one annotator mistakenly annotated a gerundive use of the verb. SYN-10 was used to measure differences between the semantic and syntactic annotation approaches, both by direct comparison of the resulting gold standards (Sections 6.3, 7.3) and investigation of how the annotation approach affected the evaluation of an acquired SCF lexicon (Sections 6.4 and 7.4).

5. BioCat: a new subcategorization resource for biomedicine

There are few existing SCF resources for biomedicine, and those that exist tend to rely on manual development or domain-specific tools. The new resource we present here, BioCat, takes a different appraoch. It uses a set of tools developed for SCF acquisition in general language which are, except for the POS tagger, domain-independent, and applies these tools to a large biomedical corpus. We consider the NLP tools to be domain-independent because they are unlexicalized; that is, they have no information about specific words, only general information about English. For example, the parser has no verb-specific information about the likelihood of different syntactic arguments.

We produced BioCat using an updated version of the tools in [13], which we will refer to as the Cambridge system, or Cambridge tools. The updating consisted of a more recent unpublished version of the SCF classifier, which re-implemented the original classifier rules in a different programming language.

An SCF acquisition system in general consists of hypothesis generation (pre-processing, parsing, and identifying potential SCFs using a classifier) followed by filtering (see [1] for an overview). In the Cambridge system an input corpus is first parsed with the RASP system [16]. A classifier consisting of manually-defined rules then matches the RASP output to the SCFs in the Cambridge inventory (Section 4), and the resulting lexicon is filtered.

5.1. Subcategorization acquisition system

For pre-processing and parsing we used RASP, a modular statistical parsing system which includes a tokenizer, tagger, lemmatizer, and a wide-coverage unification-based tag-sequence parser. We used the standard scripts supplied with RASP to output the set of grammatical relations (GRs) for the most probable analysis returned by the parser or, in the case of parse failures, the GRs for the most likely sequence of subanalyses. RASP is an unlexicalized parser, meaning that it does not have access to a lexicon of information about the behavior of specific words (as opposed to classes of words, e.g. words with particular part-of-speech tags), and thus does not already embody a notion of subcategorization.[4] To identify potential SCFs, a rule-based classifier incrementally matches GRs with SCFs. The rule set was an updated version of that used in [13]; note that it was developed for general language and not adapted for biomedical text. From the classifier output, preliminary lexical entries are constructed for each verb, containing the raw and relative frequencies of SCFs found for each verb in the data. Finally, the entries are filtered to obtain a more accurate lexicon.

5.2. Filtering

We used two filtering methods. The first method was simple relative frequency filtering. Here, an empirically determined minimum threshold is set on the relative frequencies of SCFs, so that only SCFs with per-verb relative frequencies above the threshold are retained. This simple method has been shown to yield more accurate results than more complex statistical hypothesis tests [18]. Previous work on SCF acquisition for general language using similar SCF systems found a threshold of 0.02 to give the most accurate results. In development experiments on held-out data we found 0.02 and 0.03 to give the most accurate results under different conditions, and we chose to use a threshold of 0.03 to match the threshold used by the BioLexicon (see Section 6.1).

Second, we used a novel method which we call SCF-specific filtering. The intuition behind this method is that the appropriate reliability threshold for each SCF may be different, since some SCFs are inherently much more frequent than others. We did not have information about the overall frequency of the different SCFs in biomedical text, so we used information about their overall frequency in general language from the COMLEX and ANLT dictionaries, along with empirical information about high and low frequencies from the unfiltered lexicon acquired for biomedicine, to set a specific threshold for each SCF. We tested this method to see if it could improve accuracy, even though it uses information about general language which may or may not be applicable to the biomedical domain.

