GENIA CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of 2,000 MEDLINE abstracts, collected using the three MeSH terms *human*, *blood cells* and *transcription factors*.

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

The corpus is available in three formats

- A text file containing part-of-speech (POS) annotation, based on the Penn Treebank format
- An XML file containing inline POS annotation
- A "merged" XML format, containing inline annotations, corresponding to both POS and term annotations

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 archive contains the directory *GENIAcorpus3.02p.* It contains the following files (NOTE: each file contains the *complete* corpus in the format specified).

- *GENIAcorpus3.02.pos.txt* - a plain text file containing the POS-tagged corpus, with formatting based on the Penn TreeBank (PTB) formatting.
- *GENIAcorpus3.02.pos.xml* – an XML encoded file containing the abstracts with inline POS tags.
- *GENIAcorpus3.02.merged.xml* – an XML file containing the abstracts with inline annotation corresponding both to POS tags and term annotations
- *GENIAontology.daml* – a file containing the GENIA term ontology which has been used to annotate the terms.
- *gpml.merged.dtd* – The DTD to which the 2 XML files conform.
- *gpml.css* – A stylesheet to highlight the annotated terms
- *gpml.readme.html* – a file providing brief details of the structure of the file and the term annotation format
- *gpml.css.legend.html* – a file explaining the color-coding used in the css file

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 500,000 tokens. The PTB text-based POS corpus format requires 5.1MB of disk space, while the XML version of the POS corpus requires 10.6MB of disk space. The merged XML corpus, containing both POS and term annotations, requires 16.2 MB of disk space.

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 corpus is tokenized, and token is assigned a POS tag. The merged version of the corpus additionally contains annotations corresponding to biomedical terms.

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

Each abstract was automatically tokenized using the Penn tokenizer (http://www.cis.upenn.edu/~treebank/tokenization.htm) and assigned POS tags using the JunK tagger (Kazama et al, 2001). Corrections to the automatically assigned annotations were then made by annotators (see the POS annotation guidelines: http://www-tsujii.is.s.u-tokyo.ac.jp/%7Ejdkim/publications/GENIA_Guidelines_POS.pdf).
POS tags generally follow the Pen TreeBank (PTB) format (Santorini, 1990), with some modifications:

- The NNP and NNPS (proper name) tags are not used, except for the names of journals, authors, research institutes, and initials of patients. Especially, (discoverers') names in technical terms (e.g. Epstein-Barr virus, Southern blotting) are not tagged as NNP.
- The SYM tag has been eliminated as far as possible.

The POS annotation of the GENIA corpus is reported in detail in Tateisi & Tsujii (2004). See also http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Part%2Dof%2DSpeech+Annotation.

Term annotations were manually annotated, and assigned categories from the GENIA ontology (35 categories corresponding to the terminal ontology nodes). The term annotation takes care of semi-structured coordinated clauses by recovering ellipsis. As an example, the phrase *CD2 and CD 25 receptors* refers to two terms, i.e. *CD2 receptors* and *CD25 receptors*, but *CD2 receptors* doesn't appear in the text. The annotation is carried out in such a way to allow these separate terms to be identified. Term annotation is described in more detail in Kim et al. (2003). See also http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Technical+Term+Annotation.

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)

PTB-style text format

The file contains one token/POS pair per line, and a "====================" line (20 equal signs) is put between sentences. For example:

```
====================
These/DT
findings/NNS
should/MD
be/VB
useful/JJ
for/IN
therapeutic/JJ
strategies/NNS
and/CC
the/DT
development/NN
of/IN
immunosuppressants/NNS
targeting/VBG
the/DT
CD28/NN
costimulatory/NN
pathway/NN
./.
====================
```

