Deliverable D11.1
Report on Evaluation Studies

Project acronym:	BOOTStrep
Project full title:	Bootstrapping Of Ontologies and Terminologies STrategic REsearch Project
Proposal/Contract no.:	FP6 - 028099
Duration:	April 01, 2006 – March 31, 2009
Project coordinator:	FSU Jena
Website:	www.bootstrep.eu
Author:	Su Jian, Shou Xiao Mang, Qiu Long, Upali Sathyajith Kohomban, Dilip Kumar Limbu (I2R) Vivian Lee, Jung-jae Kim, Piotr Pezik, Dietrich Rebholz-Schuhmann (EBI) Yutaka Sasaki, Paul Thompson, John Mcnaught, Sophia Ananiadou (UoM) Udo Hahn, Katrin Tomanek (FSU Jena) Stefan Schulz, Kornél Markó (UKLFR)
Date of preparation:	April 2009
Dissemination level:	PP

Table of Contents

Executive Summary

Work Package 11 targets on the evaluation of language resources constructed and available in BOOTStrep for information access purpose. Biolexicon, BioOntology and NLP tools are accessed in Information Retrieval (IR) and Information Extraction (IE) tasks.

For IR evaluation, 3 investigations have been conducted. First, an IR evaluation set focusing on Gene Regulation(GR) with 60 queries for 5 entity type, 4 GR category and 9 GR event have been constructed jointly by EBI and I2R. I2R has conducted further evaluations on the above BOOTStrep evaluation set with Named Entity Recognition and Biolexicons, which have shown promising results. Second, UoM has conducted IR evaluation on Genomics track 2007 evaluation set, which has shown slight improvement using biolexicon. Further more, UKLFR's multilingual IR evaluation leveraging on Genomics track materials, shows that using translated queries in German achieves the same level of performance as using English subword.

For IE evaluation, 4 investigations have been conducted. FSU evaluates GRE extraction with GeneReg Corpus annotated by the biologists in WP08 and also directly compares with REGULON database. EBI explores using Gene Regulation Ontology (GRO) on gene regulation event (GRE) extraction. I2R explores raw text and Stanford & Enju parsers for the GRE extraction with GeneReg corpus. UoM further extend Biology Event Linguistic Annotation done in WP04. And multi-slot GRE template extraction has further conducted using this corpus.

All the above evaluations show the positive indications with the usefulness of the various resources built in the project, which trigger the further investigation, enhancement and extend the use of these resources beyond the project.

1 Introduction

In the first two years of BOOTStrep project, various resources have been developed or collected. These include BioLexicon with a high-coverage biological terminology and a more focused sublanguage lexicon having corresponding linguistic specifications, GRO, comprehensive text analysis pipelines which incorporate term handling (tokenization, morphological normalization, lexicon look-up), syntactic analysis (POS tagging. chunking, parsing), semantic processing (term recognition, the extraction of named entities, relations, and events etc.), up to the level of discourse analysis (Coreference and anaphora resolution) and fact database capturing Regulation of Gene Expression Event information locked in the biomedical abstracts and full papers.

This report describes various activities with WP11 by I2R, UoM, EBI, FSU Jena and UKLFR teams, which target the evaluation of the above resources in the light of various use cases for biologists to access the information, including information retrieval and extraction tasks.

2 IR Evaluation

In order to evaluate the language resources in the project, EBI and I2R develop a GR IR evaluation set after studying the suitability of Genomics Track materials. I2R further evaluate the contribution of Named Entity Recognition to information retrieval with and without the incorporation of Biolexicon. In parallel with this effort, UoM evaluates the usefulness of the BioLexicon based on the TREC Genomics Track 2007 evaluation set. Meanwhile, UKLFR evaluates multilingual lexicon based on Genomics track materials as well.

2.1 GR IR Evaluation Set (EBI, I2R)

It's well known that there're full of biomedical names in biology literature. Named Entity Recognizer, an important NLP tool for mapping BioText to BioOntology is expected to be beneficial to IR task especially with entity type of queries. Yet among 36 entity type queries with Genomics track 2007, only 6 queries are relevant to Gene Regulation, the focused domain of BOOTStrep. Thus an evaluation set tailored to the language resources of BOOTStrep project is needed.

The language resources in BOOTStrep project and those available in the public are mainly developed with abstracts. Thus better performance could be achieved with those resources themselves on abstracts than full papers. Another reason that I2R proposes to use abstracts is due to less time needed for relevance judgment. This is a practical and critical factor we've to consider so that the evaluation can be done in the project period.

Among ad hoc retrieval, passage retrieval and question answering, ad hoc retrieval is recommended by I2R due to there're mature evaluation standard with ad hoc retrieval. There's no consensus on Passage Retrieval with P1 and P2 used in Genomics track 2006 and 2007. And also passage retrieval and question answering need much more efforts to do relevance judgment.

The 91k EBI K12 strain corpus generated by EBI by applying 12 rules to 2008 Medline archival is used for IR evaluation. Since there are hundreds of E. coli strains isolated from different ecosystems, to avoid confusions and redundancies of the gene/protein names among these strains, it makes sense to focus on one specific strain. So far K12, one of the cultivated strains that are well-adapted to the laboratory environment, and being the most deeply understood organism at the molecular level, is the most complex and attracts most research focus, thus is chosen to be the focus the text collection.

The 12 rules applied are as the following (rules were applied sequentially):

Rejection of scientific articles with artificial protein mentions

The first rule excludes articles that indicate that the documents report on results from experimentally generated proteins. The exclusion rule (Rule 1) used the MeSH headings "Recombinant Fusion Protein" and "Recombinant Protein" to reject Medline abstracts and full text articles that have been annotated with these headings. Thus, Recombinant (Fusion) Proteins which use of E. coli expression system to generate proteins from other organisms (e.g. human or mouse proteins) in recombinant form are excluded.

Selection of documents based on the mention of E. coli and the stains in their titles

In the next steps, all documents that refer to "E. coli" or "Escherichia coli" were further analysed. If they include in the title the mention of the strain "K12" or "K-12", then they are selected (Rule 2). If the title refers to other

strain names (e.g. “O157”, “EPEC”) or to keywords indicating other strains (e.g. “enterohemorrhagic”, "pathogenic”), then the document is again rejected (Rule 3). All remaining documents are again selected, i.e. documents with mention of E. coli but without any further sub-specification of the E. coli strain (Rule 4).

If the document title contains species names, which represent a different species from E. coli (e.g. “human”, “mouse”) in addition to the mention of E. coli, then these documents were rejected again (Rule 5).

Selection of documents based on the mention of E. coli in the abstract content

Medline abstracts are analysed to identify the mention of the species in the document. Similar to the analysis of the document title, the first two sentences are used for the selection of the documents. If these sentences refer to “E. coli” or “Escherichia coli”, then the documents are selected. Again, those documents are chosen contained mentions of “K12” or “K-12” (Rule 6) and other documents were rejected if they refer to strain names or keywords indicating other strains (Rule 7). All other documents are included (Rule 8), if the first two sentences do not contain other species names (Rule 9).

Selection of documents based on MeSH terms

Additional documents are selected, if they contained mentions of “E. coli” in the MeSH terms and are not excluded by previous rules (Rule 10).

Scientific documents that were annotated with anatomical terms in the MeSH headings were again excluded (e.g. “Appendix”, “Face”, “Tears”; Rule 11). This is also true for articles that have been annotated with cell types, i.e. with the sub-tree of the MeSH headings that is identified with the id “A11” (i.e. “Cells”; Rule 12).

These 12 rules are first applied on Medline 2007 archive to generate the 69K collection which is used by FSU Jena. After moving on to the Medline 2008 archive and fixing a bug in the program, the final 91K EBI K12 strain corpus is generated for the IR evaluation.

60 entity type queries are proposed by EBI which cover the 5 major entity types related with E. coli GR:

- Transcript Factors--20 queries
- Genes--20 queries
- RNA--10 queries

- Protein--5 queries
- Cell Component--5 queries

These queries are related with 4 core GR categories:

- Transcription factors (TF) and formation of TF complex--36 queries
- DNA binding of TFs (at TF recognition sites) --18 queries
- Gene expression (RNA, protein) –25 queries
- Regulation of gene expression (up-, down-regulation) – 52 queries

The queries also cover 9 important GR events in E. Coli functional systems:

- Carbon utilization (CU) – 5 queries
- Redox sensing (RS) – 11 queries
- Environment sensing e.g. temperature, water (ES) – 18 queries
- Ion transport (IT) – 3 queries
- Cell structure (CS) – 11 queries
- General enhancer (GE) –3 queries
- Cellular metabolic process (carbon, nitrogen, phosphate, sulfur, nucleotide, cofactor) (CM) –26 queries
- Antibiotic resistance (AR) –2 queries
- Restriction and repair (RR) – 4 queries

The queries, their categories and event types are attached in Annex 1.

To provide a golden standard for the evaluation, TREC genomics tracks have used system pooling to do relevance judgment, where the top relevant documents of many system retrieval results are checked on their relevance to the corresponding queries. This is not feasible as we only have several systems in the project. On the other hand query pooling (Sanderson and Joho, 2004) appears a much more efficient and effective new trend which only requires similar queries with related terms. For example, besides nuclear waste dumping, radioactive waste, radioactive waste storage, hazardous waste, nuclear waste storage, Utah nuclear waste, waste dump can be used as queries as well. The top retrieved documents will be judged on their relevance. Query pooling usually takes about 2 hours per topic for news articles while system pooling requires 10-20 hours per topic. Besides using it in Image CLEF, its current promoter, Mark Sanderson, general chair of SIGIR 2004 and PC chair of SIGIR 2009 also gives a Keynote speech in NTCIR, Asian TREC on the IR evaluations. I2R thus develops the query pooling web service and Vivian from EBI conducts the whole relevance judgment with average 3 topics per day working part timely, which is inline with the speed with relevance judgment on news articles. The number of relevant documents for all the queries ranged from 1 to 227 in the top 300 records.

Thus a large scale GR IR evaluation set is constructed with 60 entity type queries on 5 entities, 4 GR core categories and 9 important events in E. coli functional systems with 91 k Medline abstracts on E. coli K12.

2.2 IR Evaluation on GR IR Evaluation Set (I2R)

I2R used the Lemur language model (http://www.lemurproject.org/) retrieval as the baseline of the IR evaluation. The test retrieval using this model on genomics track 2005 with MAP 0.2497 ranked 6th among 44 autoruns, 7th among 58 manual, interactive and auto runs. Its performance with BOOTSTREP test set is MAP 0. 2851.

I2R's IR retrieval used the top 300 records for each queries and re-rank them according to the recognized NEs in these documents. The performances of the NERs are:

- TF Recognizer trained and tested on Jena E. coli TF corpus with Fscore 72.27% by I2R NER
- Gene/DNA, RNA, Protein Recognizer trained and test on IJCNLPBA test data with Fscore: Gene/DNA: 69.83%, RNA: 64.10%, Protein 73.77%
- Cell Component Recognizer trained and test on GENIA, 10 cross validation, Fscore 69.7%

The re-ranking mechanism of the top 300 records for each query boosts documents by giving additional score to the abstracts with queried entities; It also considers the sentence retrieval scores for the sentences with queried entities. The result obtained shows performance improvement in all three entity types with at least 10 queries available (**Table 1**). Three measurements used are:

Mean Average Precision (MAP): the mean of sum of precision divided by total number of relevant documents across all queries.

R-precision (R-P): the precision at rank R, where R is the number of documents relevant to the query.

Reciprocal-Rank (R-R): the reciprocal of the first relevant document's rank in the ranked list returned for a topic.

		MAP			R-P			R-R	
Cat	NER	Baseline	CNG%	NER	Baseline	CNG%	NER	Baseline	CNG%
TF	0.15	0.12	+30.8	0.21	0.16	+24.9	0.53	0.44	+21.6
Gene	0.40	0.38	+5.9	0.42	0.38	+11.3	0.72	0.65	+10.5
RNA	0.40	0.38	+2.9	0.42	0.38	+10.6	0.74	0.70	+6.1
Overrall	0.30	0.29	+3.5	0.32	0.30	+8.7	0.6	0.58	+3.6

Table 1. I2R IR performance on GR IR evaluation set.

The retrieval result shows clear improvement in TF, Gene and RNA entity groups with increment from 2.9% to 30.8% by using NER result in re-ranking. Since there are only 5 queries for Protein and Cell Component types each, thus not sufficient to be assessed independently. But they are accounted for the overall performance. There is 3.5% improvement in the overall performance with all 5 entity types.

Except for transcript factors, the recognizers for gene, protein, RNA and Cell component are all trained on Medline abstracts with GENIA (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi) on human cell signals (Zhou *et al.*, 2004), a different domain from the E. coli GR IR evaluation set we used. Likely the named entity recognition performance might be dropped to some extent in the application domain.

On the other hand, biolexicon provided by EBI covers 4860 unique E. coli gene / protein entities with about 50k names and synonyms from EcoCyc and Uniprot. We expect our NER should have better performance when incorporate this application domain lexicon. As the single word lexicon likely introduces noise to NER, we thus also compare the IR performance when NER incorporate the above biolexicon without single word entries. **Table 2** shows the corresponding IR results with 20 queries on gene.

Cat	MAP	R-Prec	Recip_RNK
NER	0.4039	0.4202	0.7151
+G/P list	0.176	0.2031	0.3929
+G/P list-sw	0.1772	0.1892	0.3667

Table 2. IR Performance with 20 gene queries when incorporate named entities recognized using NER directly and NERs with additional lexicon list for Gen/Protein with / without single word entries.

Besides EBI biolexicon also covers 323 transcript factors collected from two review papers, including additional data file 4:full data set for the E. coli transcriptional regulatory network reconstructed for our analyses http://genomebiology.com/content/supplementary/gb-2008-9-10-r154-s4.txt (Freyre-González *et al*., 2008) and a complete list of transcription factors: supplementary data- The set of 271 transcription factors and their domain assignments (now updated to 273) http://www.mrc-lmb.cam.ac.uk/genomes/cgi/ec_tf/get-domarch.pl (Babu *et al*, 2003). It turns out that only less than half of these 323 TF entries can be founded from 1027 case insensitive unique annotations. As these 323 TF entries are used in the two representative papers on E. coli, we expect that they represent important TFs in E. coli and should contribute better performance when incorporated in TF NER, as even that it has already been trained on the same application domain it might still have sufficient coverage due to the limitation of training data. As most of TFs are proteins, we thus consider the further expansion with the synonyms in Gene / Protein list. Similarly, we also consider the situation without single word entries. **Table 3** below shows the corresponding IR results with 20 TF queries.

Cat	MAP	CNG%	R-Prec	CNG%	Recip_Rnk	CNG%
NER	0.1513		0.2053		0.5329	
+list	0.1514	+0.07	0.2051	-0.10	0.5594	+4.97
+list+G/P syns	0.152	+0.46	0.2068	+0.73	0.5561	+4.35
+list +G/P syns-sw	0.1521	+0.53	0.204	-0.63	0.5594	+4.97

Table 3. IR Performance on 20 TF queries when incorporate named entities recognized using NER directly and NERs with additional lexicon list for TF and the further expansion with TF's synonym in Gene/Protein list.

Out of our expectation, the above quick attempts on incorporation of domain biolexicon are not really helpful. Our initial investigation shows that much less named entities are spotted when incorporation of biolexicon. This result triggers our further investigations of effective adaptation with and without biolexicon which are still ongoing.

To our knowledge, the most relevant work to our above attempt is Genomics Track 2007 task, which is also evaluated by UoM in BOOTStrep. There're 36 entity type queries. Comparing to our GR evaluation set, there're several differences:

1) the queries used in Genomics track are quite diverse and scarce with each type (See section 2.3.1 for details). It has 11 queries on gene, 5 queries on proteins and other 12 types having 1 to 3 queries each, which is not possible to do diagnose study as what we're able to do with GR evaluation set on the contribution of different type of entities recognized by NER as it's commonly believed that it's only meaningful to consider the IR performance with at least 20 queries.
2) Some queries are actually on quite generic terms. For example, there are 3 queries on biological substances and 2 queries on molecular functions. These kind of generic terms are quite high above in the ontology, which usually need at least several NERs / term recognizers to spot.
3) Some queries may require the information embedded in long descriptions in the text which is actually beyond the entity level. For instance, there are 2 queries on pathways.
4) Some queries are on attributes of entities. For example, there are 2 queries on cell or tissue types and 1 query on tumor types.

All the above difference show that GR evaluation set is much more suitable to do focused study on the contribution of named entity recognition to information retrieval.

With Genomics Track 2007, NLM has explored the use of synonym, entity and relation extraction (Demner-Fushman *et al*., 2007). However this does not achieve comparable performance as the top performance which is achieved by NLM and its collaborators through the fusion of multiple search engines. There're 3 other works (Stokes *et al*., 2007; Fautsch and Savoy, 2007; Jimeno and Pezik, 2007) explore synonym expansion and NER. Yet none of them achieve comparable performance as the top one.

In conclusion, a large scale IR evaluation set is constructed with extensive queries on GR category and event information. The entity queries facilitate the evaluation of contribution of NER and other related NLP technologies to IR. Our quick attempt shows NER is quite useful for IR, which is the first success demonstration of the contribution of NER to information retrieval at least with bio-literature.

Besides the further work on incorporations of biolexicon and adaptation, we'll also explore different ways of incorporating NEs recognized. For example, instead of only involving NEs in the re-ranking process, we'll use NER to process

the whole text collection, further leveraging on coreference resolution to access the information embedded in the different mentions of the same entities.
We'll also explore the incorporation of GR event identification from UoM, Verb subcategorization from CNR, Parsing by Enju parser (http://www-tsujii.is.s.u-tokyo.ac.jp/enju/) and the parsers from JENA-FSU and semantic role labeling done by UoM.

2.3 IR evaluation on Genomics Track 2007 (UoM)

Because it takes time to develop GR IR evaluation set, UoM evaluates Biolexicon in TREC Genomics track 2007 passage retrieval task in parallel.

2.3.1 TREC Genomics Track 2007 Evaluation Set

Document collection
The corpus for the TREC Genomics Track 2007 is a collection of full papers obtained from 49 biomedical journals. The corpus occupies 13.3 GB in total and contains 162,259 full papers.

Questions
Official queries for the Genomics Track 2007 are in the form of questions asking for lists of specific biomedical entities.
Targeted biomedical entities are as follows (Hersh *et al*., 2007):

ANTIBODIES Immunoglobulin molecules having a specific amino acid sequence by virtue of which they interact only with the antigen (or a very similar shape) that induced their synthesis in cells of the lymphoid series (especially plasma cells).

BIOLOGICAL SUBSTANCES Chemical compounds that are produced by a living organism.

CELL OR TISSUE A distinct morphological or functional form of cell, or the name of a collection of interconnected cells that perform a similar function within an organism.

DISEASES A definite pathologic process with a characteristic set of signs and symptoms. It may affect the whole body or any of its parts, and its etiology, pathology, and prognosis may be known or unknown.

DRUGS A pharmaceutical preparation intended for human or veterinary use.

GENES Specific sequences of nucleotides along a molecule of DNA (or, in the case of some viruses, RNA) which represent functional units of heredity.

Entity type	# of questions
ANTIBODIES	1
BIOLOGICAL SUBSTANCES	3
CELL OR TISSUE	2
DISEASES	1
DRUGS	2
GENES	11
MOLECULAR FUNCTIONS	2
MUTATIONS	1
PATHWAYS	2
PROTEINS	5
SIGNS OR SYMPTOMS	2
STRAINS	1
TOXICITIES	2
TUMOR TYPES	1
Total	36

Table 4. Classification of official queries

MOLECULAR FUNCTIONS Elemental activities, such as catalysis or binding, describing the actions of a gene product or bioactive substance at the molecular level.

MUTATIONS Any detectable and heritable change in the genetic material that causes a change in the genotype and which is transmitted to daughter cells and to succeeding generations

PATHWAYS A series of biochemical reactions occurring within a cell to modify a chemical substance or transduce an extracellular signal.

PROTEINS Linear polypeptides that are synthesized on ribosomes and may be further modified, crosslinked, cleaved, or assembled into complex proteins with several subunits.

SIGNS OR SYMPTOMS A sensation or subjective change in health function experienced by a patient, or an objective indication of some medical fact or quality that is detected by a physician during a physical examination of a patient.

STRAINS A genetic subtype or variant of a virus or bacterium.

TOXICITIES A measure of the degree and the manner in which something is toxic or poisonous to a living organism.

TUMOR TYPES An abnormal growth of tissue, originating from a specific tissue of origin or cell type, and having defined characteristic properties, such as a recognized histology.

Table 4 shows the number of official questions. The form of questions are (unintentionally) formed in the following format:

{<WH>|<PREPOSITION> <WH>} <MODIFIER>* [<ENTITY TYPE>] <WORD>+

where <WH> is “what" or “which", <ENTITY TYPE> is one of the above entities and <MODIFIER> is a noun, verb, or adjective phrase that restrict the range of entities in a question. Question types and question focuses are explicitly given in question sentences.

Due to this question style, the main task of question analysis is to find query terms that are effective to find passages relevant to questions.

Genomics Track 2007 Task

The task of the TREC Genomics Track 2007 is to retrieve passages for a full paper corpus and return ranked list of 1,000 passages for each question.

Passages are defined as follows (Hersh *et al*., 2007):

> Retrieved passages could contain any span of text that did not include any part of an HTML paragraph tag (i.e., one starting with <P or </P).

In this report, for the experiments using the TREC genomics data, retrieved passage are spans with the maximum length. Each passage can be identified by triples (PMID, offset, length), where the offset is the starting position of the passage in a document in terms of the number of bytes from the top of the document.

Gold standard relevance judgment

A total of 66 runs were submitted by 29 groups. Each of the runs returns at most 1,000 passages for each question and judges with domain expertise manually checked relevance of the submitted passages. Relevant passages are pooled in the gold standard. **Table 5** show the numbers of passages and documents relevant to 36 questions pooled in the gold standard.

QID	Entity	passages	documents
200	PROTEIN	320	193
201	MUTATIONS	37	12
202	DRUGS	53	43
203	CELL OR TISSUE TYPES	321	147
204	CELL OR TISSUE TYPES	164	74
205	SIGNS OR SYMPTOMS	93	65
206	TOXICITIES	38	19
207	TOXICITIES	15	12
208	BIOLOGICAL SUBSTANCES	22	16
209	BIOLOGICAL SUBSTANCES	78	11
210	MOLECULAR FUNCTIONS	71	57
211	ANTIBODIES	57	42
212	GENES	358	133
213	GENES	377	185
214	GENES	209	98
215	PROTEINS	137	73
216	GENES	42	34
217	PROTEINS	38	34
218	GENES	163	74
219	DISEASES	22	16
220	PROTEINS	16	6
221	PATHWAYS	183	87
222	MOLECULAR FUNCTIONS	57	42
223	STRAINS	18	8
224	GENES	3	3
225	BIOLOGICAL SUBSTANCES	1	1
226	PROTEINS	152	57
227	GENES	281	172
228	GENES	15	14
229	SIGNS OR SYMPTOMS	150	57
230	PATHWAYS	82	29
231	TUMOR TYPES	16	13
232	DRUGS	93	57
233	GENES	19	16
234	GENES	609	483
235	GENES	182	107

Table 5. Relevant passages and documents in the gold standard file

Sometimes the number of texts relevant to a question is very small. For example, the numbers of passages/documents relevant to Topic 224 and 225 are only three and one respectively. This is a typical phenomenon in entity-oriented Passage Retrieval. From our experience in passage retrieval, it is effective to use keywords and phrases that are specific to each question. UoM reports what kind of keyword generation methods are advantageous for passage retrieval for genomics track evaluation set.

Evaluation metrics

Three kinds of Mean Average Precision (MAP) are officially used in the task 2007 (Hersh *et al*., 2007)

Passage2 MAP

The original Passage MAP for the 2006 track was found to be problematic in that splitting passages into shorter units had substantial positive effects on Passage MAP. To avoid this, Passage2 MAP calculates MAP as if each character in each passage were a ranked document.

Aspect MAP

Passages in gold standard are grouped into aspects identified by one or more Medical Subject Headings (MeSH) terms. Aspect retrieval MAP is the average precision for the aspects of a topic, averaged across all topics.

Document MAP

Any document ID that had a passage associated with a topic ID in the set of gold standard passages was considered a relevant document for that topic.

2.3.2 UoM Passage Retrieval system

Passage retrieval methods

UoM adopted probabilistic IR toolkit Xapian1 for our retrieval platform. UoM has created two indexes, one for document retrieval and the other for passage retrieval. The built-in Xapian tokenizer and standard English stemmer are specified. The IR model is a variant of Okapi BM25 (Robertson *et al*., 1992).

$$\frac{(k_3 + 1)q}{(k_3 + q)} \quad \frac{(k_1 + 1)f}{(k_1 \text{ L} + f)} \quad \log \quad \frac{(r+0.5)(N-n-R+r+0.5)}{(n-r+0.5)(R-r+0.5)}$$

where

k_1, k_3: constants

K: $k_1(bL + (1-b))$

1 http://xapian.org/

q: within query frequency
f: within document frequency
n: the number of documents in the collection indexed by this term
N: the total number of documents in the collection
r: the number of relevant documents indexed by this term
R: the total number of relevant documents
L: the normalized document length

UoM used default parameter setting, k_1=1, k_3=1, b=0.5.

To capture local information in a passage and global characteristics of its full paper, retrieved passages are ranked by the score that is a weighted sum of BM25 scores of passage p and its full papers.

$BM25_{d,p} = \alpha\ BM25(p) + (1-\alpha)BM25(d)$,

The baseline passage retrieval algorithm to compare usefulness of lexical resources is as follows:

1. Analyze a question sentence using a dictionary-based Part-of-Speech (POS) tagger based on the biological lexicon, the UMLS Specialist Lexicon, or an *n*-gram collection.
2. Create a list of query terms from the question.
3. Retrieve N_d full papers using the query terms.
4. Retrieve N_p passages using the query terms.
5. Rerank the passages according to the $BM25_{d,p}$ score based on the scores of the retrieved documents and passages
6. Output the top 1,000 passages from the ranked passages.

For conciseness, technical terms extracted from a sentence based on the BioLexicon are called as the *BL term* and terms extracted from a sentence based on the Specialist Lexicon as the *SL term*. Here, N_d is set to 1,000 because it is the number of papers to be output. N_p is set to a large number, 1,000,000.

Since the goal is to estimate usefulness of lexical resources, first, UoM decided the parameter α of the baseline model without using external resources. Stop words are removed from a question and the remaining words are used as query terms. The Document, Aspect, and Passage2 MAP are measured using the TREC Genomics Track 2007 test set. **Figure 1** shows the results for α = {0.0, 0.1, 0.2,...,1.0}. As the result, α is set to 0.5 which is the peak of the Aspect and Passage2 MAP curves.

Query Analysis

UoM compared the following question analysis methods.

[w1] Query word uni-grams: The baseline question analysis method is to use tokens (i.e., uni-grams) which are not stop words.

[w12] Query word uni- and bi-grams: In addition to w1, bi-grams of consecutive non-stop words are used.

[w123] Query word uni-, bi-, and tri-grams: In addition to w12, tri-grams of consecutive which non-stop words are used.

[b1] Query lemma uni-grams: Uni-grams of lemmas (i.e., base forms) of tokens which are not stop words.

[b12] Query lemma uni- and bi-grams: In addition to lemma uni-grams, bi-grams of consecutive lemmas of non-stop words.

[b123] Query lemma uni-, bi-, and tri-grams: In addition to above, tri-grams of consecutive lemmas of words which are not stop words.

[w1∪BL/SL] Multi-word terms and query words that are not in BL/SL terms: If multi-word terms in a question are in the biological lexicon, the terms are added to the query term list. Then, word uni-grams that are not in query terms are added to the query term list.

[w1+BL/SL] Query word and BL or SL terms: First, word uni-grams are added to the query term list. Then, if lexicon terms are found in a question, the terms are added to the query term list.

[b1∪BL/SL] Multi-word terms and query lemmas that are not in BL/SL terms: If multi-word terms in a question are in the biological lexicon, the terms are added to the query term list. Then, lemma uni-grams are not in the BL terms are added to the query term list.

[b1+BL/SL] Query lemmas and BL or SL terms: First, word uni-grams are added to the query term list. Then, if multi-word terms in a question are in the biological lexicon, the terms are added to the query term list.

Experiments

Experiments on the TREC Genomics Track 2007 data have been conducted with different question analysis methods described in the previous section.

Table 6 (a), (b), and (c) show results of *n*-gram, the BioLexicon, and the Specialist Lexicon.

The best document MAP is 0.2763 when the BL terms are added to query word uni-grams. It is clear that adding bi-grams and/or tri-grams generated from a noun, verb, and adjective phrases is not effective as the MAP scores decrease to 0.2257 and 0.2156.

The document MAP of the queries consisting of the Specialist Lexicon terms and word uni-grams is 0.2759, which is better than the n-gram-based approach but is not as good as the BioLexicon-based query analysis.

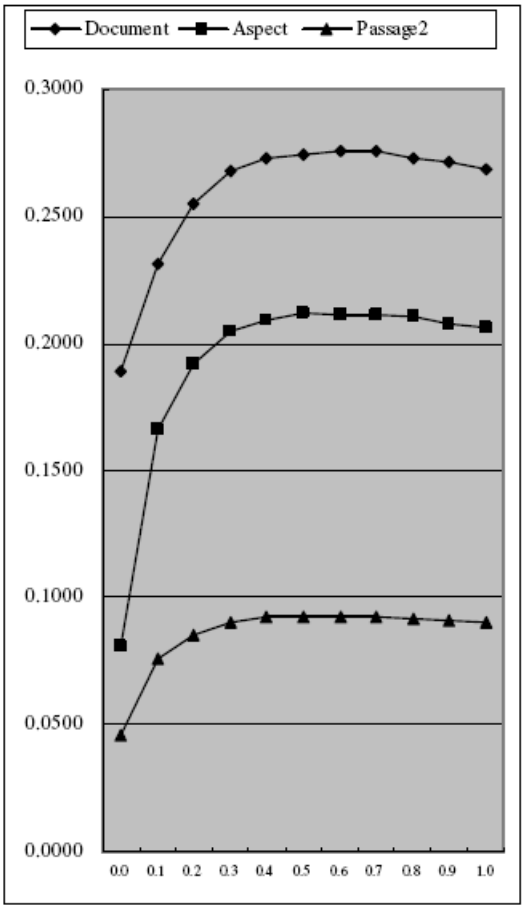


Figure 1. Balancing paragraph and document BM25

Term source	Type	Mean Average Precision Score		
		Document	Aspect	Passage2
(a) n-gram	w1	0.2744	**0.2119**	0.0924
	w12	0.2257	0.1955	0.0760
	w123	0.2156	0.1697	0.0737
	b1 (BL)	0.2272	0.1773	0.0768
	b12 (BL)	0.2190	0.1674	0.0688
	b123(BL)	0.2137	0.1601	0.0666
	b1 (SL)	0.2483	0.1811	0.0765
	b12 (SL)	0.2217	0.1707	0.0650
	b123(SL)	0.2134	0.1514	0.0628

(b) BioLexicon	w1∪BL w1+BL b1∪BL b1+BL	0.2747 **0.2763** 0.2274 0.2369	0.2069 0.2018 0.1717 0.1668	0.0923 **0.0931** 0.0766 0.0779
(c) Specialist Lexicon	w1∪SL w1+SL b1∪SL b1+SL	0.2665 0.2759 0.2440 0.2536	0.1967 0.1959 0.1667 0.1637	0.0855 0.0887 0.0722 0.0755

Table 6. Evaluation results

Related work

The top 6 official runs at the TREC Genomics Track 2007 are generated by the following approaches. NLM and it's collaborators (Demner-Fushman, 2007) experimented three models: an interactive model (NLMinter), a fusion model (NMLfusion), and a knowledge-based model (LHNCBC). Two of them, NLMinter and NLMfusion, archived excellent performance. NLMinter used manually constructed queries consisting of a conjunction of topic terms and other additional terms. NMLfusion is the equally-weighted fusion of the results of four automatic IR methods. Whereas LHNCBC attempted to exploit semantic types and synonyms, the performance was not comparable to the leading runs. Both MuMshFd and MuMshFdRsc (Stokes *et al*., 2007) employ an automatic query expansion with entities and ontological terms. In addition, MuMshFdRsc applies passage reduction and re ranking. UniNE1 is a retrieval system based on Divergence from Randomness with WordNet (Fellbaum, 1998) expansions and UniNE3 is a fusion of three IR models incorporating with WordNet expansions (Fautsch and Savoy, 2007). Some paper (Jimeno and Pezik, 2007) claimed that query expansion could degrade IR performance.

Conclusion

In this study, UoM investigates whether using an in-domain dictionary is meaningful in the IR tasks. Our study shows that adding technical terms from Biolexicons to query terms improves Document and Passage2 MAP slightly and also better then using UMLS Specialist Lexicon.

2.4 Multilingual IR Assessment (UKLFR)

For the multilingual IR assessment task, UKLFR conductes an evaluation of a search engine based on the MORPHOSAURUS which was optimized for BOOTStrep (see Deliverable 10.1). The evaluation task of the

MORPHOSAURUS-based search system (Hahn *et al.*, 2001; Markó, 2005a; Markó, 2005b; Hahn *et al.*, 2005; Schulz *et al*., 2006) is based on two evaluation sets, TREC Genomics Tracks (Hersh and Voorhees, 2008) and Ohsumed (Hersh *et al*., 1994).

2.4.1 Evaluation Sets and Related Work

TREC Genomics Tracks

TREC Genomics is a linguistic Corpus delivered from TREC (Text REtrieval Conference). The TREC Genomic Track was founded by the National Science Foundation (NSF) Information Technology Research (ITR) program and ran from 2003 to 2007 (Hersh and Voorhees, 2008).

In 2003 the track was made with modest financial resources, and at that time, a one year subset of 525,938 MEDLINE records was obtained. As topics gene names were chosen with the goal of finding all MEDLINE references that had a relation with the gene or one of the proteins it encodes. In spite of the underestimating of the number of relevant documents, the collection still did allow researchers to start working with biologically oriented content, but it was not so helpful for testing IR systems. That is why the 2003 dataset was not used it in our evaluation.

In 2004, the ad hoc retrieval task was expanded by a larger test collection with true relevance judgments. The data collection was based on a ten-year MEDLINE subset and was used both by TREC 2004 and TREC 2005 with different topics (queries). The Medline subset records were extracted based on the Date Completed (DCOM) field for all entries from 1994 to 2003. This provided a total of 4,591,008 records (about one-third of the total MEDLINE dataset). Records without abstracts constituted 26.3% of the total.

Topics for the ad hoc task were collected by real biologists. In the 2004 track, simple information needs were collected and formatted into 50 topics with the following fields:

- ID: identifier
- TITLE: abbreviated statement of information need
- NEED: full statement of information need
- CONTEXT: background knowledge to place information need in context

Here some examples of the original queries:

“ID: 1 TITLE: Ferroportin-1 in humans NEED: Find articles about Ferroportin-1, an iron transporter, in humans. CONTEXT: Ferroportin1 (also known as

SLC40A1 ; Ferroportin 1; FPN1; HFE4; IREG1; Iron regulated gene 1; Iron-regulated transporter 1; MTP1; SLC11A3 ; and Solute carrier family 11 (proton-coupled divalent metal ion transporters), member 3) may play a role in iron transport."