5.3. Input corpus

For our corpus we used the PubMed Open Access Subset (PMC OA), which is the largest publicly available corpus of full-text articles in the biomedical domain [19, downloaded 6 October, 2009]. This PMC OA collection comprises 169,338 articles drawn from 1233 medical journals indexed by the Medline citation database, totaling approximately 400 million words. Articles are formatted

[4] To test the influence of different parsers, we performed an experiment using output of the unlexicalized Stanford parser [17] as input to the subcategorization steps and found that accuracy was the same on the SCF evaluation as for RASP.

according to a standard XML tag set [20]. We used the same dataset as [21,1], composed of journals that are assigned to a single subdomain, and discarding subdomains with less than one million words of data. The resulting dataset contains a total of 342 journals in 37 biomedical subdomains. It has been shown that the open access collection is representative of the broader biomedical literature [22].

5.4. Resource evaluation

The resulting BioCat resource has two versions, each containing 3911 verbs. The version built using simple relative frequency filtering has an average of 6.3 SCFs per verb, ranging from a minimum of 1 to a maximum of 18. The version built using SCF- specific filtering has an average of 6.7 SCFs per verb, ranging from a minimum of 1 to a maximum of 23. An example of the acquired lexical entry for the verb *transcribe* is given in Figure 3. Sections 6 and 7 describe how we evaluated BioCat against our gold standards.

6. Experiments

We performed a number of experiments designed to evaluate current SCF technology and the definition of subcategorization.

6.1. Evaluation of BioCat

The purpose of this experiment was to evaluate our new resource, BioCat, against a biomedical SCF gold standard, SEM-30. We evaluated two versions of BioCat, corresponding to the two filtering methods described in Section 5.

In each case the SCF acquisition system for BioCat described in Section 5.1 was applied to the input corpus from Section 5.3, drawn from across PMC OA, and the relevant filtering method applied. From the resulting SCF lexicon, the entries corresponding to the verbs in SEM-30 were extracted, and compared to the gold standard entries in SEM-30.

We used standard evaluation measures from previous SCF evaluations for general language [18,13], namely type precision, type recall, and F-score (the harmonic mean of precision and recall). Type precision is the percentage of SCFs in the lexicon entry for a particular verb that are correct according to the gold standard, and type recall is the percentage of gold standard SCFs for a particular verb that the lexicon contains for that verb. All measures are given as a macro-average over the type precision, type recall, and F-score for the individual verbs in the gold standard. We noted the number of SCFs present in the gold standard but missing from the filtered output, i.e. not just missing for a particular verb but missing altogether, as a way of evaluating the coverage of the SCF system. We also noted the number of gold standard SCFs unseen in the *unfiltered* system output; that is, false negatives which were not detected at all by the classifier even before filtering. In this work we evaluated only on per-verb presence or absence of SCFs in the automatically acquired lexicon, not per-verb frequencies of those SCFs.

6.2. Comparative evaluation of BioCat and the BioLexicon

The purpose of this experiment was to perform a comparative evaluation of BioCat against another SCF resource, the BioLexicon [4,6,7]. BioCat was built using unadapted, general language tools applied to a multi-subdomain biomedical corpus; while the BioLexicon was built using tools adapted to a single biomedical subdomain, applied to data also drawn from that subdomain.

Both lexicons result from state-of-the-art approaches to biomedical SCF acquisition; and given the impracticability of manually adapting NLP tools to every subdomain, both approaches are natural, with BioCat having a potential advantage from the wide coverage of its source data, and BioLexicon having a potential advantage from domain adaptation. This experiment tests how the two approaches perform against an SCF gold standard drawn from a wide variety of subdomains (see [1] for further discussion).

BioCat was described in Section 5; here we give a brief description of the BioLexicon, followed by a description of the mapping between the two lexicons.

6.2.1. The BioLexicon

The BioLexicon [4,6,7] is currently the only biomedical NLP resource containing an automatically constructed SCF lexicon. It is built on data from the E. Coli subdomain, and each component used in acquisition of the lexicon – for example, the part-of-speech tagger, named entity recognizer, and parser – was manually adapted to the subdomain of molecular biology.

