The GENEREG Corpus for Gene Expression Regulation Events — An Overview of the Corpus and its In-Domain and Out-of-Domain Interoperability

Ekaterina Buyko, Elena Beisswanger, Udo Hahn

Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena
Fürstengraben 30, 07743 Jena, Germany
{ekaterina.buyko|elena.beisswanger|udo.hahn}@uni-jena.de

Abstract

Despite the large variety of corpora in the biomedical domain their annotations differ in many respects, e.g., the coverage of different, highly specialized knowledge domains, varying degrees of granularity of the targeted relations, the specificity of linguistic grounding of relations and named entities referred to in the documents, etc. We here introduce GENEREG (*Gene Regulation Corpus*), the result of an annotation campaign led by the Jena University Language & Information Engineering (JULIE) Lab. The GENEREG corpus consists of 314 abstracts dealing with the regulation of gene expression in the model organism *E. coli*. Our emphasis in this paper is on the compatibility and, thus, linkage, of the GENEREG corpus with the alternative GENIA event corpus and with several in-domain and out-of-domain lexical resources, e.g., the SPECIALIST LEXICON, FRAMENET, and WORDNET. The links we established from the GENEREG corpus to these external resources will help improve the performance of the automatic relation extraction engine JREX trained and evaluated on GENEREG.

1. Introduction

We currently witness a proliferation of semantically annotated corpora, containing named entity and relation annotations, particularly in the area of biomedical language processing. As examples, we here only mention the AIMED corpus (Bunescu et al., 2005), the LLL corpus (Nédellec, 2005), the BIOINFER corpus (Pyysalo et al., 2007), and, most recently, the *BioNLP'09 Shared Task* corpus (Kim et al., 2009) gathered from the GENIA event corpus (Kim et al., 2008) for the *BioNLP'09 Shared Task Challenge on Event Extraction*.[1]

These corpora are characterized by highly specialized named entity or relation types still covering only tiny little bits of the vast domain knowledge in the life sciences. While this seems inevitable for such supersized domains with a huge span of 'interesting' entities and relations, one might think of docking such specialized corpora to complementary resources in order to increase their usability.

We, first, introduce GENEREG (*Gene Regulation Corpus*), the result of an annotation campaign led by the Jena University Language & Information Engineering (JULIE) Lab. The corpus, a preliminary version was described by Buyko et al. (2008), is suited for the extraction of relations between entities whose focus is on the regulation of gene expression. Its structure and some quantitative characteristics will be described in Section 3.

In an effort to consolidate our work on GENEREG, we further explored ways to link the metadata it contains to other, independently developed annotated corpora or lexical resources. Such concerns for empowering the interoperability of language resources were made concrete by linking GENEREG with the GENIA event corpus on the one hand, and with lexicon resources such as the SPECIALIST LEXICON,[2] WORDNET,[3] FRAMENET,[4] etc., on the other hand. These efforts will be described in Section 4.

2. Related Work

One of the key issues in the BioNLP domain is the extraction and interpretation of relations (RE) between named entities. For the development and evaluation of RE methodologies, a large variety of RE annotated corpora were developed in different labs. Most of these corpora contain protein-protein interaction (PPI) annotations, e.g., AIMED, BIOINFER, LLL, but they also differ in many respects (Pyysalo et al., 2008). While some corpora provide untyped undirected annotations (AIMED), other corpora employ annotations based on ontological definitions (BIOINFER). Another major difference between these corpora is the level of detail, i.e., the granularity of annotations. Finally, only few corpora provide the marking of key words that stand at the textual level for the conceptualization of an interaction between named entities (e.g., BIOINFER, GREC (Thompson et al., 2009)).

In recent challenges such as the *BioNLP'09 Shared Task*, these issues gained further attention. While the granularity of relations used for corpus annotation is increasing (e.g., from *PPI* to *Binding*, *Regulation*, etc.), the annotations are also grounded in the explicit linguistic expressions being used for the denotation of relations (e.g., various interaction types are linked to the key interaction words).