"ID: 2 TITLE: Generating transgenic mice NEED: Find protocols for generating transgenic mice. CONTEXT: Determine protocols to generate transgenic mice having a single copy of the gene of interest at a specific location."
"ID: 3 TITLE: Time course for gene expression in mouse kidney NEED: What is the time course of gene expression in the murine developing kidney CONTEXT: Relevant articles describe genes involved in kidney development."

The best results in the 2004 track were obtained by a research group from the NIST (Fujita, 2004) with a MAP (medium average precision) of 0.4075. They made use of all the query fields information (title, need, and context) with a combination of Okapi weighting, Porter stemming, symbols expansion using LocusLink and MeSH records, and query expansion. Furthermore, they added a language modeling technique based on the Dirichlet-Prior smoothing and obtained a higher MAP value of 0.4264. One of the points with a negative impact on the improvement of the evaluation track is that the different groups tried a variety of approaches, without comparing their baseline results, which made it difficult to study what techniques provided better benefit and which techniques could be combined together or should not be used at all.

Similar to 2004, 2005 had also 50 topics, classified in five Generic Topic Types (GTTs) with ten topics per GTT. These GTTs consisted of semantic types, such as genes or diseases placed in a specific context. After the development of the GTTs, biologists were interviewed to obtain specific information needs that conformed to each GTT. Like almost all relevance judgment, relevance assessors judged each document as definitely relevant (DR), possibly relevant (PR), or not relevant (NR). For the official scope where a binary relevance judgment is required, both DR and PR classified documents were considered relevant. In the 2005 track, articles described specific genes, disease, therapy, mutation, etc. and not just a general overview. Moreover relevance judges were given more explicit instruction to the GTTs for example:

- An article is relevant if it explains how to improve, extend, perform, a therapy method or an experimental procedure.
- An article is relevant if it describes the specific function of a gene in specific diseases or biological processes.

In both 2004 and 2005 tracks, the main measure of performance was (MAP) .The topics were classified in three categories:
Automatic-no manual intervention in building queries

- Manual-manual construction of queries but no further human interventions.
- Interactive-full interactive construction of queries with manual interaction with system output

Here are some examples of the queries, one query for each GTT:
“Describe the procedure or methods for how to open up a cell through a process called electroporation"
“Provide information about the role of the gene APC (adenomatous polyposis coli) in the disease Colon Cancer.”
“Provide information on the role of the gene APC (adenomatous polyposis coli) in the process of actin assembly.”
“Provide information about the genes HNF4 and COUP-TF I in the suppression in the function of the liver”
“Provide information about Mutations of NM23 and its/their impact on tracheal development.”

In both 2004 and 2005 tracks, the main measure of performance was (MAP). Somewhat similar results were obtained in the 2005 track. As with 2004, the basic Okapi approach with good parameters resulted in a good baseline performance for a number of groups. Manual synonym expansion of queries gave the highest MAP of 0.302 (Huang *et al.,* 2005), whereas automated query expansion did not fare as well (Ando *et al.*, 2005).

Ohsumed
The Ohsumed data collection was created to assist information retrieval research. It constitutes a five-year (1987-1991) subset of 348,566 MEDLINE article records, covering references from 270 medical journals. The articles are clinically oriented and like all other corpora, not all information fields included in the collection files seemed relevant for the research purpose and had therefore been deleted. The relevant fields are title, abstract, MeSH terms, author, source, and publication type. The resulting document collection is about 400 megabyte in size. The physicians, who built the Ohsumed document collection – in the scope of a clinical study using MEDLINE (Hersh *et a*l., 1994), generated 106 queries. The queries were then controlled by four persons, two experienced physicians and two medical librarians. The relevance judgment was taken by multiple groups of physicians resulting in 16,140 unique query-reference pairs. The documents

were judged: definitely, possibly, or not relevant. In order to assess the errors related to humans, over 10% of the relevance judgments were made by two raters.

2.4.2 UKLFR Multilingual Search Engine

The Multilingual IR-system used for the evaluation purposes in the context of BOOTStrep mainly consists of three different components:

- the MORPHOSAURUS multilingual semantic indexing system,
- a multiword expression normalizer, and
- a standard open-source search engine.

MORPHOSAURUS

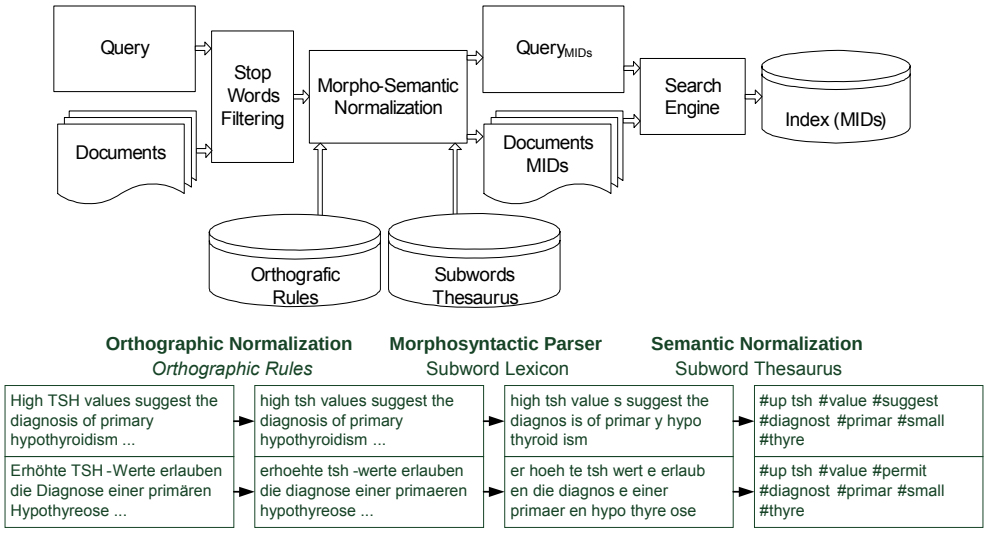


Figure 2. MorphoSaurus Processing Architecture (top), Morpho-Semantic Normalization pipeline (bottom)

As described in Deliverable 10.1, within the MORPHOSAURUS framework (an acronym for MORPHeme theSAURUS), subwords (words and word fragments that carry a non-decomposable beaning) are assembled in a multilingual lexicon and thesaurus, with the following considerations in mind:

- Subwords are listed with their attributes such as language (English, German, Portuguese, Spanish, French, and Swedish) and subword type (stem, prefix, suffix, invariant). Each lexicon entry is assigned to exactly one morpho-semantic identifier representing its equivalence class, the MID.

- Two kinds of semantic links between MIDs are added. 1. the relation *has-meaning*, which relates one ambiguous class to two or more non-ambiguous ones, and 2. the relation relation *expands-to*, which consists of predefined segmentations in case of utterly short subwords.

Figure 2 depicts how source documents are converted into an interlingual representation by a three-step procedure. First, each input word is orthographically normalized in terms of lower case characters and according to language-specific rules for the transcription of diacritics (top-right). Next, words are segmented into sequences of subwords from the lexicon (bottom-right). Finally, each meaning-bearing subword is replaced by one language-independent semantic identifier (MID), which unifies intralingual and interlingual (quasi-) synonyms, thus producing the interlingual output representation of the system (bottom-left). In **Figure 2**, bold-faced MIDs co-occur in both document fragments.

In our studies, both the document collections and queries (in different languages) were automatically transformed by the MorphoSaurus system into the language-independent MSI (morphosemantic indexing) Interlingua (plus lexical remainders). Finally, the MSI-coded queries were evaluated on the MSI-coded corpora at an interlingual representation level.

The baseline of the experiments is given by the Ohsumed and TREC Genomics 2004-2005 corpra both in terms of their Porter-stemmed English queries, as well as their Porter-stemmed English document collection.

Besides the evaluation of the cross-lingual access we also want to test the multilingual synonymy module that was developed over the last months within the MORPHOSAURUS Framework to allow the search for multiword expressions and to find synonyms of these multiword expressions.

Normalization of Multiword Expressions

In the biomedical lexicon, the use of multiword expressions (MWEs) constitutes an important part, especially gene and protein names, which could have dozens of MWE synonyms in different languages stored in multiple biomedical databases and research repositories using different conventional annotations.

Despite these constraints and difficulties, it is very motivating to recognize MWEs in text considering the biomedical context, because many linguistic aspects which MWEs could have are simply not present. Aspects like metaphors or the problem of situatedness are not present, since the document collections represent biomedical publication articles which have been written by scientists for research issues.

There are different definitions on multiword expressions which describe different aspects, but they all agree that a MWE is a sequence of words that transmit some well-defined meaning.
One of the constraints is the classification of MWEs and their quality. For example, the MWE “half a dozen eggs” is in a certain aspect synonym with “six eggs”. The classification of MWEs is very important to limit the complexity of the recognition process. In the above cited example we have recognized a semantic relation between the two MWEs, because humans, in contrast to machines, are able to semantically interpret such expressions. However there exist multiple kinds of linguistic specificities that differ from one language to another language, from one country to another country, and from an ethnic group to another ethnic group. Furthermore, not only semantically related phrases are MWEs, but sometimes there exist syntactically linked phrases which could be considered as MWEs synonyms.

To better understand this idea we can see the following MWEs classifications given by the linguistics research group at the Stanford University (http://mwe.stanford.edu):

- Institutionalization / conventionalization: The procedure which makes an expression accepted to a lexicon through continuing use over time.
- Lexico-grammatical fixity: How rigid is the relation between the words and which fixed phrases are MWEs and which are not. For example a lexicogrammatically fixed MWE: “kick the bucket” (which of course precludes variations like “the bucket was kicked”, “slowly kick the bucket”), and a lexicogrammatically fixed non-MWE: “look like”, “(to be) looked like”, “is looking like”.
- Semantic or pragmatic non-compositionality: Sometimes, semantics or pragmatics of certain parts differs from the whole; and some words could have a meaning in some context and other meaning in other context.
- Syntactic irregularity: for ex. syntactically-irregular: “all of a sudden”, syntactically regularity: “kick the bucket” and “fly off the handle”.
- Non-identifiability: when the meaning cannot be identifiable only from a part of the phrase, such a case could result due to misleading lexical clusters.
- Situatedness: The relation between MWEs is guaranteed only if there exist a pragmatic point, like the situated MWEs: “good morning” and “all aboard”, non-situated MWEs: “first off” and “to and fro”.

- Linguistic Figuration: If the phrase represents some metaphor, metonymy, hyperbole, etc...
- Phrase as proverb.

In addition to these multiple aspects, there exist more MWE classifications but the above seen examples satisfy our purpose.

The MWEs aspects help us to recognize and extract MWEs from every input text if and only if we already know which possible types of MWEs could be present in the text we have, otherwise the complexity of processing all possible representations could be very high, and the usability of such an approach would be inutile. For this reason, studying MWEs should be restricted in context and aspects to achieve motivating results.
The following is a typical example of a multiword expression provided by EBI:

- Preferred Term: Balbiani ring 2 chain
    - Synonym: Balbiani ring protein 2
    - Synonym: BR-2
    - Synonym: BR-6
    - Synonym: BR2
    - Synonym: Giant secretory protein I-B
    - Synonym: BR6

For the BOOTStrep project 6,872,790 MWEs with 356,468 synonym references were collected. This database includes only biomedical expressions which represent the main axe of the evaluation of MorphoSaurus with the TREC genomics document collection. The multiword expression normalizer, which translates all synonymous expressions to the preferred term, was evaluated in a separate condition. This evaluation was still based on a preliminary MWE lexicon.

Search Engine

For an unbiased evaluation, several experiments were run with LUCENE, a freely available open-source search engine which combines Boolean searching with a sophisticated statistical ranking model. Furthermore, this search engine has another advantage: it supports a rich query language like multi-field search, including more than ten different query operators. The files of the index were fed using different conditions: (1) stemmed text, (2) MORPHOSAURUS normalized, (3) multiword normalizer without and (4) with MORPHOSAURUS.

2.4.3 Evaluation

For the cross language IR scenarios, the queries were translated into German and French by native language domain experts.The performance measurement of an IR system is based both on the MAP measurement and the top 5 (p5), top 10 (p10) and top 20 (p20) hits of the search. If the five first retrieved documents were all relevant, we will have a higher p5 value than if they were not relevant. Then, if they occur in the right ranking as given in the gold standard, the p5 value would be equal to 1.0 that means a hundred percent precision. These performance values are important because the use of semantics, which means inclusion of all MORPHOSAURUS modules, could engender more retrieved documents in contrast to using plain text, which deals with the text baseline. Because we limit our retrieved documents to 1000, the probability that non relevant documents could be present in the top 1000 is higher using the semantics.

Results without Multiterm Normalization
The results of the evaluation using the TREC Genomics corpus are illustrated in **Table 7**.

Scenario	MAP	P5	P10	P20
en-plain-en	0.16	0.44	0.38	0.32
de-semantics-en	0.14 (87.5%)	0.37 (84.0%)	0.33 (86.8%)	0.27 (84.4%)
fr-semantics-en	0.13 (81.3%)	0.26 (59.0%)	0.25 (65.8%)	0.22 (68.6%)

Table 7. Results of TREC Genomics

Not surprisingly, with respect to mean average precision (MAP) and the quality of a few top ranked documents (top 5, 10, 20), the plain text search (English queries and English documents) yields best results. By using semantics, cross-lingual retrieval is made possible. For German-English, remarkable 88% of the monolingual baseline is reached (MAP), whilst for French-English the precision is still 81% of that for the baseline.

Table 8 presents the results of the evaluation using Ohsumed.

Scenario	MAP	P5	P10	P20
en-plain-en	0.19	0.39	0.32	0.27
de-semantics-en	0.21 (110.5%)	0.37 (94.9%)	0.33 (103.1%)	0.28 (103.7%)
fr-semantics-en	0.15 (78.9%)	0.30 (76.9%)	0.24 (75.0%)	0.20 (74.1%)

Table 8. Results of Ohsumed

The main difference in Ohsumed, compared to Genomics, is a higher MAP for the semantic enrichment in the German-English scenario, compared to the one for the monolingual baseline. One reason for this is the much smaller document collection in Ohsumed. On the other hand, Ohsumed focuses clinical terminology rather then a biomedical one (Genomics) and MORPHOSAURUS was initially designed for clinical documents only. Additionally, the main focus was on German-English cross-language retrieval, resulting in a higher-coverage German subword lexicon, which explains the good results in that scenario, which even outperforms the monolingual one by 111 percent. For French-English, the results are still encouraging by reaching 79% of the mean average precision of the English monolingual experimental setting.

Results with Multiterm Normalization

Surprisingly, the results of this evaluation step were below what we had expected, as the performance with multiword normalization was below the ones not incorporating the multiterms module (85% within the baseline condition).

This interesting result will require a thorough error analysis and subsequent modifications in the future. One reason that we have found so far concerns the quality of the multiterms lexicon. The synonymy relationship in the lexicon seems to be much too broad, incorporating upper/narrower terms, as well. For example, the multiword expression "*DNA repair protein radC homolog*" is being regarded synonymous to "*DNA repair*". Similar, "*50S ribosomal protein L31*" is set as a synonym to "*DNA repair enzyme*".

3 IE Evaluation

For IE evaluation, 4 investigations have been conducted: 1) FSU's evaluation on GRE extraction with GeneReg Corpus and real life direct comparison with REGULON database; 2) Evaluation of GRO on gene regulation event (GRE) extraction; 3) I2R's exploration with raw text and Stanford & Enju parsers for the GRE extraction with GeneReg corpus; 4) UoM's generic GRE extraction.

3.1 GRE Extraction at FSU

First of all, FSU has developed UIMA tools which is extremely useful for (semi-) manual evaluation of NLP processing results. Secondly, the gene regulation event extraction system developed by FSU as part of the activities of WP7 was

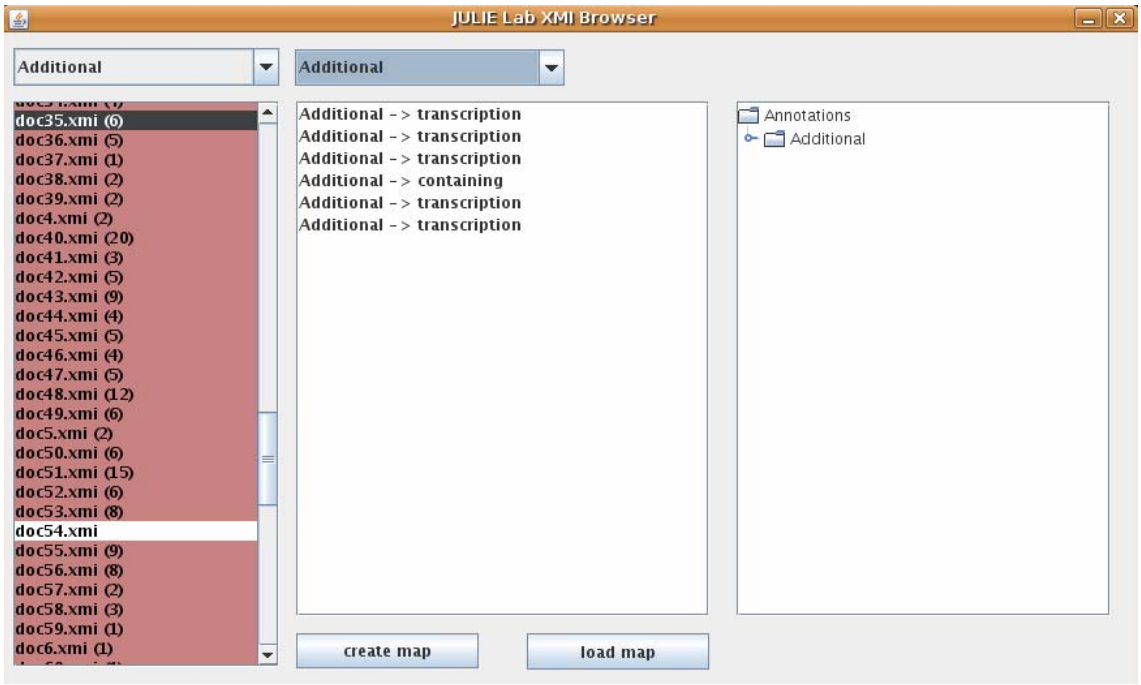


Figure 3. XMI Browser

extensively evaluated in both a clean-lab and a real-life scenario.

3.1.1 Evaluation within UIMA Framework

FSU has developed the *XMI Browser*, a browser to allow quick access and overview to the annotations added to documents after processing with the UIMA Tool-Suite (see WP 7). In the UIMA context, processed text including its annotations is stored in the XMI storage format.

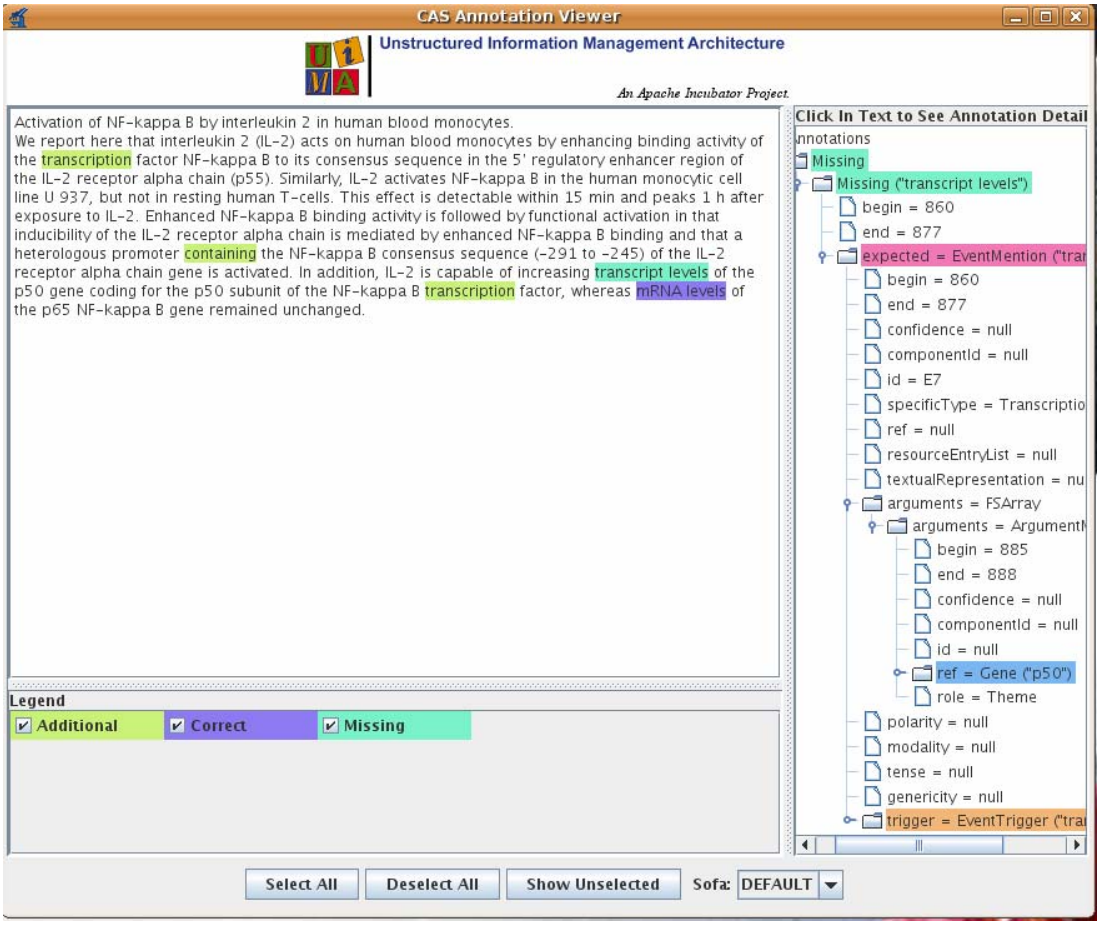


Figure 4. XMI Annotation Viewer

The XMI Browser is especially helpful when processing results are manually reviewed, e.g., for evaluation purposes. The browser allows quick access and overview to the annotations stored in the XMI files. **Figure 3** shows a screenshot of the browser: the left column shows a list of XMI files. Further, for each file, it shows how many annotations of the chosen annotation type (*Additional*, in our case here) are contained in the file. The middle column shows for a selected XMI file all the occurrences of a chosen annotation type (here, again `Additional`). Double-clicking on the respective document opens UIMA's XMI viewer, showing this specific type of annotations as can be seen in **Figure 4**.

To allow performance evaluation in terms of typical measures (F-score, Recall, Precision, etc.) *within* the UIMA framework, FSU has developed the *UIMA Evaluator*, an evaluation pipeline which allows to directly compare UIMA annotations, e.g., results of automatic processing with a gold standard.

Comparing two annotation sets (gold standard vs. automatic annotations), new annotations are added to the XMI file. These additional annotations store information about the differences of the compared documents: annotations of type *Correct* are added in case of a true positive, annotations of type `Missed` is added for a false negative error, and annotations of type `Additional` are added for false positive errors.

In a scenario-specific manner, one can define in what constitutes the different error types, i.e., what is considered a false negative error, etc. Based on these, three additional annotations, F-score, Recall, and Precision are calculated and shown.

For manual evaluation purposes, the additional "performance" annotations can be visualized using the XMI Browser described above; the XMI Browser can be use to quickly see which documents contain differences or errors and of which kind. Figure 3 shows which files contain false positive errors (annotations of type *Additional*).

The XMI Browser has been intensively used for the manual FP Analysis carried out by FSU as part of WP 11 (see Section 3.3.2?). Moreover, FSU has also systematically employed the XMI Browser and the UIMA Evaluator during the development and refinement of the FSU GRE extraction system. Domain experts (biologists, in this case) have been asked to review the different errors made by the system and categorize them. Based on this feedback, the system could be strongly improved as problematic cases unknown by that time were unveiled and their handling could be incorporated into the system.

3.1.2 FSU GRE Extraction System

FSU approaches the task of the automatic extraction of GREs with a machine learning-based (ML) system. No regularities are specified a priori by a human although, at least in the supervised scenario of our approach, this approach relies on training data supplied by human (expert) annotators who provide many instances of ground truth decisions from which regularities can automatically be learnt.

The extraction of GREs is a complex task composed of a series of subtasks. Abstracting away from lots of clerical and infrastructure services (e.g., sentence

splitting, tokenization) at the core of any GRE extraction lie the following basic steps:

- identification of pairs of gene mentions putatively the arguments of a relation -- the well-known named entity recognition and normalization task;
- decision whether the entity pairs really constitute a relation; and
- identification of the order of the arguments in the relation which implicitly amounts to characterize each arguments as either agent or patient.

Apart from the above mentioned pre-processing steps, FSU's ML-based extraction system for GREs requires several additional syntactic processing steps including POS-tagging, chunking, and full dependency- and constituency-based parsing. These tasks are accomplished by the UIMA components FSU developed as part of WP7 during the BOOTStrep project. All tools are trained on biomedical corpora to account for the sublanguage used in this special domain.

Entity Identification

For the first step, i.e, to identify gene names in the documents, FSU applied GeNo, a FSU approach to multi-organism gene name recognition and normalization (Wermter et al., 2009). GeNo was evaluated on the BioCreative II test set where it yields an overall F-score of 86.4% (precision: 87.8%, recall: 85.0%). These numbers show that GeNo is on a par with the best system on that task.

GeNo recognizes gene mentions by means of an ML-based named entity tagger trained on publicly available corpora. Then, it tries to map all identified mentions to organism-specific Uniprot identifiers. Mentions that cannot be mapped are discarded; only successfully mapped mentions are kept. GeNo is utilized in its original version, i.e., without special adjustments to the E. coli organism. For the GRE extraction, however, only those mentions detected to be genes of E. coli are fed into the relation extraction component.

Event Identification

FSU's approach to GRE is based on Maximum Entropy models. The approach is an extended variant to the approach described before in (Buyko et al., 2008) and Deliverable D.7.3. The extension includes the use of dependency parse information (e.g., dependency tree level features) and shortest dependency path information as features. In short, the complete feature set of FSU's approach consists of:

- word features (covering words before, after and between both entity mentions);

- entity features (accounting for combinations of entity types, flags indicating whether mentions have an overlap, and their mention level);
- chunking and constituency-based parsing features (concerned with head words of the phrases between two entity mentions; this class of features exploits constituency-based parsing as well and indicates, e.g., whether mentions are in the same NP, PP or VP);
- dependency parse features (analysing both the dependency levels of the arguments as discussed by Katrenko and Adriaans (2006) and dependency path structure between the arguments as described by Kim et al. (2008)); and
- relational trigger (key)words (accounting for the connection of trigger words and mentions in a full parse tree).

The FSU GRE extraction system allows for thresholding. To achieve higher recall values, the confidence threshold for the *negative class* (i.e., a pair of entity mentions does not constitute a relation) can be set to values > 0.5. Clearly, this is at the cost of precision as the system more readily assigns the *positive class*. However, as FSU's evaluations showed, significantly higher recall values can be achieved.

3.1.3 Evaluation of FSU GRE Extraction System

The FSU GRE extraction system is first ``intrinsically'' evaluated, i.e., in a cross-validation manner on our corpus annotated with respect to GREs. Second, in a more realistic scenario, the system is evaluated against REGULON, a database collecting knowledge about gene regulation in E. coli. This scenario tests which part of manually accumulated knowledge about gene regulation in E. coli can automatically be identified by the FSU GRE extraction system and at what level of quality.

Intrinsic Evaluation of Feasibility

GeneReg Corpus

The WP08 GeneReg corpus (Buyko et al., 2008) constitutes a selection of 314 Medline abstracts dealing with gene regulation in E. coli. These abstracts were randomly drawn from a set of 32,155 selected by MeSH term queries from Medline using keywords such as *Escherichia coli*, *Gene Expression* and *Transcription Factors*. These 314 abstracts were manually annotated for named entities (NEs) involved in gene regulatory processes (such as transcription factor, including co-factors and regulators, and genes) and pairwise *relations* between transcription factors (TFs) and genes, as well as triggers (e.g., clue verbs)

essential for the description of gene regulation relations, or GREs. As for the event types, the GeneReg corpus distinguishes between (a) unspecified regulation of gene expression, (b) positive, and (c) negative regulation of gene expression. The amounts of these three types of GREs in GeneReg are shown in **Table 9**. Combined, they account for 99.8% of the total number of GREs annotated. The type of the remaining GREs is annotated as *unknown*.

GRE type	**Positive**	**Negative**	**Unspecified**
Numbers of instances in GeneReg (V0.9)	636	331	548

Table 9. Numbers of Gene Regulatory Events in GeneReg (Buyko et al., 2008)

Out of the 314, a set of 65 randomly selected abstracts was annotated by a second annotator to identify inter-annotator agreement (IAA) values. For the task of correct identification of the pair of interacting named entities in gene regulation processes, an IAA of 78.4% (R), 77.3% (P), 77.8% (F) was measured, while 67% (R), 67.9% (P), 67.4\% (F) were achieved for the identification of interacting pairs plus the 3-way classification of the interaction relation. More details on the corpus can be found in (Buyko et al., 2008) and Deliverable D.7.3.

Experimental Setting

FSU GRE extraction system treats all of the above mentioned three types (unspecific, negative and positive) as one common type``relation of gene expression''. So, it either finds that there is a relation of interest between a pair of gold entity mentions or not. FSU evaluates the system by a 5-fold cross-validation on the GeneReg corpus. The fold splits were done on the abstract-level to avoid the otherwise unrealistic scenario where a system is trained on sentences from an abstract and evaluated on other sentences but from the same abstract (Pyysalo et al., 2008). As the focus is only on the performance of the GRE extraction component, gold entity mentions as annotated in the respective corpus are used.

Results

For the experimental settings given above, the system achieved an F-score of 42% with a precision of 59% and a recall of 33%. Increasing the confidence threshold for the negative class improves recall. **Table 10** summarizes performance values for different thresholds:

As expected, thresholding is at the cost of precision. It shows that using an extremely high threshold of 0.95 results in a dramatically increased recall of 73% compared to 33% with the default threshold. Although at the cost of diminished precision of 44% compared to originally 59%, the lifted threshold boosts the overall F-score by 2 points.

threshold	Recall	Precision	F-score
Default (0.5)	33%	59%	42%
0.8	54%	43%	48%
0.95	73%	32%	44%

Table 10. Generic GRE extraction task performance

Extrinsic Evaluation of Robustness

REGULON (http://regulondb.ccg.unam.mx/) is the primary and largest reference database providing manually curated knowledge of the transcriptional regulatory network of E. coli K12. On K12, approximately for one-third of K12's genes, information about their regulation is available. REGULON is updated with content from recent research papers on this issue. While REGULON contains much more information, the focus was solely on REGULON 's information about gene expression events in E. coli. In the following, the term REGULON refers to this part of the REGULON database. REGULON has the following information for each regulation event: regulatory gene (the ``agent'' in such an event, a transcription factor), the regulated gene (the ``patient''), the regulatory effect on the regulated gene (activating, suppression, dual, unknown), and evidence that supports the existence of the regulatory interaction.

Evaluation against REGULON constitutes a real-life scenario. Thus, the complete extraction system was, including gene name recognition and normalization as well as relation detection. Hence, the system's overall recall values are highly affected by the gene name identification. Gene names totally overlooked obviously permit the identification of respective relations.

Experimental Setting

To evaluate the extraction system against REGULON, FSU first processes a set of input documents (see below), collectes all unique GREs extracted and compared this set of events against the full set of known events in REGULON. A true positive (TP) hit is obtained, when an event found automatically corresponds to one in REGULON, i.e., having the same agent and patient. The type of regulation is not

considered. A false positive (FP) hit is counted, if an event was found which does not occur in the same way in REGULON, i.e., either patient or agent (or both) are wrong. False negatives (FN) are those events covered by REGULON but not found by a system automatically. From these hit values, standard precision, recall, and F-score values were calculated.

Of course, the system's performance largely depends on the size of the base corpus collection processed. Thus, for all five documents sets separate performance scores are obtained.

Table 11 gives an overview to the document collections used for evaluating the robustness of the FSU system: The ``ecoli-tf'' variants are documents filtered both with E. coli TF names and with relevance to E. coli (See 12 rules used in Section 2.1). The ``-relevant'' variants were filtered only with relevance to E. coli. These document sets were created by the EBI based on cascaded rules Abstracts are taken from Medline citations, while full texts are from a corpus of different biomedical journals. The third document set, ``regulon-ra'', is a set containing abstracts from the REGULON references.

document collection	document type	number of documents
ecoli-tf.abstracts	abstracts	4,347
ecoli-tf.fulltexts	fulltexts	1,812
ecoli-relevant.abstracts	abstracts	68,545
ecoli-relevant.fulltext	fulltexts	6,184
regulon ra	abstracts	2,704

Table 11. Document collections used for the extrinsic evaluation.

The FSU ML-based GRE extraction system is designed to recognize all types of gene regulation events. REGULON, however, contains only the subtype, i.e., regulation of transcription. Thus, the system is evaluated against REGULON in two modes: per default, all events extracted by the systems are considered; in the *TF-filtered* mode, only relations with an agent from the list of all known TFs in E. coli are considered. This list is available from REGULON's website.

Raw Performance Scores

The results of the FSU GRE extraction system are shown in **Table 12**.

mode	document set	Recall	Precision	F-score
TF-filtered	ecoli-tf.abstracts	9%	70%	16%
default	ecoli-tf.abstracts	9%	45%	15%
TF-filtered	ecoli-tf.fulltexts	10%	54%	17%
default	ecoli-tf.fulltexts	10%	29%	15%
TF-filtered	ecoli-relevant.abstracts	10%	68%	17%
default	ecoli-relevant.abstracts	10%	21%	13%
TF-filtered	ecoli-relevant.fulltexts	11%	53%	18%
default	ecoli-relevant.fulltexts	11%	14%	13%
TF-filtered	regulon ra	7%	78%	13%
default	regulon ra	7%	47%	12%

Table 12. Extrinsic evaluation results of FSU GRE extraction system

Recall values here range between 7 and 11%, while precision is between 14 and 78%, depending on both the document set as well as the application of the TF filter.

As already shown for the intrinsic evaluation, application of different confidence thresholds increases the recall of the system. This was also done for the evaluation against REGULON. **Table 13** shows the impact of increased confidence thresholds for the negative class on the regulon-ra set for the TF-filtered evaluation mode.

threshold	Recall	Precision	F-score
default (0.5)	7%	78%	13%
0.8	9%	70%	16%
0.95	11%	63%	19%

Table 13. Performance results under different threshold values

While the results for the first evaluation scenario, called intrinsic evaluation, are approximately state of the art with a best F-score of 44%, performance values in the real-life scenario are not so shiny with a best F-score on the order of 19% on

the regulon-ra set. This holds, in particular, for the comparison with the work of Rodriguez-Penagos et al. (2007). Still, the FSU approach is a much more general one than the above mentioned approach which consisted of a specifically tuned manual rule set for E. coli.

Manual Analysis of False Positives

REGULON is taken as an absolute gold standard for the evaluation described in this section. So, if the system would correctly extract an event which is not contained in REGULON for some reason, that would count as an FP. Moreover, all kinds of error (e.g., agent and patient mixed up) are subsumed as FP errors. To analyze the cause and distribution of FPs in more detail, a manual analysis of the FP errors was performed and original FP hits were assigned to one out of these four FP error categories:

Category 1: Not a GRE

This is really an FP error, as the extracted relation does not at all constitute a gene regulation event.