To create the BioLexicon, six million words of MEDLINE E. Coli abstracts and articles were parsed with the Enju deep parser [23], which was adapted to the biomedical domain as described in [24], using a variety of external resources such as GENIA [25]. Unlike RASP, Enju is a lexicalized parser, which means that it already contains a notion of subcategorization, which can be adapted to different domains. No SCF inventory was assumed in advance; rather, the set of grammatical relations for each verb instance was considered as a potential SCF. Potential SCFs were filtered using simple relative frequency filtering, at a threshold of 0.03, leading to an inventory of 136 SCFs. Further arguments and strongly-selected adjuncts were chosen according to their log-likelihood with respect to the verb.

The BioLexicon is available through ELRA.[5] We used the BioLexicon exactly as provided without additional training or adaptation. No evaluation is provided with the BioLexicon which would reveal how well the acquisition technology performs on E.Coli or on general biomedical corpus data, so our experiment represents the first such evaluation. We used SEM-26 for the evaluation, since four verbs in SEM-30 were not included in the BioLexicon.

6.2.2. Mapping between lexicons

Performing a comparative evaluation was not straightforward, since the mapping between the BioLexicon SCF inventory and the Cambridge inventory, used for the gold standard and BioCat, is many-to-many. In general, the Cambridge inventory makes some more fine-grained linguistic distinctions, whereas the BioLexicon has more lexicalized elements in the SCFs.

We first used a "best match" mapping in which we manually selected the closest match in the Cambridge inventory for each BioLexicon SCF. This mapping resulted in a set of 22 SCF types for the BioLexicon. This is far lower than the 97 SCF types reported for the BioLexicon in [5] (we found 136 SCF types when querying the BioLexicon). However, since the BioLexicon inventory differentiates SCFs containing PPs based on the lexicalized preposition, many SCFs were collapsed during the mapping: PP-*from*, PP-*to*, PP-*of*, etc. would all map to PP in the Cambridge inventory. For comparison, the SEM-26 gold standard contained 27 SCF types from the Cambridge inventory, as discussed in Section 4.1.

To make sure the mapping did not penalize the BioLexicon, we also used a "coarse" mapping which represented a common denominator between the Cambridge and BioLexicon inventories. We semi-manually created equivalence classes of SCFs such that both inventories could distinguish the classes from one another, and these classes became the SCFs in a new, coarse-grained inven-

[5] http://www.catalog.elra.info.

tory containing 14 broad SCFs. We mapped BioCat, the BioLexicon, and the gold standard to this coarse inventory.[6]

A simple relative frequency threshold of 0.03 was used for filtering in both the BioLexicon and BioCat. We did not use SCF-specific filtering in this experiment since it was not available for the BioLexicon.

6.2.3. Evaluation methods

A mapped version of BioLexicon was created using the "best match" mapping from Section 6.2.2. The SCF entries for the verbs in SEM-26 were selected from this lexicon and compared with the gold standard SCFs in SEM-26. The entries for the same verbs were selected from BioCat (filtered with relative frequency filtering) and likewise compared with the gold standard SCFs in SEM-26.

Mapped versions of BioLexicon and of BioCat (again, filtered with relative frequency filtering) were then created using the "coarse" mapping from Section 6.2.2. The SCF entries for the verbs in SEM-26 were selected from these lexicons, and compared with the gold standard SCFs in the "coarse" version of SEM-26.

We report type precision, type recall, and F-score against SEM-26 and the coarse gold standard. We also report the number of SCFs missing from the filtered lexicon, but not the number of SCFs unseen in the unfiltered lexicon, because we did not have access to the unfiltered BioLexicon.

6.3. Direct comparison of semantic and syntactic annotation

The purpose of this experiment was to investigate the semantic and syntactic definitions of subcategorization, by direct comparison of the manually annotated gold standards resulting from the two approaches.

We first compared SYN-10 and SEM-10 using the kappa measure [26]. Kappa is typically used to measure inter-annotator agreement, in the case when multiple annotators perform the same task on the same data. However, in our case the two annotators were given different instructions, so, kappa measures the difference between the two *methods* of annotation, corresponding to the two definitions of subcategorization.