Pyysalo et al. (2008) showed that the majority of annotated interactions from various RE corpora fall under the *Causal-Change* sub-tree of the BIOINFER ontology (Pyysalo et al., 2007). These interactions correspond to events occur-

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

[2] http://www.nlm.nih.gov/pubs/factsheets/umlslex.html

[3] http://wordnet.princeton.edu/

[4] http://framenet.icsi.berkeley.edu/

ring as part of biochemical processes in cells. The interaction annotations from different corpora vary thereby at the level of granularity. Let us consider an example from the AIMED corpus. *"p53 transcriptional activation mediated by co-activators TAFII40 and TAFII60."* contains an annotation of two PPI interaction relations, namely *PPI(p53, TAFII40)* and *PPI(p53, TAFII60)*. However, more specifically, *TAFII40* and *TAFII60* regulate the transcription of *p53*. The interactions mentioned in the sentence finally boil down to two molecular events, namely *Positive Regulation* and *Transcription*, which are more specific than the more general interpretation as *PPI*. It is exactly this kind of more precise information biologists are looking for.

Obviously, PPI extraction is a complex task since various molecular events and even cascades of events are involved both of which are hard to sort out. This observation holds for the GENEREG corpus as well. For example, the sentence "*XapR regulates the expression of xanthosine phosphorylase (XapA).*" contains a *Regulation of Gene Expression* relation between *XapR* and *XapA* that can be represented by means of two cascaded GENIA corpus events, i.e., *Regulation*(`Theme`:*Gene Expression*(`Theme`:XapA), `Cause`:XapR).

Given the inherent complexity of relation extraction at such a fine level of distinction RE may benefit from a methodological approach which deals with the extraction of molecular events in a bottom-up manner. This way, the general PPI problem can be decomposed into more specific and, possibly, more feasible subtasks. To put this work on firmer empirical ground, we wanted to consolidate our work on the GENEREG corpus by making it compatible with the GENIA event corpus.

In the BioNLP'09 Challenge, the tagging of key words (called triggers) has been shown to be particularly useful for the extraction and interpretation of fine-grained interactions between entities. Similarly, Buyko et al. (2008) gathered evidence that the extraction of trigger words considerably boosts an RE system's performance in terms of its f-score. Thus, the need arises to link the annotations of trigger words from different corpora in a standardized way.

Some attempts have already been made in this direction in the PASBIO (Wattarujeekrit et al., 2004) and BIOFRAMENET (Dolbey et al., 2006) projects both of which, unfortunately, cover only a very small portion of biomedical verbs. We here advocate the consideration of larger-scale lexical resources such as the SPECIALIST LEXICON and the BIOLEXICON (Sasaki et al., 2008) which offer a substantial number of subcategorization frames for biomedical verbs. Turning our attention to the general language domain, we additionally find lexical resources such as WORDNET, VERBNET, FRAMENET that (though differing in the scope of their lexical coverage) might be beneficial for the biomedical domain as well.

In this study, we consolidate the GENEREG corpus by elaborating on its compatibility with the GENIA event corpus annotation guidelines and by linking the most frequent trigger words from both corpora to selected in-domain (i.e., biological) and out-of-domain (non-biological) lexicon and corpus resources.

3. Corpus Annotation

The GENEREG corpus consists of 314 PUBMED[5] abstracts dealing with the regulation of gene expression in the model organism *E. coli*. The regulation of gene expression can be described as the process that modulates the frequency, rate or extent of gene expression. During gene expression, the coding sequence of a gene is converted into a mature gene product or products, namely proteins or RNA (taken from the definition of the Gene Ontology class *Regulation of Gene Expression*, GO:0010468).[6] GENEREG provides three types of semantic annotations: *named entities* involved in gene regulatory processes, such as transcription factors and genes, pairwise *relations* between regulators and regulated genes, and *event triggers* (clue verbs) essential for the description of, e.g., gene expression and gene regulation events.[7] For all three annotation levels, the annotation vocabulary was taken from the *Gene Regulation Ontology* (GRO) (Beisswanger et al., 2008).

