Bootstrapping a Verb Lexicon for Biomedical Information Extraction

Giulia Venturi[1], Simonetta Montemagni[1], Simone Marchi[1],
Yutaka Sasaki[2,3], Paul Thompson[2,3], John McNaught[2,3] and Sophia Ananiadou[2,3]

[1] Istituto di Linguistica Computazionale, CNR, Pisa, Italy
[2] School of Computer Science, University of Manchester, UK
[3] National Centre for Text Mining, University of Manchester, UK
{giulia.venturi, simonetta.montemagni, simone.marchi}@ilc.cnr.it
{yutaka.sasaki, paul.thompson, jock.mcnaught, sophia.ananiadou}@manchester.ac.uk

Abstract. The extraction of information from texts requires resources that contain both syntactic and semantic properties of lexical units. As the use of language in specialized domains, such as biology, can be very different to the general domain, there is a need for domain-specific resources to ensure that the information extracted is as accurate as possible. We are building a large-scale lexical resource for the biology domain, providing information about predicate-argument structure that has been bootstrapped from a biomedical corpus on the subject of E. Coli. The lexicon is currently focussed on verbs, and includes both automatically-extracted syntactic subcategorization frames, as well as semantic event frames that are based on annotation by domain experts. In addition, the lexicon contains manually-added explicit links between semantic and syntactic slots in corresponding frames. To our knowledge, this lexicon currently represents a unique resource within in the biomedical domain.

Keywords: domain-specific lexical resources, lexical acquisition, syntax-semantics linking, Information Extraction, Biological Language Processing

1 Introduction

It is well known that Information Extraction applications require sophisticated lexical resources to support their processing goals. In particular, accurate applications focussed on extraction of event information from texts require resources containing both syntactic and semantic information. Many applications could benefit from lexical resources providing an exhaustive account of the semantic and syntactic combinatorial properties of lexical units conveying event information.

The need for such resources increases when dealing with texts belonging to a specialized domain such as biology. There are several reasons for requiring domain-specific lexical resources. Even more than in general language, within specialized domains, much lexical knowledge is idiosyncratically related to the individual behavior of lexical units. In particular, it can be the case that the types of events mentioned

in domain-specific texts are described using predicates that do not feature prominently in the general language domain and may not be included in general language resources. Or, in the reverse case, predicates that do occur in the general language domain may have different syntactic or semantic properties within the specialized domain. Using information about such predicates from general language resources may result in incorrect analyses or interpretations.

The lexical component still remains a major bottleneck for current Information Extraction systems, especially when the target is event information in domain-specific collections of documents. So far, most lexical resources providing information on predicate-argument structure have been developed manually by lexicographers. It is, however, a widely acknowledged fact that manual work is costly and the resulting resources have limited coverage. Last but not least, porting to new domains is a labour-intensive task. Automatic or semi-automatic lexical acquisition is a more promising and cost-effective approach to take, and is increasingly viable given recent advances in NLP and machine learning technology, together with availability of corpora.

In the European BOOTStrep project (FP6 - 028099), we are building a large-scale domain-specific lexical resource [1] also providing information about predicate-argument structure that is bootstrapped from texts. The topic of this paper is the bootstrapping of predicate-argument structure information from biomedical corpora; in particular, we focussed on *verbs*, for which syntactic subcategorization and semantic event frames have been acquired from a biomedical corpus on the subject of E. Coli. Subcategorization extraction has been carried out through unsupervised learning operating on the dependency-annotated text without relying on any previous lexico-syntactic knowledge about subcategorization frames. Semantic frames are currently based on a subset of the corpus used for subcategorization extraction, which has been manually annotated with gene regulation bio-events by domain experts. The two sets of frames were obtained independently, resulting in two different and unrelated sets of subcategorization and semantic event frames. On the two sets of frames acquired for the same verbs, the syntax-semantics linking was performed manually. The resulting verb lexicon thus includes subcategorization and semantic frames information as well as the explicit linking between semantic and syntactic slots in corresponding frames. To our knowledge, such a lexicon currently represents a unique resource in the biomedical domain, which has the potential to effectively support event extraction from biomedical texts.