Category 2: GRE but other than transcription

Unlike REGULON which contains only one subtype of GREs, namely transcriptions, our system identifies all kinds of GREs. Therefore, it identifies events which, by definition, cannot be contained in REGULON and, therefore, are not really FPs.

Category 3: Partially correct transcription event

This category deals with incorrect arguments of GREs. We distinguish three types of FPs:

(a) the patient and the agent role are interchanged,

(b) the patient is wrong, while the agent is right, and

(c) the agent is wrong, while the patient is right.

In all these three cases, though errors were committed human curators might find the partially incorrect information useful to speed up a curation process.

Category 4: Relation missing in REGULON

Those are relations which should be contained in REGULON but are missing for some reason. The agent is a correct TF and the sentence contains a mention of an transcription event. There are several reasons why this relation was not found in REGULON as we will discuss in the following.

Table 14 shows the results of the manual FP analysis of the system (no TF filter applied) on the ecoli-tf-abstracts and ecoli-tf-fulltexts.

category	percentage of FP errors in ecoli-tf.abstracts	percentage of FP errors in ecoli-tf.fulltexts
1	44.5	54.5
2	11.2	10.9
3a	3.8	3.9
3b	8.5	4.4
3c	8.2	5.4
4	23.8	21.0

Table 14. Results of manual analysis of False Positives

It can be seen, that the largest source of error is due to category 1, i.e., an identified relation is completely wrong. As full text documents are generally more complex, the relative amount of this kind of errors is higher here than on abstracts (54.5% compared to 44.5%). However, on abstracts and full texts, a bit more than 10% of the FP are because the system found too general gene regulation events which by definition are not contained in REGULON.. GREs identified that were partially correct (category 3) constitute 20.4% (abstracts) or 13.4% (full texts) of the FP errors.

And finally, another 20% of the FPs are correct transcription events but could not be found in REGULON (category 4). There might be several reasons for it: identified gene names were incorrectly normalized so that they could not be found in REGULON: REGULON have not yet added a relation or overlooked it; relations are correctly identified as such in the narrow context of a paragraph of a document but were actually of speculative nature, this includes also if a document states only that it is ``likely'' or ``possibly'' that something is a relation.
To summarize the manual FP analysis, it shows that about 50% of all FPs are not completely erroneously identified relations. These numbers must clearly kept in mind when interpreting the performance scores reported on in the previous subsection. So, the analysis of false positives reveals that the strict criteria we applied for our evaluation may appear in another light for human curators. Confounded agents and patients (21% on the abstracts, 14% on full texts) and information not contained in Regulon (24% on the abstracts, 21% on full texts)

might still be useful from a heuristic perspective to focus on interesting data during the curation process.

3.2 GRE Extraction at EBI

Ongoing work at the EBI is concerned with the use of the gene regulation ontology (GRO) to identify gene regulatory events from the scientific literature. This work should lead to improvements for the information extraction from the scientific literature.

GRO represents types as well as relations between types. GRO covers not only is-a relations and part-of relations but also other relations such as has-agent and has-patient. The rich representation of relations used in GRO enables identification of complex events from the scientific literature. The conceptual knowledge in GRO has been used to shape an information extraction solution that embeds the rules of the GRO at different levels into the IT solution. The main benefit is to identify events that are given as an underspecified representation on gene regulatory events in the text. The solution is benchmarked against the annotated corpora and against bioinformatics data resources.

EBI has developed a rule-based system that first identifies syntactic structures of sentences by using the Enju parser (http://www-tsujii.is.s.u-tokyo.ac.jp/enju/), converts them into semantic structures with a pattern-based method, and deduces deep meaning from the semantic structures with an inference module. It contains 1,123 patterns for the semantic analysis in the domain of gene regulation and 28 inference rules representing the domain knowledge.

This rule-based system has focused on extracting gene transcription regulation events of E. coli for populating RegulonDB (http://regulondb.ccg.unam.mx/). It has extracted 992 unique events from the ecoli-tf.abstracts (4,347 abstracts), which cover 24% of RegulonDB events. It is also found that this rule-based system and the statistical system developed by FSU are complementary: While the later extracted 705 unique events, only 285 events are extracted by both systems.

3.3 GRE Extraction at I2R

I2R also adopts a supervised ML approach to the GRE extraction task on GeneReg corpus. Regarding the semantic category of the participating agents, the annotated GREs can be divided into two categories. One category contains the *core* events, whose agents are *transcription regulator*, *transcription factor*, or *transcription cofactor,* which FSU works on as described in section 3.1. The other category contains the rest of the annotated events that involve *polymerase* or *ligand* as agents. While it is biologically reasonable to distinguish these two categories of GREs, such a distinction between GREs is linguistically less discernible. Consequently, I2R ignored this distinction so the I2R GRE extraction system is able to learn from as many annotated GRE instances as possible.

As the supervised GRE extraction system needs the negative class of pairs of entities – those that are *not* involved in a GRE – for training and test, these pairs of entities were created by simply collecting pairs of annotated entities from within each sentence in the corpus. Therefore the relevant subtask of Named Entity Recognition (NER) is no longer a potential source of noise in I2R's system evaluation. Thus I2R conducts the intrinsic evaluation similar as FSU (Section 3.1.2).

The numbers of successfully collected positive class of entity pairs (those involved in positive, negative, or unspecified GREs) and negative entity pairs (those not involved in any GREs) after preprocessing, including parsing, are listed in **Table 15**. While successful parsing every sentence remains as a challenge, only a few GREs (less than 6% for any of the three major types) are missed after the preprocessing is completed.

Entity pair class		**Positive Entity Pairs**			**Negative Entity Pairs**
	GRE type	**Positive**	**Negative**	**Unspecified**	**Null**
Numbers of entity pairs	Enju	612	319	518	11239
	Stanford	615	324	523	11441

Table 15. Numbers of collected GREs in GeneReg

Besides generic GRE extraction as evaluated by FSU earlier in this report, I2R further accesses the extraction performances of each subcategory with positive, negative or unspecified regulation. Standard ten-fold cross validations are used to evaluate the I2R GRE extraction system.

Core Features

The I2R GRE extraction system for GRE extraction is adapted from the relation extraction system described in (Zhou et al. 2005). Similar to the (unextended) FSU system described in D.7.3, the annotated *triggers* that are intuitively highly relevant to GREs (Section 3.3, D.7.3) are incorporated as well. Together, these features are denoted as the *core* features in the remaining part of this section.

Effects of Different Parsers

Instead of using the Collins' parser (Zhou et al. 2005), I2R has compared the results of using two different parsers: the Stanford parser and the Enju parser in the project. The Stanford parser is mainly trained on the Penn Treebank but also includes "some GENIA training material", according to Christopher Manning's reply to the mailing list of the parser. Meanwhile, the Enju parser comes with a biomedical parsing model. While both parsers have been adapted to the biomedical domain, it is interesting to see how their different adaptation approaches might affect a GRE extraction system.

	Positive	Negative	Unspecific	Generic
Core (by Stanford)	48.4	31.5	37.7	68.8
Core (by Enju)	48.1	**38.2**	36.1	68.8

Table 16. Enju parser performs better at Negative GRE extraction

The performance results in F-score are listed in **Table 16**, show that the Enju parser is particularly helpful for extraction of negative GREs. The F-score achieved by Enju for negative GRE extraction is 38.2, or 6.7 higher than its counterpart, the Stanford parser. Whether the boost in this particular negative GRE type is related to the smaller amount of negative GREs is to be confirmed by further investigation. Besides as Enju parser provides more information beyond the parse tree, we'll further explore the other information such as phrase head as well.

Incorporation of Information from Raw Text

Supervised learning always suffers from not enough training data. Extra training instances gathered from the web appears very useful for the relation extraction task (Yong and Su, 2008), which is able to boost up the performance up to 31% F-score for those relations with limited training data. In this project, I2R explores the extra entity pairs from the E. coli K12 Strain corpus distributed by EBI

(Section 2.1 *GR IR Evaluation Set (EBI, I2R)*), which contains 68,545 MedLine abstracts without the manual annotations of GRE.

For each positive entity pair from a positive, negative, or unspecified GRE of GeneReg Corpus, the search engine Lucene is used to retrieve from the E. coli K12 Strain corpus up to 150 individual sentences with both entities. As it's possible that the same entity pairs may hold different relations or even no semantic relation at all in the raw corpus, thematic clustering is used to further identify those pairs with the same relations in GeneReg corpus. I2R GRE extraction system extracts features from these automatic extracted similarly as from the original GeneReg sentence where the entity pair is found. These features are used as auxiliary features accordingly.

	Positive	**Negative**	**Unspecific**	**Generic**
Core	48.4	31.5	37.7	68.8
Core+RawText	48.3	**36.8**	**38.3**	**70.2**

Table 17. Extra examples effectively improve the GRE extraction with Stanford parser

	Positive	**Negative**	**Unspecific**	**Generic**
Core	48.1	38.2	36.1	68.8
Core+RawText	**48.5**	**39.7**	**38.0**	**69.6**

Table 18. Extra examples effectively improve the GRE extraction with Enju parser

Table 17 and **18** show that adding such information from extra examples generally increases the extraction accuracy when compared to the results obtained by using only the core features, regardless of which parser is used. This is in line with the hypothesis that extra examples can be helpful when the provided ones are limited.

Meanwhile, compared to *Core (by Stanford)* in **Table 17**, *Core+RawText (by Stanford)* has higher F-scores at extracting negative, unspecific (and also generic) GREs but its accuracy at extracting positive GREs (the type with the most instances) drops. This reflects the fact that there are still noises with the entity pairs from the raw text, which we may try to reduce in more effective ways.

Under-sampling

In D.7.3, FSU reported that under-sampling improved the GRE extraction performance for all GRE types but negative GREs. I2R carries out experiments to further estimate to what extent GRE extraction can benefit from under-sampling. The finding is that making the training instances more balanced between positive class and negative class by under-sampling does improve the performance of the GRE extraction system. However, the experimental results also show that it is not a straightforward task to find out the best possible training set, or the optimal under-sampling ratio.

As can be seen from **Table 15**, the negative entity pairs dominant GeneReg's GRE training set (for instance, the positive entity pairs only account for less than 5% of all the entity pairs). This unbalanced training set may cause deteriorated system performance because it differs from the expected – balanced – training set. One way to alleviate this problem is to artificially remove some of the instances in the dominant negative entity pairs by under-sampling.

Initially in I2R experiments, 30%, 50% and 70% of the entity pairs of the negative class were randomly dropped from the training set, respectively; and no entity pairs were dropped from the test set for these three runs. The results are shown in **Table 19**. Clearly, the I2R GRE extraction system benefits from a more balanced training set after removing some negative entity pairs. For instance, the F-scores obtained for positive GRE extraction range from 52.8 to 55.6, at least 4.3 points higher than the best result achieved without under-sampling (Row 1 in the table, a copy of the "Core+RawText" row in **Table 18**). The F-scores obtained for negative and unspecified GREs are higher than previous best results, as well.

Percentage removed	Positive	Negative	Unspecific	Generic
0% negative entity pair	48.5	39.7	38.0	69.6
30% negative entity pairs	54.1	40.0	38.3	70.0
50% negative entity pairs	52.8	40.7	**46.6**	70.2
70% negative entity pairs	**55.6**	**41.8**	45.9	**70.8**

Table 19. Results with under-sampling

Just looking at **Table 19**, one may have the impression that as long as there are more negative entity pairs than positive ones, the more negative entity pairs are removed (the more comparable the two classes are in size), the better the

extraction results would be. Removing 70% of the negative entity pairs apparently leads to better results than removing only 30% does. However, further experiments show that this could be an arguable conclusion.

	Positive	Negative	Unspecific	Generic
Core+RawText (Under-sampleing Run 1)	**54.1**	40.0	**38.3**	70.0
Core+RawText (Under-sampleing Run 2)	**48.8**	**44.7**	41.4	70.5
Core+RawText (Under-sampleing Run 3)	50.5	41.1	41.4	70.8
Core+RawText (Under-sampleing Run 4)	52.3	**39.7**	**43.2**	71.0
Core+RawText (Under-sampleing Run 5)	52.8	41.8	**43.2**	71.0

Table 20. Performance fluctuates with 30% negative entity pairs randomly removed in multiple under-sampling runs

Here is the reason. During under-sampling, the negative entity pairs are removed randomly. This means if under-sampling experiment is repeated, different training sets are produced each time even when the same amount of negative entity pairs are removed. Consequently, different extraction results can be expected. **Table 20** shows the F-scores of multiple (5) runs of under-sampling experiments, each run with a 30% removal (The first row repeats the *second* row in **Table 19**). The extraction accuracy for each GRE type changes from run to run. The associated fluctuations are rather high – around 5 points for all the three GRE types: positive, negative and unspecific. Considering the extraction accuracy differences between different removal amounts shown in **Table 19** are usually less than 5 points, it is not safe to conclude which removal amount is better by just comparing the results of a single run of them. How to select the optimal under-sampling ratio and remove negative entity pairs thus remains as an open question.

On the other hand, the performances of generic GRE extraction appears quite stable with different under-sampling ratios and different negative entity pairs being removed. This seems to tell that under-sampling is mainly useful for the cases when training data is too limited.

3.4 GRE Extraction at UoM

UoM takes an approach to GRE extraction based on a corpus of MEDLINE abstracts that has been annotated with instances of GREs by a group of domain experts. In particular, being different from FSU, EBI and I2R, UoM targets multi-

slot template extraction for GRE extraction. Instead of identifying whether a specific pair of entities is involved in a Regulation of Gene Expression event, UoM extract slots of a pre-defined template from the abstracts.
In the general domain, a number of pioneers in the 1990s investigated methods for learning IE rules from annotated corpora (Riloff, 1996; Soderland *et al.,* 1995; Kim *et el.,* 1995; Huffman, 1996; Califf and Mooney, 1997). Following these studies, UoM takes an approach to inducing textual rules that extract biological events from text.

3.4.1 GREs

Table 21 shows the IE template for GREs.

Slot name	**Description**
Agent	Drives/instigates event
Action (Verb)	Action/process
Theme	a) Affected by/results from event b) Focus of events describing states
Manner	Method/way in which event is carried out
Instrument	Used to carry out event
Location	Where *complete* event takes place
Source	Start point of event
Destination	End point of event
Temporal	Situates event in time w.r.t another event
Condition	Environmental conditions/changes in conditions
Rate	Change of level or rate
Descriptive-Agent	Provides descriptive information about the AGENT of the event
Descriptive-Theme	Provides descriptive information about the AGENT of the event
Purpose	Purpose/reason for the event occurring

Table 21. GRE template

Only a subset of these slots may be present in a target text for some event instances. For example, consider the sentence *Cytotoxic T lymphocyte antigen-4 plays a critical role in negatively regulating T cell responses*. This sentence

contains slot values corresponding to the *Semantic Roles* of *Agent*, *Action*, *Theme*, and *Manner* as in **Table 22:**

Slot	**Value**
Agent	Cytotoxic T lymphocyte antigen-4
Action	regulating
Theme	T cell responses
Manner	negatively

Table 22. Slot can be filled from "*Cytotoxic T lymphocyte antigen-4 plays a critical role in negatively regulating T cell responses*."

Extracted event instances are represented as *feature terms* with the following form:

event($slot_1$ => $value_1$, … , $slot_n$ => $value_n$,).

where

- $slot_i$ are the names of the slots in **Table 21**.
- $value_i$ are a sequence of consecutive words.

For example, the event in the above table is represented as follows:

event(Agent=> " Cytotoxic T lymphocyte antigen-4",
Action=> "regulating",
Theme=> "T cell responses",
Manner=> "negatively").

To generate IE rules from biological event annotated corpus, UoM represents events in an XML-style inline format consisting of three different types of element, rather than a stand-off format:

EVENT – surrounds text spans containing the *action* of the event, *e.g. regulating.* This can either be a VP or an NP, depending on whether the action is described by a verb or nominalised verb (*e.g. regulation*)

SLOT – surrounds text spans corresponding to event slot values. The *eventid* attribute links each slot with its respective event, whilst the *Role* attribute indicates the *semantic role* (*e.g. Agent*) .The verb/nominalised verb describing the action of the event is annotated using the *Role* value of *Verb.*

NE – surrounds text spans annotated as named entities. The *cat* attribute stores the NE category assigned.

For example, the annotation of the sentence *Cytotoxic T lymphocyte antigen-4 plays a critical role in negatively regulating T cell responses* is represented as follows:

<SLOT eventid="1" Role="Agent"> <NE cat=”Proteins”>Cytotoxic T lymphocyte antigen-4</NE> </SLOT> plays a critical role in <SLOT eventid="1" Role= "Manner"> negatively</SLOT> <EVENT id=”1”> <SLOT eventid=”1” Role="Verb"> regulating </SLOT> </EVENT> <SLOT eventid=”1” Role=”Theme”>T cell responses</SLOT>.

The Action slot is annotated as the SLOT tag whose role attribute is "Verb". Other template slots, *i.e.* Agent, Manner, and Theme, are directly linked to the "Role" attribute of SLOT tags.

3.4.2 Event Extraction Rules

Event extraction rules take the following form:

event($slot_1$ =>X_1, … $slot_n$ =>X_n,) :- *event constraint clause.*

For each event, an event constraint clause is constructed by considering the *event annotation span*. This span begins with the earliest SLOT span associated with the event, and ends with the latest SLOT span. The event annotation span is split into single-word tokens, except for event slot values, which are treated as multi-word tokens.

Constraint clause generation from annotations

For each word in the span, the following information is included in the event constraint clause:

- Every word *w* has a unique reference variable *X* and is constrained as *X:w*.
- If the part-of-speech of word *w* is *p*, the part-of-speech constraint is denoted as *X:p,* where X is the reference variable of *w*.
- If word *w* is annotated as a named entity, *i.e.* <NE type=”*NE*”>*w*</NE>, the NE constraint is denoted as *X:ne;* where *X* is the reference variable of *w.*
- If word *w* is annotated as a slot filler, *i.e.*, <SLOT eventid=”*n*” Role=”*role*”>*w* </SLOT>, the semantic role constraint is denoted as *X:role*; otherwise *X*:o where *X* is the reference variable of *w.*
- Word order is designated by X_1>X_2, which means that the word referenced by X_2 follows the word referenced by X_1.

The event annotation shown in Example 1 is thus converted into the following event constraint clause:

X1:"Cytotoxic T lymphocyte antigen-4", X1:'Agent',X1:'NN',X1:'Protein',X1>X2,
X2:"plays", X2:o, X2:'VVZ',X2>X3,
X3:"a", X3:o, X3:'DT', X3>X4,
X4:"critical", X4:o, X4:'JJ', X4>X5
X5:"role", X5:o X5:'NN',X5>X6,
X6:"in", X6:o, X6:'IN',X6>X7,
X7:"negatively",X7:'Manner',X7:'RB', X7>X8,
X8:"regulating", X8:'Action', X8:'VVG', X8>X9
X9:"T cell responses", X9:'Theme', X9:'NNS'.

An event extraction rule is a generalized form of event constraints extracted from the annotated corpus. A set of constraint clause *C'* is more general than *C* if $C \supseteq C'$.

The following is an example of an event extraction rule:

event(Agent=>Y1,Action=>Y8,Theme=>Y9, Manner=>Y7) :-
Y1:'Agent',Y1:'Protein',Y1>Y2,
Y2:o,Y2:"plays",Y2>Y3,
Y3:o,Y3:"a",Y3>Y4,
Y4:o,Y4:"critical",Y4>Y5
Y5:o,Y5:"role",Y5>Y6,
Y6:o,Y6:"in",Y6>Y7,
Y7:'Manner',Y7:'RB',Y7>Y8,
Y8:'Action',Y8:'VVG',Y8>Y9
Y9:'Theme', Y9:'NNS'.

If the rule can be successfully unified with a constraint clause generated from an event annotation span, then the fillers of the *Agent, Action, Theme*, and *Manner* slots are extracted as words that are referenced by Y1, Y8, Y9, and Y7, respectively.

The method of inducing event extraction rules is presented in Section 3.4.4.

3.4.3 Event Extraction Method

The UoM approach to biological IE is text-based. It combines statistical and symbolic approaches.

The event extraction algorithm consists of the following steps.

1. Analyze a sentence with a POS tagger.
2. Automatically annotate gene/protein names with an in-house general gene/protein name recognizer (anonymous, 2008).
3. Apply gene regulation NER.
4. Apply gene regulation SRL.
5. Apply event extraction rules to a word sequence that is constrained by syntactic and semantic information.

The gene regulation *named entity recognition (NER)* and *semantic role labelling (SRL)* tools have been newly developed by UoM for the purposes of event extraction. Sequential labelling is employed to identify named entities and semantic roles. UoM employed Conditional Random Fields (CRFs) (Lafferty *et al.*, 2001), as CRFs have previously been applied to NER and SRL tasks successfully. (*e.g.,* McCallum and Li, 2003; Settles, 2004).

Since it is very costly to create a sufficiently large corpus with a wide variety of general biomedical named entities and semantic role annotations, UoM has produced the NER and SRL tools using the UoM GRE corpus, which is annotated only with named entities and semantic roles that are directly related to GREs. This means, for example, that gene/protein names that are irrelevant to gene regulation are not annotated in the corpus.

a) Gene Regulation NER

Gene regulation NER generates named entity annotations which can be used as constraints in event extraction rules. Gene regulation NER models are trained using the standard IOB2 labeling method (Sang, 2000). That is, the label "B-*NE*" is given to the first token of the target *NE* sequence, "I-*NE*" to each remaining token in the target sequence, and "o" to other tokens.

Features used are as follows:

- Word feature: surface word
- POS feature
- Suffix feature: last two and four letters
- Word shape: capital letters in a word are normalised to "A", lower case letters are normalised to "a", and digits are replaced by "0". For example, the word form "IL-2" is normalised to "AA-0".
    - The first letter
    - The last four letters
- General purpose gene/protein labels

UoM applied first-order CRFs using the above features for the tokens within a window size of −2, −1,0,+1,+2 positions of the current word.

b) Gene Regulation SRL

Gene regulation semantic roles are annotated to facilitate the application of semantic role constraints by event extraction rules. The semantic role labelling models are trained using CRFs in a similar way to NER. That is, the label ``B-*Role*'' is given to the first token of the target *Role* sequence, "I-*Role*" to each remaining token in the target sequence, and "o" to other tokens. Features used here are the same as for NE labelling, except general gene/protein NER labels.

c) GRE Extraction

Provided with an input sentence, sequential labelling models determine NE and semantic role labels of tokenized input sentences. The word sequences are subsequently converted into ordered words with syntactic and semantic constraints.

Using the output of NER and semantic role labelling, a *constraint clause* is generated for each unannotated input sentence, using the same format as the constraint clauses generated from the annotated corpus.

Constraint clause generation from text

Similar to construction of constraint clauses from an annotated corpus, a set of constraints will be generated from an unannotated sentence.

- Every word *w* has a unique reference variable *X*, which is denoted as: *X:w.*
- If the part-of-speech of word *w* is *p*, the part-of-speech constraint is denoted as *X:p,* where X is the reference variable of *w*.
- Each word labelled as an NE by the named entity recognizer is constrained with named entity *ne* as *X:ne*, where X is the reference variable of *w*.
- Each word to which an SRL is assigned by the semantic role labeller is constrained with semantic role label *role* as *X:role*, where X is the reference variable of *w*.
- Word order is designated by $X_1 > X_2$.

Finally, each event extraction rule is applied to the constraint clauses generated from a given sentence one by one. Matching of an event extraction constraint clause and a constraint clause from a sentence is straightforward: (1) align two reference variable sequences in order (*i.e.* one from the rule and the other from the sentence) and (2) check if all the constraints from the extraction rule match the constraints from the sentence. If several identical events are extracted by some rules, the redundant events are removed.

3.4.4 Textual Induction of Extraction Rules

Biological information extraction is in its early stages. As domain experts have different views on biological events, the type of information to be extracted is often ill-defined when compared to the domains that conventional IE studies have targeted, such as job postings and corporate merger domains. Moreover, the bio-event extraction task defined in this report is more complex than conventional bio-IE tasks, such as the extraction of protein-protein interactions. Following close consultation with biologists in order to elicit their precise requirements concerning GRE extraction, UoM determined that fourteen kinds of slots should be extracted.

UoM has therefore chosen to investigate whether conventional IE rule generation techniques are effective in the biological domain. In this respect, UoM took an approach to generating IE rules in a bottom-up manner similar to CRYSTAL (Soderland *et al.,* 1995). The bottom-up IE rule induction is defined as follows.

1. Extract an event constraint clause from the annotated training corpus as described in **Section 3.4.2**.
2. Create a set of constraint clauses from unannotated texts as described in **Section 3.4.3**.
3. Remove all the word constraints *X:w* where *x* is constrained by an NE or a semantic role. (**baseline rule**)
4. For each constraint *X:w* in the current clause
   4-1 Remove *X:w*
   4-2 Use the remaining constraints to extract event instances from constraint clauses generated from the unannotated training text.
   4-2 Count the number of true positive *tp* events and false positive *fp* events.
   4-3. Calculate the Laplace estimate (Cestnik, 1990) (*tp*+1*)*/(*tp*+*fp*+2).
5. If the best score is lower than a threshold2, return the current clause.

---

2 The threshold is set to 2/3, considering that *tp*=1 and *fp*=0 for expected common generalization situation.

6. Remove the constraint with the best score
7. Repeat from Step 4.

3.4.5 Corpus Annotation

In order to facilitate evaluation of the Bio-Lexicon for information extraction, a further phase of Bio-Event Linguistic Annotation (BELA) was carried at UoM. The original phase of this annotation was carried in WP4 to facilitate extraction of semantic event frames to be included within the Bio-Lexicon. As the WP4 annotation was geared towards this specific goal, the annotations produced did not necessarily correspond to complete linguistic *event instances.* However, for the purposes of WP11, a corpus in which the annotations correspond to linguistic event instances was required. This would allow evaluation of an information extraction (IE) system trained to recognize linguistic event instances with the aid of the semantic frames contained within the Bio-Lexicon.

Many of the details of the annotation remain the same as the annotation performed during WP4. As the details of this annotation are explained in detail in D4.1, this section focuses only on the changes that were made to the annotation scheme, guidelines and software prior to the second phase of annotation. Certain changes to the scheme and software were carried out as part of complementary annotation work in WP7 and WP8, as well as in WP11. However, for the sake of clarity, all changes made since WP4 are reported in this section. The updated annotation guidelines are also included in Annex 2.

The main changes made for the WP11 annotation by UoM are as follows:

- **Scheme** - The WP11 annotations constitute *event instances*, which are the target annotations of the IE system. This is in contrast to the WP4 annotation, where the annotation was geared towards the extraction of event frames. One of the main differences is that for event instances, all items in lists should be annotated.
- **Guidelines** – Errors made by annotators during WP4 were analysed and used to identify potential weaknesses in the first version of the guidelines. As a result, changes were made to sections on argument text span selection, semantic role assignment and NE category assignment. The categories within the NE hierarchies were also reorganized.
- **Software**- A number of updates were made to the WordFreak annotation software. These included reducing the number of automatically higlhlighted biologically-relevant verbs from 700 to 323, according to the

verbs that were actually annotated during the WP4 annotation. In addition, updates were made to solve some problems with syntactic chunks encountered during WP4. Finally, the part of the interface for assigning NE categories was redesigned, to make the hierarchical structure of the concepts more explicit and easier to follow.

- **Abstract selection** - The WP4 annotation consisted of 2 different levels of Bio-Event annotation. In addition to the linguistically-oriented BELA annotation, there was also the higher-level Bio-Event Biological Annotation (BEBA), produced by EBI, which considered events from a more biological viewpoint, possibly drawing information from several sentences. As it is envisaged that the recognition of BELA events may constitute a first step in the recognition of BEBA events, the WP11 BELA annotation was focussed on the annotation of abstracts that had already undergone BEBA annotation. It was felt that the production of a corpus consisting of 2 levels of event instance annotations would provide a valuable resource for IE training and evaluation. Prior to undergoing WP11 BELA annotation, each abstracts was reviewed by a biology expert at UoM for relevance to gene regulation. As a result of this process, a total of 167 *E. coli* abstracts and 77 human abstracts were selected for annotation.

Annotator Recruitment and Training

At UoM, 7 PhD students were recruited, of whom 2 also carried out the WP4 annotation, all with at least some experience of gene regulation, and with native or near-native competency in English. This last requirement was imposed due to the complexities of the task.

Following an initial training session, a training program was begun. Based of the experiences of the WP4 annotation, a more well-structured training program was devised. In addition to regular group meetings, it was felt that the production of regular, individual feedback reports for annotators would be advantageous, due to the fact that many errors made are annotator-specific. Training proceeded in fortnightly cycles, with annotation in the 1st week and production of feedback reports in the 2nd week by 2 researchers (a computational linguist and a biologist). This schedule allowed annotators to review their feedback prior to carrying out further annotation, as well as providing more time for the researchers to review the annotations prior to providing feedback.

Calculating Agreement

Prior to providing information about inter-annotator agreement during the training period, we provide in this section a description of the way in which UoM has calculated inter-annotator agreement; this has changed from the direct agreement rates which were reported in D4.1 concerning the WP4 annotation.

The Kappa statistic is a standard way of calculating inter-annotator agreement for classification tasks. However, its calculation is problematic for most of the annotation tasks described in this report. Calculation of Kappa requires classifications to correspond to mutually exclusive and discrete categories. The annotation tasks for which agreement has been calculated are as follows:

1. Event identification (how frequently annotators agree on which events to annotate)
2. Argument identification (for agreed events, how frequently the same arguments are chosen by each annotator)
3. Semantic role assignment (for agreed arguments, how often the same semantic roles are assigned by each annotator)
4. Biological concept identification (within agreed arguments, how often do annotators identify the same biological concepts)
5. Biological concept category assignment (for agreed biological concepts, how often are the assigned categories agreed upon by each annotator)

The only one of these tasks to which Kappa can be straightforwardly applied is semantic role assignment, where each semantic argument is assigned one of 13 different role types. Whilst biological concept category assignment is also a classification task, calculation of Kappa is more problematic due to the hierarchical structure of the categories, meaning that they are not mutually exclusive.

Due to these problems, UoM has chosen to follow Hripcsac & Rothschild (2005) in choosing the *F-Measure* to calculate inter-annotator agreement. Its use means that we can straightforwardly compare the performance of information extraction systems trained using the annotated data with the human annotator performance. The F-measure can be calculated straightforwardly for all annotation tasks described above, allowing annotators' performance in various tasks to be compared easily. Unlike Kappa, there is no requirement for categories to be mutually exclusive. For the purposes of calculating inter-annotator agreement, precision and recall between two annotators can be calculated by treating one set of annotations as the gold standard. The F-measure is the same whichever set of annotations is used as the gold standard (Brants, 2000).

Inter-annotator Agreement During Training

Table 23 illustrates the changes in the inter-annotator agreement rates as the training period progressed. Each column of the table provides the inter-annotator agreement figures calculated after each cycle of the annotator training (e.g. C1 represents the first cycle of training). For comparison purposes, the final column of the table contains the inter-annotator agreement scores achieved in the final corpus of the WP4 annotation. It should also be noted that the first 4 cycles of training were concerned with annotation of *E.coli* abstracts. For the final cycle of training, UoM switched to human abstracts, as part of the final corpus would also consist of these. We thus wanted to verify to what extent annotation quality could be maintained following a change of subject, and also identify any potential problems that this may cause.

AGREEMENT RATE	C1[3]	C2	C3	**C4**	C5[4]	**WP4 final**
Event identification	58.35	56.01	68.26	**77.07**	71.94	57.31
Arg. identification (partial span match)	80.45	85.05	91.45	**89.39**	91.09	90.12
Arg. identification (exact span match)	61.92	63.98	73.96	**79.84**	79.17	75.58
Semantic role assignment	67.27	75.21	93.91	**84.89**	86.59	81.17
NE identification	71.35	78.65	78.29	**88.55**	82.36	73.20
NE category assignment (exact category)	72.34	72.05	71.61	**68.84**	59.76	67.94
NE supercategory assignment	89.21	89.32	93.45	**90.57**	84.09	93.46
NE cat assignment (inc. parent)	77.53	76.74	75.11	**71.58**	63.65	73.27

Table 23. Inter-annotator agreement scores achieved during each cycle of training

For most annotation sub-tasks, annotator performance at the end of the training period either equals of exceeds the performance achieved in the fnal WP4 corpus. Particularly of note are the figures in the *event identification* row. Even in cycle C1, the agreement rates are higher than the agreement rate achieved in the final WP4 corpus. This shows that the updated guidelines and extra emphasis of the correct events to annotate during the initial training sessions had the desired effect. Further training and feedback caused this agreement to rise by almost 19% to 77.07% by cycle C4.

[3] Annotation of *E.coli* abstracts
[4] Annotation of human abstracts

In many cases, the highest agreement rates were achieved in training cycle C4, which was the last cycle using *E.coli* abstracts. When the abstract subject was changed to *Human* in cycle C5, many of the agreement rates dropped slightly, suggesting that an adjustment period is required when the subject changes. However, for semantic role assignment and identification of arguments, the agreement rates stay constant or continue to rise, even when the subject of the abstracts is changed. This suggests that these tasks are more domain-independent, once annotators have got to grips with them.
The results for the assignment of NE categories show a different trend. Although an agreement rate of around 70% seems respectable, given the complexity of the task (73 possible NE categories), there was no discernible improvement, and even a slight decline, during the training period. Despite this, agreement achieved is slightly higher than during the WP4 annotation, suggesting that the reorganisation of the terms may have had some effect.

There are a number of possible reasons for the lower agreement of NE categories. Firstly, as the training period progressed, it seemed that the annotators were taking less care over this task than others, despite encouragement. However, agreement figures for the E.coli portion of the final WP11 corpus show a slight improvement. Additionally, as our biologists had differing levels of experience, their ability to accurately assign more specific types of NE labels may have been variable. However, higher levels of agreement are achieved if the hierarchical structure of the NE categories is taken into account. If all NE categories are mapped to their top level supercategories, then agreement rates of up to 90% are achieved.

Final Corpus Statistics

Following the training period, the final corpus was collected within a period of 3 weeks. As mentioned above, this corpus consists of a total of 244 abstracts, of which 167 are on the subject of E. coli, and the remaining 77 relate to *H.sapiens*. General statistics regarding the final corpus are shown in **Table 24**.

The statistics also reinforce the importance of considering events that are described by nominalised verbs as well as verbs. In the *E.coli* corpus, events that are centred on nominalised verbs are almost as common as those centred on verbs, although the range of different words that are used to describe events is much greater for verbs than for nouns. The human corpus shows slightly different characteristics in this respect, although the proportion of events described by nominalised verbs is still significant.

