GREC CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of 240 MEDLINE abstracts on the subject of gene regulation. 167 of these abstracts concern the *E. coli* species, while the remaining 73 abstracts concern the *Human* species.

1.2 Representation of the corpora (flat files, database, markup)

The corpus is available in two formats

- standoff annotation
- XML-encoded annotation

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Sophia Ananiadou
Address: Manchester Interdisciplinary Biocentre,131 Princess Street, Manchester M1 7DN, IK
Affiliation: National Centre for Text Mining, School of Computer Science, University of Manchester
Position: Director
Telephone: +44 161 306 3092
Fax: +44 161 306 5201
e-mail: Sophia.Ananiadou@manchester.ac.uk

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource is available on the MetaShare platform as an archive.

2.3 Copyright statement and information on IPR

The resource is freely available for research purposes.

3 TECHNICAL INFORMATION

3.1 Directories and files

The corpus is available in 2 different formats, contained within 2 directories:

- GREC_Standoff - contains plain text files (.txt) containing abstract texts and associated standoff annotations files (.a1 and .a2). Split into two sub-directories:
  - Ecoli – cooresponds to *E. coli* abstracts
  - Human – corresponds to Human abstracts
- GREC_XML – contains the annotated abstracts in XML format. There are three subdirectories:
  - GRECResources – contains the DTD file to which the XML files conform
  - Ecoli – contains the annotated *E. coli* abstracts in XML format
  - Human – contains the Human abstracts in XML format

3.2 Data structure of an entry

This is not relevant as the corpus is a set of text files.

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

The corpus contains approximately 52,000 tokens. The standoff version of the corpus requires approximately 3 MB on disk, while the XML version of the corpus requires approximately 2.3 MB on disk.

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is a monolingual, annotated corpus.

4.2 The natural language(s) of the corpus

The language of the corpus is English.

4. 3 Domain(s)/register(s) of the corpus

The corpus contains abstracts of biomedical research articles.

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

The standoff version of the corpus is annotated with biomedical event structures and named entities associated with these event structures. The XML version of the corpus is additionally annotated with sentence boundaries. Further information is provided at: http://www.nactem.ac.uk/GREC/.

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

Each abstract was automatically split into sentences using the GENIA tagger (Tsuruoka et al., 2005). 6 biologist annotators then annotated the following information (see Thompson et al, 2009)

- Verbs and nominalised verbs describing gene regulation events (event triggers)
- Semantically-related arguments of these event triggers within the same sentence; each argument was assigned a semantic role from a set of 13 possible roles
- Named entities occurring within sematic arguments were annotated and assigned appropriate named entity types. The hierarchy contains around 70 NE categories, arranged with 5 supertypes,i.e. PROTEINS, NUCLEIC_ACIDS, LIVING_SYSTEMS, PROCESSES and EXPERIMENTAL. Further details, with the full set of NEs used, can be found in the annotation guidelines: http://www.nactem.ac.uk/download.php?target=GREC/Event_annotation_guidelines.pdf

As a simple example, consider the following sentence:

The narL gene product ***activates*** *the nitrate reductase operon*

The sentence contains a single event, with the trigger *activates*. There are two arguments:

- The narL gene product
- the nitrate reductase operon

The argument *The narL gene product* is assigned the semantic role *AGENT* and the biological concept *Protein*, whilst the argument *the nitrate reductase operon* is assigned the semantic role *THEME* and the biological concept *Operon*.

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not applicable – this is a monolingual corpus.

4.4.4 Attributes and their values (if annotated)

Standoff annotations

For the standoff version of the corpus, each text file (with extension ".txt") is accompanied by two files (with extensions containing the annotations. The format is based largely on the one used in the BioNLP Shared Task'09 (Kim et al., 2009) (see also http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/) with some modifications.

".a1" files

These files contains text spans that constitute event arguments and named entities occurring within these spans. The format of these files is as follows:

```
T1      Activator 0 4          EnvZ
T2      SPAN 15 27             through OmpR
T3      Regulator 23 27        OmpR
```

Each span/NE is assigned as id beginning with "T". The NE type (or "SPAN" – for event argument text spans that do correspond directly to NEs) is accompanied by start and end offsets in the abstract and the corresponding text.

".a2" files

These files contain text spans that constitute event triggers, and their associated event triggers. The format is as follows:

```
T13     Gene_Activation 263 273       activation
T14     GRE 296 304     requires
T15     GRE 309 317     function
E1      Gene_Activation:T13 Theme:T1,T4
E2      GRE:T14 Agent:E1 Theme:E3
```

Lines with an id starting with "T" correspond to event triggers, and follow the same format as the lines in the ".a1" files. Instead of NE types, a category of the event is shown in addition to offsets and text spans.