We also compared the number of SCFs per verb in SEM-10 versus SYN-10, to check whether the semantic style of annotation produces a wider variety of SCFs. Since SYN-10/SEM-10 verbs were chosen for their higher number of adjuncts, this measure might over-represent the number of SCFs found in semantic annotation, so we also compared the number of SCFs per verb in all of SEM-30. Finally, we compared these values with the number of SCFs per verb in the general language gold standard of [13], for those verbs in SEM-30 also appearing in [13] (see Table 2, rightmost column).

6.4. Evaluation of BioCat using SEM-10 and SYN-10

The purpose of this experiment was to investigate how the definition of subcategorization used in the gold standard affects the perceived accuracy of an automatically acquired SCF lexicon. We used BioCat for this experiment since its inventory matches that of the gold standards.

For this experiment we used the version of BioCat created with SCF-specific filtering, as described in Section 5.2, since it achieved the best performance and since we were not comparing to another lexicon. The SCF entries for each verb in SYN-10 and SEM-10 were selected from BioCat and compared with the gold standard SCFs in SYN-10 and SEM-10. As in the previous experiments, we report type precision, type recall, and F-score.

[6] To aid future experimentation, the "best match" mapping and the resources mapped to the coarse inventory are included in the public release of materials accompanying this paper.

Table 3
Accuracy of BioCat on SEM-30. Missing SCFs were missing altogether from the filtered lexicon.

Filtering	F-score	Precision	Recall	Missing
0.03 Threshold	44.96	39.37	52.41	13
SCF-specific	59.94	60.87	59.04	11

Table 4
Accuracy of BioCat (threshold 0.03) and the BioLexicon, using best-match, on SEM-26.

Lexicon	F-score	Precision	Recall	Missing
BioCat	46.20	40.00	54.68	11
BioLexicon	58.37	87.14	43.88	20

7. Results and discussion

7.1. Evaluation of BioCat

The accuracy of BioCat on SEM-30 is shown in Table 3. With relative frequency filtering, the system achieves an overall F-score of about 45, with recall favored over precision. Using SCF-specific filtering, the system achieves an F-score of nearly 60 with precision slightly favored over recall. This improvement demonstrates that knowledge about general language SCFs can be useful for filtering in biomedicine. The number of missing SCFs also decreases slightly when using SCF-specific filtering, indicating that this filtering method is more successful at retaining SCFs which are rare but correct. We note that no SCFs were completely unseen in the *unfiltered* lexicon, meaning that the system is capable of finding all the SCFs in the gold standard.

The result for SCF-specific filtering is about 9 points lower than state of the art methods for general language, e.g. [13]. It is a respectable result considering that no adaptations were made to the SCF acquisition system besides applying it to a large biomedical corpus, but it does show that there is a need for adaptation to the biomedical domain.

7.2. Comparative evaluation of BioCat and the BioLexicon

The accuracy of BioCat and the BioLexicon on SEM-26 using "best match" (our first strategy for mapping the disparate SCF inventories of the two resources; Section 6.2.2) is shown in Table 4.[7] We can see that the BioLexicon has a much higher F-score than BioCat even though it uses simple relative frequency filtering, approaching the F-score achieved by BioCat with SCF-specific filtering. Interestingly, we can also see that the BioLexicon strongly favors precision over recall, while BioCat is stronger on recall. The high precision of the BioLexicon is a result of the fact that it is produced with a deep, lexicalized parser already adapted to the biomedical domain, including a POS tagger trained on biomedical text. The input to the SCF classifier thus already takes into account some subcategorization information specific to the biomedical domain. This results in a high-precision system for biomedical text, but relies on up-front domain adaptation, whereas the Cambridge system is less precise but can be ported to new domains as long as there is a large corpus of raw data available.

[7] Note that the figures for BioCat differ from those in the first row of Table 3 because they are for only 26 verbs.

NP-ING:
This study indicates that all treatment protocols seemed to be sufficiently effective and safe and that [$_{NP}$cheyletiellosis in rabbits] can be successfully treated [$_{ING}$using ivermectin or selamectin] in clinical practice.
While the AT immunologic activity is normal in this deficiency, [$_{NP}$plasma AT functional activity] is markedly reduced [$_{ING}$leading to risk of thrombosis].