The paper is organized as follows: section 2 provides the background and the motivation of our work, whilst section 3 outlines our approach to lexicon construction. Sections 4 and 5 report respectively on the processes of subcategorization induction and event frame extraction. Section 6 concerns the linking of the acquired syntactic and semantic frames. Conclusions and further work are reported in section 7.

2 Background

Various research groups are currently concerned with the creation of corpus-based general-purpose lexical semantic resources providing information on predicate-argument structure; see for instance the FrameNet [2] and PropBank [3] projects.

The FrameNet project, following Fillmore's theory of *frames semantics* [4], is creating an on-line lexical resource supported by corpus evidence. It documents the range of semantic and syntactic combinatory possibilities of each word in each of its senses, through computer-assisted annotation of example sentences and automatic tabulation and display of the annotation results. One of the major outcomes of this work is represented by the FrameNet lexical database, in which each predicative lexical unit (i.e. verb, noun or adjective) is paired with a semantic frame, i.e. a conceptual structure describing a particular type of situation or event along with its participants. For example, the lexical entry for the verb *construct* identifies the semantic frame underlying its meaning, which is "Building", and whose core frame elements are Agent, Created_entity, Components. The lexical entry also specifies the ways in which frame elements are syntactically realised in texts.

A slightly different approach has been followed within the PropBank project. Both a corpus of one million words of English text, annotated with argument role labels for verbs on the top of the Penn-II syntax trees, together with a lexicon defining those argument roles on a per-verb basis, have been created. For example, the predicate-argument structure of the verb *construct* has been annotated with the following numbered arguments: ARG0 (i.e. builder), ARG1 (i.e. construction), ARG2 (i.e. material), ARG3 (i.e. end state of ARG1).

In response to the requirement for domain-specific lexical resources, a number of attempts have been made to produce domain-specific extensions of the resources described above, e.g. BioFrameNet [5] and PASBio [6]. BioFrameNet is a domain-specific FrameNet extension, mainly focused on the domain concepts of intracellular transport. PASBio, extending a model based on PropBank to molecular-biology domain, takes the role of a reference resource in the stage of corpus annotation for creating training examples for machine learning (i.e. Event Extraction). Currently, these resources are reasonably small-scale (PASBio currently contains 30 predicates, whilst BioFrameNet was carried out as dissertation work).

To our knowledge, the only existing computational lexicon specifically developed for the biomedical domain is the SPECIALIST lexicon [7]. Unlike the previously mentioned cases, the lexicon is built and maintained manually and is not corpus-driven. It is a large lexicon of general English words and biomedical vocabulary, designed to provide the lexical information needed for the SPECIALIST Natural Language Processing System (NLP). Lexical entries in this lexicon also include verb complementation patterns providing important syntactic information.

3 Our Approach

We are building a *verb lexicon* to address the requirement for a large-scale resource that is specific to the biomedical domain, and includes *both* syntactic subcategorization *and* semantic event frame information. Our approach to the construction of the lexicon has a number of defining features, which set it apart from the other resources described above.

Firstly, in contrast to the SPECIALIST lexicon, our own lexicon construction technique is corpus-based. This ensures that the most relevant verbs are included within the lexicon, and their encoded behaviour is domain-specific.

Secondly, in contrast to the purely manual construction method of many other lexical semantic resources, the information in our lexicon has been derived semi-automatically, using different techniques and different sizes of corpora to obtain each type of information. The extraction of subcategorization frames was carried out using an unsupervised learning technique, using a dependency annotated corpus of approximately 6 million tokens (consisting of both MEDLINE abstracts on the subject of E.Coli, in addition to full papers). In contrast, event extraction was carried out based on a subset of this corpus (677 abstracts), which was manually annotated with bio-event information. This annotation was carried out on top of linguistic annotations covering morphosyntax and shallow syntax ("chunking"). The final step of the process was to link the syntactic arguments of predicates to their semantic counterparts in the event frames, thus facilitating the automatic labelling of syntactic arguments of verbs with semantic roles. In the current work, this linking step has been carried out manually.

In the following sections, we discuss the different techniques of obtaining syntactic and semantic information for inclusion within the lexicon, together with the merging and linking of the results.