Lines with an id stating with "E" correspond to the event structures. The type of the event is separated by a colon from the id of the event trigger. This is followed by a list of the semantic arguments associated with the event. For each argument, the semantic role assigned to the argument is separated by a colon from the id of the argument. The argument can correspond to either:

- one or more simple text spans (contained within the associated ".a1" file), with ids begins with "T". Event arguments can consist of discontinuous text spans. In this case, each part of the argument is identified separately in the ".a1" file, and the event argument in the ".a2" file is specified using a comma-separated list of ids.
- another event structure described within the same ".a2" file. In this case, the id will start with "E", and will refer to (starting with "E"): an event argument can be another event.

Further information about the format of the standoff annotations can be found here: http://www.nactem.ac.uk/GREC/standoff.html

XML annotations

In the XML version of the corpus, a single file is present for each abstract, containing both the abstract text and the annotations. This format is based on the one used to represent the GENIA event corpus (Kim et al, 2008), with a small number of additions/modifications. See also:
http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Event+Annotation

An example of the annotation is shown below:

```
<sentence id="S7">In contrast, <term sem="SPAN"
id="T15" lex="sites_upstream_of_the_promoters">sites
upstream of the promoters</term> did not appear to be
necessary for repression, but were required for
activation by <term sem="Activator" id="T16"
lex="Lrp">Lrp</term> plus <term sem="Amino_Acids"
id="T17" lex="alanine">alanine</term> or <term
sem="Amino_Acids" id="T18" lex="leucine">leucine</term>
of one of the major dad promoters, <term sem="Promoter"
id="T19" lex="P2">P2</term>.</sentence>

<event id="E10">
<type class="Gene_Repression" />
<Condition idref="T15" />
<clue>In contrast, sites upstream of the promoters did
not appear to be necessary for
<clueType>repression</clueType>, but were required for
activation by Lrp plus alanine or leucine of one of the
major dad promoters, P2.</clue>
</event>
```

The following tags and attributes are used:

- *sentence* tag - represents a sentence
    - *id* attribute – the id of the sentence

- *term* tag – represents a NE or event argument within a sentence
    - *id* attribute – an id assigned to the term tag
    - *sem* attribute – the NE category assigned to the term span, or *SPAN* for event arguments that do constitute NEs.
    - *lex* – the textual representation of the contents of the term tag, with spaces replaced by underscores.
- *event* tag - represents an event structure
    - *id* attribute – an id assigned to the event
- *type* tag– the type of the event
    - *class* attribute – semantic class assigned to the event
- *Agent, Theme, Manner, Instrument, Location, Source, Destination, Temporal, Condition, Rate, Descriptive-Theme, Descriptive-Agent, Purpose* tags – One or more of these will be present within the event tag, according to the semantic roles assigned to the identified event arguments (see Thompson et al. (2009) for more details about the semantic roles used).
    - *Idref* attribute – stores the id of the event argument – either a *term* tag id or an *event* tag id. As mentioned above, arguments may consist of multiple, discontinuous spans. The attributes *idref1*, *idref2,* etc. may also be present if the argument consists of a discontinuous text span, to store the ids of other *term* tags that constitute the complete argument.
- *clue* tag – contains the complete sentence in which the event is contained. Expressions that constitute clues for identifying the event in the text are annotated within the sentence. In the current version of the corpus, only *clueType* is annotated.
- *clueType* tag – surrounds the event trigger within the *clue* text.

Further information about the format of the standoff annotations can be found here: http://www.nactem.ac.uk/GREC/xml.html

4.5 Intended application of the corpus

The corpus is intended to allow to the training of advanced, domain-specific semantic search systems that allow searches to be carried out over documents using structured semantic queries using named entities, semantic roles etc. as search constraints.

4.6 Reliability of the annotations (automatically/manually assigned) – if any

The reliability of the annotations was verified by calculating inter-annotator agreement rates for a portion of the corpus. Since the event annotation task consists of a number of sub-tasks, agreement rates were calculated for each of these. These are shown in Table 1.

Agreement Type	**F-Score**	
	E.coli	**Human**
Event identification	72.27%	76.37%
Argument identification (relaxed span match)	90.23%	91.27%
Argument identification (exact span match)	75.10%	77.48%
Semantic role assignment	88.96%	88.30%
Biological concept identification	82.55%	82.03%
Bio-concept category assignment (exact)	71.02%	66.03%
Bio-concept assignment (considering parent)	75.38%	68.97%
Bio-concept supercategory assignment	95.52%	94.75%

Table 1: Inter-annotator agreement figures for the GREC corpus

5 RELEVANT REFERENCES AND OTHER INFORMATION

Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y. and Tsujii, J.. (2009). Overview of BioNLP'09 Shared Task on Event Extraction. *In Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pp. 19.

Kim, J.-D., Ohta, T. and Tsujii, J.. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10

Thompson, P., Iqbal, S. A., McNaught, J. and Ananiadou, S.. (2009). Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10:349

Tsuruoka, Y., Tateisi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S. and Tsujii, J.. (2005). Developing a Robust Part-of-Speech Tagger for Biomedical Text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, pages 382-392, Springer-Verlag