NP-PP-PP:
[$_{NP}$ The constitutive TRPV1t activity] is inhibited [$_{PP}$to baseline] [$_{PP}$in the presence of SB].
Unlike the Tetrahymena ribozyme, [$_{NP}$the changes] induced [$_{PP}$in precursor RNA] by incubation [$_{PP}$in the absence of divalent cations] result in activation of the ribozyme.

Fig. 5. Examples of SCFs in SEM-26 and BioCat but missing from the BioLexicon.

The higher recall of BioCat likely reflects the fact that it is built from across PMC OA, while the BioLexicon is based on only a single subdomain of biomedicine. The Cambridge system is able to hypothesize SCFs which are likely to be important for interpretation of the text; the trade-off, however, is that the Cambridge system hypothesizes more frames overall, resulting in relatively low precision. This may be overcome in the future, however, with more sophisticated filtering methods, as suggested by the results in Section 7.1.

Fig. 5 shows examples of SCFs found in the SEM-26 gold standard and in BioCat but not in the BioLexicon. Such frames are potentially important for information extraction, demonstrating the importance of recall in SCF acquisition. Note that the BioLexicon may include these frames for other verbs, but at least for the verbs in SEM-26 they were either filtered out or not present to begin with.

The accuracy of BioCat and the BioLexicon using the coarse-grained inventory (our second strategy for mapping SCF inventories; Section 6.2.2) is shown in Table 5. As expected, both lexicons show higher accuracy when evaluated using this more forgiving inventory. The same general trends still hold, however, with the BioLexicon favoring precision while BioCat favors recall.

Note that even using the coarse-grained SCF inventory, the BioLexicon is missing more SCFs from the filtered lexicon than BioCat.

Table 5
Accuracy of BioCat (threshold 0.03) and the BioLexicon using coarse-grained inventory, on SEM-26.

Lexicon	F-score	Precision	Recall	Missing
BioCat	65.38	55.43	79.69	2
BioLexicon	69.23	90.00	56.25	4

THAT-S:
Additionally, our image analysis allowed us to detect [$_{S}$that FTG mice also ventured further into the open arm compared to FNTG controls].
All the caregivers expressed [$_{S}$that the feeling of safety for the patient and the caregiver was essential], emphasizing that professional backup 24 hours a day was important.

ING:
All of these stimuli activated [$_{ING}$signaling] through the MAP kinase/ERK pathway and led to the induction of P-YB-1S102.
Although none of the mutations increased [$_{ING}$binding] to the same degree as removing the entire USH, they had little effect on the solubility of the protein compared to removal of the entire USH.

Fig. 6. Examples of SCFs in the coarse-grained version of SEM-26 and BioCat but missing from the BioLexicon.

Table 6
Agreement between methods using instructions for syntactic and semantic gold standards.

Verb	Kappa score	Instances
activate	0.204022	152
analy (s|z) e	0.214015	227
associate	0.803061	203
compare	0.390602	224
decrease	0.498399	173
express	0.479512	223
improve	0.619959	239
mutate	0.548926	108
occur	0.044539	242
predict	0.311750	172
overall	0.586751	1963

Table 7
Number of SCFs per verb in the different gold standards.

	SEM-10	SYN-10	SEM-30	General language [13]
Low	4	3	1	1
High	10	9	10	25
Average	6.6	5.9	5.4	9.4

Fig. 6 shows examples of coarse-grained frames that were in the coarse-grained gold standard and BioCat, but missing from the BioLexicon.

7.3. Direct comparison of semantic and syntactic annotation

Table 6 shows the results of the kappa agreement test. The overall kappa was 0.58, well below the 0.67 threshold which is considered a minimum for moderate agreement on NLP annotation tasks [27]. The low kappa indicates that the definition of subcategorization has a significant effect on how the resulting gold standards will look. The kappa for some verbs was well below 0.5. Recall, however, that the ten verbs in SYN-10 and SEM-10 were chosen in part for their large number of adjuncts, so the agreement between the syntactic and semantic methods of annotation might be lower for this set of verbs than for others in SEM-30.