3.1. Event Triggers

After an extensive manual analysis of biomedical texts, events classified as relevant for the process of gene expression regulation were grouped into nine categories based on GRO concepts, *viz.* *Gene Expression*, *Transcription*, *Regulation of Gene Expression (ROGE)*, *Positive ROGE*, *Negative ROGE*, and *Experimental Intervention* with subtypes *Genetic Modification*, *Artificial Increase*, and *Artificial Decrease*.[8] Trigger words indicating textual mentions of the listed events were annotated with the corresponding categories. An event trigger is any literal verbal form that explicitly signals the occurrence of a particular molecular event. Trigger words are basically main verbs, verb nominalizations and adjectives. For example, the sentence "*H-NS and StpA proteins* ***stimulate expression*** *of the maltose regulon in Escherichia coli.*" contains two triggers: first, *'stimulate'* is a trigger for a process in the category *Positive ROGE*, second, *'expression'* is a trigger for a process belonging to the category *Gene Expression*.

3.2. *Gene Expression Regulation* Relations

Omitting the direct linking of arguments to atomic events anchored by trigger words in the text, in the second step, a domain expert directly annotated pairwise relations between genes and regulators affecting the expression of the genes. This annotation was based on the GRO class *ROGE* with its two subclasses *Positive ROGE* and *Negative ROGE*. A relation instance contains two arguments, *viz.* `Agent`, the entity that plays the role of modifying gene expression, and `Patient`, the entity whose expression is modified. Agents can be occupied by transcription factors (in core regulatory relations), or by polymerases and chemicals (in

[5] `http://www.pubmed.org`

[6] `http://www.geneontology.org/`

[7] The annotation process and IAA figures for named entities, event triggers and semantic relations are described by Buyko et al. (2008).

[8] *Experimental Intervention* events extend the original GENEREG version presented by Buyko et al. (2008).

auxiliary regulatory relations).[9] The sentence *“**H-NS** and **StpA** proteins stimulate expression of the **maltose regulon** in Escherichia coli.”* contains two *Positive ROGE* instances with *H-NS* and *StpA* as regulators and *maltose regulon* as regulated gene group. Table 1 summarizes the overall annotation results.

Semantic Category	Core	Auxiliary	TOTAL
ROGE	417	192	609
Positive ROGE	465	325	790
Negative ROGE	282	89	371
TOTAL	1164	606	1770

Table 1: Number of regulation of gene expression (ROGE) relation annotations per semantic category

GeneReg relations are represented by at least thirteen stable semantico-syntactic patterns (Buyko et al., 2008). The patterns range from the most frequent ones containing the mention of regulation verbs, adjectives and nominalizations (e.g., *‘regulator’*), to uncertain expressions such as *‘be essential’*, *‘be involved in’*, adjectives indicating requirements or dependencies (e.g., *‘dependent’*), and causal relation constructions between molecular processes in which gene and transcription factors are involved.

4. Studies on the Interoperability of GeneReg

4.1. Linking to the Genia Event Corpus

The *BioNLP’09 Shared Task* focused on the extraction of detailed behavior of proteins, characterized as biomolecular events, and provided annotations of selected events from the Genia events corpus (Kim et al., 2008). The main difference between Genia events and GeneReg relations is the level of detail of annotations. While GeneReg provides only shallow event annotations anchored through trigger words in the text, Genia links the event arguments to their triggers.[10] The two GeneReg annotation layers, *viz.*, *event triggers* and *relations* are currently not inter-linked. The second main difference is the use of biomedical knowledge for inference in GeneReg (in contrast to the Genia guidelines). GeneReg annotators frequently used the knowledge about experimental conditions for drawing conclusions about the role of the transcription factor in the gene regulation processes. The second difference can be explained considering the sentence *“In absence of H-NS and StpA proteins we detected the loss of the maltose regulon expression in Escherichia coli.”* Here an annotator will infer that *H-NS* and *StpA* proteins positively regulate *maltose operon* genes.

We explored about 150 (about 13%) of the relations annotated in the GeneReg corpus. For 142 relations (approximately 90%), we can provide Genia event annotations and automatically infer the corresponding GeneReg relations. Thereby, the participants of GeneReg relations have to be linked to various events, e.g., *Transcription* or *Gene expression* events (see the example in Section 2.).

