The ITI TXM corpora

Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., Roebuck, S., Tobin, R., & Wang, X. (2008). The ITI TXM corpora: Tissue expressions and protein-protein interactions. In LREC 2008 Workshop.

The ITI TXM Corpora: Tissue Expressions and Protein-Protein Interactions

**Bea Alex, Claire Grover, Barry Haddow, Mijail Kabadjov,
Ewan Klein, Michael Matthews, Stuart Roebuck, Richard Tobin and Xinglong Wang**

University of Edinburgh, School of Informatics
2 Buccleuch Place, Edinburgh, EH8 9LW, Scotland, UK
txm-researchers@inf.ed.ac.uk

Abstract

We report on two large corpora of semantically annotated full-text biomedical research papers created in order to develop information extraction (IE) tools for the TXM project. Both corpora have been annotated with a range of entities (CellLine, Complex, DevelopmentalStage, Disease, DrugCompound, ExperimentalMethod, Fragment, Fusion, GOMOP, Gene, Modification, mRNAcDNA, Mutant, Protein, Tissue), normalisations of selected entities to the NCBI Taxonomy, RefSeq, EntrezGene, ChEBI and MeSH and enriched relations (protein-protein interactions, tissue expressions and fragment- or mutant-protein relations). While one corpus targets protein-protein interactions (PPIs), the focus of other is on tissue expressions (TEs). This paper describes the selected markables and the annotation process of the ITI TXM corpora, and provides a detailed breakdown of the inter-annotator agreement (IAA).

1 Introduction

This paper describes two corpora constructed and annotated for the TXM project. The aim of the TXM project was to develop tools for assisting in the curation of biomedical research papers. The ITI TXM corpora were used to train and test machine learning based NLP components which were interfaced with a curation tool.

There already exist several corpora of annotated biomedical texts (Section 2), all with individual design and annotation characteristics. The ITI TXM corpora combine a number of attractive characteristics of such available corpora, thus making them a valuable resource for NLP research. We annotated full-text papers since our intended target application (the curation tool) worked with such documents. Furthermore, it has been shown in previous research that there is valuable information in full-text articles that cannot be obtained from their abstracts alone (e.g. by Shah et al., 2003 and McIntosh & Curran, 2007). The markables used in the ITI TXM corpora included not only a range of named entities and relations, but also extensive, multi-species normalisation of proteins, genes and other entities, to standard publicly available databases.[1] Furthermore, some of the relations were enriched with additional biomedical information enabling finer-grained classification, and connecting the relations with other entities in the text. At around 200 full-text papers each, the corpora are relatively large in size. In addition, we will release multiple annotations of many of the papers, enabling the comparison of different annotators' views of the corpus. The set of markables chosen for both corpora arose out of extensive discussions between biologists managing the curation, and NLP researchers creating the NLP components. The biologists were consulted to determine what information they wanted to be extracted. At the same time, their ideas had to be balanced against what was possible using the state-of-the-art in NLP technology, and what could be reliably annotated. The final set of markables resulted out of several iterations of piloting and measurements of IAA.

This paper is organised as follows: after discussing related work on biomedical corpus design and annotation in the next section, a description of how the documents were selected for the corpora is provided in Section 3. An overview of both corpora, a description of the markables, the annotation process and details of the IAA are presented in full in Section 4. Finally Section 5 offers some conclusions and lessons learnt from the annotation project.

2 Related Work

In recent years, there have been numerous efforts in constructing and annotating biomedical corpora. Comprehensive lists of publicly available corpora are maintained by Cohen et al.[2] as well as Hakenberg[3]. This related work section does not provide an all-inclusive list of biomedical corpora but rather presents different characteristics of corpus design and annotation illustrated by typical examples.

Existing resources vary in size, type of data, markables and levels of annotation, the way the annotation is applied, their distributed formats and their domains. The GENIA corpus (Ohta et al., 2002), for example, is one of the largest and most widely used data sets in the text mining community. It consists of 2,000 Medline abstracts and is manually annotated with a series of semantic classes defined in the GENIA ontology. Other corpora are made up of sets of sentences from biomedical research articles, as is the case for BioInfer (Pyysalo et al., 2007) and GENETAG (Tanabe et al., 2005). The latter is a collection of 20,000 Medline sentences annotated for gene and protein names in one semantic class. Parts of this corpus were used in the BioCreAtIvE I and II competitions that, amongst other tasks, enabled different text mining research groups to evaluate how well their systems perform at extracting gene/protein names from biomedical literature.

Although there have been a series of corpus construction efforts for the purpose of biomedical text mining, only a small number of groups (e.g. Wilbur et al., 2006 and Krallinger et al., 2006) report IAA figures. In other words, it is rare to find information about how consistent two independent

[1] Normalisation refers to the task of grounding a biomedical term in text to a specific identifier in a referent database. See Table 3 for the publicly available databases used.