4 Extraction of Subcategorization Frames

For the purposes of the extraction of subcategorization frames (hereafter referred as SCFs), we adopted a "discovery" approach to SCF acquisition, based on a looser notion of subcategorization frame, which includes typical verb modifiers in addition to strongly selected arguments. Such an approach took into account the desideratum within the biomedical field that subcategorization patterns should also include strongly selected modifiers (such as location, manner and timing), as these are deemed to be essential for the correct interpretation of texts [8].

In order to meet this basic requirement, we used the Enju syntactic parser for English [9][1], characterised by a wide-coverage probabilistic HPSG grammar and an efficient parsing algorithm, and whose output is returned in terms of predicate-argument relations. In particular, we used the Enju version adapted to biomedical texts [10]. The SCF induction process was performed through the following steps:

- syntactic annotation of the acquisition corpus with Enju (v2.2). The acquisition corpus included both MEDLINE abstracts and full papers containing a total of approximately 6 million word tokens;
- for each verbal occurrence, extraction of the observed dependency sets (ODSs). Each ODS is represented as a set of dependencies described in terms of relation type (e.g. ARG1, ARG2, etc.) complemented in some cases with information concerning the morpo-syntactic category of the head (this information type is useful to further specify generic dependency relations like MOD). For what concerns prepo-

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/enju/

sitional and sentential complements, rather than using the general Enju labels (i.e. ARG1, ARG2), a representation was reconstructed in which the preposition or conjunction introducing the complement was made explicit: due to its crucial role in the subcategorization induction process, this information type is part of the dependency label (e.g. PP-in or that-CL) used in the ODS. The order of the dependencies in each ODS is normalised and does not reflect their order of occurrence in context;

- induction of relevant SCF information associated with a given verb. For each observed dependency set, the conditional probability given the verb type *v* was computed: thresholding was used to filter out noisy frames (i.e. frames containing not only arguments and strongly selected modifiers, but also adjuncts) as well as possible errors of either parsing or ODS extraction. After careful examination of the results obtained with different thresholds, ODSs with an associated probability score >= 0.03 were selected as eligible SCFs to be included in the resulting verb lexicon.

For each acquired SCF, the following information types are specified: its conditional probability given the verb (i.e. "p(subcat|v)") and the percentage of times it occurs with the verb in the passive voice (i.e. "Pass"). It should be noticed that each SCF has been extracted for one *normalised verb token*, i.e. the extraction process makes abstraction from the passive usages. Thus, the latter information is particularly useful to account for SCFs typically associated with the verb used in the passive voice; this is the case, for instance, of the SCFs ARG1#ARG2#TO-INF# and ARG1#ARG2#that-CL# frames which with the verb *find* appear to be typically associated with the verb used in the passive voice (e.g. *This was found to be interesting*). Such information has been exploited during the syntax-semantics linking in order to reconstruct the full syntactic realisations of bio-verb arguments even though some of them do not have any semantic counterpart explicitly mentioned in the text.

Table 1. Subcategorization frame examples

Verb	SFC	p(subcat|v)	Pass
abolish	ARG1#ARG2#	0.8669767	0.1437768
abolish	ARG1#ARG2#MOD@VBG#	0.0390697	0.1904761
abolish	ARG1#ARG2#PP-in#	0.0939534	0.7029702
accumulate	ARG1#ARG2#	0.2940677	0.0403458
accumulate	ARG1#	0.4627118	0
accumulate	ARG1#ARG2#PP-in#	0.1084745	0.140625
accumulate	ARG1#PP-in#	0.1347457	0

5 Event Frame Extraction

This section briefly describes the automatic extraction of semantic event frames based on a corpus of 677 MEDLINE abstracts. The abstracts have been annotated with Gene Regulation events by a group of domain experts [11]. Annotation is centered on both

verbs and nominalised verbs that describe relevant events within the corpus. For each event, semantic arguments that occur within the same sentence are labelled with semantic roles (see Table 2) and Named Entity types.