	Complete Corpus	E.coli abstracts	Human abstracts
No of abstracts	244	167	77
No of events	3091	2436	680
Av. Events per abstract	12.66	14.59	8.83
Distinct nom. verbs annotated	90	84	36
Events centred on nom. verbs	1293 (42%)	1091 (45%)	204 (30%)
Distinct verbs annotated	181	154	108
Events centred on verbs	1799 (58%)	1345 (55%)	476 (70%)

Table 24. General corpus statistics

AGREEMENT RATE	**F-Score**		
	E. coli	**Human**	**WP4**
Event identification	72.27%	76.37%	57.31%
Argument identification (relaxed span match)	90.23%	91.27%	90.12%
Argument identification (exact span match)	75.10%	77.48%	75.58%
Semantic role assignment	88.96%	88.30%	81.17%
NE identification	82.55%	82.03%	73.20%
NE supercategory assignment	95.52%	94.75%	N/A
NE category assignment (exact)	71.02%	66.03%	67.94%
NE category assignment (considering parent)	75.38%	68.97%	N/A

Table 25. General agreement statistics

Inter-annotator Agreement

To facilitate calculation of IAA scores for the WP11 corpus, a portion of the corpus (57 abstracts, approximately a quarter of the complete size) was annotated by all six annotators. Pairwise comparisons were calculated between each different pair of annotators, and averages are shown in **Table 25** for each

portion of the corpus (E. coli and human), together with a comparison of the IAA scores achieved in the final corpus of the WP4 annotation. Note that the figures for the WP4 annotation have been converted into F-scores from the direct agreement rates originally reported in D.4.1

Table 25 shows that in general, standards achieved during the training period have been maintained or exceeded in the final corpus collection. The only exceptions to this are event identification and NE identification. During the training period, UoM noted that these tasks appear to require some period of adaptation when annotating in a different domain. As annotation in the final corpus collection phase changed back to *E.coli* (after the annotation of human abstracts in the final training cycle), this could have caused problems for some annotators.

The table shows that for most subtasks of the annotation process, agreement levels are above 70%, which we believe is very respectable. Standards of annotation have achieved in the WP4 annotation have been at least maintained, or, for a number a subtasks, exceeded, in the WP11 annotation.
Particularly high levels of agreement (88% or above) are achieved for both the identification of semantic arguments and the assignment of semantic roles to these arguments. As these are the tasks that UoM originally identified as being more linguistically-oriented, our results suggest that a detailed set of guidelines, together with an intensive training program allow these tasks to be carried out by biologists to a fairly high degree of accuracy.

In terms of the assignment of categories to biological concepts, a very high level of reliability can be attained (approximately 95% agreement) if only the 5 most coarse grained categories, i.e. *Nucleic_Acids, Proteins, Living_Systems, Processes* and *Experimental,* are considered. Exact agreement of sub-concept labels within these categories is somewhat lower, although this increases slightly if matching is extended to include the parent category. As mentioned in the section on training, higher agreement rates for such a fine grained NE scheme may be difficult to achieve, due to the differing amounts of experience of the annotators within the field of gene regulation.

3.4.6 Experimental Results

The aim of this section is to evaluate semantic frame extraction performance, given a set of annotated training data.

NE Type	Recall	Precision	F
Nucleic_Acid	0.581	0.708	0.638
Protein	0.534	0.646	0.585
Experimental	0.191	0.479	0.273
Process	0.542	0.682	0.604
Living_System	0.432	0.721	0.540
Total	0.535	0.679	0.599

Table 26. NER performance (overall)

NE Type	Recall	Precision	F
Nucleic_Acid	0.601	0.717	0.654
Protein	0.585	0.690	0.633
Experimental	0.222	0.471	0.302
Process	0.546	0.689	0.609
Living_System	0.466	0.682	0.554
Total	0.561	0.695	0.621

Table 27. NER performance (E. coli)

NE Type	Recall	Precision	F
Nucleic_Acid	0.333	0.669	0.445
Protein	0.302	0.538	0.387
Experimental	0.000	0.000	0.000
Process	0.379	0.715	0.496
Living_System	0.333	0.727	0.457
Total	0.325	0.634	0.430

Table 28. NER performance (Human)

The annotated corpus was randomly separated into 10 document groups. UoM conducted 10-fold cross validation based on the 10 document groups. The named entity recognizer, the semantic role labeller, and the event extraction rules were constructed using 9 groups of annotated documents. Event instances

were then extracted from the remaining group of documents using the event extraction rules. Evaluation metrics are precision, recall, and the F-score.

a) Gene Regulation NER Results

Tables 26-28 show the performance of Named Entity recognition. **Table 26** shows overall performance of the detection of five NE categories relevant to gene regulation. The overall NER performance has an F-score of 59.9.

As only the NEs relevant to GREs are annotated, an estimated IAA for NEs is around 60%.

In the cross-validation, 7 folds are from E. coli gene regulation abstracts and 3 folds are from Human abstracts. **Tables 27** and **28** show separate NER performance statistics for the two species. **Table 28** shows that NER in Human abstracts is more difficult than in E. coli abstracts.

Semantic role	Recall	Precision	F
Agent	0.4454	0.6311	0.5222
Action	0.7101	0.8466	0.7724
Theme	0.4849	0.6436	0.5531
Manner	0.2951	0.5600	0.3865
Instrument	0.000	0.000	0.000
Location	0.3077	0.5786	0.4017
Source	0.3404	0.8000	0.4776
Destination	0.2952	0.6526	0.4066
Temporal	0.0755	0.3636	0.1250
Condition	0.1637	0.4444	0.2393
Rate	0.1765	0.6923	0.2812
Descriptive-Agent	0.1667	0.4800	0.2424
Descriptive-Theme	0.0667	0.4762	0.1170
Purpose	0.1622	0.5455	0.2500
Total	**0.5254**	**0.7189**	**0.6071**

Table 29. Semantic role labelling performance

b) Gene regulation SRL results

Table 29 shows the semantic role labelling performance. The overall performance has an F-score of 60.71%. As IAA of the event argument identification and the semantic role assignments are 75% and 89%, respectively, the overall semantic role labelling quality is sufficiently high.

UoM also evaluated the performance of SRL in the E. coli and Human folds separately. E. coli SRL achieves an F-score of 62.66%, compared to 53.62% for Human SRL.

method	Species	N-best		Recall	Precision	F
		NER	SRL			
Baseline rules	Overall	1	1	0.0847	0.3400	0.1357
		3	3	0.1155	0.3003	0.1669
		5	5	0.1320	0.2754	0.1785
	E. coli	1	1	0.0945	0.3613	0.1501
		3	3	0.1256	0.3217	0.1806
		5	5	0.1435	0.2974	0.1936
	Human	1	1	0.0506	0.2345	0.0832
		3	3	0.0803	0.2204	0.1178
		5	5	0.0923	0.1962	0.1255
Induced rules	Overall	1	1	0.1354	0.3262	0.1913
		3	3	0.1672	0.2756	**0.2081**
		5	5	0.1804	0.2362	0.2045
	E. coli	1	1	0.1490	0.3398	0.2072
		3	3	0.1818	0.2915	**0.2239**
		5	5	0.1946	0.2514	0.2193
	Human	1	1	0.0878	0.2634	0.1317
		3	3	0.1161	0.2120	0.1500
		5	5	0.1310	0.1800	**0.1516**

Table 30. Event extraction performance (exact event match)

c) Gene Regulation Event Extraction Results

Table 30 shows the performance of *complete* event extraction using 10-fold cross validation. To obtain better recall, UoM used n-best results from CRF labellers for both NER and SRL. The induced IE rules outperformed the baseline method whereas the overall performance has an F-score of 20.81%. This result should not, however, be taken pessimistically. Bio-event extraction is a challenging task, illustrated by the fact that the IAA of complete event annotation is 38-40%.

method	Species	N-best		Recall	Precision	F
		NER	SRL			
Baseline rules	Overall	1	1	0.1147	0.5248	0.1882
		3	3	0.1565	0.4647	0.2341
		5	5	0.1788	0.4276	0.2521
	E. coli	1	1	0.1270	0.5556	0.2067
		3	3	0.1700	0.4966	0.2532
		5	5	0.1930	0.4583	0.2716
	Human	1	1	0.0734	0.3971	0.1239
		3	3	0.1114	0.3498	0.1690
		5	5	0.1311	0.3215	0.1862
Induced rules	Overall	1	1	0.1724	0.4720	0.2526
		3	3	0.2164	0.4062	0.2823
		5	5	0.2365	0.3550	**0.2839**
	E. coli	1	1	0.1902	0.4890	0.2739
		3	3	0.2359	0.4284	**0.3043**
		5	5	0.2555	0.3759	0.3042
	Human	1	1	0.1127	0.3945	0.1753
		3	3	0.1507	0.3194	0.2048
		5	5	0.1730	0.2785	**0.2134**

Table 31. Event extraction performance (essential slot match)

If UoM focuses only on the extraction performance of the *core* event slots of *Agent*, *Action*, and *Theme*, the F-score is 21-30% (**Table 31**).

3.4.7 Related Work

For general English texts, there have been several pioneering studies regarding the generation of information extraction patterns, including AutoSlog-TS (Riloff,1996), CRYSTAL (Soderland *et al.,* 1995), PALKA (Kim *et al.,* 1995), LIEP (Huffman, 1996) and RAPIER (Califf and Mooney,1997). CRYSTAL (Soderland *et al.*, 1995) induced a restricted class of regular expressions that extract information from text. Kushmerick *et al.* (1997) proposed the generation of patterns that extract information from HTML documents.

Recently, biological information extraction has been investigated using AImed (Bunescu and Mooney, 2005) for protein-protein interaction, BioInfer (Pyysalo *et al*., 2007), and the Genia event corpus (Kim *et al*, 2008). The bio-event extraction task handled in this report targets more complex than conventional bio-IE tasks. Our task requires the filling of templates with fourteen kinds of slots.

In GRE extraction, multiple templates should sometimes be extracted from a single sentence. For example, “Thy-1 expression regulates expression of TGF-β

modulatory proteins." contains three events: two *expression* events and one *regulation* event. Thus, in order to ensure that slot fillers are associated with the correct event templates, UoM needed to generate information extraction rules that fill multiple slots of a given template.

Our method utilises CRF-based NER and semantic role labelling tuned to the gene regulation domain. This means that only NEs and SRLs that are relevant to gene regulation are detected. Thanks to these domain-adapted recognition tools, it is very likely that when SRLs and/or NEs are assigned to a sentence, the labelled parts describe GREs. In this respect, not only event extraction rules but also automatically assigned SRLs and NEs function as a guide to the detection of GREs.

3.4.8 Summary

This section has presented the UoM approach to automatic event instance extraction for GREs in the biology domain. Event extraction rules are inductively generated from a sequence of words that are constrained by syntactic and semantic information. Gene regulation NER and SRL models are trained on a corpus annotated with GREs.

The gene regulation NER and SRL performances were close to the inter-annotator agreement rates amongst human annotators. A combination of statistical NER and SRL with symbolic event frame extraction outperformed the baseline method and achieved an F-score of 15-22% for all template slots and 21-30% for essential Agent-Action-Theme slots. These results are promising due to the inherent complexity deep-semantic event annotation, as illustrated by the IAA rate of 38-40% for exact event identification by domain experts.

4 Conclusion

This report describes various activities with WP11 by I2R, UoM, EBI, FSU Jena and UKLFR teams, which target the evaluation of the various language resources including Biolexicon, Bioontology and NLP tools. The evaluations are in the light of various use cases for biologists to access the information, including information retrieval and extraction tasks. These use cases have been defined and annotated by the biologists in the domain.

The evaluations show the positive indications with the usefulness of the various resources in the project, which trigger the further investigation, enhancement and extend the use of these resources beyond the project.

Bibliography

R.K. Ando, et al.. 2005. TREC 2005 Genomics Track Experiments at IBM Watson, in TREC 2005 proceeding

Madan Babu *et al.*. 2003. Evolution of transcription factors and the gene regulatory network in Escherichia coli. Nucleic Acids Research, 2003, Vol. 31, No. 4 1234-1244

Beisswanger, E., Lee,V., Kim,J.J., Rebholz-Schuhmann,D., Splendiani,A., Dameron,O., Schulz,S., Hahn,U. Gene Regulation Ontology (GRO): Design Principles and Use Cases. Stud Health Technol Inform. 2008;136:9-14. PMID: 18487700

Brants T: Inter-annotator agreement for a German newspaper corpus. In: *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000): 2000*; 2000.

Razvan C. Bunescu and Raymond J. Mooney. 2005. Subsequence Kernels for Relation Extraction. *In Proc. of Advances in Neural Information Processing Systems (NIPS-2005).*

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

Mary E. Califf, and Raymond J. Mooney. 1997. Relational Learning of Pattern-Match Rules for Information Extraction, *In Proceedings of the ACL-97 Workshop in Natural Language Learning*, pp 9–15.

Bojan Cestnik, Estimating probabilities: A crucial task in machine learning, *In Proc. of the Ninth European Conference on Artificial Intelligence (ECAI-90),* pp. 147-149, Stockholm, 1990.

Dina Demner-Fushman, Susanne M. Humphrey, C. Ide, Russel F. Loane, James G. Mork, Patrick Ruch, Miguel E. ruiz, Lawrence H. Smith, W. John Wilbur, Alan R. Aronson, Combining resources to find answers to biomedical questions, *Proc. of REC-16*, NIST Special Publication, 2007.

Claire Fautsch, Jacques Savoy IR-Specific Searches at TREC 2007: Genomics and Blog Experiments, *Proc. of TREC-16*, NIST Special Publication, 2007.

Fellbaum, C., editor, *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA., 1998.

Francopoulo, G., M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria, Lexical Markup Framework (LMF), *Proc. of LREC 2006*, Genova, Italy.

S. Fujita. 2004. Revisiting Again Document Length Hypotheses TREC 2004 Genomics Track Experiments at Patolis, in TREC 2004 proceeding

Udo Hahn *et al*.. 2001. Subword segmentation: Leveling out morphological variations for medical document retrieval. In Suzanne Bakken (Ed.), AMIA 2001 - Proceedings of the Annual Symposium of the American Medical Informatics Association. A Medical Informatics Odyssey: Visions of the Future and Lessons from the Past, pp. 229-233. Washington, D.C., November 3-7, 2001. Philadelphia, PA: Hanley & Belfus.

Udo Hahn *et al*.. 2005. Subword clusters as light-weight interlingua for multilingual document retrieval. In MT Summit X - Proceedings of the 10th Machine Translation Summit of the International Association for Machine Translation. Phuket, Thailand, September 12-16, 2005.

Willam Hersh *et al*.. 1994. OHSUMED: An interactive retrieval evaluation and new large test collection for research. Proceedings of the 17th Annual ACM SIGI Conference, 192-201

William Hersh, Aaron Cohen, Lynn Ruslen, Phoebe Roberts, TREC 2007 Genomics Track Overview, *Proc. of TREC 2007*, 2007.

William Hersh and Ellen Voorhees. 2008. TREC special issue overview.

George Hripcsak G and Adam S. Rothschild. 2005. Agreement, the F-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association* 2005, 12(3), 296-298.

Lars Juhl Jensen, Jasmin Saric, and Peer Bork. 2006. Literature mining for the biologist:from information retrieval to biological discovery, Nat.Rev. Genet. 7, pp. 119-129.

Antonio Jimeno and Piotr Pezik, Information Retrieval and Information Extraction in TREC Genomics 2007, *Proc. of TREC-16*, NIST Special Publication, 2007.

Sophia Katrenko and P. Adriaans. 2006. Learning relations from biomedical corpora using dependency trees. In KDECB 2006 -- Knowledge Discovery and Emergent Complexity in Bioinformatics, pages 61-80.

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

Jun-Tae Kim and Dan I. Moldovan. 1995. Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction. *IEEE Transaction on Knowledge and Data Engineering (IEEE TKDE)*, 7(5), pp.713–724.

Seon-Ho Kim, Juntae Yoon, and Jihoon Yang. 2008. Kernel approaches for genic interaction extraction. Bioinformatics, pages 118-126.

Nicholas Kushmerick, Daniel S. Weld and Robert Doorenbos. 1997. Wrapper induction for information extraction. In: *Proc. Of International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, Japan*, pp. 729–735.

John Lafferty, Andrew McCallum and Fernando Pereira. 2001 Conditional Random Fields: Probabilistic Models for Segmenting and Labelling

Sequence Data. *In Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001)*, pp 282–289.

Kornél Markó *et al*.. 2005a. Automatic lexicon acquisition for a medical cross-language information retrieval system. In MIE 2005 - Proceedings of the XIX International Congress of the European Federation for Medical Informatics, pp. 829-834. Geneva, Switzerland, August 28-31, 2005.

Kornél Markó *et al*.. 2005b. Multilingual lexical acquisition by bootstrapping cognate seed lexicons. In RANLP 2005 - Proceedings of the International Conference on 'Recent Advances in Natural Language Processing', pp. 301-307. Borovets, Bulgaria, 21-23 September, 2005.L. Huang et al.. 2005. UM-D at TREC 2005: Genomics Track, in TREC 2005 proceeding.

Andrew McCallum and Wei Li. 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons, *In Proc of the Seventh Conference on Natural Language Learning (CoNLL-03).*

McCray, A.T., Srinivasan, S. and Browne, A.C., Lexical methods for managing variation in biomedical terminologies, *SCAMC'94*, pp. 235-239, 1994.

Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter, and Tapio Salakoski. 2009. Comparative analysis of five protein-protein interaction corpora. BMC Bioinformatics.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen and Tapio Salakoski. 2007. BioInfer: a corpus for information extraction in the biomedical domain, BMC *Bioinformatics* 8:50.

Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau, Okapi at TREC, *Proc. of Text REtrieval Conference*, pp. 21-30, 1992.

Carlos Rodriguez-Penagos, Heladia Salgado, Irma Martinez-Flores, and Julio Collado-Vides. 2007. Automatic reconstruction of a bacterial regulatory network using Natural Language Processing. BMC Bioinformatics.

M. Sanderson and H. Joho. 2004. Forming Test Collections with No System Pooling; in the proceedings of the 27th ACM SIGIR conference.

Erik F. Tjong Kim Sang. 2000. Noun Phrase Recognition by System Combination, *In Proc. of the First Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2000)*, pp. 50-55, Seattle.

S. Schulz *et al*.. 2006. Semantic atomicity and multilinguality in the medical domain: Design considerations for the MORPHOSAURUS subword lexicon. In LREC 2006 - Proceedings of the 5th International Conference on Language Resources and Evaluation. Genua, Italy, May 24-26.

B. Settles. 2004. Biomedical Named Entity Recognition using Conditional Random Fields and Novel Feature Sets. *In Proc. of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, pages 104–107, Geneva, Switzerland.

Smith, L., T. Rindflesch, and W. J. Wilbur, MedPost: a Part-of-Speech Tagger for BioMedical Text, *Bioinformatics*, 20:2320-2321, 2004.

Steven Soderland, David Fisher, Jonathan Aseltine and Wendy Lenert. 1995. CRYSTAL: Inducing a Conceptual Dictionary, *In Proc. of the 13th International Joint Conference on Artificial Intelligence (IJCAI-95).* pp.1314–1319.

Nicola Stokes, Yi Li, Lawrence Cavedon, Eric Huang, Jiawen Rong and Justin Zobel, Entity-based Relevance Feedback for Genomic List Answer Retrieval, *Proc. of TREC-16, NIST Special Publication*, 2007.

Paul Thompson, Philip Cotter, John McNaught, Sophia Ananiadou, Simonetta Montemagni, Andrea Trabucco, and Giulia Venturi. 2008. Building a Bio-Event Annotated Corpus for the Acquisition of Semantic Frames from Biomedical Corpora. *In Proc. of the Sixth International Conference on Language Resources and Evaluation (LREC 2008).*

Richard T.H Tsai, Wen-Chi Chou, Ying-San Su, Yu-Chun Lin, Chen-Lung Sung, Hong-Jie Dai, Irene T.H Yeh, Wei Ku, Ting-Yi Sung and Wen-Lian Hsu. 2007. BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features, *BMC Bioinformatics* 8:325.

len M. Voorhees, The TREC-8 Question Answering Track Report, *Proc. of Eighth Text REtrieval Conference (TREC-8)*, pp. 77-82, 1999.

Joachim Wermter, Katrin Tomanek, and Udo Hahn. 2009. High-performance gene name normalization with GeNO. Bioinformatics.

Wai Keong Yong and Jian Su. 2008. An Effective Method of Using Web-based Information for Relation Extraction. In Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008), Hyderabad, India.

GuoDong Zhou et al.. 2004. Recognition names in biomedical texts: a machine leaning approach; in BIOINFORMATICS vol. 20 no. 7 pp 1178-1190.

GuoDong Zhou, Jian Su, Jie Zhang, Min Zhang. 2005. Exploring Various Knowledge in Relation Extraction. In Proceedings of the 43rd Annual meeting of the Association for Computational Linguistics (ACL05), Ann Arbor, Michigan, US. Pages 427-434.

Annex 1: 60 Queries, their GR categories and event types in GR IR evaluation set

Category:
1. Transcription factors (TF) and formation of TF complex
2. DNA binding of TFs (at TF recognition sites)
3. Gene expression (RNA, protein)
4. Regulation of gene expression (up-, down-regulation)

Event type:
(CU) Carbon utilization
(RS) Redox sensing
(ES) Environment sensing e.g. temperature, water.
(IT) Ion transport
(CS) Cell structure
(GE) General enhancer
(CM) Cellular metabolic process (carbon, nitrogen, phosphate, sulfur, nucleotide, cofactor)
(AR) Antibiotic resistance
(RR) Restriction and repair

Transcription regulators (20 queries)

TR1. What [TRANSCRIPTION REGULATORS] control genes involved in sugar utilization?
Category: 1, 4 Event type: CU

TR2. What [TRANSCRIPTION REGULATORS] control the phosphotransferase system (PTS)?
Category: 1, 4 Event type: CU, CM

TR3. What [TRANSCRIPTION REGULATORS] control genes involved in amino acid catabolism during carbon starvation?
Category: 1, 4 Event type: CU, CM

TR4. What [TRANSCRIPTION REGULATORS] control genes involved in aerobic respiratory control?
Category: 1, 4 Event type: RS

TR5. What [TRANSCRIPTION REGULATORS] control genes involved in nitrate and nitrite regulation and anaerobic respiration?
Category: 1, 4 Event type: RS

TR6. What [TRANSCRIPTION REGULATORS] control genes involved in nitrogen metabolism?

Category: 1, 4 Event type: RS, CM

TR7. What [TRANSCRIPTION REGULATORS] control genes involved in iron transport?
Category: 1, 4 Event type: IT

TR8. What [TRANSCRIPTION REGULATORS] regulate the fimbrial operons?
Category: 1, 4 Event type: CS

TR9. What [TRANSCRIPTION REGULATORS] control the initiation of DNA synthesis?
Category: 1, 2, 4 Event type: CM, RR

TR10. What [TRANSCRIPTION REGULATORS] regulate the heat shock stress response?
Category: 1, 3, 4 Event type: ES

TR11. What [TRANSCRIPTION REGULATORS] are involved in the transcription controlled by RNA polymerase sigma S factor (RpoS) upon entry into stationary phase?
Category: 1, 3, 4 Event type: ES

TR12. What [TRANSCRIPTION REGULATORS] are involved in the transcription controlled by integration host factor (IHF)?
Category: 1, 4 Event type: ES, GE

TR13. What [TRANSCRIPTION REGULATORS] are involved in the transcription controlled by cold shock protein A (CspA)?
Category: 1, 3, 4 Event type: ES

TR14. What [TRANSCRIPTION REGULATORS] are involved in the transcription controlled by the response regulator for osmoregulation, OmpR?
Category: 1, 4 Event type: ES

TR15. What [TRANSCRIPTION REGULATORS] are involved in the transcription controlled by the LysR-type regulator protein, ArgP?
Category: 1, 2, 4 Event type: CM, RR

TR16. What [TRANSCRIPTION REGULATORS] are involved in the PhoB-dependent transcriptional activation during starvation for phosphate?
Category: 1, 4 Event type: ES, CM

TR17. What [TRANSCRIPTION REGULATORS] are involved in the superoxide sensor SoxR-mediated transcriptional regulation in response to oxidative stress?
Category: 1, 4 Event type: RS

TR18. What [TRANSCRIPTION REGULATORS] are involved in the LexA-regulated SOS repair system?
Category: 1, 2, 4 Event type: RR

TR19. What [TRANSCRIPTION REGULATORS] sense DNA supercoiling and thus indirectly sense many environmental conditions (growth phase, energy level, osmolarity, temperature, pH, and so on) that affect this DNA property?
Category: 1, 2, 4 Event type: ES, RR

TR20. What [TRANSCRIPTION REGULATORS] are involved in complex transcription activation of the mal (encoding genes for maltose catabolism) and mel (encoding genes for melibiose catabolism) operons, including the operon specific activators and their co-dependent global regulator?
Category: 1, 4 Event type: CU, CM

Genes (20 queries)

G1. What [GENES] are regulated by the CreBC two-component system that responds to growth in minimal media?
Category: 1, 2, 4 Event type: ES, CM

G2. What [GENES] are regulated by the redox-sensitive transcription regulator, OxyR?
Category: 1, 2, 4 Event type: RS

G3. What [GENES] change expression (i.e. increase or decrease of the gene expression) in association with the oxygen level, glucose treatment and appearance of transcription regulator ArcA?
Category: 1, 2, 4 Event type: RS, CM

G4. What [GENES] are the targets of RutR, the master regulator of pyrimidine catabolism?
Category: 1, 2, 4 Event type: CM

G5. What [GENES] are involved in the regulation of novobiocin resistance?
Category: 4 Event type: AR

G6. What [GENES] are regulated by the transcription factor pair of FlhDC-FliA, which forms part of the genetic network controlling the temporal program of flagellar assembly, with FlhDC being its principal regulator, and FliA the flagellum-specific sigma factor?
Category: 1, 2, 4 Event type: CS

G7. What [GENES] are regulated by the transcription regulator pairs, FNR-NarL and FNR-ArcA, which regulate anaerobic respiration and fermentation, with ArcA and NarL determine the type of respiration mode under the coordination of FNR?
Category: 1, 2, 4 Event type: RS

G8. What [GENES] are regulated by the two-component regulatory system CpxA/CpxR, which senses the stresses of misfolded proteins and degrading factors?
Category: 1, 2, 4 Event type: CS

G9. What [GENES] are regulated by the transcription factor pair of MarA and SoxS?
Category: 1, 2, 4 Event type: AR

G10. What [GENE] expression is affected by the antagonistic regulatory interaction between FIS and H-NS proteins?
Category: 1, 2, 4 Event type: CS, GE

G11. What [GENES] are regulated by RpoD (sigma70), the housekeeping sigma factor involved in the cellular machinery of growth phase?
Category: 3, 4 Event type: ES

G12. What [GENES] are regulated by the DcuS-DcuR two-component sensor-regulator in response to external C4 dicarboxylates and citrate?
Category: 1, 2, 4 Event type: RS, CM

G13. What [GENES] are regulated by FadR, which is involved in fatty acid metabolism, including negative regulation of fatty acid degradation and positive regulation of the biosynthesis of unsaturated fatty acids in a concerted manner with negative regulation of FabR?
Category: 1, 2, 4 Event type: CM

G14. What [GENES] are involved in nucleosides uptake and usage and are regulated by two complex control systems governed by CytR and DeoR?
Category: 1, 2, 4 Event type: CM

G15. What [GENES] are regulated by NhaR, which is involved in cation transport and intracellular pH regulation?
Category: 1, 2, 4 Event type: IT

G16. What [GENES] are involved in the regulation of curli synthesis, which plays a role on adhesion to surfaces, cell aggregation, and biofilm formation?
Category: 3 Event type: CS

G17. What [GENES] are involved in biotin synthesis, which is regulated by the rate of protein biotination?

Category: 3, 4 Event type: CM

G18. What [GENES] encode nitric oxide (NO)-detoxifying enzymes (i.e. NO defense genes) that are induced and coordinately controlled in response to NO stress?
Category: 3, 4 Event type: RS, CM

G19. What [GENES] are regulated by UxuR/ExuR and UidR in utilization of hexuronide?
Category: 1, 2, 4 Event type: CU, CM

G20. What [GENES] are autoregulated by its own gene products?
Category: 3, 4 Event type: CM

Proteins (5 queries)

P1. What [PROTEINS] interact with cAMP receptor protein (CRP) in CRP-mediated transcriptional regulation?
Category: 1, 4 Event type: ES

P2. What [PROTEINS] increase the mRNA stability and the level of gene expression?
Category: 3, 4 Event type: ES, CM

P3. What [PROTEINS] interact with the hemolysin expression modulating protein, HHA?
Category: 3, 4 Event type: CS, CM

P4. What [PROTEINS] connect the two-component systems, EvgS/EvgA and PhoQ/PhoP, and in turn promote the expression of PhoP-activated genes?
Category: 1, 4 Event type: ES

P5. What [PROTEINS] affect assembly of transcription elongation complexes?
Category: 3 Event type: CM

RNA (10 queries)

R1. What [RNA] transcription is activated by factor-for-inversion stimulation (FIS) protein?
Category: 3, 4 Event type: GE

R2. What [RNA] are recognized and bound by the carbon storage regulator, CsrA, which inhibits the translation?
Category: 3, 4 Event type: CM

R3. What [RNA] region of cspE is associated with its transcript stability and inducibility at both the transcript and the protein level upon cold shock?
Category: 3 Event type: ES

R4. What [RNA] binds sigma70 RNA polymerase and downregulates transcription at many sigma70-dependent promoters during stationary phase when the majority of the transcription machinery is bound by the RNA?
Category: 3, 4 Event type: ES

R5. What [RNA] encoded by the gcvB gene regulates the genes involved in transport of amino acids and peptides (including sstT, oppA and dppA)?
Category: 3, 4 Event type: CM

R6. What [RNA] are regulated by DksA and its co-factor ppGpp?
Category: 3, 4 Event type: ES

R7. What [RNA] acts as regulatory signals in sensing and responding to tryptophan, and stalls the translation of tandem Trp codons as well as prevents transcription termination?
Category: 3, 4 Event type: ES, CM

R8. What [RNA] form the hairpin structure in response to a redefined stop codon, which enables UGA-directed selenocysteine incorporation and tethers the specialized translation elongation factor?
Category: 3 Event type: CM

R9. What [RNA] motif modulates ribonuclease action and affect RNA decay by degradosomes?
Category: 3 Event type: CM

R10. What small [RNA] use the RNA chaperone Hfq and act as regulators of translation and message stability by pairing to target messenger RNAs?
Category: 3 Event type: ES, CM

Cell components (5 queries)

CC1. What [CELL COMPONENTS] harbor the ferrous iron transport system (Feo), which contribute to the iron supply of the cell under anaerobic conditions?
Category: 4 Event type: IT, CS

CC2. What [CELL COMPONENTS] are protected or repaired by the proteins produced in the RpoE (sigma24)-mediated response to stress signals?
Category: 3, 4 Event type: ES, CS

CC3. What [CELL COMPONENTS] are appendages anchored to the outer membrane to facilitate bacterial movement and their synthesis/ or assembly depends on the sequential and temporal order of structural gene expression control?
Category: 3 Event type: CS

CC4. What [CELL COMPONENTS] form an extracellular polysaccharide layer, governed by a complex network of regulators, e.g. the Rcs-system (RcsA and RcsB) that responds to environmental stimuli?
Category: 3, 4 Event type: ES, CS

CC5. What [CELL COMPONENTS] undergo morphological and physiological changes during the growth transition from the exponential growth to the stationary phase, while the pattern of gene expression changes in such a way that the growth-related genes are mostly switched off and the stationary-phase-specific genes are expressed?
Category: 3 Event type: ES, CS

Annex 2: Updated Event Annotation Guidelines

Introduction

We are in the process of building a machine-readable dictionary of biological terms and verbs which can help with automatically finding important facts that are contained within biological texts. This document describes a task called *annotation* which will help us in the construction of suitable dictionary entries for verbs. It begins with an explanation of the types of information that we wish to include within these dictionary entries, followed by a description of the task that will be undertaken to collect this information. Finally, a set of guidelines that explain exactly how the task should be carried out are presented.

Events and variables

Verbs typically represent different kinds of events. Details of these events, i.e. the variables that are involved in them, are introduced by a set of phrases that accompany the verb in the sentence. The simple sentence shown in (1) helps to illustrate this.

(1) *The narL gene product activates the nitrate reductase operon*

In (1), there is a verb, *activates,* that is surrounded by 2 phrases i.e. *the narL gene product* and *the nitrate reductase operon.* These phrases can be seen to *belong* to the verb, in that they are used to describe the variables involved in the *activation* event. Each phrase represents a different variable that is involved in the event: the phrase *the narL gene product* represents the thing that *causes* or *drives* the event, whilst *the nitrate reductase operon* is the thing *affected by* the event.

In (1), the phrases that denote the variables of the event correspond to the subject and object of the verb, but it is also possible for verbs to have more than 2 variables associated with them, as shown in (2).

(2) *The LysR-type transcriptional regulator CysB controls the repression of hslJ transcription in Escherichia coli*

In (2), the event is represented by the verb *controls.* As with (1), the subject of the verb, i.e. *the LysR-type transcriptional regulator CysB* is what *instigates* that event. Likewise, the object of the verb, i.e. *the repression of hslJ transcription*, is what is *affected by* the event. In (2), however, there is a further phrase associated with this *controls* event, i.e. *in Escherichia coli*. This indicates *where* the described *control* action takes place.

The above examples illustrate that, when considered at a general level, the same types of variables occur with different types of events. In sentences (1) and (2), for example, we have seen that the subjects of both verbs describe what *causes* or *instigates* the event, whilst the objects describe what was *affected by* or *acted upon* during the event. The general type of information that a particular phrase provides about an event is called its *semantic role.*

Each semantic role has a name associated with it. For example, phrases that describe what instigates or drives are assigned the AGENT semantic role, whilst things that are affected by events are assigned the THEME role. AGENT and THEME are considered as *core* roles, in that they provide what is normally the most important information about the event, and at least one of them is present in the descriptions of the majority of events. Phrases corresponding to the AGENT and THEME normally occur in closest proximity to the verb that represents the event.

Other phrases can provide a number of other types of information about event, including where the event took place, i.e. the LOCATION role. These phrases normally occur further away from the verb, but are also relevant to the description of the event. Altogether, we have defined 12 semantic roles that seem to characterise the majority of variables involved in biological events. A full list of these roles, together with descriptions and examples, is provided in the section 8 of this document.

Different verbs typically occur with different patterns of variable-denoting phrases. That is to say, the number of phrases that contribute to the description of the event, and the semantic roles that these phrases correspond to, varies from verb to verb. This is because different verbs represent different events, and the description of each event requires a particular set of variables. In order to interpret these events automatically, the dictionary entry for each verb should indicate the patterns of variable-denoting phases that most typically accompany it in biological texts.

Nominalised verbs

Whilst events are most typically represented by verbs, it is also possible for them to be represented by nouns. Such nouns are called *nominalised verbs*. They are so called because they convey the same meaning as a related verb, but behave as a noun, in that, for example, they can be preceded by adjectives and/or determiners such as *a* or *the*. Nominalised verbs often have a similar, but different form to their related verbs. **Examples include *transduction* (related verb: *transduce*), *expression* (related verb: *express*), *analysis* (related verb: *analyse*)** Other nominalised verbs have the same form as the verbs from which they are derived, e.g. ***control* and *increase.***

Nominalised verbs are interesting in that, like verbs, they can be accompanied by phrases that correspond to variables involved in the event that they represent. **As nominalised verbs are very common in the biological literature, we want to create similar dictionary entries as for verbs, i.e. to describe the different patterns of variable-denoting phrases that can occur with them.**

In example (2) above, the THEME of the *control* event is actually the description of a further event, i.e. *the repression of hslJ transcription.* The event is represented using a nominalised verb, *repression* (related to the verb *repress*). The THEME of the *repression* event, i.e. what is *being repressed*, follows the preposition *of*, i.e. *hslJ transcription.*

The THEME of the *repression* event, i.e. *hslJ transcription* represents yet another event using the nominalised verb *transcription* (from the verb *transcribe*). In this case, the thing that directly precedes the nominalised verb, i.e. *hslJ,* is a variable in the event. This is the thing that is *undergoing* the transcription, and hence is the THEME of the event.