The average number of SCFs per verb in the different gold standards is shown in Table 7. SEM-10 had an average of 6.6 SCFs per verb, ranging from a low of 4 to a high of 10; while SYN-10 had an average of 5.9 SCFs per verb, ranging from a low of 3 to a high of 9. This observation suggests that the syntactic method results in annotating slightly fewer frames per verb than the semantic method, which is not surprising since the semantic method takes into account a broader range of phrases that the syntactic method might consider adjuncts. However, the difference is not large.

The average number of SCFs per verb is slightly higher for SEM-10 than SEM-30, which may reflect the fact that the ten verbs were manually selected based on their large number of adjuncts. More interestingly, the average and the maximum number of SCFs per verb is much lower for SEM-30 than for the general language gold standard of [13]. This observation suggests that verb usage becomes specialized in biomedical text, with the range of SCFs being only a limited subset of those observed in general language. Inter-

Table 8
Accuracy of BioCat (with SCF-specific filtering) on semantic and syntactic gold standards, for the ten verbs in the syntactic gold standard.

Gold standard	F-score	Precision	Recall
SEM-10	53.19	40.98	75.76
SYN-10	47.67	36.28	69.49

estingly, this was the case even though the semantic definition of subcategorization was used for SEM-30, and the syntactic for the general language gold standard.

7.4. Evaluation of BioCat using SEM-10 and SYN-10

Table 8 shows the results of evaluating BioCat against SEM-10 and SYN-10. Interestingly, BioCat is more accurate on SEM-10 than SYN-10, despite the fact that it uses syntactic information (parser output) as the input to hypothesis generation. This reflects the fact that the Cambridge system hypothesizes a wide variety of phrases as parts of the SCFs, including some that are considered adjuncts by the linguistic definition of subcategorization, but not by the biomedical definition.

Note that the F-score for SEM-10 is lower than for the full set of 30 verbs in SEM-30 (Table 3); precision in particular is much lower. This is because the small number of verbs provides insufficient evidence across SCFs for the SCF-specific filtering to perform at its best (although it still slightly out-performs threshold filtering).

8. Conclusions

Our study has provided some insights into the current state of verb subcategorization frame acquisition for biomedicine. We have made available the first set of biomedical SCF gold standards suitable for performing quantitative and qualitative evaluation of automatically acquired biomedical SCF resources. Using the Cambridge SCF acquisition system, which was not specifically adapted for biomedicine but applied to a large biomedical corpus, we acquired BioCat, a biomedical SCF lexicon which achieved reasonable results using simple relative frequency filtering. A new method of SCF-specific filtering was found to offer improved accuracy even though it depended on SCF frequency information from general language. Still, SCF acquisition performance drops off considerably compared to general language, losing more than 10 points on F-score, indicating that there is room for adaptation of SCF systems to biomedicine.

We compared two biomedical SCF lexicons, each representing a different aspect of the state of the art in SCF acquisition. We found that the BioLexicon, built with a SCF acquisition system in which each component has been adapted to biomedical text using manually annotated data in the molecular biology subdomain, favored precision over recall when evaluated against our SCF gold standard drawn from across PMC OA. On the other hand, BioCat, built using a state of the art system for general language SCF acquisition and unadapted to biomedical text save for the input corpus, favored recall over precision. The contrast between the two highlights the need for techniques that can acquire SCF information from a broader range of subdomains.