Table 2. Semantic roles

Role Name	**Description**	**Example** (**bold** = semantic argument, *italics* = focussed verb)
AGENT	Drives/instigates event	**The narL gene product** *activates* the nitrate reductase operon
THEME	a) Affected by/results from event b) Focus of events describing states	**recA protein** was *induced* by UV radiation The **FNR protein** *resembles* CRP
MANNER	Method/way in which event is carried out	cpxA gene *increases* the levels of csgA transcription by **dephosphorylation** of CpxR
INSTRUMENT	Used to carry out event	EnvZ *functions* through **OmpR** to control NP porin gene expression in E. Coli.
LOCATION	Where *complete* event takes place	Phosphorylation of OmpR *modulates* expression of the ompF and ompC genes in **Escherichia coli**
SOURCE	Start point of event	A transducing lambda phage was *isolated* from **a strain** harboring a glpD''lacZ fusion
DESTINATION	End point of event	Transcription is activated by *binding* of the cyclic AMP (cAMP)-cAMP receptor protein (CRP) complex to **a CRP binding site**
TEMPORAL	Situates event in time/ w.r.t another event	The Alp protease activity is *detected* in cells **after introduction** of plasmids
CONDITION	Environmental conditions/changes in conditions	Strains carrying a mutation in the crp structural gene fail to *repress* ODC and ADC activities in response to **increased cAMP**
RATE	Change of level or rate	marR mutations *elevated* inaA expression by **10- to 20-fold** over that of the wild-type.
DESCRIPTIVE-AGENT	Descriptive information about AGENT of event	HyfR *acts* as **a formate-dependent regulator**
DESCRIPTIVE-THEME	Descriptive information about THEME of event	The FNR protein *resembles* **CRP**.
PURPOSE	Purpose/reason for the event occurring	The fusion strains were *used* **to study** the regulation of the cysB gene

We chose to use a set of 13 *event-independent* semantic roles, which were defined specifically for the task though the examination of a large number of relevant events

in E. Coli abstracts. Event-independent semantic roles have previously been used in large-scale projects involving the production of semantic frames for general language verbs, e.g. VerbNet [12] and SIMPLE [13]. However, to our knowledge, our work is the first to propose a set of event-independent roles for use within the biological domain.

We used VerbNet and SIMPLE as a starting point for the definition of our role set, with the assumption that certain semantic roles are common across all domains. This assumption was confirmed through examination of examples within our corpus, resulting in our use of roles such as AGENT, THEME, and SOURCE. Whilst some general language roles do not seem relevant to the description of biological events (such as BENEFICIARY or EXPERIENCER), others are particularly important to the precise definition of complex biological relations, even though not necessarily specific to the field, e.g. LOCATION and TEMPORAL (see [8]). To the subset of relevant roles identified from VerbNet and SIMPLE, we added the role CONDITION. This corresponds to descriptions of environmental conditions, which are highly important within the domain.

5.1 Event Annotation Spans

An event annotation span is a continuous annotation associated with the same event id within an abstract. An event annotation span begins with the text span covered by the earliest semantic argument, and ends with the latest semantic argument associated with the event within the text.

For example, given the sentence "transfer operon expresses F-like plasmids", its event annotation span is as follows:

```
<SLOT eventid="9" Role="Agent"> <NE cat="DNA"> transfer
operon</NE></SLOT> <EVENT id="9"><SLOT eventid="9"
Role="Verb"> expresses </SLOT></EVENT> </SLOT> <SLOT
eventid="9" Role="Theme"> <NE cat="DNA"> F-like plas-
mids </NE></SLOT>
```

5.2 Syntactic Analysis of Event Annotation Spans

For each event, each event annotation span is syntactically analyzed as follows:

- Tokenize the span into XML tags and words where named entities (NEs) are treated as single words.
- Decide on the POS tags and lemmas of tokens. For words occurring outside of NE spans, “O” is assigned as the value of the NE category field. NEs are assigned “NN” as the value of the POS field.
- Add semantic role labels to words and NEs based on the IOB labelling scheme. That is, add B-*role* to the first word in the *role* annotation, and I-*role* to the following words in the annotation.

For example, the sentence introduced above is analyzed as shown in Table 3.