For 19 relations (12%), we annotated *Mutagenesis* events. A *Mutagenesis* event denotes the process by which genetic material undergoes a detectable and heritable structural change. Eexperimental environments for gene regulation detection often involve genetic modifications of genetic material. By means of these genetic modifications and the expression levels of other genes, researchers implicitly draw conclusions about the role of the transcription factor in the gene regulation processes. The sentence “*Transcription of the chromosomal asr was abolished in the presence of a phoB-phoR deletion mutant.*” contains such an inferred relation of *asr* with *phoB* and *phoR*, respectively. This relation would be represented according to the Genia guidelines, e.g., for *phoB* as *Negative Regulation*(`Theme`:*Transcription*(`Theme`:asr), `Cause`(*Mutagenesis*(`Theme`:phoB))).

For 15 relations (approximately 10%), we annotated a cascade of *Regulation* events. For example, the sentence “*Fis protein is a major factor responsible for catabolite repression at the nrf promoter*.” contains a *Regulation of Gene Expression* relation between *Fis* and *nrf* that can be represented by means of two cascaded *Regulation* events, i.e., *Regulation*(`Theme`:*Negative Regulation*(`Theme`:nrf), `Cause`:Fis).

In 10% of the annotations we could not provide the Genia annotations, where 7 relations (about 5%) are statements and not events, and 8 relations (again, about 5%) are too complex and cannot be represented as Genia events. For example the sentence “*Primer extension analysis of the asr transcript revealed a region similar to the Pho box (the consensus sequence found in promoters transcriptionally activated by the PhoB protein) upstream from the determined transcription start.*” contains such a relation annotation between *asr* and *PhoB* that would not be annotated in the Genia event corpus according to its guidelines.

4.2. Linking to Biomedical and General Language Lexicon and Corpus Resources

4.2.1. Linking to Biomedical Lexicons

Although event trigger annotation and collection is crucial for biomedical information extraction, there is still a need for lexical resources containing specific verbs and nouns which express molecular events in texts. Up until now, some home-grown verb lists have been compiled (e.g., by Fundel et al. (2007)), while the BioLexicon and the Specialist Lexicon currently constitute perhaps the most comprehensive repositories for ‘biological’ verbs. The most frequent triggers of the GeneReg and *BioNLP’09 Shared Task* corpora can almost completely be found in both of these lexicons. The extracted frames of trigger verbs are either just syntactic (as with the Specialist Lexicon) or automatically gathered syntactico-semantic ones (as with the BioLexicon).

[9] Auxiliary regulatory relations (606 instances) extend the original GeneReg corpus.

[10] In GeneReg the additional linkage annotation costs could be avoided by automatically connecting triggers to relations.

Event (lemma)	Verbs	Nouns	Other	FrameNet	VerbNet/PropBank	NomBank	WordNet	Total
Transcription (14)	3	10	1	2 (14%)	1 (7%)	6 (42%)	7 (50%)	7 (50%)
Gene expression (17)	5	11	1	6 (35%)	2 (11%)	2 (11%)	10 (59%)	10 (59%)
Regulation (21)	10	6	5	13 (62%)	8 (38%)	4 (19%)	18 (86%)	18 (86%)
Positive regulation (43)	19	13	11	22 (51%)	12 (28%)	5(11%)	31 (72%)	32 (74%)
Negative regulation (29)	15	13	1	20 (69%)	9 (31%)	5 (17%)	19 (66%)	19 (66%)
TOTAL (124)	52	53	19	63 (50.1%)	32 (25.8%)	22 (17.7%)	85 (68.5%)	86 (69.3%)

Table 2: Number of trigger words matching general language resources per event category