Overall, it can be seen that the accuracy of both BioCat and the BioLexicon against a biomedical gold standard is lower than for general language SCF acquisition against general language SCF gold standards [18,28,13]. We believe the lower accuracy arises from different sources for the two lexicons. BioCat is insufficiently adapted to biomedical text, and hypothesizes a wide variety of SCFs inappropriate for the domain, resulting in low precision. The BioLexicon, on the other hand, suffers from lower recall, which may mean that a system whose components have been manually adapted to a single subdomain does not generalize well enough to the variety of subdomains in PMC OA. New methods for biomedical SCF acquisition are clearly needed in order to create accurate, scalable SCF lexicons to help with downstream NLP tasks, but an approach which relies heavily on manual work will not port easily between different domains. New, minimally supervised SCF acquisition methods such as [29] have recently become available and can be used for acquiring domain- and subdomain-specific SCF lexicons. In addition, some of the best results on SCF acquisition for general language have used information about verb semantic classes to smooth conditional SCF distributions [18], based on the linguistic fact that semantically similar verbs tend to have syntactically similar behavior. This avenue needs further exploration in biomedicine. Incorporating word sense disambiguation may also improve accuracy and understanding of subcategorization in biomedicine, especially since verb behavior in different subdomains may involve overlays of general and specialized senses.

We observed that using two different definitions of subcategorization – the "semantic" definition, which collapses the argument-adjunct distinction, and the "syntactic" definition, which retains it – results in very different styles of annotation, and therefore different evaluation results for an SCF system depending on the definition used in the gold standard. Interestingly, because the Cambridge system, based on a sophisticated SCF inventory, readily hypothesizes many phrase types co-occurring with verbs as part of the SCF, it is more consistent with the semantic definition of subcategorization and achieved higher accuracy on the semantic gold standard than the syntactic one. This behavior may or may not be desirable depending on the application, but needs to be taken into consideration.

Finally, we make the new resources we have created and presented in this article, including our different gold standards and the large BioCat lexicon, publicly available so that they can benefit further research in this area.

Acknowledgments

This work was funded by the EU FP7 Project 'PANACEA', EPSRC (UK) grant EP/G051070/1, and the Royal Society (UK). We gratefully acknowledge Yuval Krymolowski for programming the SCF-specific filtering, and Diane Nicholls for syntactic annotation.

References

[1] Lippincott T, Rimell L, Verspoor K, Korhonen A. Approaches to verb subcategorization for biomedicine. J Biomed Inform, in press.
[2] Ananiadou S, Thompson P, Nawaz R. Improving search through event-based biomedical text mining. In: Proceedings o the first international worksho on automated motif discovery in cultural heritage and scientific communication texts (AMICUS 2010), CLARIN/DARIAH 2010. Vienna, Austria; 2010.
[3] Rupp C, Thompson P, Black W, McNaught J. A specialised verb lexicon as the basis of fact extraction in the biomedical domain. In: Proceedings of interdisciplinary workshop on verbs: the identification and representation of verb features (Verb 2010). Pisa, Italy; 2010.
[4] Sasaki Y, Montemagni S, Pezik P, Rebholz-Schuhmann D, McNaught J, Ananiadou S. BioLexicon: a lexical resource for the biology domain. In: Proc. of the third international symposium on semantic mining in biomedicine (SMBM 2008); 2008.
[5] Rebholz-Schuhmann D, Pezik P, Lee V, Kim JJ, del Gratta R, McNaught YSJ, et al. Towards a reference terminological resource in the biomedical domain. In: Proc. of 16th ann. int. conf. on intelligent systems for molecular biology (ISMB-2008). Toronto, Canada: Oxford University Press; 2008.
[6] Venturi G, Montemagni S, Marchi S, Sasaki Y, Thompson P, McNaught J, et al. Bootstrapping a verb lexicon for biomedical information extraction. In: Gelbukh A, editor. Computational linguistics and intelligent text processing. Lecture notes in computer science, vol. 5449. Berlin/Heidelberg: Springer; 2009. p. 137–48. doi: 10.1007/978-3-642-00382-0-1, URL http://dx.doi.org/10.1007/978-3-642-00382-0-11..
[7] Thompson P, McNaught J, Montemagni S, Calzolari N, Gratta RD, Lee V, et al. The biolexicon: a large-scale terminological resource for biomedical text mining. BMC Bioinformatics 2011;12. 397-397.
[8] Grishman R, Macleod C, Meyers A. COMLEX syntax: building a computational lexicon. In: Proceedings of COLING. Kyoto; 1994.
[9] Wattarujeekrit T, Shah P, Collier N. PASBio: predicate-argument structures for event extraction in molecular biology. BMC Bioinformatics 2004;5.
[10] Tsai RTH, Chou WC, Lin YC, Sung CL, Ku W. BIOSMILE: adapting semantic role labeling for biomedical verbs: an exponential model coupled with automatically generated template features. In: Proceedings of the BioNLP'06 workshop on linking natural language processing and biology. Association for Computational Linguistics; 2005. p. 57–64.