XML-format

An example of the XML format of the merged corpus (*GENIAcorpus3.02.merged.xml*) is shown below. Note that the format of the XML file containing only POS tags (*GENIAcorpus3.02.pos.xml*) is the same, except for that the *cons* tags (corresponding to term annotations) are not present.

```
<article>
<articleinfo>
<bibliomisc>MEDLINE:95369245</bibliomisc>
</articleinfo>
<title>
<sentence><cons lex="IL-2_gene_expression"
sem="G#other_name"><cons lex="IL-2_gene"
```

```
sem="G#DNA_domain_or_region"><w c="NN">IL-2</w> <w
c="NN">gene</w></cons> <w
c="NN">expression</w></cons> <w c="CC">and</w>
<cons lex="NF-kappa_B_activation"
sem="G#other_name"><cons lex="NF-kappa_B"
sem="G#protein_molecule"><w c="NN">NF-kappa</w> <w
c="NN">B</w></cons> <w
c="NN">activation</w></cons> <w c="IN">through</w>
<cons lex="CD28" sem="G#protein_molecule"><w
c="NN">CD28</w></cons> <w c="VBZ">requires</w> <w
c="JJ">reactive</w> <w c="NN">oxygen</w> <w
c="NN">production</w> <w c="IN">by</w> <cons
lex="5-lipoxygenase" sem="G#protein_molecule"><w
c="NN">5-lipoxygenase</w></cons><w
c=".">.</w></sentence>
</title>
<abstract> …
</abstract>
```

The tags and attributes are as follows:

- *article* tag – surrounds each abstract in the corpus
- *articleinfo* tag – contains information about the article
- *bibliomisc* tag– contains the MEDLINE id of the article
- *title* tag– contains the title of the article
- *abstract* tag – contains the main text of the abstract
- *sentence* tag – surrounds the text of each sentence in the title/abstract
- *w* tag – surrounds each token
    - *c* attribute – the part of speech assigned to the token
- *cons* tag – corresponds to a term annotation – surrounds one of more *w* tags. These can be embedded, as is the case with *IL-2 gene* and *IL-2 gene expression* in the example above.
    - *lex* attribute – the complete lexical representation of the term, with spaces replaced by underscores
    - *sem* attribute - the semantic category assigned to the term. The *G* prefix denotes that the category has been assigned from the GENIA term ontology.

Co-ordinated structures containing 2 or more terms that are not both completely specified in the conjunction (due to ellipsis) are annotated as shown below. The coordinated phrase in this case is *hematopoietic and trophoblast cells*, which refers to the two terms *hematopoietic cells* and *trophoblast cells.*

```
<cons lex="(AND hematopoietic_cell
trophoblast_cell)" sem="(AND G#cell_type
G#cell_type)"><cons lex="hematopoietic*"><w
c="JJ">hematopoietic</w></cons> <w c="CC">and</w>
```

```
<cons                              lex="trophoblast*"><w
c="NN">trophoblast</w></cons> <cons lex="*cell"><w
c="NNS">cells</w></cons></cons><w
c=".">.</w></sentence>
```

A *cons* tag is placed around the whole phrase, and the use of *AND* in both the *lex* and *sem* attribute values shows that there are 2 distinct terms involved. An embedded *cons* tag is placed around both the unique and common parts of each of the two terms. The use of the * symbol in *lex* attribute of these embedded *cons* elements indicates that each *cons* element contains only part of the term, with the remainder in another *cons* element.

4.5 Intended application of the corpus

The corpus is intended to allow to facilitate the building of domain-specific term recognition systems, and to help to adapt existing POS taggers and other applications to the biomedical domain.

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

In terms of the manually-corrected POS tags, inter-annotator agreement scores were 0.985 Kappa.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Kazama, J., Y. Miyao, and J. Tsujii, 2001. A maximum entropy tagger with unsupervised hidden markov models. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium.*

Kim, J-D., T. Ohta, Y. Tateisi, and J. Tsujii (2003). Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics,* 19:180i–182i.

Santorini, B., 1990. Part-of-speech tagging guidelines for the Penn Treebank project. Technical Report MS-CIS- 90-47, Department of Computer and Information Science, University of Pennsylvania.

Tateisi, Yuka and Jun'ichi Tsujii. Part-of-Speech Annotation of Biology Research Abstracts. In *Proceedings of 4th International Conference on Language Resource and Evaluation (LREC2004).* pp. 1267-127