Event	Frames in FrameNet
Transcription	`Causation` (*induction*), `Becoming aware` (*detect*)
Gene Expression	`Causation` (*induction*), `Becoming aware` (*detect*), `Creation` (*produce*), `Presence` (*present*)
Regulation	`Objective influence` (*effect, affect, influence*), `Control` (*control*), `Participation` (*involve, involvement*), `Cause change` (*change, alter*), `Contingency` (*dependent*), `Response` (*response*).
Positive regulation	`Causation` *(induce, lead, result, cause*), `Cause change position on a scale` (*increase, enhance, promoter*), `Being necessary` (*require, essential, necessary*), `Contingency` (*dependent*), `Cause to start` (*stimulate*), `Amassing` (*accumulation*), `Relative time` (*after*), `Time vector` (*through*), `Importance` (*important*), `Extreme value` (*high*), `Being active` (*active*)
Negative regulation	`Hindering` (*inhibit*), `Cause change position on a scale` (*decrease, reduce, reduction, diminish*), `Change position on a scale` (*decline*), `Preventing` (*prevent*), `Possession` (*lack*)

Table 3: Frames in FrameNet corresponding to GeneReg's event categories

4.2.2. Linking to General Language Lexicons and Corpora

Directing our attention now to the non-biomedical domain we find large manually created lexico-semantic resources such as WordNet, VerbNet (Kipper et al., 2000), FrameNet (Baker et al., 2003) that might be useful in the biomedical domain as well. We explore, in this section, the linking of event triggers (similar to the work of Uematsu et al. (2009)) from GeneReg and *BioNLP'09 Shared Task* annotations for *Transcription*, *Gene expression*, *Regulation*, *Positive regulation* and *Negative regulation* events.[11]

We manually linked the most frequent triggers to the most prominent general language lexical resources, *viz.*, WordNet, FrameNet, and VerbNet, but also included the lexical patterns used for general language proposition annotations as contained in the PropBank (Palmer et al., 2005) and the NomBank (Meyers et al., 2004). The results are summarized in Table 2.[12]

The resource with the highest number of matches (68.5%) is WordNet where we found between 50% (*Transcription*) to 86% (*Regulation*) of all triggers. WordNet is followed by FrameNet with 50.1% matches, and VerbNet/Propbank with 25.8% matches. At the bottom of the list appears NomBank with 17.7% matches. The most difficult event to link is the *Transcription* event as it is expressed through compounds such as *”mrna levels“*, *”transcriptional activity”*, *“mrna expression”*. *Transcription* and *Gene expression* events share a set of frames, e.g., `Causation` and `Becoming aware`, that represent different views on the production of proteins from the DNA strain: the view of regulation by proteins and the view of a biologist in experiments (see Table 3). The sharing of frames can be explained by the fact that the *Transcription* event is a part of the *Gene expression* event (see Gro (Beisswanger et al., 2008)).

Regulation and *Positive regulation* triggers are the ones with the highest coverage in general language lexicon resources. They could be successfully linked in particular to FrameNet (see Table 3). These events are usually expressed by words that describe general regulation, influence or control. *Regulation* events are expressed by frames such as `Objective influence`, `Causation`, and `Control`. *Positive regulation* and *Negative regulation* correspond to frames that express more outspoken influence such as `Cause change position on a scale`, and `Hindering`. Still, many triggers could not be connected to FrameNet. The linkage ratio lies between 14% (for *Transcription*) to 69% (for *Negative regulation*). Very specific biomedical triggers such as *‘down-regulation’* or *‘up-regulation’* are not at all represented in any of the lexical resources we explored.

[11] *ROGE*, Positive and Negative *ROGE* in GeneReg are subtypes of *Regulation*, *Positive regulation* and *Negative regulation* events in Genia, respectively.

[12] For many triggers, we could not find a corresponding lemma or its sense in the screened resources. Accordingly, in Table 2, we only counted the lemmata with correctly traceable and identified senses.

5. Conclusion and Outlook

We described and discussed the semantic complexity of the GENEREG corpus by linking its metadata to an alternative event corpus in the life sciences (GENIA), as well as to biomedical and general language domain lexicons including lexical specifications from general language corpus annotations. We gave evidence for the coverage of resources such as FRAMENET and WORDNET related to biomedical triggers which are relevant for the extraction of gene expression regulation events. While some overlap exists for all of the considered resources, WORDNET exhibits the highest degree of coverage way beyond the 50% line.