Concepts

The majority of phrases that denote the variables of events fall into one of two categories:

a) *Things*, such as genes, operons, proteins or species. We refer to these as *entities*
b) *Events,* often expressed using a nominalised verb such as *repression*, *transcription* or *increase.*

Entities and events can be classified according to the *concept* that they represent, whether this is a gene, species, biological process etc. Part of the annotation task concerns assigning *concept types* to variables, in addition to the more general semantic roles, which were mentioned earlier.

Let us consider again the sentences from section 2.

(1) *The narL gene product activates the nitrate reductase operon*
(2) *The LysR-type transcriptional regulator CysB controls the repression of hslJ transcription in Escherichia coli*

For example, an AGENT was identified for the each of events that were identified in sentences (1) and (2) in section 2 above. However, the *concept type* of the AGENT was different in each case. In the *activates* event of sentence (1), the AGENT was *The narL gene product,* which is a *protein.* On the other hand, the AGENT of the *controls* event in sentence (2), i.e. *The LysR-type transcriptional regulator CysB* represents a different type of concept, i.e. a *regulator.* It may be that in different occurrences of events described by the same verbs, the AGENT can correspond to varying concept types.

In order to assist with the automatic extraction of important facts from biological texts, we would like our dictionary entries to specify the *type(s)* of concepts than can occur as the values of variables. The type(s) specified for each variable will come from a hierarchy of concepts that we have defined for the biological field.

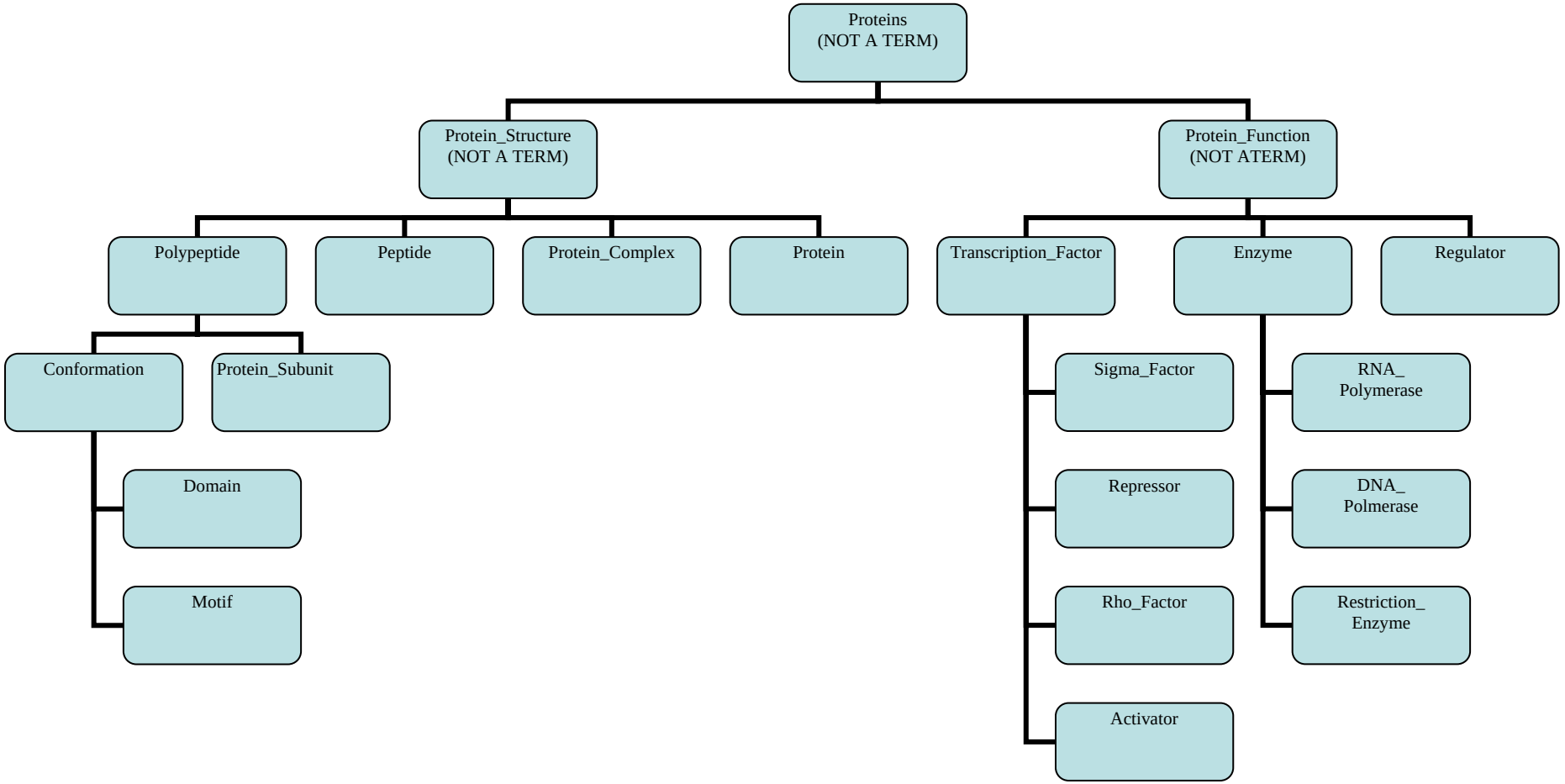
The section headings **PROTEINS**, **NUCLEIC_ACID**, **LIVING_SYSTEM**, **PROCESSES** and **EXPERIMENTAL** are the concept groupings that are used to sub-

divide the term-list into a more manageable set of options. Most of these concepts, or classes, are intended to be specific, but unfortunately they are not mutually exclusive. Indeed many can be considered subsets of other classes listed in the term list, such as REGULATOR potentially being a member of PROTEIN_COMPLEX, PROTEIN, BIOLOGICAL_PROCESS.

However, the following general rule applies: **always apply the most specific type from the hierarchy that is applicable to the concept in question.**

In the following section, the hierarchical structure of each group of concepts is shown in the form of a tree, followed by brief definitions of each term.

PROTEINS


Proteins
(NOT A TERM)
Protein_Structure
(NOT A TERM)
Protein_Function
(NOT ATERM)
Polypeptide
Peptide
Protein_Complex
Protein
Transcription_Factor
Enzyme
Regulator
Conformation
Protein_Subunit
Domain
Motif
Sigma_Factor
Repressor
Rho_Factor
Activator
RNA_
Polymerase
DNA_
Polmerase
Restriction_
Enzyme

Proteins

Complex chemical substances chiefly composed of amino acids and their positional references. This includes the physical structure and functional roles associated with each type.

Protein_Structure

Protein structure specifies the sequence of amino acids that a protein consists of and how these chains of amino acids form a 3-D structure in space. There are 4 levels of protein structure: Primary level is the sequence of amino acids, secondary level is the folding of parts of protein into alpha helix and beta sheets, the tertiary structure signifies the complete conformation of protein in 3-D and quaternary structure is only present when more than one polypeptide chains comprise a protein.

Peptide

Short polymer of amino acids containing 2 to 50 amino acids. May not have an associated function or may be a fragment of a protein.

Polypeptide

A polymer of amino acids usually longer than 50 amino acids. Also identified as protein when it can act alone to perform a biological function.

Protein_Complex

The structure formed by the association of two or more individual polypeptides through non-covalent bonding. A Protein_Complex can perform more than one functions, e.g. include ‘The cyclic AMP (cAMP)-cAMP receptor protein complex’, ‘mutagenic UmuD'C protein complex’.

Conformation

Refers to the 3-D structure of a polypeptide in space. This is the highest level structure and may also be referred in literature as protein configuration or protein 3-D model.

Protein_Subunit

A functional part of a Protein which is derived from a process of gene expression independent to the rest of the protein, e.g. ‘iron-sulphur protein subunit’, ‘acyl-carrier subunit’

Domain

Part of a Protein or Protein_Subunit, usually associated with protein primary structure. e.g. ‘superimposable dinucleotide fold domains’, ‘transmembrane domain II’ and assigned a specific function.

Motif

Motif, or more accurately a structural motif is a local structure in a protein chain which manifests itself as a fold or loop, like ‘helix-turn-helix loop’.

Protein_Function

This specifies the role of protein in vivo or in vitro. A protein could perform a catalytic role where it is referred to as an enzyme or it may be a part of cell membrane or act as a transport protein. E.g. ATP-ion channel, ATPase dependent pump.

Transcription_Factor

Any DNA binding protein that binds to a protein binding site on DNA with the aim of regulating gene expression, e.g. ‘FabR’ or ‘RNA polymerase II transcription factor’.

Sigma_Factor

Any of the reported catalogue of RNA polymerase co-factors, e.g. ‘heat shock sigma factor 32’, ‘eubacterial sigma factor’.

Repressor

A protein or chemical whose observed effect is to either directly or indirectly lessen or obliterate the rate of gene expression, e.g. ‘DNA-binding protein H-NS (represses...)’. Similar verbs would include: ‘shrink’, ‘weaken’, ‘attenuate’, ‘ease’, ‘minimize’, etc.

Rho-factor

A protein found in bacteria which takes part in termination of transcription. It is a part of a much larger termination complex.

Activator

A protein or chemical whose observed effect is to either directly or indirectly initiate the process of gene expression, e.g. ‘Phosphorylation of OmpF (an activator of...)’
Similar verbs would include: ‘start’, ‘actuate’, ‘stimulate’, ‘trigger’, ‘initiate’, ‘begin’, ‘mount’, etc.

Enzyme

All proteins performing catalytic functions are classified as enzymes. The suffix ‘ase’ is attached at the end of the name of an enzyme to distinguish from other proteins, e.g. ‘beta galactosidase’ or lactose permease’.

RNA_Polymerase

An enzyme which transcribes RNA from a DNA template. When the template is RNA, as in some viruses, the enzyme is referred to as reverse transcriptase. All classes of RNA polymerases should be annotated as RNA_Polymerase.

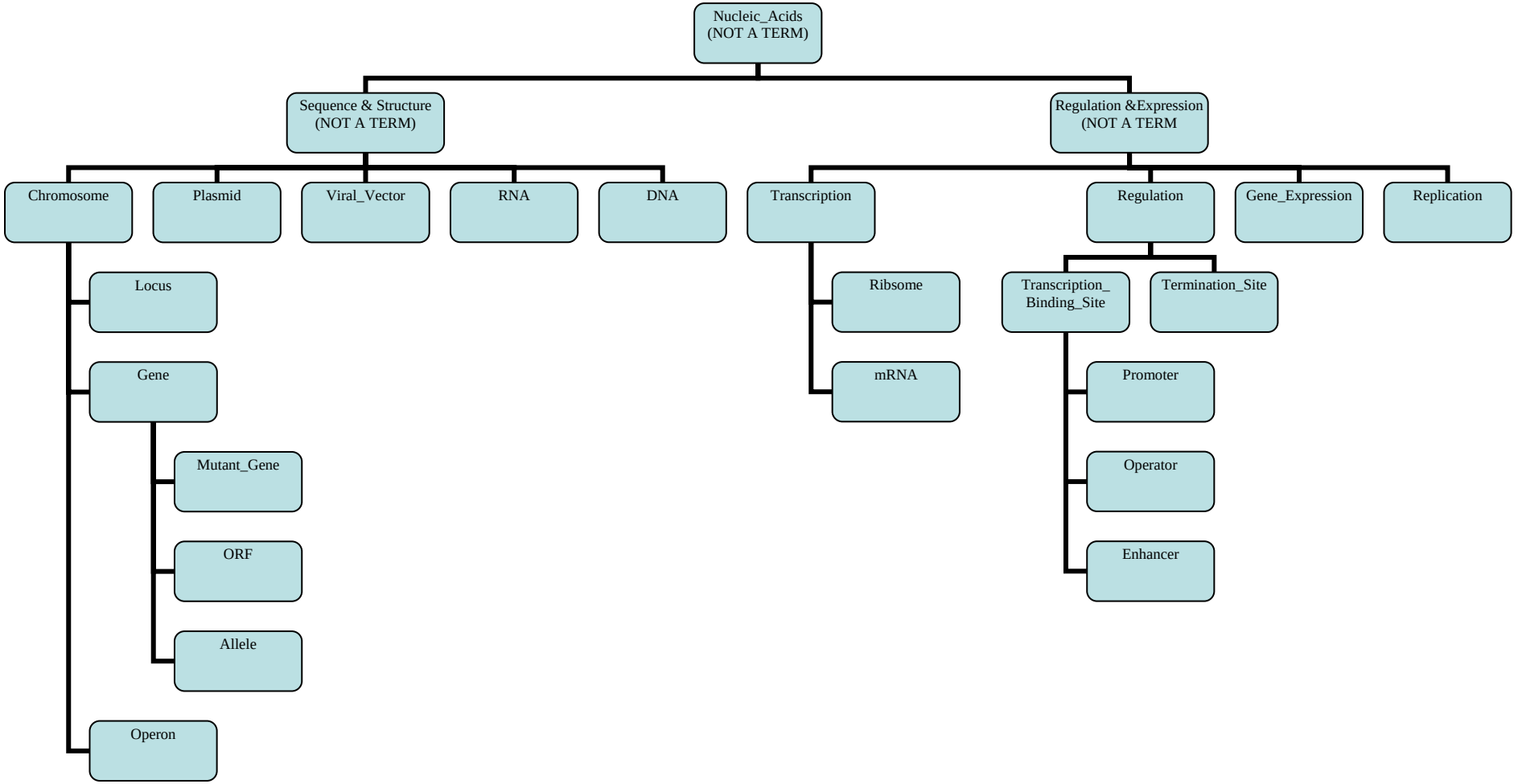
DNA_Polymerase

An enzyme that is involved in the replication of DNA. Different types of DNA polymerases exist in nature performing polymerization of DNA along with 5’ to 3’ or 3’ to 5’ proof reading function. All instances of these enzymes, whether intact enzyme or parts of it should be annotated as DNA_Polymerase. eg. Klenow fragment.

Restriction_Enzyme

An enzyme that cuts double-stranded DNA at specific sites. The sites are 4 to 8 bp long and are called restriction sites. Examples include ‘EcoR1’, ‘BamH1’, ‘HindIII’ etc.

NUCLEIC_ACIDS



Nucleic_Acids

Chromosome

A chromosome is a single long sequence of nucleotides (DNA) which is compacted into a dense structure with the aid of histone proteins. This structure is not visible as such except during mitosis. The two arms of chromosomes are p and q and they should be annotated as Chromosome. In bacteria, the single long chain of genomic DNA is sometime referred to as chromosome and does not contain histone proteins.

Locus

The reported chromosomal location of a gene, e.g. ‘(at) srl’, ‘(at) recA200’

Gene

A locatable region of DNA which contains a sequence of bases that encode for the amino acid strings required to build proteins, e.g. ‘lipoprotein gene’, ‘phoE gene’ through the formation of mRNA. Also several genes express rRNA and tRNA.

Mutant_Gene

Any alteration in the sequence of nucleotides in a gene whether brought on by natural factors or those brought on through the manipulation of the organisms genome, e.g. ‘K-12 lamB mutants’, ‘dnaAts mutants’. The class Mutant_Gene should be applied to all forms of Gene mutation, where the Gene is the term linked to the semantic role.

ORF (Open Reading Frame)

A region of DNA containing a sequence of bases that could potentially encode a protein, e.g. ‘2754-bp open reading frame’, ‘short-component open reading frames’

Allele

An allele is an alternative form of a gene (one member of a pair) that is located at a specific position on a specific chromosome. Eg. Blood group A, B and O are allelic forms of a single gene.

Operon

A functioning unit of DNA composed of an operator, a common promoter, and one or more genes, e.g. ‘lactose operon’.

Plasmid

A copy of circular DNA found in bacteria and yeast. For example ‘lacZ containing plasmid pBR322’.

Viral_Vector

A molecule of viral DNA or RNA that is used as a vector for carrying gene segments.

RNA

Polypeptide of ribonucleotides. For all types of RNA (rRNA, tRNA, siRNA etc.) other than mRNA use the RNA category for annotation.

DNA

The polypeptide of deoxynucleotides. Any entity comprising of DNA should be assigned to an appropriate category in under the Nucleic_Acids categories. If it cannot be assigned appropriately to any other category, then use DNA category as a last resort. Example 'the conserved DNA region on chromosome 21'.

Transcription

The process of RNA formation from a DNA template.

Ribosome

Any compositional rRNA or subunit structures of a functioning Ribosome.

mRNA (messenger RNA)

The transcribed RNA from an ORF. mRNA in eukaryotes is further processed into introns an exons.

Regulator

A protein or chemical whose observed effect is to either directly or indirectly alter the rate of gene expression without a stated bias, e.g. 'FlhD/FlhC (a regulator of...)' where 'a regulator of' is excluded from the span of the term tagged, but used to indicate the type of NER tag used or by the verbs agentive form 'the FlhD/FlhC *regulator'*. Similar verbs would include: 'modulate', 'control', 'govern', 'coordinate', 'guide', etc.

Transcription_Factor_Binding_Site

The type associated with a sequence of bases which form the binding sites for proteins involved in the initiation or regulation of gene expression, e.g. 'TFBS', 'TATA binding protein'.

Promoter

The regulatory region of DNA located upstream of a gene, providing a control point for regulated gene transcription, e.g. 'promoter of the uxuR', 'promoter region'.

Operator

A regulatory sequence usually found upstream of an ORF where activator or inhibitor of that gene binds.

Enhancer

A sequence of DNA found usually upstream of an ORF where an activator protein binds. This results in enhanced expression of the gene.

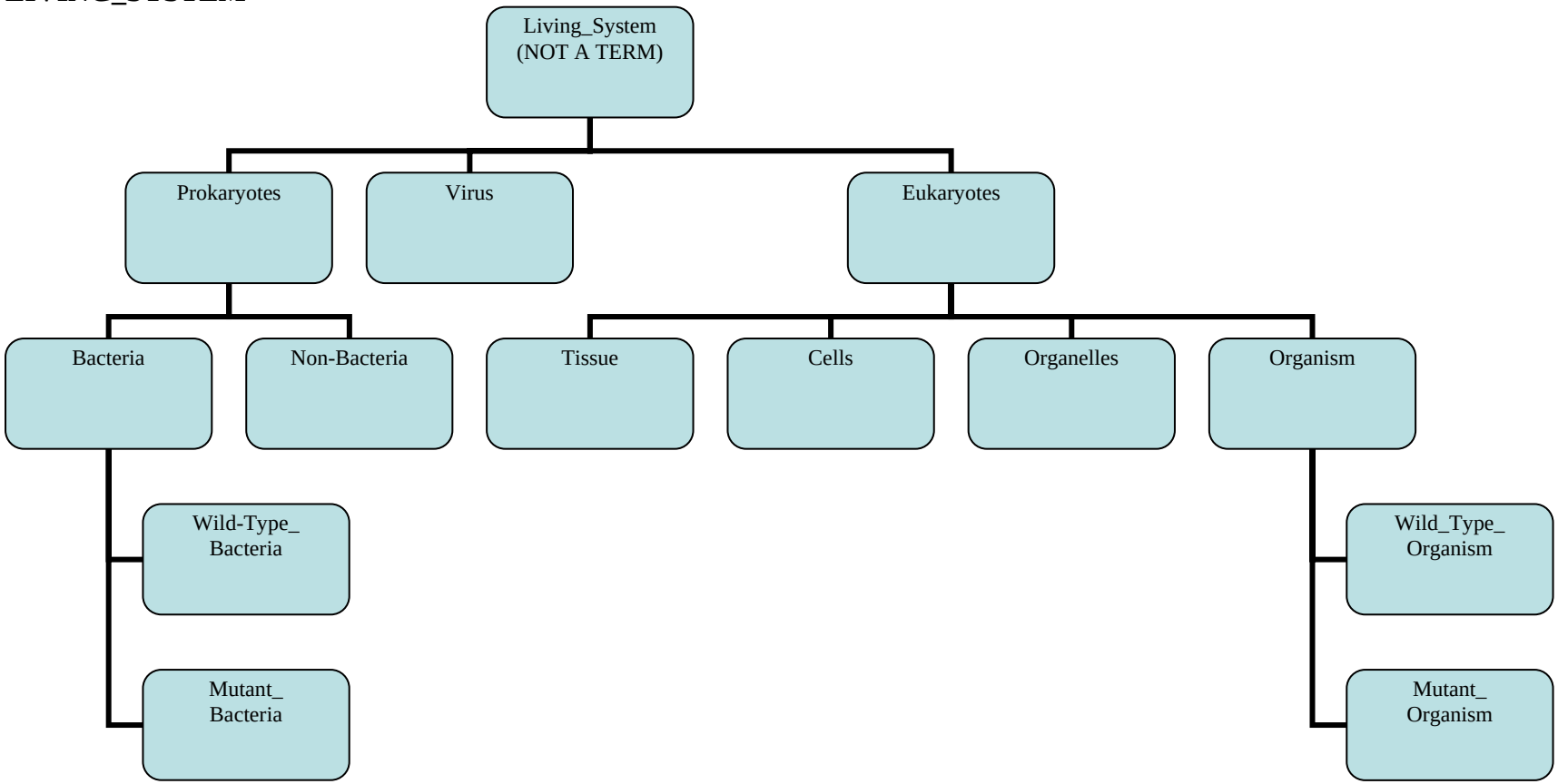
Gene_Expression

The process of formation of a protein from a gene. It includes transcription and translation.

Replication

Specifically DNA replication. The process of copying DNA. Here it applies to both choromosome replication in eukaryotes and plasmid replication in bacteria.

LIVING_SYSTEM



Living_Systems

Here mainly refers to living cells, tissues and organisms.

Prokaryotes

They are a group of organisms, usually single-celled, that lack a nucleus and usually divide through non-sexual binary fission. Examples include bacteria and blue green algae.

Bacteria

A group of unicellular organisms that are found all over the planet. They are characterized by the absence of nucleus and contain a single long (sometimes circular) DNA molecule. All strains of bacteria should be annotated with category Bacteria.

Non_Bacteria

Any single celled prokaryote other than bacteria.

Wild_Type_Bacteria

The bacteria found in nature. These are non-modified bacteria found naturally in environment and inside the bodies of living organisms. Only annotate as Wild_Type_Bacteria when the context is clear.

Mutant_Bacteria

The bacteria whose DNA has been modified structurally by deletion, insertion or point mutation. Only annotate bacteria as mutant when the context specifies it to be so.

Virus

Virus is a infectious agent that cannot grow outside the body of an organism it infects. Usually referred as virus, eg. Polio virus, ebola virus etc., but could also appear as bacteriophage and as viral vector.

Eukaryotes

These are cells which have distinct nucleus and contain various organelles for specilized functions. All plants, fungi and animals fall into this category.

Tissues

A complex organization of one or more types of cells. Tissues form the structural basis of organs and systems in complex organisms. Eg. includes bone tissue, muscle tissue etc.

Cells

The fundamental structural and functional unit of life. Also called the building block of life. Should only be annotated when no choice is found among the other categories.

Organelles

Subcellular compartmentalized bodies found in eukaryotic cells. Mitochondria and chloroplasts are good examples.

Organism

It is an individual living system which function independently. The hierarchy of cells and tissues perform specialized functions. All multicellular living systems fall into this category.

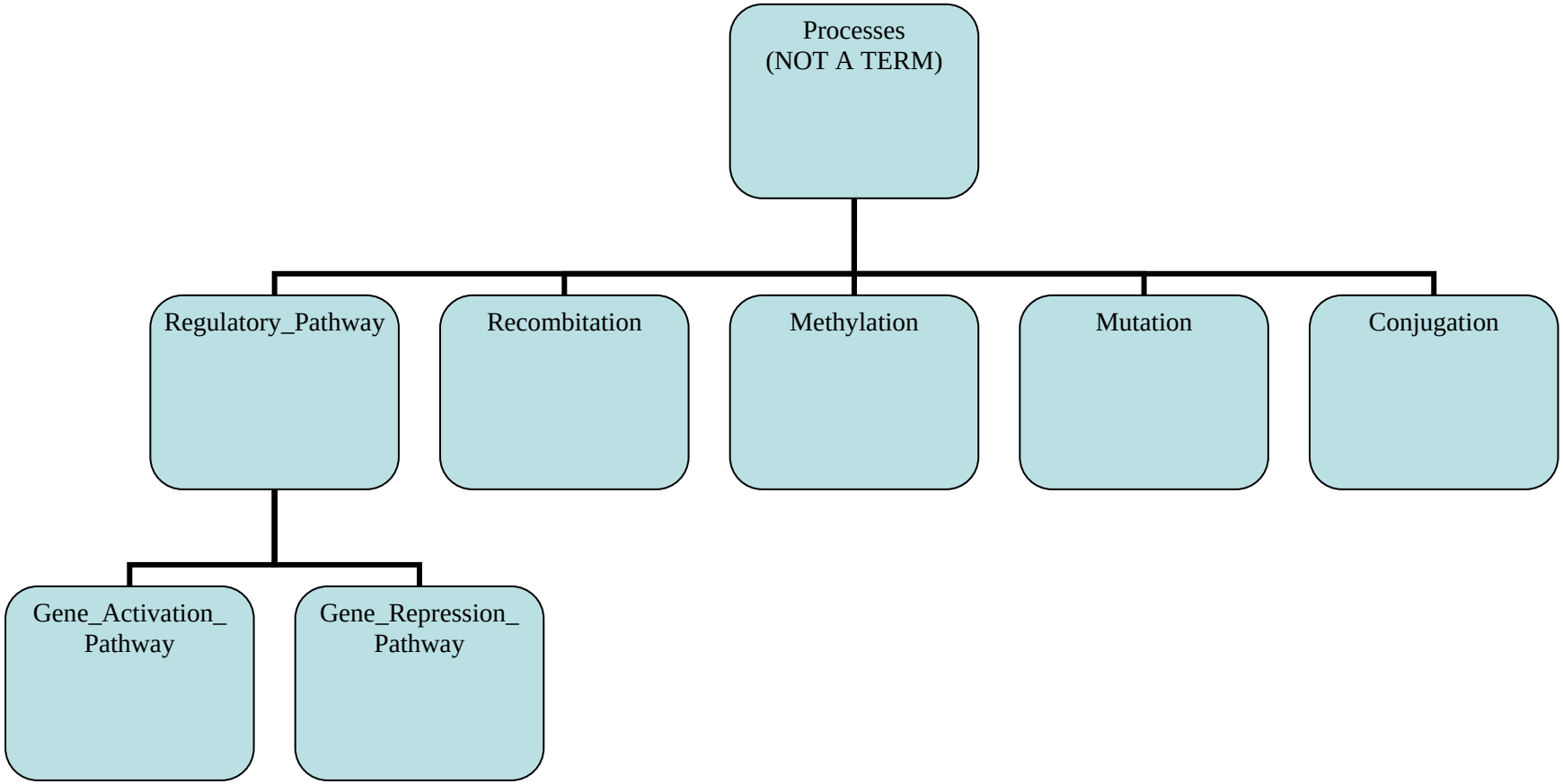
Wild_Type_Organism

An organism that is found in nature. Any organism used in an experiment which is not mutant will go in this category.

Mutant_Organism

An organism that has been altered genetically to insert or delete a certain function. All transgenic organism like NUDE mice are also included in this category.

PROCESSES


Processes
(NOT A TERM)
Regulatory_Pathway
Recombitation
Methylation
Mutation
Conjugation
Gene_Activation_
Pathway
Gene_Repression_
Pathway

Processes

A set of concept classes used to label biological processes described in text. The specifics of the interactions described can be inferred from the terms SEMANTIC ROLE TYPES that will be labelled during the document curation process.

Regulatory_Pathway

Any sequence of interactions that take parts in activation or inhibition of genes.

Gene_Activation_Pathway

Implied series of interactions (containing nominalised verbs) ranging from signal transduction leading to the initiation of transcription to the final post-transcriptional modifications of the protein product. In other words, everything to do with the synthesis of a protein, named or otherwise, e.g. ‘the gene pathway’ or ‘kinase activation pathway’.

Gene_Repression_Pathway

The series of interactions leading to the inhibition of a gene.

Recombitation

Process by which genetic material is exchange between the two homologous sister chromatids during synapse formation in prophase stage of mitosis. This term is also used in prokayotes like bacteria when interchange of DNA is taken place.

Methylation

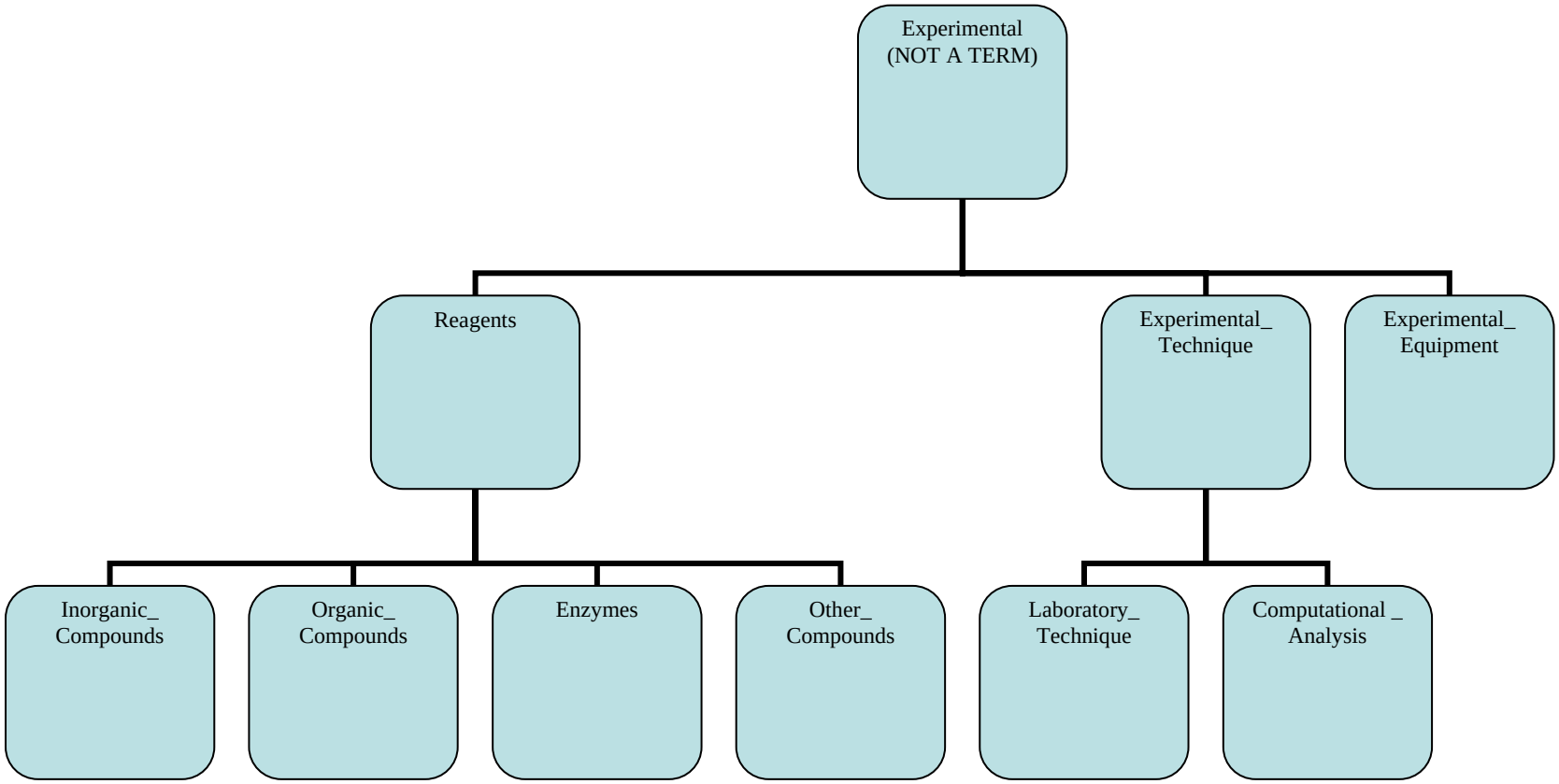
Specifically in biological systems refers to the addition of methyl group by enzymes to lysine or arginine of histone proteins in the chromosomes. The methylation has wide implications in gene regulation and transcription.

Mutation

Any alteration in sequence of DNA either by deletion or insertion of nucleotides or through conversion of one nucleotides into other (point mutation). During assigning an event as mutation make it sure the context clarifies the type of mutation. If it is not clear from the context then assign category Mutation.

Conjugation

EXPERIMENTAL



Experimental

Reagents

Reagent is a general term implied to a chemical substance that is consumed during a chemical reaction.

Inorganic_Compounds

Non-carbon based compounds such as salts or other minerals.

Organic_Compounds

For the purposes of this annotation scheme these are the set of carbon based compounds produced through biosynthesis, e.g. lipids, drugs, metabolites.

Other_Compounds

Compounds that could not be assigned to any of the above categories. This will be a rare situation as all compounds are organic or in-organic. So use this category when every other option has been exhausted.

Experimental_Technique

The generalised class referring to techniques or SOPs. Most should be covered by either of the two classes ‘LABORATORY_TECHNIQUE’ or ‘COMPUTATIONAL_PROCESS’.

Laboratory_Technique

The wet-work associated with experimentation. All forms of laboratory-based technique for recording observations, altering conditions or physical forms of the subject under analysis e.g. mass spec, western blot, gene splicing, restriction digests, etc.

Computational_Analysis

In-silico analysis e.g. BLAST homology search, all forms of statistical analysis and any of the dry work associated with experimentation.

Experimental_Equipment

Laboratory equipment used in the execution of experimentation, e.g. laboratory consumables, machines, glassware, etc.

The task

We have identified a list biologically-relevant verbs which *potentially* describe gene regulation events. Firstly, we want to discover which of these verbs *actually* represent such events in biomedical abstracts. For those that *do* represent gene regulation events, we would like to construct dictionary entries that characterize their behaviour. In order to do this, we need to discover the following:

a) The *patterns* of variable-denoting phrases that can occur with these verbs.
b) The *semantic roles* of each phrase in the pattern
c) The *type of concept* that best characterizes each phase (if the phrase corresponds to a concept in our hierarchy)

This information will be discovered by examining the usage of these verbs within a large number of biomedical abstracts. For each occurrence in a text of one of the verbs in our list, we wish to carry out a number of steps. This process described is called *annotation*, and will be carried using a program called *WordFreak*, which has been adapted for the task to make it as simple as possible. The tool marks the verbs contained in the biologically relevant list, and allows them to be located automatically. A separate document explains the use of this tool. The steps to be undertaken during annotation are as follows:

1) Determine if the event described by the verb is related to *gene regulation*. If so, steps 2-6 are carried out
2) Identify/locate variables in the event denoted by the verb.
3) Mark appropriate text spans to represent the event variables.
4) Determine the correct semantic role for each of the variable-denoting phrases identified.

NOTE: All variables in the sentence that are associated with the event should be annotated, regardless of whether one of the existing semantic roles seems appropriate. If none of the semantic roles seems to characterize the variable, then UNDERSPECIFIED role may be applied, together with a comment

5) If the phrase corresponds to a concept in the concept hierarchy (either an entity or a process), assign an appropriate label
6) If a variable-denoting phrase describes a further event, represented using another verb or nominalised verb, then steps 1), 2) and 3) are repeated to identify and characterize the variable-denoting phrases that are used to describe this "embedded" event.