Table 3. Example syntactic analysis of event annotation span

word	POS	lemma	NE	Role
transfer operon	NN	transfer operon	DNA	B-Agent
expresses	VBZ	express	O	B-Verb
F-like plasmids	NN	F-like plasmids	DNA	B-Theme

5.3 Event Frames

Event frames take the following general form:

```
event_frame_name(
      slot_name => slot_value,
      …
      slot_name => slot_value),
```

where

- `event_frame_name` is the base form of the event verb or nominalized verb;
- `slot_names` are the names of the semantic roles within the event pattern;
- `slot_values` are NE categories, if they have been assigned within the event pattern.

5.4 Event Frame Extraction

Converting syntactically analyzed event annotation spans to semantic event frames is straightforward.

- the event frame name is the lemma of the verb;
- for each semantic role (starting with a B-*role* label and followed by I-*role* labels), use its NE as the slot value, if an NE has been assigned.

For example, the event frame corresponding to the above event annotation span example is as follows:

```
express( Agent=>DNA,
            Theme=>DNA ).
```

6 Syntax-Semantics Linking

The syntax-semantics linking was carried out manually on the basis of different information types. The starting point of this process was represented by:

- the list of 1760 subcategorization frames, acquired from the Enju annotated corpus (see section 4);
- the list of 856 verbal bio-event frames based on annotations in the Gene Regulation corpus (see section 5); it should be noticed that for the linking purposes we took

into account bio-event frames including both slots which specify a named entity category, as well as those slots which do not specify such information.

The linking focussed on 168 verbs for which both subcategorization and event frame information was available, in particular on the 628 subcategorization frames and the 486 bio-event frames extracted for those verbs.

The linking process was carried out manually and it was defined by simultaneously taking into account different information types, in particular:

- we considered that a syntax-semantic mapping process is controlled by strategies which presuppose hierarchies of semantic roles and grammatical functions.
- we made use of a list of 'prototypic' syntactic realisations of semantic arguments, as provided in the annotation guidelines followed by annotators during the manual annotation of bio-event frames (provided in [14]).
- we exploited general language repositories of semantic frames containing both syntactic and semantic information as possible benchmarks,
- we also referred to the manually annotated Gene Regulation Corpus, when the evidence of the other information sources was not sufficient to perform the syntax-semantics mapping.

Firstly, we analysed the literature regarding syntax-semantics linking, according to which "Thematic Hierarchies" appear to be by far the most widely used method to explain the mapping from semantic representation to syntax. A hierarchy of "cases" (semantic relations) was first formulated by Fillmore [15] to help determine subject selection. After him, most theories make use of a mapping between an ordered list of semantic roles and an ordered list of grammatical relations. Thus, rather than having invariable correspondence relations, these approaches suggest that, given a thematic role hierarchy (agent>theme ...) and a syntactic functions hierarchy (subject>object ...), the mapping usually proceeds from left to right, mapping the semantic role further to the left onto the first available position in the syntactic hierarchy. Several proposals have been made for what concerns the thematic role hierarchy which widely differ a) with respect to the theoretical stands and b) in what is being hierarchisized. If on the one hand there is general agreement on the fact that the Agent role should be the highest ranking role, on the other hand no consensus is found in the literature (see [16] for a survey of the wide range of proposals) for what concerns the relative ordering of the remaining roles.

Another important source of information was represented by the 'prototypic' syntactic realizations of semantic arguments as defined in the annotation guidelines for event annotation in the Gene Regulation Corpus, especially for what concerns less prominent roles, typically expressed as prepositional phrases. In order to solve doubtful mapping cases, general language repositories of semantic frames containing both syntactic and semantic information were also consulted. Amongst others, we choose to exploit VerbNet [12] because, similarly to our own work (see section 5), it uses a set of frame-independent thematic roles. The Gene Regulation corpus was also taken as a further source of evidence: in particular, it was useful in dealing with verbs that do not feature in a general language repository of frames or that may have different syntactic realisations or different semantic properties within the biomedical domain.