[11] Tsai RTH, Dai HJ, Huang CH, Hsu WL. Semi-automatic conversion of BioProp semantic annotation to PASBio annotation. BMC Bioinformatics 2008;9(Suppl. 12). S18-1.
[12] Cohen KB, Hunter L. A critical review of pasbio's argument structures for biomedical verbs. BMC Bioinformatics 2006;7(Suppl. 3):S5–1.
[13] Preiss J, Briscoe T, Korhonen A. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In: Proceedings of the 45th annual meeting of the association for computational linguistics. Prague, Czech Republic; 2007.
[14] Boguraev B, Carroll J, Briscoe E, Carter D, Grover C. The derivation of a grammatically-indexed lexicon from the longman dictionary of contemporary english. In: Proceedings of the 25th annual meeting of ACL. Stanford, CA; 1987. p. 193–200.
[15] Palmer M, Gildea D, Kingsbury P. The proposition bank: an annotated corpus of semantic roles. Comput Linguist 2005;31(1):71–106. http://dx.doi.org/10.1162/0891201053630264. <http://www.mitpressjournals.org/doi/pdf/10.1162/0891201053630264; URL http://www.mitpressjournals.org/doi/abs/10.1162/0891201053630264>.
[16] Briscoe E, Carrol J, Watson R. The second release of the RASP system. In: Proceedings of the COLING/ACL 2006 interactive presentation sessions. Sydney, Australia; 2006.
[17] Klein D, Manning CD. Accurate unlexicalized parsing. In: Proceedings of ACL; 2003. p. 423–30.
[18] Korhonen A. Subcategorization acquisition. Ph.D. thesis, University of Cambridge Computer Laboratory; 2002.
[19] NIH. The pubmed central open access subset; 2009a. <http://www.pubmedcentral.nih.gov/about/openftlist.html>.
[20] NIH. Journal publishing tag set; 2009b. <http://dtd.nlm.nih.gov/publishing/>.
[21] Lippincott T, O Seaghdha D, Korhonen A. Exploring subdomain variation in biomedical language. BMC Bioinformatics 2011;12(1).
[22] Verspoor K, Cohen KB, Hunter L. The textual characteristics of traditional and open access scientific journals are similar. BMC Bioinformatics 2009;10. 183-1.
[23] Miyao Y, Tsujii J. Feature forest models for probabilistic HPSG parsing. Comput Linguist 2008;34:35–80.
[24] Ohta T, Tsuruoka Y, Takeuchi J, Kim JD, Miyao Y, Yakushiji A, et al. An intelligent search engine and gui-based efficient medline search tool based on deep syntactic parsing. In: Proceedings of the COLING/ACL on interactive presentation sessions (COLING-ACL '06). Stroudsburg, PA, USA: Association for Computational Linguistics; 2006. p. 17–20. doi:http://dx.doi.org/10.3115/1225403.1225408.
[25] Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus – a semantically annotated corpus for bio-textmining. Bioinformatics 2003;19(Suppl. 1):i180–2.
[26] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
[27] Krippendorff K. Content analysis: an introduction to its methodology. Beverly Hills, CA: Sage Publications; 1980.
[28] Korhonen A, Krymolowski Y, Briscoe T. A large subcategorization lexicon for natural language processing applications. In: Proceedings of LREC; 2006.
[29] Lippincott T, O Seaghdha D, Korhonen A. Learning syntactic verb frames using graphical models. In: Proceedings of the 50th annual meeting of the association for computational linguistics (ACL). Jeju, Korea; 2012.