   For example:

   OmpF reduction *required* a mutation in the marA region

For the verb *required*, the two variables are *OmpF reduction* and *a mutation* both correspond to nominalised verbs. These both contain nominalised verbs, and so their own variables should be annotated. *Reduction* has one variable, a THEME, i.e. *OmpF,* whilst *mutation* also has a single variable, a LOCATION, i.e. *in the marA region.*

NOTES:

a) Nominalised verbs should ***only*** be considered if they occur **within a variable-denoting phrase of one of the verbs of interest.**

b) Nominalised verbs **should NOT as be annotated as separate events** if they specify **no variables of their own**

c) Nominalised verbs **should NOT be confused with other types of nouns that also seem to have variables associated with them**. Nominalised verbs ***only*** correspond to nouns that have the same meanings as verbs (i.e. they represent actions or states). **Some common confusions are as follows**:

i. "The UV light *inducibility* of the uvrB operon was demonstrated …"

Here, *inducibility* represents a *property* of the *uvrB operon*, rather than an event, but such properties should ***not*** be marked for their variables. Other examples are *sensitivity* and *absence.*

ii. the open reading frame ybbI encodes the *regulator* of expression of the copper-exporting ATPase, CopA.

Words ending with *–or* or *–er* such as *regulator* correspond to entities perform for a particular action (here an entity that performs *regulation*). Like nominalised verbs, they can have variables associated with them (e.g. what is being regulated). However, such words ***should not*** be considered

The basic steps of the annotation task are relatively straightforward. There are, however, a number of challenges to the task, some of which should be made more straightforward by the guidelines that are provided in the sections below. Some of these challenges are as follows:

a) Determining whether each pre-marked verb denotes an event related to *gene regulation.* Annotation should **only** be carried out if this is the case.

b) Identifying/locating event variables. In many cases, sentences to be annotated can be fairly complex and require careful reading in order to correctly identify the variables.

c) Marking appropriate lengths of variable-denoting phrases. Generally, we want these to be as short and consistent as possible, to aid in the machine-learning of dictionary entries. A set of guidelines provided in the *Marking Spans* section below aims to outline more precisely what should and should not be included within the marked phrases.

d) Assigning appropriate semantic roles to variable denoting phrases. Each semantic roles can generally appear in a range of positions or contexts with respect to the verb or nominalised verb that represents the event. The *Semantic Roles* section

aims to help with this by providing a clear description of each semantic role, together with examples of different contexts in which variable-denoting phrases corresponding to the role can appear.

e) Determining which verb or nominalised verb a particular variable-denoting phrase belongs to. If there are multiple verbs or nominalised verbs within a sentence, it is important to consider carefully which of these each phrase actually belongs to, i.e. to which of the event descriptions the phase is contributing. Let us reconsider sentence (2) from above:

   The LysR-type transcriptional regulator CysB controls the repression of hslJ transcription in Escherichia coli.

   We determined above that there are 3 events described in this sentence, i.e. *control, repression* and *transcription.* At the end of this sentence is the location *in Escherichia coli*. The challenge is thus the decision of which of the event(s) this location belongs to as a variable-denoting phrase. If the location could feasibly apply to more than one event, then it is possible for a variable to be specified as belonging to multiple events.

f) Assigning concept categories to event variable. There are approximately 60 categories, which are organized in a hierarchy. Careful consideration may be required to determine the most appropriate category to assign. It is always the case that the most specific category that can apply to the concept should be assigned. If there is doubt, then a concept further up the hierarchy may be assigned.

Determining appropriate events to annotate

As mentioned above, each abstract to be annotated contains a number of pre-marked verbs which have biologically relevant meanings. However, **only those verbs that are relevant to gene regulation should be annotated.**

To put this in clearer terms, the types of events that should be annotated are those that describe any interaction which leads, either directly or indirectly to the production of a protein. This general rule should, however, be restricted to sentences that contain some mechanical description of transcription, translation or post-transcriptional modifications and/or their controls. Some examples include:

- indirect activation of protein production through environmental stimulus
- the finalisation of protein through post-translational modifications including all naturally occuring processes and those manipulated experimentally.
- DNA alterations, mutations, and chimera creation, ***if*** they describe modifications to the process of gene expression or the proteins expressed

Here are some other rules general rules:

- Generally speaking protein-protein interactions are not to be annotated when the result of their interactions does not lead to the expression of a gene.
- Alterations to DNA (structural, compositional), kinetics, that do not lead to gene expression should also **not** be annotated.
- Even if the abstract relies on underlying gene expression, protein finalisation, etc. but does not describe any such reaction in detail, do not annotate. For example, growth of cancerous tissue is obviously the result of aberrant gene expression, but unless the mechanism is described, ignore it.
- Do not annotate events relating to the function of the protein, rather than the processes resulting in the creation of the protein.

For example, in sentence 1 below, both the "binds" and "activates" events should be annotated. However, in sentence 2, the "binds" event should not be annotated, as it is unclear whether or not the interaction leads to an expression event.

1. Protein X binds to Protein Y which activates promoter Z.
2. Gene X expresses Protein Y which binds to the Protein Z.

However, if Protein Z is described as playing a regulatory role in the same text, then the binding of Protein Y to Protein Z CAN be annotated.

Identifying variables

After it has been determined whether a verb relates to a gene regulation event, the next step is to identify the variables involved in the event. **An important point to note here is that variables should be annotated whether or not they correspond to concepts in the hierarchy. If the variable corresponds to a concept in the hierarchy, then the concept should be assigned. Otherwise, the variable should still be annotated and assigned only an appropriate semantic role.** For example:

We ***employed*** oligonucleotide-directed site-specific mutagenesis to dissect the promoter region of the gene

For the event denoted by the verb *employed*, the AGENT is *we* i.e. the authors. Although *we* does not correspond to a biologically interesting entity, it should still be annotated as a variable of the *employed* event and assigned the semantic role of AGENT.

Identifying variables can often be quite straightforward, the task can be more complex for sentences containing multiple verbs. Normally one of these verbs is referred to as the *main* verb, in that it describes the main or most important event in the sentence, i.e. it characterises what the sentence is about. Other verbs denote secondary events in the sentence.

During the annotation process **it is required that *all* verbs in the biologically relevant list are annotated with their variables (provided that the verb describes a gene regulation event**, see section 6), *regardless of whether these verbs are main verbs or secondary verbs within the sentence.*

In the following sections, we provide examples and discussion of how variables can be identified in various types of more complex sentences.

Descriptive events

Sentences sometimes contain descriptive information about an entity or event that is involved in the main event of the sentence. An example is shown in (a).

(a) Expression of the ompF and ompC genes, which *encode* the major outer membrane proteins, OmpF and OmpC, respectively, is *affected* in a reciprocal manner by the osmolarity of the growth medium.

This sentence contains 2 events, namely:

- The main event, denoted by the verb *affected.*
- A secondary event, denoted by the verb, *encode,* which proves descriptive information about some of the entities involved in the main event, i.e. *the ompF and ompC genes.*

Whenever a sentence contains multiple verbs that are marked for annotation, the ones which relate to gene regulation should all be annotated, regardless of their position in the sentence (i.e. a main or secondary event).

Where there are two or more verbs in a sentence, their variables can be "intertwined" and sometimes well separated from the verb that denotes the event. In (a), for example, the THEME of the *affected* event is the event denoted by the nominalised verb *expression.* This is well separated from the verb *affected* by the *encode* event and its variables. It is thus important to think carefully about which parts of the sentence belong to which event. It may be helpful to consider how the different events could be separated out into different sentences, containing the variables associated with a particular event. For (a), this results in the following:

i) Expression of the ompF and ompC genes is *affected* in a reciprocal manner by the osmolarity of the growth medium.
ii) The ompF and ompC genes *encode* the major outer membrane proteins, OmpF and OmpC, respectively.

In ii), the word *which* that precedes the verb *encode* in (a) has been replaced by *The ompF and ompC genes*. Verbs preceded by *which* provide a description or extra information about something that has already been mentioned in the sentence; the word

which is a sort of placeholder for the thing that has previously been mentioned. If the verb following *which* is a verb to be annotated, then it must be determined which other phrase in the sentence the word *which* is referring to: it is this phrase, and not the word *which*, that should be annotated as the variable of the event. In (a), the phrase that *which* is referring to is *The ompF and ompC genes*, an so it is this chunk that should be annotated as the AGENT of the *encode* event.

A final point to note with this example is that the phrase *the ompF and ompC genes* is both the AGENT of the *encodes* event and the THEME of the event denoted by the nominalised verb *expression.* There is no problem with this – the same phrase can be annotated as being a variable of more than one event.

Similar types of constructions can occur with the word *that*, as shown in (b)

(b) Analysis of mutants with deletions that were *derived* from the uxuR::Mud1 insertion strain *confirmed* the counterclockwise transcription direction of the uxuR gene.

In (b), the verb that denotes the main event of the sentence is *confirmed*, whilst a secondary event is denoted by the verb *derived.* This secondary event provides extra information about the mutants. Sentences can be created that contain the variables that are relevant to each event. These are shown in i) and ii).

i) Analysis of mutants with deletions *confirmed* the counterclockwise transcription direction of the uxuR gene.
ii) Mutants with deletions were *derived* from the uxuR::Mud1 insertion strain.

Sometimes, secondary events of the type shown in (b) can be expressed without using *that* but instead using the –*ing* form of the verb. An example is shown in (c).

(c) A mutant strain *displaying* altered regulation of the recA gene was *isolated* as a revertant of a lexA3 recA200 double mutant

In (c), the variables of both *isolated* (the main verb) and *displaying* should to be annotated. The verb *displaying* is providing a description of the mutant strain; it is a shortened form of *that displays* or *that displayed.* Here, therefore, the phrase *a mutant strain* should be annotated as both the THEME of the *isolated* event and the AGENT of the event denoted by *displaying.*

Events specifying evidence or certainty level

A certain type of sentence construction is reasonably common when the author wishes to mention explicitly the type of evidence that exists for a mentioned event. An example is shown in (d).

(d) Normal expression of fimA was *shown* to *require* the integration host factor (IHF).

In (d), there are 2 verbs, i.e. *shown* and *require*. Only *require* should be annotated as an event. The main event in this sentence is the one denoted by *require*, whilst the verb *shown* is one of a set of verbs that can be used in this type of sentence construction to indicate the type of evidence for the event. We will refer to these as "evidential" verbs. The verb *shown* indicates that there is strong evidence to back up the specified main event. Replacing *shown* with *believed* would indicate that there is no evidence to back up the *require* event; it would be just a conjecture.

In terms of the syntactic structure of sentence (d), the event denoted by *normal expression* belongs to the verb *shown*. However, in terms of meaning, this *expression* event should be marked as a variable of the event denoted by the verb *require,* i.e. the AGENT. This emphasizes the fact that meaning as well as the structure of the sentence should be taken into account during annotation. It is possible to rephrase (d) so that the structure makes it easier to determine that the *expression* event is a variable of the *required* event. This rephrasing is shown in (e).

(e) It was shown that normal expression of fimA *requires* the integration host factor (IHF).

Both the sentence structures shown in (d) and (e) occur in the biological literature. In constructions of the type shown in (d), the subject of the "evidential" verb will almost always be a variable of the verb in the infinitive form (i.e. the one preceded by "to", *require* in (d)). In most cases, it will be the AGENT of the event denoted by the infinitive verb.

A list of "evidential" verbs that can occur in constructions such as (d) and (e) are shown below. This list should be taken as indicative rather than exhaustive.

predict, assume, think, suspect, believe, expect, claim, hypothesise, suggest, claim, indicate, suggest, deduce, argue, infer, show, reveal, demonstrate, confirm, prove, report, find, conclude, observe

NOTE: The majority of the above verbs are purely "evidential" and will not be marked for possible annotation in WordFreak. ***However,*** a small number of the verbs do not have purely evidential uses (i.e. specifying evidence related to another event). The verb *show* is one such example, which can be used to specify an event in its own right. An example is shown below in (f):

(f) A strain containing a deletion of the sbcB gene *showed* little dRpase activity

Here, *showed* is being used to describe a property of the strain, and hence it should be annotated as an event.

Another construction similar to (d) can occur with verbs such as *seem* and *appear*. These can also be considered as "evidential" verbs, such as the ones in the list above, but occur in slightly different sentence constructions. The construction shown in (e), which is possible for all the verbs shown in the list above, puts the evidential verb in a passive construction, i.e. a form of the verb *to be*, followed by the past tense form of the evidential verb, e.g. *were assumed, was inferred,* etc. However, *seem* and *appear* occur in active constructions. An example is shown in (g).

(g) oxyS RNA *seems to* <u>*modulate*</u> the stability of a region of secondary structure in the ribosome-binding region of the gene's mRNA

Other than the use of the evidential verb in the active form rather than the passive, this sentence behaves in the same way as other ones containing evidential verbs: the subject of the evidential verb is the AGENT of the verb in the infinitive form. Here, therefore, *oxyS RNA* is the AGENT of *modulate.* It can also be rephrased to make the link between *oxyS RNA* and *modulate* clearer, as shown in (h).

(h) It *seems* that oxyS RNA <u>*modulates*</u> the stability of a region of secondary structure in the ribosome-binding region of the gene's mRNA

A final class of verbs that can co-occur with ones that denote events to explicitly indicate the author's level of certainty towards the event are the modal verbs, such as *could, may* or *might.* An example is shown in (i).

(i) Pseudo-HPr activity *could* <u>*replace*</u> HPr in PEP-dependent phosphorylation of PTS carbohydrates.

In (i), Pseudo-HPr activity is structurally the subject of the modal verb *could,* but in terms of meaning, it is also the AGENT of the event denoted by *replace.* Although the sentence expresses uncertainly as to the truth of the event denoted by *replace,* we are not concerned with the truth of the event when performing annotation; we just want to find out the types of variables that can occur with the verb in different contexts.

Negative events

Following on from what was said in the previous section**, we wish to annotate the variables of events even if the text specifies that the event <u>*did not*</u> happen.** Typically, events are negated using *do + not* in active sentences and *be + not* in passive sentences. An example is shown in (i).

(j) Several transgenic lines *did* not <u>*express*</u> the lacZ transgene.

Although (i) conveys the fact that the *express* event did not actually happen, for the purposes of annotation, we consider the event as though it was positive. So, *several*

transgenic lines is annotated as the AGENT and *the lacZ transgene* is annotated as the THEME.

Another way in which events can be negated is through the use of the verb *fail,* as shown in (j).

(k) Strains carrying a mutation in the crp structural gene *fail* to *repress* ODC and ADC activities.

Once again, for the purposes of annotation, the variables of the *repress* event should be labelled with semantic roles as though the event was positive. So, for example, the strains that are the subject of *fail* should be annotated as the AGENT of *repress.*

Events specified using nominalised verbs

It was described above that variables of events that are denoted by verbs can be further "embedded" events, which are often described using nominalised verbs. In this case, we wish to annotate the phrases that correspond to the variables of the nominalised verb. An example is shown below:

(l) **Phosphorylation** of OmpR by the osmosensor EnvZ *modulates* **expression** of the ompF and ompC genes in Escherichia coli.

In (k), we initially consider the verb *modulates.* The AGENT and THEME of the event denoted by *modulates* are both "embedded" events that are denoted by nominalised verbs, i.e. the AGENT is the *phosphorylation* event and the THEME is the *expression* event. The verb *modulates* also has a third variable, i.e. the LOCATION of the event, *in Escherichia coli.* Having identified the "embedded" events, we then proceed to identify their own variables. The *phosphorylation* event, for example, specifies a THEME, i.e. *OmpR* and an AGENT, i.e. *the osmosensor EnvZ.* The *expression* event specifies a THEME, i.e. *the ompF and ompC genes.*

It can be noticed that the verb in (k), i.e. *modulates,* acts as a sort of boundary for the variables of the events specified by nominalised verbs. The variables specified of both of these events occur on the same side of the verbs as the event itself. So, both variables of the *phosphorylation* event occur *before* the verb *modulates,* whilst the variables involved in the *expression* event occurs *after* the verb.

When variables involved in events denoted by nominalised verbs are being identified, we impose the restriction that they must *always* occur on the same side of the verb for which the nominalised verb has been identified as a variable. This also applies when phrases on the other side of the verb seem to relate to the nominalised verb, when meaning is considered. An example of this is shown in (l).

(m) **Overproduction** of the exuR repressor also *caused* **a decrease** of the beta-galactosidase level.

In this example, the variables of the verb *caused* are both events expressed by nominalised verbs. The AGENT is the *overproduction* event and the THEME is the *decrease* event. When considering the *decrease* event, the THEME is easily identifiable as *the beta-galactosidase level.* However, if the meaning of the sentence is considered, then the *overproduction* event could be seen as the AGENT of the *decrease* event: the meaning of the sentence is actually that the overproduction of the exuR repressor *decreased* the beta-galactosidase level. However, as the *overproduction* event is separated from the *decrease* event by the verb *caused,* it should *not* be annotated as one of its variables.

Marking variable spans

After event variables have been identified, the next step in the annotation process is to mark appropriate text spans to represent each variable. In general, we want these text spans to be as short and consistent as possible. However, determining how much text to annotate can sometimes be a tricky process. In order to help with this, we provide in this section a set of guidelines that aim to help with this consistency by defining the kinds of things that should and should not be included within the marked phrases.

Chunks

To help generally with consistent marking of phrases, the biological texts are automatically split into chunks before the annotation is begun. Chunks can be considered as the "building blocks" of the sentence, and so it makes sense that these should be the units we consider when determining the variable-denoting phrases to mark. A simple example of a chunked phrase is shown below.

[NP The narL gene product] [VP activates] [NP the nitrate reductase operon] [PP in] [NP Escherichia coli]

In this example, there are 3 types of chunks. NP (noun phrase) chunks contain sequences of nouns, together with any accompanying adjectives and determiners (e.g. *a, the, that* etc). VP (verb phase) chunks contain verbs or groups of verbs that occur together (e.g. *has activated, were activated,* etc.) whilst PP (prepositional phrase) chunks contain prepositions. Other types of chunks that may be identified include ADVP (adverb phrase) which contain adverbs such as *osmotically* or *aerobically.* The three phrases that correspond to the variables involved in the *activation* event are show below:

AGENT: *The narL gene product*
THEME: *the nitrate reductase operon*

LOCATION: *in Escherichia coli*

By comparing these with the chunked text, it can be seen that each variable-denoting phrase is contained within its own chunk. Indeed, it is normally the case that individual chunks, or in some cases sequences of chunks, correspond to event variables. Thus, in order help maintain consistency between different variable-denoting phases, we impose the guideline that variable-denoting phrases should normally consist of whole chunks. This and other guidelines are explained more fully below.

Instructions of how to correctly select text spans using WordFreak are provided in the WordFreak user manual.

General guidelines

The guidelines in this section apply to all kinds of event variables.

1) **Normally, phrases that denote event variables should cover complete chunks**.

   Consider the following chunked sentence:

   [NP The Klebsiella rcsA gene] [VP encoded] [NP a polypeptide] [PP of] [NP 23 kDa].

   If we consider what should be the AGENT of the event denoted by the verb *encoded*, there are several stretches of text that could seem appropriate to represent this entity, e.g.

   a) Klebsiella rcsA gene, or just
   b) rcsA

   However, for consistency of annotation, *all* words within the NP chunk should be annotated as the event variable, i.e. *The Klebsiella rcsA gene*. **If possible, the event variable should only span a single chunk.** However, there may be cases where multiple chunks must be spanned in order to fully capture the event variable. Some examples are shown in the more detailed guidelines below.

   In a few special cases, it is permitted for PARTS of chunks to be spanned as event variables. Mainly, this applies to annotating *variables* of *nominalised verbs* which may occur in the same chunk, as detailed below:

   Sometimes, nominalised verbs are directly preceded by an argument, e.g. *hslJ transcription*, where *hslJ* is the thing being transcribed, and hence the THEME of the *transcription* event. In terms of chunking, both the nominalised verb and its THEME occur in the same NP chunk, i.e. *[NP hslJ transcription]*. In order to mark *hslJ* as the THEME of the event, it is necessary to use only part of the chunk.

HOWEVER:

When a chunk containing a nominalised verb occurs as a variable of another event, then the *whole* of the chunk should be marked, regardless of whether it contains any variables that belong to the nominalised verb. For example, consider the following sentence:

[NP marR mutations] ***elevated*** *[NP inaA expression]*

If we consider the verb *elevated*, then both of its variables are chunks containing nominalised verbs (i.e. *mutations* and *expression*). At this stage of the annotation, the variables should only be considered as "unanalyzed" units. So, the AGENT of *elevated* is the chunk *marR mutations*, and the THEME is *inaA expression*. Once the variables of *elevated* have been annotated, their internal structure can be considered, if any of them contain nominalised verbs. In this case, *mutations* has the THEME *marR*, whilst *expression* has the THEME *inaA.*

2) **Annotations may consist of *discontinuous* chunks of text**

It is possible for a single annotation to consist of discontinuous chunks of text, i.e. chunks that are not located next to each other. This may be necessary be comply with some of the more specific guidelines below, where examples are given. Instructions of how to create annotations consisting of discontinuous chunks are provided in the *WordFreak* annotation tool manual.

3) **For most role types, event variables should not begin with prepositions.**

It is often the case that phrases denoting event variables are preceded by prepositions. In most cases, such prepositions should *not* be included within the text span covered by the event variables – although they can be fairly reliable indicators of the semantic role of the phrase, they do not contribute to the meaning of the variable. For example, in passive sentences, AGENTs are preceded by the preposition *by*, as illustrated below.

a) *The polyamine biosynthetic enzymes are negatively controlled [PP by] [NP* ***cAMP****] in Escherichia coli.*

Here, the event is denoted by the verb *controlled.* In passive sentence such as this, the subject of the verb (in this case *The polyamine biosynthetic enzymes)* is the THEME of the event, whilst the AGENT (*cAMP*) is preceded by the preposition *by.* The fact that *by* precedes *cAMP* is a fairly reliable indicator that it corresponds to the AGENT role. However, *by* does not actually contribute to the *meaning* of the event variable.

Other types of phrases that include prepositions (e.g. *in response to*) may precede arguments playing particular roles, and these should normally also be excluded from the argument text spans. Further details of prepositions and other phrases that typically precede arguments playing semantic roles are found in the descriptions of individual semantic roles in the *Semantic Roles* section below.

4) **Event variables that are assigned the LOCATION and TEMPORAL roles should *always* begin with prepositions, if a preposition is present**

The LOCATION semantic role has previously been briefly mentioned. In contrast to other role types, prepositions that precede LOCATIONS are an integral part of the variable, as they contribute towards its interpretation. Consider the following sentence:

*Dam methylation alters binding of Lrp [PP **at**] [NP **the GATC1130 site**]*.

In this example, the preposition *at* does more than just indicating the role played by the phrase that follows. The entity *the GATC1130 site* would be interpreted differently if *at* was replaced by another location-indicating preposition, e.g. *in* or *near*. Thus, for locations, the preposition at the beginning contributes to the meaning of the location, and thus should *always* be included within the annotated text span, if present. The same is true for the TEMPORAL role, which is fully section 7.1.8.

Type-specific guidelines

In the *Concepts* section above, it was described how the majority of phrases that denote variables are either:

a) Entities
b) Events, usually expressed using nominalised verbs, but may also be expressed using another verb

Other categories of phrases, e.g. adverbs, are possible, and are detailed in the descriptions of individual semantic roles in the *Semantic Roles* section where appropriate. The guidelines that follow, however, relate specifically to variable-denoting phrases that correspond to either entities or events.

Entity phrases

Entities can be expressed with various degrees of specificity. Some examples are as follows:

- A general type, e.g. *a positive regulator*
- An name and type, e.g. *the OmpR protein*
- A name only, e.g. *OmpF*

All of these may be marked as event variable phrases in different contexts. However, the general rule that should be followed when marking phrases that correspond to entities is the following:

1) **Only the chunk(s) containing most specific characterization of the entit(ie)s should be marked as the event variable.**

Exactly what constitutes the most specific characterization will vary from sentence to sentence. The most specific characterizations possible are *names* of entities, e.g. *OmpR*, and if chunks containing names are present, then these are the ones that should be annotated. In some cases, entities are referred to only be their names, as in a):

a) *[NP **EnvZ**] functions through [NP **OmpR**] to control porin gene expression in Escherichia coli K-12.*

In other cases, the entity name is accompanied by its type, but they both occur in the same chunk. In this case, the whole chunk should be marked as the event variable. An example is shown in b).

b) *It was concluded that expression of [NP **the uxuR gene**] itself is repressed by its own product.*

It is often the case that entities represented by names are either preceded or followed by a more general characterization of their type, as shown in c) and d). In such cases, *only* the chunks that contain the name of the entity should be marked as the event variable.

c) *[NP a chromosomal locus], [NP **slpA**], ….*
d) *[NP **the OmpR protein**] , [NP a positive regulator] [PP of] [NP both genes], ….*

If, however, a general characterisation or type of an entity is present *without* an accompanying entity name, then this general characterisation should be marked as the event variable phrase. An example is shown in e).

e) *[NP **This operon**] is negatively controlled …*

Sometimes, the name of an entity is accompanied by a shorter name or acronym, often in brackets. In this case, it is the shorter name that should be annotated. Examples are shown in f), g) and h).

f) *[NP the trp promoter] ([NP **trpPO**]) ….*

g) *[NP the integration host factor] ([NP **IHF**]) ….*
h) *[NP the fumarate reductase] ([NP **frdABCD**]) [NP operon].*

It may be the case that the name of an entity spans more than one chunk; in this case, all chunks that contain the name should be spanned, as shown in i) and j).

i) *[NP **marA**] **: :** [NP **Tn5**]*
j) *[NP **the uxuR**] **: :** [NP **Mud1 insertion strain**]*

The next guideline refers to lists of entities:

2) When list of entities occur, the general rule to follow is that a *single, discontinuous* annotation should be created, consisting *only* of the items in the list, excluding punctuation marks (e.g. commas) and other words such as *and, or* etc.

An example is shown in k).

k) *A transducing lambda phage carrying [NP **glpD''lacZ**], [NP **glpR**], and [NP **malT**] was isolated from a strain harboring a glpD''lacZ fusion.*

Here, the actual variable annotated consists of the three separate spans *glpD''lacZ, glpR* and *malt,* excluding the comma and the word *and.* Instructions of how to create such a discontinuous span can be found in the manual for the *WordFreak* annotation tool. **Concept types should be assigned to each item in the list whether the items represent the same concept or different concepts**. **It is suggested that concept types are assigned to individual entities in the list *prior* to creating the variable annotation.**

As with single entities, lists of entities may be preceded or followed by a general characterization of the entities. The same rule applies about only annotating the most specific characterizations of the entities. In l) m) and n), the general characterizations or long names are followed by shorter entity names, and so, following guidelines 1) above, it is these shorter entity names that should be annotated. As with example k), the annotated spans consist of discontinuous chunks, corresponding to the individual items in the lists.

l) *[NP Escherichia coli superoxide dismutase] ([NP **sodA**] and [NP **sodB**]) [NP genes] ....*
m) *[NP the fumarate reductase] ([NP **frdABCD**]) [NP operon] and [NP the aerobic C4-dicarboxylate transporter] ([NP **dctA**]) [NP gene] ….*
n) *The Escherichia coli Ada protein activates sigma(70)-dependent transcription [PP **at**] [NP three different promoters] ([NP **ada**] , [NP **aidB**], and [NP **alkA**]) ….*

In m), the marked entities, i.e. *ada, aidB* and *alkA* specify the LOCATION of the *activates* event. The preposition used to specify the location, i.e. *at*, precedes the more general characterisation of the list of entities, i.e. *three different promoters.* However, according to guideline 4) of the general guidelines for marking entities, prepositions should be included in LOCATION spans if they are present.

Lists of entities that consist only of two items that are conjoined with *and* or *or* many be contained within the same chunk. In this case, a discontinuous span should still be used, by selecting the appropriate parts of the chunk, minus the conjoining word. Examples are shown in o) and p).

o) *[NP the major outer membrane proteins], [NP* ***OmpF*** *and* ***OmpC****]….*
p) *[NP The regulatory proteins* ***OmpR*** *and* ***EnvZ*** *]….*

In some cases, the full form of lists of items is "reduced", in that a word or phrase at the end of the list applies to all items in the list, for example:

q) the csrB-lacZ expression defects were caused by ***[NP uvrY], [NP csrA], or [NP barA mutations]***

In this case, the list is "shortened", in that the individual items are actually *urvY mutations, csrA mutations* and *barA mutations*. In such cases, where the individual items in the list do not have "complete" meanings on their own, the span to be annotated is the ***complete*** span, starting with the earliest item on the list with the incomplete meaning, and ending with the last. In this case, punctuation marks and *and/or* etc, *should* be included within the span annotated.

3) **Negative items in lists should be dealt with in the same way as positive items**

Some lists can include negative as well as positive members, i.e. some members of the list are explicitly marked as not playing the role in the event that the positive members of the list play. Negatively marked items in list are normally preceded *but not*, following the positive items in the list. An example is shown in r).

r) *[NP* ***Iron****], [NP but not] [NP* ***manganese****], acted as a corepressor …*

In r), there are 2 items in the list, i.e. *iron* and *manganese*, with *manganese* being negated. In terms of annotation, the list should be treated as though both items are positive, and annotation of the list should proceed according to guideline 2) above. So, the chunks *iron* and *managene* are both annotated as a discontinuous span.

4) **Only the chunk(s) corresponding to the entity itself, and not any additional information, should be annotated**

Entities are frequently accompanied by extra information or descriptions of some kind. However, only the chunk(s) corresponding to the entity or entities themselves should be marked. The examples s) to u) help to clarify this.

s) *[NP **a transcriptional repressor**] [PP of] [NP Soda]* ***…***
t) *[NP **Strains**] [VP carrying] [NP a mutation] [PP in] [NP the crp structural gene] …*
u) *[NP **The uxuA-uxuB operon**] [PP of] [NP the glucuronate pathway]…*

In s), the entity itself is the transcriptional repressor and so this is what should be marked. The chunks following it show what the repressor is acting upon, i.e. *Soda*. This is extra information about the repressor and so should not be marked. In t), the entity itself is just *strains*. The remaining chunks give more specific information about *which* strains are being discussed. In u), the entity itself is *the uxuA-uxuB operon*, whilst the remaining chunks indicate that this operon is part of the glucuronate pathway.

Entities may also be preceded by quantifications (e.g. *some of*, *many of* etc), as shown in v). These are also considered as extra information and should be excluded from the variable-denoting phrase.

v) *[NP some] [PP of] [NP **the novel CsgD-regulated genes**]….*

In some cases, when an entity name or general type is not explicitly mentioned, it may be necessary for the variable-denoting phrase to cover several chunks in order to correctly characterise the entity. In w), for example, the THEME of *affects* is not *the arcA modulon* as the THEME of *affects*, but rather *members* of this modulon.

w) *It is possible that Fnr also indirectly [VP affects] [NP some] [PP of] [NP **the other members**] [PP **of**] [NP **the arcA modulon**].*

Event phrases

A variable involved in an event may correspond to a further event or process. This may be represented using either a verb or nominalised verb, and thus may occur in either an NP or a VP chunk.

1) ***Only* the chunk that contains the verb or nominalised verb should be marked as the variable-denoting phrase**.

Chunks that follow the one that contains the verb or nominalised verb may correspond to variables involved in the "embedded" event or process, but these should not be included within the marked variable-denoting phrase. The following examples help to illustrate this.

a) *[VP **assaying**] the fused lacZ gene product*
b) *[NP **binding**] of the cyclic AMP (cAMP)-cAMP receptor protein (CRP) complex to a CRP binding site*
c) *[NP **The introduction**] of a cysB allele*

Semantic roles

Each variable-denoting phrase that contributes to the description of a particular event should be assigned a semantic role. The role labels proposed are general enough to apply to a wide range of variables in different events.

HOWEVER:

In certain cases, it may be that none of the 12 roles seem suitable to characterise a particular event variable. If this is the case, then a 13th role, called UNDERSPECIFIED, may be assigned to an argument. Whenever this role is assigned, it **must** be accompanied by a comment which characterises the role being played by the argument. This will allow us to determine whether further roles must be added to our scheme. *Please also inform us if you encounter such variables.*

Description of semantic roles

Below are descriptions of our proposed semantic roles. For each role, the following information is provided:

- a general characterisation of the role
- types of arguments that can fill the role (e.g. entities, events, etc.).
- typical clues in the sentence structure or context, e.g. position with respect to the verb or common preceding prepositions. It is important to note that such contexts are only *clues*. Variable-denoting phrases can also occur in other contexts than those detailed, meaning that it is always important to consider the general meaning that each variable contributes towards the event before assigning a role
- a number of illustrative examples. In each example, the verb or nominalised verb of interest is shown in italics, whilst the text span that corresponds to the semantic role being discussed is shown in bold type. The chunking of these text spans is also shown.

Before beginning the description of the semantic roles, there are a couple of important points that should be noted.

1) **In certain circumstances, it is possible for a particular semantic role to be assigned to *more than one variable* in an event.**

Consider the following example, where we focus on the nominalised verb *ions:*

*DNase I footprinting assays were used to study the **interactions** of these regulatory proteins with the tsx-p2 promoter region*

There are 2 variables associated with the *interactions* event, i.e. *these regulatory proteins* and *the tsx-p2 promoter region.* Both of these entities can be seen to be *responsible* for the event occurring, and so it is appropriate that they should both be labelled with the AGENT role.

Depending on the meaning of the verb/nominalised verb, it is sometimes possible for 2 *separate* event variables to occur in the form of a *list.* An example is shown below, where we concentrate on the nominalised verb *combinations*:

*Complementation was carried out with **combinations** of a host strain and a plasmid.*

The nominalised verb *combinations* is followed by a list of 2 items i.e. *a host strain and a plasmid*. The decision to be made is whether this is a list of items that constitute a *single* event variable, or whether the items in the list represent *separate* event variables.

A simple test to determine this is whether items can be removed without changing the overall meaning of the event (i.e. the event would still make sense). If this is the case, then it is likely that the list corresponds to a single event variable.

*A transducing lambda phage **carrying** glpD''lacZ, glpR and malT was isolated …*

In the above example, there are 3 things being carried, i.e. *glpD''lacZ, glpR and malT*. However, if one of more of these items is removed from the list, the event still makes sense, and the overall meaning of the event remains the same (i.e. the lambda phage is carrying *something*). In this case, therefore, the list as a whole corresponds to a *single* event variable.

However, if removing an item from the list changes the meaning of the event, or means that it no longer makes sense, then it is likely that the items in the list correspond to *separate* variables of the event. This is the case for the *combinations* example above. The meaning of this nominalised verb is such that we expect there to be two or more things that are combined together. Hence, if one of the items is removed from the list, then the event no longer makes sense. We can thus conclude that in the above example *a host strain* and *a plasmid* correspond to 2 separate variables of the event. In this case, they are both THEMEs.

2) **The *same* phrase can be annotated as a variable in multiple events,** if it sems that a variable belongs to more than one event in a sentence. The following sentence serves to illustrate this:

The LysR-type transcriptional regulator CysB controls the repression of hslJ transcription in Escherichia coli.

In this sentence, it is not easy to determine to which of the three events described (i.e. *control*, *repression* or *transcription*) the location *in Escherichia coli* applies. Reading surrounding sentences may help to make it clearer which event the location is linked to, but otherwise it is permissible for the phrase to be marked as the LOCATION of more than one event in the sentence.