The linking process resulted in 668 linked frames. Different types of mapping were performed, namely full and partial mapping. In full mapping cases, the arity of the

subcategorization and bio-event frames is the same; that is to say that all semantic arguments of the bio-event frame have a syntactic counterpart at the level of the subcategorization frame. For what concerns partial mapping, we distinguished the following sub-cases:

1. the semantic frame contains more slots (i.e. semantic roles) than the corresponding subcategorization frame. In these cases, a mapping could only be defined for a subset of the semantic roles in the bio-event frame. For example, for the verb *express*, for which the semantic frame Agent#Theme#Location#Condition# and the subcategorization frame ARG1#ARG2#PP-in# have been acquired the following mapping has been defined:
   `AGENT>ARG1#THEME>ARG2#LOCATION>PP-in#`**`CONDITION>0`**
2. subcategorized slots do not find a semantic counterpart in the corresponding bio-event frame. This is typically the case of event frames which did not contain explicit mention of an AGENT role, which however has been reconstructed as ARG1 at the level of the subcategorization frame: this applies most frequently to passive sentences such as *The wild-type pcnB gene was cloned into a low-copy-number* plasmid, whose Enju normalised syntactic representation includes a reconstructed ARG1 which does not correspond to any filled semantic argument of the corresponding bio-event frame. Consider as an example the verb *introduce*, for which the semantic frame Theme#Destination# and the subcategorization frame ARG1#ARG2#PP-into# have been extracted; in this case the mapping presents itself as follows:
   `0>ARG1``#THEME>ARG2#DESTINATION>PP-into`
3. a combination of cases 1) and 2) above, i.e. where the semantic frame contains more slots than the corresponding subcategorization frame on the one hand, and a reconstructed ARG1 does not have any counterpart at the level of the semantic frame on the other hand. Consider as an example the verb *delete*, for which the following mapping has been defined, operating respectively on the ARG1#ARG2#PP-from# and Theme#Source#Condition# subcategorization and event frames:
   `0>ARG1``#THEME>ARG2#SOURCE>PP-from#`**`CONDITION>0`**

Table 4 below summarises the results of the linking process. Note that 28 extracted bio-event frames were discarded since they turned out to originate from errors during the semantic annotation process.

Table 4. Syntax-semantics linking results

Type of mapping		Number of cases	%
Full mapping		239	35.77
Partial mapping	Sub-case 1	123	18.42
	Sub-case 2	166	24.86
	Sub-case 3	140	20.95
	TOTAL	668	100.00

7 Conclusion

In this paper, we have described the bootstrapping of a verb lexicon for Biomedical information extraction. The verb lexicon includes both syntactic subcategorization frames and semantic event frames, together with a bridge between the two levels.

The information within the lexicon is the result of integrating information extracted from corpora of different sizes and using different techniques. Syntactic subcategorization frames were acquired from an automatically annotated corpus (dependency annotation) of 6 million word tokens, using unsupervised learning. On the other hand, event frames were extracted from a subset of this corpus (677 MEDLINE abstracts) that was manually annotated by biologists. The link between the syntactic and semantic levels of information was also carried out manually.

The syntax-semantics linking was carried out on 168 biologically relevant verbs, for which both subcategorization and event frame information was available. A total of 628 subcategorization frames and the 486 bio-event frames had been extracted for those verbs. As a result of this linking process, 668 event frames have been fully or partially linked to subcategorization frames.

To our knowledge, the number of verbs covered by our lexicon, together with the typology of information that is available for each verb, make our resource unique amongst large-scale computational lexicons within the biomedical domain.

We are currently working on an extrinsic evaluation of the syntactic/semantic frames in bio-event IE tasks. The verb lexicon is an essential resource in these IE tasks, and is utilized as follows:

- Analyze bio-event text using the Enju full parser;
- Find predicate-argument structures that match subcategorization frames in the verb lexicon;
- Using the linking tables, map the matched predicate-argument structures to semantic event frames;
- Finally, by applying event frames to these semantic frames, event instances can be extracted.

In addition to events that are centred on verbs, our event frame corpus includes annotations corresponding to events that are centred on *nominalised verbs* such as *regulation* and *expression.* As events expressed in such a way play an important and possibly dominant role within biomedical texts [17], we plan to acquire subcategorization frame information for the annotated nominalised verbs, and link them to the event frames in the same way as for verbs. Further future work will include the investigation automatic or semi-automatic methods of linking together the syntactic subcategorization frames and semantic event frames.