Interlinking these resources to GENEREG will help us improve the RE performance on this corpus. Experiments on JREX, our most recent RE extraction pipeline (Buyko et al., 2009), are under way.

The corpus and annotation guidelines are freely available for academic purposes at `http://www.julielab.de`.[13]

Acknowledgements

We thank Jin-Dong Kim (Tsujii Lab, University of Tokyo) for fruitful communication concerning the GENIA event corpus annotation guidelines. This research was partially funded by the EC within the CALBC project (FP7-231727) and within the BOOTSTREP project (FP6-028099).

6. References

Collin F. Baker, Charles J. Fillmore, and Beau Cronin. 2003. The structure of the FRAMENET database. *International Journal of Lexicography*, 16(3):281–296.

Elena Beisswanger, Vivian Lee, Jung-jae Kim, Dietrich Rebholz-Schuhmann, Andrea Splendiani, Olivier Dameron, Stefan Schulz, and Udo Hahn. 2008. Gene Regulation Ontology (GRO): Design principles and use cases. In *MIE 2008 – Proceedings of the 21st International Congress of the European Federation for Medical Informatics*, pages 9–14. Gothenburg, Sweden.

Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, and Yuk Wah Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155.

Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. 2008. Testing different ACE-style feature sets for the extraction of gene regulation relations from MEDLINE abstracts. In *SMBM 2008 – Proceedings of the 3rd International Symposium on Semantic Mining in Biomedicine*, pages 21–28. Turku, Finland.

Ekaterina Buyko, Erik Faessler, Joachim Wermter, and Udo Hahn. 2009. Event extraction from trimmed dependency graphs. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 19–27, Boulder, Colorado, June. Association for Computational Linguistics.

Andrew E. Dolbey, Michael Ellsworth, and Jan Scheffczyk. 2006. BIOFRAMENET: A domain-specific FRAMENET extension with links to biomedical ontologies. In *Proceedings of the "Biomedical Ontology in Action" Workshop at KR-MED 2006*, pages 87–94. Baltimore, MD, USA.

Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2007. RELEX – relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(10).

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BIONLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop – Companion Volume for Shared Task*, pages 1–9. Boulder, CO, USA.

Karin Kipper, Hoa Trang Dang, and Martha Stone Palmer. 2000. Class-based construction of a verb lexicon. In *AAAI/IAAI 2000 – Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, pages 691–696. Austin, TX, USA.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NOMBANK project: An interim report. In *Proceedings of the Workshop on "Frontiers in Corpus Annotation" at HLT-NAACL 2004*, pages 24–31. Boston, MA, USA.

Claire Nédellec. 2005. Learning language in logic – genic interaction extraction challenge. In LLL'05 – *Proceedings of the 4th Learning Language in Logic Workshop*, pages 31–37.

Martha S. Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Jarvinen, and Tapio Salakoski. 2007. BIOINFER: A corpus for information extraction in the biomedical domain. *Bioinformatics*, 8(50).

Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter, and Tapio Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(3), April.

Yutaka Sasaki, Simonetta Montemagni, Piotr Pezik, Dietrich Rebholz-Schuhmann, John McNaught, and Sophia Ananiadou. 2008. BIOLEXICON: A lexical resource for the biology domain. In *SMBM 2008 – Proceedings of the 3rd International Symposium on Semantic Mining in Biomedicine*, pages 109–116. Turku, Finland.

Paul Thompson, Syed A Iqbal, John McNaught, and Sophia Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10(349).

Sumire Uematsu, Jin-Dong Kim, and Jun'ichi Tsujii. 2009. Bridging the gap between domain-oriented and linguistically-oriented semantics. In *Proceedings of the BioNLP 2009 Workshop*, pages 162–170. Boulder, CO, USA.

Tuangthong Wattarujeekrit, Parantu Shah, and Nigel Collier. 2004. PASBIO: Predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5(155).

[13] Licensed under a Creative Commons Attribution-Noncommercial 3.0 Germany.