3) **Care should be taken to determine whether or not a particular phrase constitutes an event variable.**

As a general rule, each annotated variable should contribute a different type of information towards the description of the event. An exception is where 2 or more variables share the same semantic role, as described above. The usual types of information that variables can contribute (i.e. their semantic roles) are described within this section.

An important distinction to make here is between those phrases that actually correspond to event variables, as those that are simply providing extra descriptive information about a variable. In normal circumstances, such descriptive phrases SHOULD NOT be annotated as event variables. An example is shown below:

Expression of narL *requires* the fnr gene product, a pleiotropic activator.

Here, the event denoted by *requires* has two variables, i.e. an AGENT (the nominalised verb *expression*) and a THEME (*the fnr gene product*). There is additionally additional descriptive information about the THEME, i.e. it is *a pleiotropic activator*. However, descriptive information does not constitute a separate *type* of information relating to the *requires* event. Hence, it is NOT marked as a variable.

An exception to the above rule is when the *event itself* concerns the provision of descriptive information about one of the other event variables. In this case, the variable providing the descriptive information will be assigned either the DESCRIPTIVE-AGENT or DESCRIPTIVE-THEME role (see section 9.1.11). Some examples are shown below:

YjfQ *acts* as a repressor

Here, the purpose of the event is to provide descriptive information about *YjfQ*. Therefore, *a repressor* SHOULD be annotated as a separate variable (in this case, DESCRIPTIVE-AGENT).

This region *contains* a m7G.

Again, the purpose of the *contains* event is to provide descriptive information about *this region.* As such, both the subject and object should be annotated as event variables. In this case, *a m7G* is assigned the DESCRIPTIVE-THEME role.

Where descriptive information SHOULD be annotated as a separate variable, it is normally the case that it is marked with certain prepositions, or it occurs as the object of a verb. More information is provided in section 9.1.11. Other possible cases where descriptive information can be confused with other semantic roles are described in the sections below.

There now follows descriptions of the 12 semantic role types that have been defined for this task.

AGENT

Below are some general features of variables that correspond to the AGENT semantic role:

- They are ***core* variables**, in that they are very often present, or at least implied, in the description of events.
- They are **responsible for an event occurring**, in that it instigate, drive or triggers the event.
- It follows that **the AGENT role should only be assigned when the event denotes an action of some kind**
- AGENTS are **typically** either **an entity** (see (a)) or **a further event** (see (b))
- Most **typically**, they occur as **the subject of the verb representing the event** (see (a) and (b))
- In any case, they **normally occur in close proximity to the verb or nominalised verb that represents the event**.

(a) [NP **The narL gene product**] *activates* the nitrate reductase operon
(b) [NP **Phosphorylation**] of OmpR by the osmosensor EnvZ *modulates* expression of the ompF and ompC genes in Escherichia coli

In (b), the marked AGENT phrase occurs further away from the verb of interest, i.e. *modulates*. This is because the intervening phrases correspond to variables involved in the *phosphorylation* event, i.e. the AGENT (*the osmosensor EnvZ*) and the THEME (*OmpR*). Only the chunk containing the word that represents the event, i.e. *phosphorylation,* should be marked as the AGENT of the *modulates* event. The variables involved in the *phosphorylation* event are identified separately.

- **Not all subjects of verbs are AGENTs**. In some cases, events do not have agents at all. This is the case for verbs that describe states rather than actions, where

nothing is actually responsible for triggering the event. An example is shown in (c).

(c) The FNR protein *resembles* CRP (the cyclic-AMP receptor protein)

The verb *resembles* is not describing an action. Rather, it is used to describe a characteristic of the FNR protein. The protein is not actually *doing* anything as part of this event, and so cannot be responsible for it occurring. Therefore, it should not be classed as an AGENT. In such events, the subject of the verb is normally classed as the THEME; more examples are provided in the next section, where the THEME role is more fully described.

In passive sentences, it is also the case that the subject will not normally be the AGENT of the event. In a passive sentence, the subject and object are "switched". The verb is in the past tense, and is preceded by a form of the verb *to be,* possibly separated by an adverb, as in (e). An example is shown in (d):

(d) The transcription of clyA was positively *controlled* by [NP **slyA**]

In (d), the subject of *controlled* is *The transcription*. However, this is the THEME of the event, as it is what is ***being*** *controlled* rather than what is ***doing*** the controlling.

- **AGENTs *can* occur in positions other than as the subjects of a verb.** One such case is illustrated in (e).

(e) The control of uvrB was found to *result* from [NP **direct repression]** by the lexA gene product

The underlying meaning of this sentence is that the direct repression by the lexA gene product *causes* the control of uvrB. Therefore, it is the repression that is driving the *result* event, and hence *direct repression* should be marked as the AGENT, even though it is the object of *results*. This emphasizes the need to carefully consider the meaning of the verb and how the variable-denoting phrases relate to it.

NOTE: Prepositions following a verb can affect its meaning, or at least the interpretation of variables in particular positions. The verb *result* is one such case.

For example:

1) X results *from* Y
2) X* results *in Y

In 1), Y is the variable that is responsible for the action, and X is the thing that results. Hence, Y is the AGENT and X is the THEME. However, in 2), the roles are swapped, so that X is the AGENT and Y is the THEME.

- **AGENTs are normally preceded by the preposition *by* in passive sentences**. An example is shown in (f).

  (f) This operon is negatively *controlled* by [NP **the uxuR regulatory gene product**].

  If AGENTs are present in such sentences, they follow the verb, preceded by the preposition *by.*

  IMPORTANT NOTE: The preposition *by* can also precede arguments playing the role of MANNER, which can occur in similar positions with respect to the verb. Care should thus be taken to distinguish between them. Further explanation is provided in the description of the MANNER role (section 9.1.3)

- **AGENTs are often omitted in passive sentences,** i.e. an agent is understood to be causing the action or event, but is not actually specified

  (g) Two types of Escherichia coli were *isolated* and *analyzed* enzymatically.

  In (g), there is no AGENT. The phrase *two types of Escherichia coli* correspond to the THEME of both *isolated* and *analysed.* In both cases, the types of Escherichia coli are the entities affected during the italicised events. Although there is an *implicit* causer of these events (most probably the authors), there is no mention of in this sentence, and hence no AGENT variable is present.

- **Nominalised verbs can also specify AGENTs**. A common way of doing this is shown in (h). The agent follows the nominalised verb, and is preceded by the preposition *by.*

  (h) *Phosphorylation* of OmpR by [NP **the osmosensor EnvZ**] modulates expression of the ompF and ompC genes in Escherichia coli.

- **An event may have more than one AGENT.** This is the case if more than one of the variables in the event can be considered to be responsible for causing the event. An example is shown in (i).

  (i) The results suggest a control circuit whereby [NP **GadW**] *interacts* with [NP **the gadA promoter**].

  Here, there are two variables in the *interacts* event, i.e. *GadW* and *the gadA promotor.* When two or more entities interact, they are normally both somehow

responsible for the interaction occurring and so in this case, both *GadW* and *gadA* should be assigned the AGENT role.

THEME

Below are some characteristics of variables that correspond to the THEME semantic role:

- They are ***core* variables**, in that they are almost always present
- They are **directly involved in events,** but are **NOT responsible for the events occurring**
- Most THEMES are **entities or further events**
- They normally occur **in close proximity to the verb or nominalised verb that represents the event**

THEMEs can be split into two basic types:

1. In **events describing some sort of action**, denoted by verbs such as *activate, transcribe,* or *induce,* THEMEs correspond to **variables that are acted upon, affected by, or resulting from the event described by the verb or nominalised verb**. In these cases, **the THEME is very often the object of the verb**. Some examples of this type of THEME are shown in (a) and (b). In (a), the THEME is an entity, whilst in (b), it is an embedded event.

(a) The narL gene product *activates* [NP **the nitrate reductase operon**]
(b) Phosphorylation of OmpR by the osmosensor EnvZ *modulates* [NP **expression**] of the ompF and ompC genes in Escherichia coli

2. In **events that describe states**, denoted by verbs such as *occupy, harbour* or *exhibit*, THEMEs correspond to **the "focus" of the event, i.e. the thing whose state is being described**. In such situations, **the THEME is normally the subject of the verb**. Examples of this type of THEME are shown in (c) - (f), where the italicised verbs describe states rather than actions. In these cases, the subjects of the verbs are marked as the THEME, as they cannot be seen to be responsible for the events occurring.

(c) In addition, [NP **the ompR-lacZ fusion**] *exhibits* a dominant OmpR- phenotype.
(d) [NP **The genes**] encoding ribosomal protein S15 (rpsO) and polynucleotide phosphorylase (pnp) *occupy* adjacent positions
(e) [NP **The recA430 protein**] *possesses* ssDNA-dependent rATP activity
(f) [NP **The PhoR1159 protein**] *lacks* the 83 and 158 N-terminal amino acids

- **In some cases, THEMEs *can* be quite far removed from the verb representing the event.** An example is shown in (g).

(g) [NP **Expression**] of the Escherichia coli torCAD operon, which encodes the trimethylamine N-oxide reductase system, is *regulated* by the presence of trimethylamine N-oxide through the action of the TorR response regulator.

In (g), the THEME of the event denoted by *regulated* is *expression*, although there are a large number of words that separate them. This is because *expression* is followed by a specification of its own theme, i.e. *the Escherichia coli torCAD operon*, after which is a description of this operon, in the clause beginning with *which.* This highlights the importance of reading the *complete* sentence before beginning annotation, in order to gain a full understanding of the event denoted by the verb, and to locate more distant event-denoting phrases.

- **THEMEs can occur in positions other than the object of the verb, even when the verb denotes an action.** An example is shown in (h).

(h) [NP **The control**] of uvrB was found to *result* from direct repression by the lexA gene product

This type of construction was introduced in section 7, where it was stated that the subject of verbs such as *found* will normally be the AGENT of the verb in the infinitive form (in this case *result*). However, the meaning of this infinitive must also be carefully considered in order to correctly assign the roles. In (h), the control of uvrB occurs *in response to* direct repression by the lexA gene product. This means that the *repression* is the AGENT and the *control* is the THEME.

- **In passive sentences, the THEME is normally the subject of the verb**, as illustrated in (i):

(i) [NP **recA protein**] was *induced* by UV radiation

- **In passive sentences, THEMEs should not be confused with AGENTs if the AGENT is omitted.** It is possible for the AGENT of an event to be omitted in passive sentences. If this is the case, care must be taken not to confuse THEMEs with AGENTs. If the verb of interest is in the past tense, and preceded by a form of the verb *to be*, then the subject is normally the THEME rather than the AGENT. An example is shown in (j).

(j) [NP **Two types**] [PP **of**] [NP **Escherichia coli**] were *isolated* and *analyzed* enzymatically.

Here, there are 2 verbs, i.e. *isolated* and *analyzed*, and the THEME of them both is *Two types of Escherichia coli.* The types of Escherichia coli were not *responsible for* the *isolating* and *analysing* events. Rather, they were *affected by* them. The two events were instigated by some unknown agent, presumably human in this case, as experimental methodology is being described.

- **Be careful of "reduced relative clauses".** In these cases, the verb is in the passive form, but this is not obvious from the surface structure of the sentence. An example is shown in (k).

(k) [NP **The region**] *required* for the activation of putP by CAP was within 234 bp upstream of the translational initiation site

The meaning of the sentence would be more explicit if it began "The region THAT WAS required …". **However, the sentence format shown in (k) requires careful attention to ensure that the correct role of THEME is assigned to *The region.*** By only looking at the structure of the sentence, *The region* looks more like an AGENT.

A further example is shown in (l):

(l) The operator region controls the production of [NP **several proteins**] *involved* in DNA repair, including protein X

The meaning here is that several proteins ARE involved in DNA repair, and hence this NP chunk corresponds to the THEME rather than the AGENT: the proteins are *not* responsible for the involvement.

- **THEMEs are also frequently specified for nominalised verbs**. The most common context in which they occur is after the nominalised verb, preceded by the preposition *of.* In (m), 2 examples of this are shown, with the nominalised verbs *phosphorylation* and *expression.*

(m) *Phosphorylation* of [NP **OmpR**] by the osmosensor EnvZ modulates *expression* of [NP the **ompF** and ompC genes] in Escherichia coli

A further example is shown in (n):

(n) A steep *rise* in the [NP ***synthesis***] of [NP polypeptide] encoded by the model template containing rare codons was demonstrated

In (n), the THEME of *demonstrated* is *A steep rise*. The thing that rose (i.e. the THEME of the *rise* event) was the *synthesis of polypeptide.* As *synthesis* is also a nominalised verb, it is just this NP chunk that gets annotated as the THEME of "rise". A variable of the *synthesis* event is also specified (i.e. *polypeptide).* This is the thing being synthesised, and hence should be annotated as the THEME of *synthesis.*

It is also possible for themes to immediately precede the nominalised verb, within the same chunk. However, as mentioned above, AGENTSs of nominalised verbs may also appear in this position, and so care must be taken that the correct semantic role is assigned. Examples involving THEMEs are shown in (o) and (p).

(o) EnvZ and OmpR act in sequential fashion to activate [NP **porin gene** *expression*].

Here, *porin gene* is the thing that is *being expressed*, i.e. the thing affected by the *expression* event, and hence it should be marked as the THEME.

(p) The release of 4.5 S RNA from polysomes is affected by antibiotics that inhibit [NP **protein** *synthesis*]

In (p), *protein* is the entity being synthesized (this could be rephrased as *synthesis of protein)* and hence protein is annotated as the THEME.

- **An event may have more than one THEME variable**, as illustrated in (q).

(q) [NP **The coding region**] of the ompF gene was *linked* with the trp promoter ([NP **trpPO**]) preceding ompF.

There are two variables specified for the *linked* event, i.e. *the coding region* and *trpPO*. Note that *trpPO* rather than *the trp promoter* is marked as the second variable because, according to the *Marking Phrases* guidelines, shorter names should be annotated when they are present. The meaning of the event is that the coding region and trpPO were linked together by some unspecified AGENT. They are thus both being affected in some way by the event and so should both be marked with the THEME role.

MANNER

Variables corresponding to the MANNER semantic role have the following characteristics:

- They describe ***the method or way*** **in which a particular event is carried out**.
- **Less central to the basic event description** than THEME or AGENT
- **Frequently occur further away from the verb or nominalised verb representing the event**
- They should **NOT be confused with the INSTRUMENT semantic role, which corresponds to *entities* used to carry out the event**.

The MANNER role can apply to **a number of different variable-denoting phrases:**

1. **Processes or methods (either biological or experimental) that are employed by the agent to bring about the event**.
   Manners of this sort have the following characteristics:
   - **Normally expressed using verbs or nominalised verbs**, (see (a) –(d)).
   - **Most often preceded by the preposition *by***, but ***via*** **and *through*** are also possible (see (a) – (c))

- In some cases, **the verb *using* can also precede MANNER phrases in the same way as prepositions**. (see (d)).
- **Typically occur *after* the verb representing the event**, as in (a) – (c).
- **May also precede the verb**, as in (d).

NOTE: Phrases corresponding to other semantic roles can precede the verb, in the same way as (d).

(a) cpxA gene *increases* the levels of csgA transcription by [NP **dephosphorylation**] of CpxR
(b) Transcription of gntT is *activated* by [NP **binding**] of the cyclic AMP (cAMP)-cAMP receptor protein (CRP) complex to a CRP binding site
(c) Structural and functional properties of this regulatory protein were *studied* through [NP **complementation analysis**] of the wild-type and five mutant ompR genes
(d) Using [NP **random Tn10 insertion mutagenesis**], we *isolated* an Escherichia coli mutant strain affected in the regulation of lysU.

Some types of nouns other than nominalised verbs can represent MANNERs, if some kind of action/technique is implied. Examples are shown in (e) – (g)

(e) CsrA *stimulates* UvrY-dependent activation of csrB expression by [NP **BarA-dependent and -independent mechanisms**].
(f) The mechanism underlying feedback inhibition of tufB expression has been *studied* by [NP **gene-dosage experiments**].
(g) Two FadR operator sites of the fadD gene were *identified* at positions -13 to -29 (OD1) and positions -99 to -115 (OD2) by [NP **DNase I footprinting**].

NOTE: An important point to note here is that the same type of phrase can be assigned different semantic roles according to the context and the meaning of the event. For example, *mechanism* is mentioned in both (e) and (f). The context and meaning of the *stimulates* event in (e) means that the phrase is assigned the MANNER role, whilst in (f), *the mechanism* is assigned the THEME role in the context of the *studied* event.

TAKE CARE:

Special care must be taken when assigning a semantic role to phrases preceded by the preposition *by*. This preposition can also be used to indicate AGENTs in passive sentences. The problem arises in passive sentences when the phrase following the preposition *by* refers to some kind of process. This is the case in (b) above. In order to make a decision on the correct role to assign, it must be considered whether a) the *binding* process is what causes the *activation* event to take place, in which case it is the AGENT, or b) the

binding process refers the way in which the *activation* event is carried out, by some unspecified agent, in which case it is the MANNER.

2. **Adverbs relating to a process that describes how the event is carried out.** Some examples are shown in (h).

   (h) These results suggest that transcription of the fadL gene [VP is **osmotically** *regulated*] by the OmpR-EnvZ two-component system

   Depending on the position of the adverb, it may occur within the VP chunk of the verb that corresponding to the event, or else in a separate adverb chunk (marked ADVP).

   NOTE: adverbs should *only* be annotated with the MANNER role if they correspond to a *process* that describes the way in which the event occurs. Adverbs may also correspond to the CONDITION role, as described in section 9.1.9. Other types of adverb, such as those that relate to judgements (e.g. *unexpectedly* or *comparatively*) should not be annotated with any type of semantic role.

3. **Certain NP chunks other than those that refer to methods or processes, e.g. NP chunks that end with the word *manner*, or a synonym,** as shown in (i) and (j).

   - These types of manner are **normally preceded by the preposition *in.***

   (i) Expression of the ompF and ompC genes, which encode the major outer membrane proteins, OmpF and OmpC, respectively, is *affected* in [NP **a reciprocal manner**] by the osmolarity of the growth medium
   (j) These results lead us to conclude that EnvZ and OmpR *act* in [NP **sequential fashion**] to activate porin gene expression; i.e., EnvZ modifies or in some way directs OmpR, which in turn acts at the appropriate porin gene promoter.

4. **Information about the direction of the event**
   - **Expressed either by adverb or NP chunks**, as shown in (k) and (l).

   (k) The gene is *transcribed* [ADVP **counterclockwise**] on the standard Escherichia coli map, as is the uxuAB operon.
   (l) The fhlA gene resides next to the hydB gene at 59 min in the E. coli chromosome, and the two genes are *transcribed* [PP **in**] [NP **opposite directions**].

5. **Fixed set of phrases of latin origin that describe the way in which experiments are carried out.**

- **These include *in vitro, in vivo, in trans* and *in sys***. Examples are shown in (m) and (n).

  (m) Furthermore, [NP **in vitro** *transcription*] of the fadL gene was strongly repressed by the addition of OmpR and EnvZ proteins.
  (n) *Introduction* [PP **in**] [NP **trans**] of a compatible plasmid carrying a wild-type uxuR gene in the lac fusion plasmid containing strain resulted in a decrease of beta-galactosidase synthesis

INSTRUMENT

Variables corresponding to the INSTRUMENT semantic role have the following characteristics:

- **They *always* correspond to *entities* that are used by the AGENT in order to carry out the event.**
- **Typically, INSTRUMENTs are preceded by prepositions or other "fixed" phrases.** The most common are ***with, with the aid of, through, using, via* or *by.*** Examples are shown in (a) – (c).
- **INSTRUMENTs should NOT be confused with the MANNER semantic role**. Like instruments, MANNERs can be thought of as describing *how* an event is carried out, but MANNERs *never* corresponds to entities.

  (a) We have isolated a strain *carrying* a fusion of the beta-galactosidase structural gene to the promoter of the uxuR regulatory gene with the aid of the Casadaban Mud ([NP **Aprlac**]) phage .
  (b) EnvZ VP *functions* through [NP **OmpR**] to control NP porin gene expression PP in NP Escherichia coli K-12 .
  (c) Using [NP **MacConkey maltose indicator plates**] we *isolated* an insertion mutation

- **Where the event is denoted using a nominalised verb, it is also possible for the INSTRUMENT to precede the nominalised verb, within the same chunk**. This is the case in (d), where P1 is the entity that is used to carry out the transduction, by some unspecified agent.

  (d) [NP **P1** *transduction*] of marA::Tn5 into a Mar mutant partially restored OmpF levels.

LOCATION

There are **three types of semantic roles that specify information about locations, i.e. LOCATION, SOURCE and DESTINATION**.

- **The LOCATION role is appropriate to assign to phrases that specify where the *whole* event takes place.**
- **They almost always begin with a preposition**
- **In contrast to most other role types, prepositions that occur at the beginning of locations should be included *within* the annotated text span.** This is because prepositions play an important part in the interpretation of locations, as illustrated in the examples that follow.
- **LOCATIONs are normally entities**
- **LOCATIONs can have varying degrees of specificity according to the preposition used.**
- **LOCATIONs should NOT be confused with SOURCE and DESTINATION variables. Such variables can also be considered as locations, but correspond to *start/end* points of events, rather than where the *whole* event takes place.**

Specific locations

Locations specified using ***in, on* and *at* are quite specific locations**; they are the actual places in which the event took place. Examples are shown in (a), (b) and (c).

(a) The Escherichia coli Ada protein *activates* sigma(70)-dependent transcription [PP **at**] three different promoters ([NP **ada**], aidB, and alkA) in response to alkylation damage of DNA.
(b) Phosphorylation of OmpR by the osmosensor EnvZ *modulates* expression of the ompF and ompC genes [PP **in**] [NP **Escherichia coli**].
(c) These fusions were *formed* [PP **on**] [NP **plasmid cloning vectors**].

NOTE: In (a), the preposition *at* does not directly precede the highlighted location. Rather, it precedes *three different promoters.* This is a general characterisation of the location, but a list of more specific entities follows, and we annotate the first of these, according to the *Marking Phrases* guidelines.

Vague locations

The prepositions *near* and *between* specify more vague locations.

In (d), the entity that following the preposition *near* is not the actual location where the event took place. Rather, it is the specification of some entity (in this case rpsL) that is in the *vicinity* of the actual location of the gene. When placed together, the preposition and the entity specify a location, but this is a more vague location than the ones specified in (a), (b) and (c).

(d) The fic gene was *located* [PP **near**] [NP **rpsL**] (formerly strA) on the E. coli K-12 map

In the case of *between*, there are normally two entities that follow. As with *near*, neither of these entities specify the exact location of the mutant. Rather, it is located somewhere in space bounded by these two entities. **The text span covered by locations specified with *between* should cover both entities that specify the bounding points of this location.** An example is shown in (e).

(e) The mutant (alc-24) was *located* [PP **between**] [NP **srl and recA200**] and caused synthesis of high levels of recA protein in both lexA+ and lexA3 strains.

Vague and specific locations

In some cases, 2 locations are specified, i.e. a vague one and a more specific one. **In this case, *both* locations should be annotated as single span assigned the LOCATION semantic role.**

For example, locations on a chromosome may be specified vaguely as being *near* to some other entity, as well as more specifically as the number of minutes on the chromosome. An example is shown in (f).

(f) The gene for ribosomal protein L13, rplM, is *located* [PP **near**] [NP **argR**], [PP **at**] [NP **70 minutes**] on the Escherichia coli chromosomal linkage map.

In this case, *near argR* and *at 70 minutes* should be annotated as a single span, having the LOCATION semantic role.

- **IMPORTANT NOTE: Entities corresponding to locations should normally be assigned a *concept type.*** However, the concept should be assigned ***only* to the entity itself and not to the complete LOCATION span.** For example, the span *in E. Coli* corresponds to the LOCATION variable, but **only the chunk *E. Coli* should be assigned a concept type**

SOURCE

Biological events frequently involve a movement or shift from one location to another. The start and/ or end points are locations, but are distinct from the types of location that should be assigned the LOCATION semantic role.

- **The SOURCE role corresponds to phrases that specify where the event *begins.***
- **SOURCE variables normally correspond to entities**

- **They are locations, but should not be confused with the LOCATION role,** which corresponds to where the *whole* event takes place.
- **They are normally preceded by the preposition *from***
- **Unlike LOCATIONs, the preceding preposition SHOULD NOT be included in the annotated span of the variable**

An example of SOURCE role is shown below in (a).

(a) *Transduction* of the marA region from [NP **a Mar strain**], but not a wild-type strain, led to loss of OmpF.

Here, we are focussing on the nominalised verb *transduction.* The THEME of this event, i.e. what is being transduced, is the *marA region*. The marked NP chunk following *from* specifies where the *transduction* event began, i.e. *a Mar strain.* This phrase does not describe where the whole of the *transduction* event took place, and so it is correct to label it as SOURCE rather than LOCATION. A further example is shown in (b), where *a strain* is the start point of the *isolation* event.

(b) To determine the expression of BMI1, a BMI1-LacZ construct was *extracted* from [NP **pBR322 plasmid**] and inserted into E.coli chromosomal DNA.
(c) A transducing lambda phage carrying glpD"lacZ, glpR, and malT was *isolated* from [NP **a strain**] harboring a glpD"lacZ fusion

- **The SOURCE role can also apply to more abstract types of phrases, particularly those with a more "psychological" nature.** An example is shown in (c).

(d) The transcriptional direction of the uxuR gene was *deduced* from [NP **the restriction pattern**] and the phenotypic properties of the new plasmids.

In (c), *the transcriptional direction* is the THEME of the *deduced* event, whilst *the restriction pattern* can be seen as the SOURCE, in an abstract way, as it is a sort of "starting point" of the deduction. Note that *the restriction pattern* is in a list with *the phenotypic properties of the new plasmids* but, according to the *Marking Phrases* guidelines, only the first item in the list is marked as the variable-denoting phrase.

TAKE CARE:

Not every phrase preceded by the preposition *from* constitutes a SOURCE variable.

Consider the example (d)

(e) That the two divergent transcripts from **the identified promoters** *represent* the kdtA and rfaQ transcripts was confirmed

If we consider the event denoted by the verb *represent*, **the phrase *the identified promoters* DOES NOT constitute a SOURCE variable for this event.**

Firstly, the verb *represent* corresponds more to a state rather than an *action* event; only the latter type of event can have a SOURCE.

Secondly, the phrase *from the identified promotors* is not actually a separate variable of the *represent* event at all; it merely provides additional information about the THEME of the event, i.e. *the two divergent transcripts.*

DESTINATION

This is the "companion" role to SOURCE. Variables assigned to this role have the following general characteristics:

- **The DESTINATION role corresponds phrases that specify to the *end point* of an event**
- **They are normally entities**
- **They are locations, but should not be confused with the LOCATION role,** which corresponds to where the *whole* event takes place.
- **They are typically preceded by the prepositions *to or into***
- **Unlike LOCATIONs, the preceding preposition SHOULD NOT be included in the annotated span of the variable**

Some examples are shown in examples (a) – (d).

(a) Transcription of gntT is activated by *binding* of the cyclic AMP (cAMP)-cAMP receptor protein (CRP) complex to [NP **a CRP binding site**]
(b) The *introduction* of a cysB allele, either on a plasmid or on an episome to [NP **the fusion strains**], resulted in the decrease of beta-galactosidase activity
(c) The repression is initiated by autophosphorylation of the sensor protein ArcB, followed by phosphoryl group *transfer* to [NP **the regulator ArcA**]
(d) P1 *transduction* of marA::Tn5 into [NP **a Mar mutant**] partially restored OmpF levels.

TEMPORAL

The TEMPORAL semantic role should be assigned to phrases with the following characteristics:

- **They situate the event in time**
- **They situate the event with respect to another event.**
- **They often begin with prepositions that indicate time or ordering of events, such as *during, following, before, after* or *at***

- **The preceding preposition (PP chunk) SHOULD BE INCLUDED within the annotated span of the variable,** as it is important to the interpretation of the phrase.

There are several types of temporal expression:

a) **Specification of the duration of an event**, as shown in (a).

(a) Analyses to quantitate the induction of this system show that derepression of the operon is first detectable 5 min after UV exposure, with the rate of synthesis *increasing* to four to six times the uninduced rate [PP **during**] [NP **the subsequent 30 min**].

2) **Situation of the event in time with respect to another event**. Some examples are shown in (b), (c) and (d).

(b) The Alp protease activity is *detected* in cells [PP **after**] [NP **introduction**] of plasmids carrying the alpA gene, which encodes an open reading frame of 70 amino acids.
(c) PhoB is known to be a transcriptional activator of the Pho regulon, expression of which is *activated* [PP **during**] [NP **phosphate starvation**].
(d) E . coli NM81 transformed with pJB22 had enhanced membrane Na+/H+ antiporter activity that was cold labile and that *decreased* very rapidly [PP **following**] [NP **isolation**] of everted vesicles.

3) **Specification that 2 or more things happen in parallel**, as illustrated in (e).

(e) Complementation of such a mutant with the cloned fragments *reversed* both phenotypes [PP **at**] [NP **the same time**]**.**

If the temporal situation of an event with respect to another event also specifies a more precise timing, this timing should also be included within the annotated span. Two examples of this are shown in (f).

(f) Upon return to permissive temperature (30 degrees C), the transcripts *reappeared* coordinately [NP **about 15 min**] [PP **after**] [NP **the first synchronized initiation**] and then *declined* sharply again [NP **10 min**] [ADVP **later**].

Firstly, the *reappeared* event happens after the *initiation* event. There is also a specification of the amount of time that elapsed between these two events, i.e. 15 min. Hence, the complete phrase *about 15 min after the first synchronized initiation* is annotated as the TEMPORAL phrase associated with the *reappaeared* event. Secondly, the *declined* event occurs 10 minutes after the *reappeared* event, and so *10 min later* is marked as a TEMPORAL phrase associated with the *declined* event.

CONDITION

The CONDITION semantic role is appropriate for:

- **Phrases describing the environmental conditions which must hold in order for the event to take place**.

Environmental conditions can take a number of forms. **Three of the most common types are the following:**

1. **Changes in conditions that trigger the event.**
   - **Frequently preceded by the phrase *in response to,* but this should be excluded from annotated text span.** Examples are shown in (a) and (b)

   (a) Strains carrying a mutation in the crp structural gene fail to *repress* ODC and ADC activities in response to [NP **increased cAMP**] obtained by carbon source manipulation or cAMP supplementation of the growth medium
   (b) The Escherichia coli Ada protein *activates* sigma(70)-dependent transcription at three different promoters (ada, aidB, and alkA) in response to [NP **alkylation damage**] of DNA.

2. **Presence or absence of particular substances in the environment.**
   - **Frequently preceded by the phrases *in the presence of* or *in the absence of*, but these should be excluded from the text span that is marked to represent the variable.** Examples are shown in (c), (d) and (e).

   (c) The dcuB gene of Escherichia coli encodes an anaerobic C4-dicarboxylate transporter that is induced anaerobically by FNR, activated by the cyclic AMP receptor protein, and *repressed* in the presence of [NP **nitrate**] by NarL
   (d) A chromosomal deletion of gcvA resulted in the inability of cells to *activate* the expression of a gcvT-lacZ gene fusion when grown in the presence of [NP **glycine**] and an inability to *repress* gcvT-lacZ expression when grown in the presence of [NP **inosine**].
   (e) Here we show that OmpR, under certain conditions, could activate porin expression in the complete absence of [NP **EnvZ**].

3. **Characterisations of the conditions under which the event takes place. These may take the form of an adverb, but also often take the form of *Under x condtions,* where *x* characterizes the conditions in which the event takes place. In this case, *under* should be omitted from the annotated text span.** An example of a condition expressed by an adverb is shown in (f), whilst a condition in the form of *under x conditions* is shown in (g).

(f) Expression of sdhCDAB (encoding succinate dehydrogenase) and lctD (encoding the flavin-linked L-lactate dehydrogenase) is *elevated* [ADVP **aerobically**] and *repressed* [ADVP **anaerobically**] in Escherichia coli
(g) [PP Under] [NP **anaerobic conditions**] the narL gene product, in the presence of [NP **nitrate**], is known to *activate* transcription of the narC operon.

In (g), there are actually 2 conditions specified: there is *nitrate* in addition to *anaerobic conditions.* In such case, the different conditions should be treated in the same way as lists. That is to say, both conditions should be assigned concept types and annotated as a single, discontinuous annotation marked with the CONDITION role. RATE

RATE

The RATE semantic role corresponds to phrases that have the following characteristics:

- **They describe changes in rates or levels that occur as part of the event.**
- **They normally have one of the following formats: *n-fold, n times* or *n %*.**
- **In most cases, the change described by a RATE variable will apply to the THEME of the event.**
- **RATE variables are often preceded by prepositions, but this SHOULD NOT be included within the annotated span**
- **Rate *changes* are often preceded by the preposition *by*.** An example is shown in (a).

In (a), the rate change applies to the THEME of the *elevated* event.

(a) marR mutations that elevate marRAB transcription and engender multiple antibiotic resistance *elevated* inaA expression by [NP **10-**] [PP **to**] [ADVP **20-fold**] over that of the wild-type.

- **RATE variables may also correspond to the level *to which* one of the other variables has been increased or decreased during the event. In this case, the preposition *to* typically precedes the variable.** An example is shown in (b).

(b) Analyses to quantitate the induction of this system show that derepression of the operon is first detectable 5 min after UV exposure, with the rate of synthesis *increasing* to [ADVP **four to**] [NP **six times the uninduced rate**] during the subsequent 30 min

- **In other cases, RATE variable phrases can stand alone, without any preceding preposition.** This is illustrated in (c) and (d).

(c) Furthermore, in a delta envZ strain of E. coli, containing the envZ Val-243 plasmid, ompC expression is *elevated* [ADVP **7-fold**] relative to that found in cells carrying the wild-type envZ plasmid.

- **RATE variables can also apply to nominalised verbs. In these cases, the amount should *only* be identified as a separate variable-denoting phrase if a specific rate of change is specified**. If the change is less precise, e.g. if *10-fold* in (d) was replaced by *small,* then *small* should not be separately identified a variable-denoting phrase.

(d) Overexpression of the sfs1 gene in MK2001 resulted in a [NP **10-fold** *increase*] of amylomaltase.

IMPORTANT NOTE: Not all phrases of the forms *n-fold, n times* or *n %* should be marked as RATE variable, e.g. if they merely express a quantity of another variable. RATE variables normally only occur with verbs or nominalised verbs that imply some sort of *change* in rate or level e.g. "increase", "decrease" etc. In (f), although the emboldened phrases express percentages, they DO NOT correspond to RATE variables. This is because they are expressing *quantities* of the THEMEs of the *expressed* events. They are *not* describing rate or level changes that occur as part of the *expressed* events.

(e) Mar mutants of an ompF-lacZ operon fusion strain *expressed* **50 to 75%** of the beta-galactosidase activity of the isogenic non-Mar parental strain, while Mar mutants of a protein fusion strain *expressed* **less than 10%** of the enzyme activity in the non-Mar strain.

However, it should be noted that percentages *can* act as RATE variables in other types of sentence. For example, in a sentence of the form *X increased Y by 10%*, the RATE of the *increased* event would be *10%*.

Care should also be taken that the RATE variable is associated with the *correct* event, if there is more than one event in the sentence. An example is shown in (g):

(f) Induction at 42 degrees C led to **100-fold** *overproduction* of EIImtl.