8 Acknowledgements

The work described in this paper has been funded by the European BOOTStrep project (FP6 - 028099). The National Centre for Text Mining is sponsored by the JISC/BBSRC/EPSRC.

References

1. Rebholz-Schuhmann, D., Pezik, P., Lee, V., Kim, J-J., del Gratta, R., Sasaki, Y., McNaught, J., Montemagni, S., Monachini, M., Calzolari, N., Ananiadou, S.: BioLexicon: Towards a Reference Terminological Resource in the Biomedical Domain. Proc. of 16th Ann. Int. Conf. on Intelligent Systems for Molecular Biology (ISMB-2008), Toronto, Canada (2008)
2. Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C. Scheffczyk, J.: FrameNet II: Extended Theory and Practice, Available online at: http://framenet.icsi.berkeley.edu/ (2006)
3. Palmer, M., Kingsbury, P., Gildea, D.: The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics, 31(1), 71-106 (2005)
4. Fillmore, C.J.:Frame semantics and the nature of language. In: Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, Volume 280, pp. 20-32 (1976)
5. Dolbey A., Ellsworth M., Scheffczykx J:. BioFrameNet: A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies. In: O. Bodenreider (Ed.), Proceedings of KR-MED, pp 87-94 (2006)
6. Wattarujeekrit, T., Shah, P., Collier, N.: PASBio: predicate-argument structures for event extraction in molecular biology. BMC Bioinformatics 5:155 (2004)
7. Browne A.C., Divita, G., Aronson, A.R and McCray, A.T : UMLS Language and Vocabulary Tools. In: Proceedings of AMIA Annual Symposium, p.798 (2003)
8. Tsai R.T.H, Chou W.C., Su Y.S., Lin Y.C., Sung C.L., Dai H.J, Yeh I.T.H., Ku W, Sung T.Y . Hsu W.L.: BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. *BMC Bioinformatics* 8:325 (2006)
9. Miyao, Y, Ninomiya, T., Tsujii, J.: Corpus-oriented Grammar Development for Acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank**.** In: Proceedings of IJCNLP-04, pp 684 -693 (2004)
10. Hara, T., Miyao, Y., Tsujii, J.: Adapting a probabilistic disambiguation model of an HPSG parser to a new domain . In: Proceedings of IJCNLP, pp 199-210 (2005)
11. Thompson, P., Cotter, P., Ananiadou, S., McNaught, J., Montemagni, S., Trabucco, A., Venturi, G.: Building a Bio-Event Annotated Corpus for the Acquisition of Semantic Frames from Biomedical Corpora. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008) (2008)
12. Kipper-Schuler, K.: VerbNet: A broad-coverage, comprehensive verb lexicon. PhD. Thesis. Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA. (2005)
13. Lenci A., Busa F., Ruimy N., Gola E., Monachini M., Calzolari N., Zampolli A. et al.: SIMPLE Linguistic Specifications LE-SIMPLE (LE4-8346), Deliverable D2.1 & D2.2. ILC and University of Pisa (2000)
14. Montemagni, S., Trabucco, A., Venturi, G., Thompson, P., Cotter, P., Ananiadou, S., McNaught, J. Kim, J.-J., Rebholz-Schuhmann, D., Pezik, P.: Event annotation of domain corpora, BOOTStrep (FP6 – 028099), Deliverable 4.1. University of Manchester, ILC-CNR and European Bioinformatics Institute (2007)
15. Fillmore, C.J.: The case for case. In: Bach, E., Harms, R.T. (eds.) Universals in Linguistic Theory, pp. 1-88. New York: Holt, Rinehart, and Winston (1968)
16. Levin, B, Rappaport Hovav, M.: Lexical Semantics and Syntactic Structure. In Lappin, S. (ed.) The Handbook of Contemporary Semantic Theory, pp. 487-507. Blackwell, Oxford (1996)
17. Cohen, K.B, Hunter, L.: A critical review of PASBio's argument structures for biomedical verbs. BMC Bioinformatics 7 (Suppl. 3), S5 (2006)