In (g), there are 2 events, one denoted by the verb *led* and the other denoted by the nominalised verb *overproduction.* The RATE variable *100-fold* belongs to the *overproduction* event, rather than the *led* event.

DESCRIPTIVE

Variables of the DESCRIPTIVE category can be best characterized as follows:

- **They describe *characteristics or behaviour* of one of the other variables in the event.**
- **Normally apply to either the AGENT or the THEME of the event. We thus distinguish two separate sub-roles, i.e. DESCRIPTIVE-AGENT and DESCRIPTIVE-THEME.**

There are 2 main contexts in which the DESCRIPTIVE role should be assigned

1) **Descriptions of characteristics or behaviour that normally follow the preposition *as***. **Such descriptions can apply either to the AGENT or THEME of the event.** Examples are shown in (a) and (b).

In (a), the descriptive phrase (*a formate-dependent regulator*) refers to the AGENT of the verb *acts* (i.e. HyfR), hence the role assigned should be DESCRIPTIVE-AGENT.

(a) It is likely that HyfR *acts* as [NP **a formate-dependent regulator**] of the hyf operon

Another type of sentence where the use of the DESCRIPTIVE-AGENT role is appropriate is shown in b):

(b) Mucous cells *participate* in [NP **the interaction**] with enteropathogens

In (b), *Mucous cells* is the AGENT and *the interaction* is the DESCRIPTIVE-AGENT, as it is providing descriptive information about what the AGENT is doing.

In example (c), the phrase *a revertant* is describing a characteristic of *the recA gene*, which is the THEME of the *isolated* event. Therefore, the phrase *a revertant* is annotated as the DESCRPTIVE-THEME.

(c) A mutant strain of E. coli displaying altered regulation of the recA gene was *isolated* as [NP **a revertant**] of a lexA3 recA200 double mutant which showed improved DNA repair and recombination functions.

A further type of sentence where the behaviour of the THEME is being described is illustrated in (d).

(d) Uridine is *involved* in [NP **the recognition**] of tRNA substances.

Here, *uridine* is the THEME of *involved*: it is not *doing* the *involving*, but rather it *is involved.* The rest of the sentence described *what* the theme is involved in, i.e. the recognition of tRNA substances. This can be seen as information about behaviour, and hence it is appropriate to assign the role DESCRIPTIVE-THEME to the chunk *the recognition.*

2) **Descriptions in events that correspond to states, rather than actions**. Such events have the following characteristics:

- **There is no AGENT**
- **The subject of the verb corresponds to the THEME of the event**
- **The DESCRIPTIVE-THEME is assigned to variables that correspond to characteristics or attributes of the THEME.**
- **The DESCRIPTIVE-THEME variable is normally the object of the verb**

Examples of such sentences are shown in (e) and (f). In both cases, the emboldened phrase corresponds to the DESCRIPTIVE-THEME.

(e) In addition, the ompR-lacZ fusion *exhibits* [NP **a dominant OmpR-phenotype**].
(f) An Escherichia coli genomic library *composed* of [NP **large DNA fragments**] (10-15 kb) was constructed using the plasmid pBR322 as vector.

The *meaning* of certain verbs/nominalised verbs (such as those in (a) – (d) above) is such that the event itself is focused on providing a description of the AGENT or THEME. In this case the descriptive phrases are treated as actual variables of the event; the descriptive phrases are *required* if the event is to make sense (or at least the meaning of the event would be different if they are not present).

In other cases, the event itself is **NOT** focussed on providing descriptive information about the AGENT or THEME, but it is still possible to include extra descriptive information within the sentence, as shown in (e), (f) and (g). **NONE** of the **emboldened** phrases correspond to DESCRIPTIVE-AGENT or DESCRIPTIVE-THEME

(g) The global regulator CsrA, **an RNA binding protein**, *coordinates* central carbon metabolism.

The event described by the verb *coordinates* has an AGENT (*The global regulator CsrA*) and a THEME (*central carbon metabolism*). **There is additionally a descriptive phrase relating to the AGENT (i.e. *an RNA binding protein*). However, this phrase should NOT be annotated as DESCRIPTIVE-AGENT, as it does not constitute a separate piece of information about the *coordinates* event.** It merely provides extra information about the AGENT, and there is no difference in the meaning of the even if this descriptive phrase is omitted.

A further example is shown in (h)

(h) These promoters *generated* transcripts **with 5' ends separated by 289 bases**

Here, *These promoters* is the AGENT of *mediated,* whilst *transcripts* is the THEME. **The phrase** *with 5' ends separated by 289 bases* **provides descriptive information about the THEME, but does NOT contribute new information about the description of the event. Hence, it should NOT be annotated as the DESCRIPTIVE-THEME.**

Descriptive information in brackets also does NOT constitute a separate variable of the event, e.g.

(i) The FIS protein (**factor for inversion stimulation**) is known to *activate* the transcription of rRNA and tRNA operons in Escherichia coli

In (i), *The FIS protein* is the AGENT of *activate.* The information in brackets (factor for inversion stimulation) simply explains the FIS acronym, but does not constitute a separate event variable.

PURPOSE

The semantic role PURPOSE is appropriate to assign to variables that have the following characteristics:

- **Variables that specify *why* the event occurred, i.e. specifications of some sort of aim, purpose, goal or reason for the event occurring.**
- **The PURPOSE role always corresponds to an event of some kind, using either a verb, (see (a) and (c)), or a nominalised verb, (see (b)).**
- **Verbs that correspond to the PURPOSE role are normally in infinitive form (i.e. preceded by *to*)**
- **Nominalised verbs that correspond to purposes are often preceded by the preposition *for*.**

In (a), some unspecified (human) agent is using the fusion strains, and the *purpose* or *reason* for using them is to study the regulation of the cysB gene.

(a) The fusion strains were *used* [VP **to study**] the regulation of the cysB gene by assaying the fused lacZ gene product

In (b), the focus is the verb *required.* Some unspecified agent requires a chromosomal locus, and the reason for this requirement is to allow the alpA+ suppression to take place.

(b) We have used Tn10 and lambda placMu mutagenesis to identify a chromosomal locus, slpA, that is *required* for [NP **alpA+ suppression**] of delta lon.

Note that *suppression* is a nominalised verb and so should subsequently be annotated with its own variables.

In (c), the *purpose* of isolating the fragment was to complete the sequence of the cadA homolog.

> (c) A 6.0-kb fragment overlapping the pJB22 insert was *isolated* [VP **to complete**] the sequence of the cadA homolog

Worked example

Armed with the above guidelines and procedure for annotating event variables, we now have all the information that we need to begin annotation of biological texts. In this section, we study an abstract and discuss in detail the annotations that should be added to it. Below is the complete abstract that we are going to consider. The 8 sentences in the abstract are displayed separately, and the verbs that are to be annotated are underlined.
.

We have isolated a strain carrying a fusion of the beta-galactosidase structural gene to the promoter of the uxuR regulatory gene with the aid of the Casadaban Mud (Aprlac) phage.

Analysis of mutants with deletions that were derived from the uxuR::Mud1 insertion strain confirmed the counterclockwise transcription direction of the uxuR gene.

The uxuR-lacZ fusion strain was also used to examine the regulation of expression from the uxuR promoter.

It was observed that an increase in the copy number of the uxuR gene results in an increased repression of beta-galactosidase synthesis.

Overproduction of the exuR repressor also caused a decrease of the beta-galactosidase level.

In all cases, the repression of beta-galactosidase synthesis was accompanied by a stronger repression of uxuB gene product synthesis.

These results indicate that the expression of the uxuR gene is repressed by its own product but also by the exuR repressor.

The different types of regulation of the two uxu operons are thus identical.

It can be seen that 6 out of the 8 sentences contain verbs to be annotated. For each of these 6 sentences, we show below the automatically identified chunks. Lists of event variables for each verb are then displayed and discussed.

Sentence 1

Below is a representation of the automatically identified chunks of the first sentence. The verbs whose variables are to be annotated are underlined.

[NP We] [VP have <u>isolated</u>] [NP a strain] [VP <u>carrying</u>] [NP a fusion] [PP of] [NP the beta-galactosidase structural gene] [PP to] [NP the promoter] [PP of] [NP the uxuR regulatory gene] [PP with] [NP the aid] [PP of] [NP the Casadaban Mud] (Aprlac) [NP phage].

We consider annotation according to the recommended annotation procedure that is described in Appendix 1. In this procedure, step 1) is to locate the first or next verb to be annotated, which in this case is *isolated.* As this is the first verb to be annotated in the sentence, we proceed to step 2) of the annotation procedure, in which the sentence is carefully read through to fully understand it. As part of this step, we can verify that *isolated* denotes the main event described by the sentence, and so it should be annotated before the other underlined verb, i.e. *carrying.*

Moving on to step 3) of the annotation procedure, the sentence should be read through again, this time concentrating on locating the variables involved in the event denoted by the verb *isolated.* Once this has been done, we can move on to step 4) of the annotation procedure, in which the annotation of event variables is carried out. Appropriate text spans for each variable are marked, and then assigned appropriate semantic roles. The outcome of this step of the annotation process is shown below.

isolated
AGENT: we
THEME: a strain
INSTRUMENT: Aprlac

The AGENT and THEME of this event occupy their typical positions with respect to the verb, i.e. the subject and object of the verb, respectively. The authors (represented by *we*) are performing the isolation and so correspond to the AGENT. The thing that they are isolating is *a strain*; hence this is the THEME. The chunks that follow provide further information about this strain, i.e. what it is carrying. According the *marking phrases* guidelines for entities, such information should not be included within the marked variable phrases; only the chunk(s) corresponding to the entity itself should be marked.

The third variable associated with this event appears at the end of the sentence, following the secondary event that describes what the strain is carrying. This highlights the importance of reading and understanding the *complete* sentence to identify *all* variables

involved in the event; if only the phrases that immediately surround the verb are considered, more distantly located variables can easily be missed. This final variable is *the Casadaban Mud (Aprlac) phage.* According to the *Marking Phrases* guidelines, we just mark the short name of this, i.e. *Aprlac*. The chunks preceding the variable, i.e. *with the aid of,* help us to determine that this variable should be assigned the INSTRUMENT role: it is being used by the authors to carry out the isolation.

Having completed step 4) of the annotation process, we can move on to step 5). This requires further action to be taken if any of the event variables correspond to either entities or events. In fact, all three of the variables of *isolated* event correspond to entities, and so we need to consider whether categories from the entity hierarchy detailed in section 4 of this document can be assigned to them. The entity corresponding to the AGENT of the event, i.e. *we*, is not of biological relevance, and so a category does not need to be assigned. However, the THEME and the INSTRUMENT do correspond to biologically relevant entities, and so categories from the hierarchy should be assigned to them.

The annotation of the variables of the *isolated* event is then complete, and we can move on to the second verb of the sentence, i.e. *carrying*. As we have already carried out step 2) of the annotation process when annotating the variables of *isolated* (i.e. reading the sentence to fully understand it) we can move straight on to step 3), in which the sentence is read again to identify the variables involved in the event denoted by the second verb of the sentence to be annotated, i.e. *carrying.* The outcome of step 4) of the annotation procedure, i.e. marking the variable-denoting phrases and assigning semantic roles, is shown below.

carrying
AGENT: a strain
THEME: a fusion

The verb *carrying* denotes a secondary event in the sentence, providing additional information about the strain that has been isolated. The strain is the thing responsible for the carrying, and hence this is marked as the AGENT. The THEME is *a fusion*; this is the thing being carried. The noun *fusion* is a nominalised verb: it represents an event in which entities are fused together. Chunks that follow this specify variables that are involved in the fusion event. According to the *Marking Phrases* guidelines, only the chunk containing the nominalised verb that represents the event is marked, and hence the THEME is simply the chunk *a fusion.*

Moving on to step 5) of the annotation procedure, entities and nominalised verbs must be further considered. In the variables of *carrying,* we have one of each of these. The AGENT, i.e. *a strain*, is a biologically relevant entity, and so should be assigned a type from the entity hierarchy. The THEME, i.e. *a fusion,* contains a nominalised verb, and so we return to step 3) of the annotation procedure, this time trying to locate the variables that are involved in the event denoted by this nominalised verb, in this case *fusion.* By

reading though the sentence with the *fusion* event in mind, we can move on to step 4) and annotate two variables of this event, as follows:

fusion
THEME: the beta-galactosidase structural gene
DESINATION: the promoter

There is no specification of what caused the fusion to take place, and hence no AGENT can be identified. However, the thing that is being fused, i.e. the THEME, is *the beta-galactosidase structural gene.* This occurs in the typical position for themes of nominalised verbs, in that it follows the nominalised verb and is preceded by the preposition *of.* The second identified variable, i.e. *the promoter*, is marked with the DESTINATION semantic role. This is because it is the end point of the *fusion* event: the entity to which the beta-galactosidase structural gene was fused. The preceding preposition *to* helps to determine this semantic role. A further point to note about this phrase is that we have only marked *the promoter* rather than *the promoter of the uxuR regulatory gene.* This is because the chunk corresponding to the entity itself is just *the promoter*. The next two chunks indicate what the promoter is acting upon, which is extra information. According to the *Marking Phrase* guidelines, this should not be included within the variable-denoting phrase.

Having marked and assigned roles to these phrases, we move on to step 5). Both the THEME and the DESTINATION are entities of biological interest, and so should be assigned categories from the hierarchy of entities.

Sentence 2

Let us now move onto the second sentence. The automatic chunking is shown below, and the verbs to be annotated are underlined.

[NP Analysis] [PP of] [NP mutants] [PP with] [NP deletions] [NP that] [VP were <u>derived</u>] [PP from] [NP the uxuR] : : [NP Mud1 insertion strain] [VP <u>confirmed</u>] [NP the counterclockwise transcription direction] [PP of] [NP the uxuR gene]

Again, there are 2 verbs that we must consider, i.e. *derived* and *confirmed.* In step 2) of the annotation process, it is suggested that the verb that denotes the main event in the sentence is located and annotated before verbs that denote secondary events. In this sentence, the verb that denotes what the sentence is about, and hence the main verb, is *confirmed.* This is sequentially the second verb in the sentence to be annotated, but should be considered first, before moving back to annotate *derived.* The results of performing stage 4) of the annotation process on the verb *confirmed* are as follows:

confirmed
AGENT: Analysis

THEME: the counterclockwise transcription direction

In this sentence, the *analysis* event is responsible for confirming the counterclockwise transcription direction of the uxuR gene, and is hence the AGENT of the *confirmed* event. The chunks that follow *Analysis* are specifying further details of the analysis event, but are not of relevance whilst considering the variables of the *confirmed* event. The only other variable involved in the *confirmed* event is a THEME, i.e. *the counterclockwise transcription direction*, which is the thing being confirmed. It is not necessary to mark as part of the phrase what this direction relates to, i.e. *the uxuR gene,* as this is considered additional information about the direction.

Moving on to stage 5) of the annotation, we reconsider the AGENT and THEME that we have just identified. Firstly, the AGENT, i.e. *Analysis,* is a nominalised verb, as it describes the event of *analyzing.* Therefore, its variables must be annotated. The results of stage 4) of the annotation are shown below.

Analysis
THEME: mutants

There only variable specified for this event is the things undergoing analysis, i.e. the THEME. The actual entity that corresponds to the THEME is just *mutants,* and so this is the only chunk that should be annotated. The chunks between *mutants* and *confirmed* provide a description of these mutants, i.e. a characteristic (*with deletions*) and where they are derived from (*derived from the uxuR::Mud1 insertion strain*). These phrases are not of interest within the context of the *analysis* event as they do not denote further other variables involved in the event.

Finally, to finish off the annotation of the *confirmed* event, we consider the THEME, i.e. *the counterclockwise transcription direction.* This is an entity, and so must be assigned a type from the hierarchy of entities.

We now move back to annotate the variables of the verb *derived,* which denotes a secondary event in the sentence, providing details about the mutants. The result of marking and assigning roles to the variables of this event in stage 4) of the annotation is shown below.

derived
THEME: mutants
SOURCE: the uxuR::Mud1 insertion strain

The VP chunk *were derived* indicates a passive construction, and hence the subject, i.e. *mutants*, is the THEME of the event. There is no specification of the instigator *derive* event, and hence there is no AGENT. The second specified variable of this event, i.e. *the uxuR::Mud1 insertion strain*, is marked as with the SOURCE role, as it corresponds to the start point of the event, that is to say from where the mutants were derived.

Finally, in stage 5) of the annotation process, we determine that both the THEME and the SOURCE are biologically relevant entities, and so we assign types to them from the entity type hierarchy.

Sentence 3

We now move on to the third sentence of the abstract. The automatic chunking is shown below:

[NP The uxuR-lacZ fusion strain] [VP was also used] [VP to examine] [NP the regulation] [PP of] [NP expression] [PP from] [NP the uxuR promoter]

Only one verb of biological relevance is marked for annotation in this sentence, i.e. *used.* The results of stage 4) of the annotation process are shown below:

used
THEME: The uxuR-lacZ fusion strain
PURPOSE: to examine

The verb *used* occurs in the following VP chunk: *was also used*. The verb is in the past tense, preceded by a form of the verb *to be*, indicating a passive. Therefore the THEME is the subject, i.e. *The uxuR-lacZ fusion strain.* As there is no phrase in the sentence that is preceded by the preposition *by,* we can be sure that no agent is specified.

The sequence of chunks following the VP that contains *used* explains the *why* the fusion strains were being used, i.e. *to examine the regulation of expression from the uxuR promoter.* Such event variables correspond to the PURPOSE role. The purpose is itself an event, and so we only mark the chunk that contains the verb or nominalised verb that characterizes the event, in this case *to examine.* The remainder of the sentence is concerned with the characterization of the *examine* event, and hence *used* has no more variable-denoting phrases associated with it.

Moving on to stage 5) of the annotation of *used,* we determine that the THEME, i.e. *The uxuR-lacZ fusion strain* is an entity and assign a type from the hierarchy. The PURPOSE, i.e. *examine,* is a verb, and so the sentence should be read again to determine its variable-denoting phrases. The result is as follows:

examine
THEME: the regulation

The verb *examine* actually has only a single variable specified, i.e. its THEME, which is itself a nominalised verb, i.e. *regulation.* The remainder of the sentence further specifies information about this regulation event, but no more variables relate to the *examine* event.

As *regulation* is a nominalised verb, the thing to do is examine whether there are any variable-denoting phrases specified for it. The outcome of this is as follows:

regulation
THEME: expression

The nominalised verb *regulation* also only specifies a THEME, i.e. *expression.* This yet another nominalised verb, and the remaining chunks of the sentence specify a variable of the *expression* event. The next step is thus to mark and label this variable of the *expression* event, as follows:

expression
SOURCE: the uxuR promoter

A single variable is specified for the *expression* event, i.e. *the uxuR promoter.* The preceding preposition *from* helps us to determine that this is the SOURCE of the expression. As this SOURCE variable is an entity, it must be assigned a type from the entity type hierarchy.

Sentence 4

The chunking of the fourth sentence in the abstract is shown below:

[NP It] [VP was observed] [SBAR that] [NP an increase] [PP in] [NP the copy number] [PP of] [NP the uxuR gene] [VP results] [PP in] [NP an increased repression] [PP of] [NP beta-galactosidase synthesis]

The verb marked for annotation in this sentence is *results.* The outcome of stage 4) of the annotation process is shown below:

results
AGENT: an increase
THEME: an increased repression

The meaning behind this sentence is that some event *causes* another event to happen. The causer, and hence the AGENT, is the event represented by the nominalised verb *increase.* The chunk corresponding to the event that is affected by the *increase* event, and hence the THEME of the verb *results,* is *an increased repression*.

Both the AGENT and the THEME of the *results* event contain nominalised verbs, and so their variable-denoting phrases must be identified. For *increase*, there is only one variable specified, as follows:

increase
THEME: the copy number

The copy number corresponds to the thing being increased, and is hence the THEME of the *increase* event. Let us move on to the variables associated with the nominalised verb in the THEME of the *results* event, i.e. *repression:*

repression
THEME: beta-galactosidase synthesis

Note that, according to the meaning of the sentence, *repression* could be considered to have an AGENT, i.e. the increase in the copy number. This is because the underlying meaning of the sentence is as follows: A*n increase in the copy number of the uxuR* ***represses*** *beta-galactosidase synthesis*. However, according to section 6.4, when variables of nominalised verbs are being identified, there is a restriction that variables of the nominalised verb can only appear on the same side of the verb which denotes the "higher level" event. In the case of sentence 4, the higher level event is denoted by the verb *results,* and the nominalised verb *repression* occurs on the right hand side of the verb. Thus, variables of *repression* can only be identified from the text to the right of *results.* This means that only a THEME is present, i.e. *beta-galactosidase synthesis*.

S*ynthesis* is a nominalised verb, as it represents the event of *synthesizing.* Therefore, we must annotate the variables involved in the *synthesis* event, as follows:

synthesis
THEME: *beta-galactosidase*

There is just one variable, which is contained within the same chunk as *synthesis,* i.e. *beta-galactosidase*. This is the thing being synthesized and hence the THEME of the event. As *beta-galactosidase* is an entity, it should be assigned a type from the entity hierarchy.

Sentence 5

Let us move on to sentence 5, whose chunking is shown below.

[NP Overproduction] [PP of] [NP the exuR repressor] [ADVP also] [VP caused] [NP a decrease] [PP of] [NP the beta-galactosidase level]

The only verb to be considered here is *caused,* and the annotation of the variable-denoting phrases, together with their semantic roles, is shown below

caused
AGENT: *overproduction*
THEME: a *decrease*

The structure of this sentence is very similar to sentence 4. Again, we have one type of event, *overproduction*, causing another type of event, *a decrease.* Both of these events are expressed using nominalised verbs. *Overproduction* has a THEME, as shown below:

overproduction
THEME: the exuR repressor

This THEME, i.e. *the exuR repressor*, is an entity, and so must be assigned an entity type from the hierarchy. Let us now consider the variables involved in the THEME of the *caused* event, i.e. *decrease.*

decrease
THEME: the beta-galactosidase level

In a similar way to sentence 4, according to meaning, the AGENT of the *cause* event, could also be seen to be the AGENT of the *decrease* event. However, as *decrease* is a nominalised verb, the variables of the *decrease* event can only occur of the same side of the verb *caused* as the word *decrease*, i.e. to its right. Thus, *decrease* is annotated as having a single variable, i.e. a THEME, *the beta-galactosidase level.* As this is an entity, it must be assigned a named entity type.

Sentence 7

Sentence 7 of the abstract is the last one that needs to be annotated. Neither sentence 5 nor sentence 8 have any verbs of biological relevance. The chunking of sentence 7 is shown below:

[NP These results] [VP indicate] [SBAR that] [NP the expression] [PP of] [NP the uxuR gene] [VP is repressed] [PP by] [NP its own product] [CONJP but also] [PP by] [NP the exuR repressor]

The results of annotating this verb are shown below:

repressed
AGENT: its own product
THEME: the expression

The verb *repressed* is contained within a passive construction. The subject, and thus the THEME of this verb is *the expression*. The AGENT of the *repressed* event, preceded by the preposition *by*, is *its own product.* There is a second AGENT, i.e. *the exuR repressor*, also preceded by the preposition *by.* However, as the two agents are conjoined with *but also*, we can consider then to be in a list. According to the *Marking Phrases* guidelines, we only annotate the first item in a list, which in this case is *its own product.* This is an entity and so must be assigned an entity type.

The final step of annotating this verb is to check whether there are any variables associated with the THEME of the *repressed* event, i.e. the nominalised verb expression. There is just a THEME, as shown below.

expression
THEME: *the uxuR gene*

This THEME is an entity, and so must have an entity type assigned to it.

Appendix 1: Annotation Procedure

The main subtasks of the annotation process are as follows:

- Identifying variable-denoting phrases associated with each event
- Marking appropriate spans of text to represent these variables
- Assigning appropriate semantic roles to the variables
- Assigning categories to entities from the hierarchy

In order that the annotation of event variables is carried out as consistently and accurately as possible, it is recommended that a certain procedure or workflow for carrying out the annotation is adopted. This appendix provides a set of suggested steps to constitute this workflow.

1) **Locate the first or next verb whose variables are to be annotated**. In WordFreak, these are marked as VP-BIO chunks, and the *Prev VP-Bio* and *Next VP-Bio* buttons allow for sequential movement between these verbs.
    - If the selected verb is the first or only verb in the sentence to be annotated, move to step 2).
    - If one or more verbs in the same sentence have already been annotated, skip to step 3).

2) **If the current verb is the first or only verb to be annotated within the sentence, then the sentence should be carefully read, with the following in mind**:
    - **Is the sentence on the topic of gene regulation**? If not, repeat step 1) until a verb in a new sentence is located.
    - **If the sentence is on the topic of gene regulation, ensure that all parts of the sentence are fully understood.** If anything in the sentence is unclear, surrounding sentences should be read to help put the events described in the sentence into context.
    - **Locate the *main* verb of the sentence**. This is the verb that describes the main or most important event in the sentence.
    - **If the main verb is within a VP-BIO chunk, and is not the currently selected verb, it is suggested to select this verb and annotate its variables *before* annotating other verbs in the sentence.** This may mean that verbs in the sentence are not annotated in sequential order. However, annotating the main event before annotating secondary events is easier and more intuitive.

    REMEMBER: As long as the sentence is related to gene regulation, ALL verbs in the sentence within VP-BIO chunks should be annotated, along with any nominalised verbs that occur within the variables of these verbs.

3) **Before beginning annotation, read through the *complete* sentence again, concentrating on the event denoted by the selected verb and trying to locate the phrases in the sentence that correspond to variables in the event.** This step is important to ensure that *all* variables of the event are located, and to ensure that no misinterpretations of variables occur. Consider sentence (a).

(a) *Mutations affecting the BarA/UvrY two-component signal transduction system* ***decreased*** *csrB transcription*

If the complete sentence is not read correctly when considering the *decreased* event, it would be easy to mistakenly mark *the BarA/UvrY two-component signal transduction system* as the AGENT of the event, rather than the correct agent, i.e. *Mutations.* Section 7 of this document discusses how to identify variables in more complex sentences.

4) **Annotate each variable-denoting phrase of the event.** It is suggested that the AGENT and THEME of the event are annotated first, if they are present in the sentence, followed by phrases corresponding to other roles.
REMEMBER:
   - ***All* variables of the event *within the same sentence* should to be annotated. This includes:**
     a) Variables that don't correspond to one of the existing semantic roles (The UNDERSPECIFIED role should be assigned, together with a comment)
     b) Variables that don't correspond to a concept in the *concept hierarchy*, e.g. *we*
   - **Each variable should generally contribute a different type of information towards the description of the event. This means, for example that lists of items generally correspond to a SINGLE event variable.**

The annotation should proceed as follows:

i) **Mark an appropriate span of text to represent the variable**, according to the *Marking Phrases* guidelines in section 6.

REMEMBER:
- **Spans should normally consists of complete chunks (single chunks wherever possible)**
- **Short entity names are to be favoured over longer names or characterisations, if both are present within the sentence**
- **Descriptive information about entities should not be included within the span**
- **Where a variable consists of a list of entities, the span should consist of all items in the list, excluding commas and conjunctions etc.**

- **LOCATION and TEMPORAL spans should include the preposition that precedes them, e.g. *in, after* etc.**

ii) **Assign an appropriate semantic role to the marked phrase**, or UNDERSPECIFIED if none of the roles in the current set seems appropriate. In the case, it is important to include a comment that explains the perceived function of the phrase in the event.

- Section 7 provides a detailed description of the various roles, which should be read carefully before beginning annotation. **However, it is suggested that Appendix 2 of this document, "Quick Role Guide" is used as an aid when carrying out semantic role assignment.** It provides a tabular, quick reference guide to the semantic roles with useful reminders about typical phrase features, clues in the surrounding text etc.

5) Re-examine each of the variable phrases marked during step 4). Further action is required if the variable corresponds either to an entity or an event, as follows:

- If the variable corresponds to an entity that is a biological concept, an appropriate category should be assigned from concept hierarchy (see section 4)
- If the variable corresponds to another, embedded event (denoted by a verb or nominalised verb), annotation of the variables of this event should be carried out by returning to step 3), but this time considering the embedded event.

REMEMBER:

- **There may be more than one level of event "embedding"; in this case, the variables of events at all levels of embedding should be annotated.** Consider sentence (b):

(b) *It was observed that an increase in the copy number of the uxuR gene* ***results*** *in an increased repression of beta-galactosidase synthesis*

In this sentence, the THEME of results is *an increased repression.* As *repression* is a nominalised verb, its own variables should be identified. The THEME of *repression* is another event, i.e. *synthesis*, whose own THEME is *beta-galactosidase.*

6) **Returning to step 1), in order to consider the next verb to be annotated.** If the verb just annotated was the main verb in the sentence, it should be verified whether there are any verbs to annotate in the sentence *before* the main verb, before moving on to look at verbs after the main verb. The annotation process ends when the variables of all VP-BIO chunks in the file have been annotated.

Appendix 2 : Quick Semantic Reference Role Guide

Role Name	Description	Phrase Type(s)	Clues
AGENT	Reponsible for event; Only assigned when event denotes action	Entity or event	Typically subject of verb, Follows *by* in passive sentences
1) **The narL gene product** *activates* the nitrate reductase operon 2) This operon is negatively *controlled* by **the uxuR regulatory gene product.** 3) The control of uvrB was found to *result* from **direct repression** by the lexA gene product 4) *Phosphorylation* of OmpR by **the osmosensor EnvZ** modulates …			
THEME	Directly involved in event but not responsible for it. Either: 1)Affected by or results from "action" event; ***or*** 2)Focus of descriptions of states	Entity or event	Object of verb in "action" events, subject in descriptions of states, subject in passive sentences
1) The narL gene product *activates* **the nitrate reductase operon** 2) **recA protein** was *induced* by UV radiation 3) The release of 4.5 S RNA from polysomes is affected by antibiotics that *inhibit* **protein *synthesis*** 4) **The recA430 protein** *possesses* ssDNA-dependent rATP activity			
MANNER	***Method or way*** in which event is carried out, normally biological or experimental process. **Don't confuse with INSTRUMENT**	Event (process), adverb, direction, *in vitro, in vivo* etc.	Events typically follow *by, through, via* or *using*
1) Using **random Tn10 insertion mutagenesis**, we *isolated* an Escherichia coli mutant strain affected in the regulation of lysU 2) CsrA *stimulates* UvrY-dependent activation of csrB expression by **BarA-dependent and -independent mechanisms**. 3) These results suggest that transcription of the fadL gene is **osmotically** *regulated* by the OmpR-EnvZ two-component system 4) The gene is *transcribed* **counterclockwise** on the standard Escherichia coli map 5) These results lead us to conclude that EnvZ and OmpR *act* in **sequential fashion**			
INSTRUMENT	***Entity* used by agent** to carry out event. **Don't confuse with MANNER**	Entity	Typically follows *with, with the aid of, via, by, through, using*
1) EnvZ *functions* through **OmpR** to control porin gene expression in Escherichia coli K-12 2) We have *isolated* a strain … with the aid of **the Casadaban Mud phage** .			
LOCATION	Where the ***complete*** event takes place. **Don't confuse with SOURCE/DESTINATION**	Entity	Typically begins with *in, at, on, near* or *between*
1) Phosphorylation of OmpR by the osmosensor EnvZ *modulates* expression of the ompF and ompC genes **in Escherichia coli** 2) The fic gene was *located* **near rpsL** on the E. coli K-12 map			

3) The mutant (alc-24) was *<u>located</u>* **<u>between srl and recA200</u>**

SOURCE	Where the event ***starts*** **Don't confuse with LOCATION**	Entity	Typically follows *from*
1) To determine the expression of BMI1, a BMI1-LacZ construct was *<u>extracted</u>* from **<u>pBR322 plasmid</u>** and inserted into E.coli chromosomal DNA. 2) *<u>Transduction</u>* of the marA region from **<u>a Mar strain</u>** ... 3) The transcriptional direction of the uxuR gene was *<u>deduced</u>* from **<u>the restriction pattern</u>**			
DESTINATION	Where the event ***ends*** **Don't confuse with LOCATION**	Entity	Typically follows *to* or *into*
1) Transcription of gntT is activated by *<u>binding</u>* of the cyclic AMP (cAMP)-cAMP receptor protein (CRP) complex to **<u>a CRP binding site</u>** 2) To determine the expression of BMI1, a BMI1-LacZ construct was extracted from pBR322 plasmid and *<u>inserted</u>* into **<u>E.coli chromosomal DNA</u>**.			

TEMPORAL	Situates event in time possibly with respect to another event	Normally an event or time interval	Often preceded by *during, before* or *after*
1) The Alp protease activity is *<u>detected</u>* in cells **<u>after introduction</u>** of plasmids carrying the alpA gene 2) The rate of synthesis *<u>increased</u>* to four to six times the uninduced rate **<u>during the subsequent 30 minutes</u>** 3) Complementation of such a mutant with the cloned fragments *<u>reversed</u>* both phenotypes **<u>at the same time</u>**			
CONDITION	Conditions or changes in conditions under which the event takes place; presence or absence of substances in environment	Entity (e.g. substance present in environment), event (e.g. change in conditions) or adverb	Conditions often in the form of *under x conditions* or adverb Substances typically follow *in the presence/absence of.* Changes in conditions typically follow *in response to.*
1) Strains carrying a mutation in the crp structural gene fail to *<u>repress</u>* ODC and ADC activities in response to **<u>increased cAMP</u>** 2) The dcuB gene of E. Coli encodes an anaerobic C4-dicarboxylate transporter that is *<u>repressed</u>* in the presence of **<u>nitrate</u>** by NarL 3) Under **<u>anaerobic conditions</u>**, the narL gene product is known to *<u>activate</u>* transcription of the narC operon			
RATE	Change in rate or level occurring as part of event. Normally applies to the THEME.	Typically of the form *n-fold, n times* or *n %*	May follow *by*
1) marR mutations that elevate marRAB transcription and engender multiple antibiotic resistance *<u>elevated</u>* inaA expression by **<u>10- to 20-fold</u>** over that of the wild-type. 2) The rate of synthesis *<u>increases</u>* to **<u>four to six times the uninduced rate</u>** during the subsequent 30 minutes			

3) ompC expression is *elevated* **7-fold**

DESCRIPTIVE-AGENT	Describe characteristics or behaviour of AGENT of event	Entity or Event	Often follows *as*; object of verb in descriptions of states
1) It is likely that HyfR *acts* as **a formate-dependent regulator** of the hyf operon 2) Mucous cells *participate* in **the interaction** with enteropathogen.			
DESCRIPTIVE-THEME	Describe characteristics or behaviour of THEME of event	Entity or Event	Often follows *as;* Object in descriptions of states
1) A mutant strain of E.Coli *was isolated* as **a revertant** of lexA3 recA200 double mutant 2) Uridine is *involved* in **the recognition** of tRNA substances 3) The ompR-lacZ fusion *exhibits* **a dominant OmpR-phenotype**			
PURPOSE	Specifies *why* the event occurs, i.e. an aim, purpose, goal or reason for the event occuring	Event	Typically a verb in infinitive form or a nominalised verb following *for*.
1) The fusion strains were *used* **to study** the regulation of the cysB gene by assaying the fused lacZ gene product 2) We have used Tn10 and lambda placMu mutagenesis to identify a chromosomal locus, slpA, that is *required* for **alpA+ suppression** of delta lon.			