[2] http://compbio.uchsc.edu/ccp/corpora/obtaining.shtml

[3] http://www2.informatik.hu-berlin.de/~hakenber/links/benchmarks.html

annotators are when marking up a representative sample of a data set. The assumption is that the level of IAA provides insights into how challenging a particular task is to a human expert, providing an upper bound for an automated system is and how appropriate the task in itself is. Lu et al. (2006) show an increase in IAA over time as annotators become more familiar with their task of marking up GeneRIFs with 31 semantic classes in the protein transport domain. Figures of IAA also help to determine weaknesses in the annotation guidelines. Mani et al. (2005) measured IAA based on a first set of annotation guidelines for marking up protein names.[4] After analysing the annotation differences, they revised their guidelines which resulted in an improvement of IAA in a second annotation round and simultaneously in better annotation quality overall. Alex et al. (2006) have shown that consistency in the annotation of named entity boundaries is crucial to obtain high accuracy for biomedical named entity recognition. The need for both clear annotation guidelines to achieve such consistency and comprehensive annotation guidelines to capture complex information in unstructured text data is often highlighted (e.g. see Wilbur et al., 2006 and Piao et al., 2007). Making such guidelines available to the research community and publishing figures of IAA is recommended by Cohen et al. (2005) who analysed the characteristics of different biomedical corpora. They also conclude that distributing data in standard formats (e.g. XML) is vital to guarantee high corpus usage.

As mentioned earlier, publicly available corpora differ in the type of textual data, i.e. a corpus can be made up of sentences, abstracts or full-text papers. McIntosh & Curran (2007) and Shah et al. (2003) indicate a clear need for biological IE from full-text articles. The former study shows that only a small proportion of identified fact instances appears in abstracts. The latter found that although abstracts contain the best ratio of keywords, other sections of articles are a better source of biologically relevant data. As a result, they advocate IE systems that are tuned to specific sections. As much of the important information is not present in the abstract but the main paper, Cohen et al. (2005) suggest that abstracts and isolated sentences are inadequate and unsuited to the opportunities that are available for text mining. Sometimes, the most relevant information in a paper is found in figure captions (Shatkay and Feldman, 2003). Currently, only few available resources contain full-text publications, one example of such a corpus being FetchProt (2005). Its annotation includes specific experiments and results, the proteins involved in the experiments and related information. Exploiting such full-text resources is vital to develop text mining systems that will be used in practice, e.g. by biologists, clinicians or curators. Publicly available biomedical corpora also often differ in their markables and levels of annotation. Some are annotated with part-of-speech tags (e.g. GENIA) and named entities, most often gene/protein names (e.g. GENETAG) that are sometimes normalised to identifiers (e.g. FetchProt). In other cases, the annotation includes binary relations between entities such as PPIs (e.g. AImed described in Bunescu et al., 2005) or non-binary relations (e.g. BioInfer). Several corpora are distributed with syntactic annotation such as phrase-based or dependency-based structures, e.g. BioIE (Kulick et al., 2004), GENIA treebank (2005), LLL (Nedellec, 2005) and BioInfer.

In this paper, we introduce two large biomedical corpora in the sub-domains of PPIs and TEs which will be distributed in one collection as the ITI TXM corpora. Both corpora are made up of full-text papers that are annotated with a series of relevant named entities, some of which are normalised. Furthermore, the annotations include various types of relations as well as relation attributes and properties (see Section 4.2). Domain experts used extensive curation guidelines that were devised based on several rounds of piloting (see Section 4.3). We provide figures of IAA for all types of semantic annotation for a representative corpus sample (see Section 4.4). Moreover, the data is distributed in XML with semantic annotations in standoff format (Carletta et al., 2005). In the future, the ITI TXM corpora will serve as a valuable resource to train IE methods for mining facts from biomedical literature.

3 Document Selection

Document selection for the PPI corpus was performed in two stages. The initial plan was to annotate only full-text articles available in XML. Therefore, 12,704 full-text XML files were downloaded from PubMedCentral OpenAccess.[5] The documents were filtered by selecting those articles that contained at least 1 of 13 terms either directly associated with PPIs or with biological concepts representative of typical curation tasks.[6] The abstracts and, if necessary, full texts of the remaining 7,720 documents were all examined by trained biologists and selected if they contained interactions that were experimentally proven within the paper, resulting in a total of 213 documents.[7] In order to ensure that enough documents were available for annotation, the same queries were performed against PubMed and additional documents were selected from the resulting list using the same criteria.[8] Several of the documents were excluded from the final set because they were used during the piloting or were rejected by the annotators as not being suitable for annotation. The resulting corpus consists of 217 documents, 133 selected from PubMedCentral and 84 documents selected from the whole of PubMed.

Document selection for the TE corpus was performed against PubMed. This was partially to ensure that enough documents were selected, and partially to address the concern that in practice, many important documents would not be available in XML and the annotations would be more representative if they accounted for this reality. The initial pool of documents was selected from PubMed using terms designed to capture documents representative of typical TE and PPI curation tasks.[9] The abstracts of the resulting 12,060 documents were randomised and examined

[4] Mani et al. (2005) refer to IAA as inter-coder reliability.

[5] `http://www.pubmedcentral.nih.gov/`
The 12,704 articles represented the complete set of available documents on 17/08/2005.

[6] The terms were: *bind*, *complex*, *interact*, *apoptosis*, *ubiquitination*, *mitosis*, *nuclear envelope*, *cell cycle*, *phosphorylation*, *glycosylation*, *signal transduction* and *nuclear receptors*.

[7] Clinical articles on drug or patient trials were excluded.

[8] `http://www.ncbi.nlm.nih.gov/PubMed/`

[9] The queries were: *"Gene Expression Regulation"[MeSH], Development, "Signal Transduction"[MeSH], "Protein Biosynthesis"[MeSH], "Cell Differentiation"[MeSH], Apoptosis, Mitosis, Cell cycle* and *Phosphorylation*

Annotations	PPI				TE			
	TRAIN	DEVTEST	TEST	All	TRAIN	DEVTEST	TEST	All
1	65	25	35	125	82	34	34	150
2	48	9	8	65	68	7	11	86
3	20	5	2	27	1	0	1	2
Total documents	133	39	45	**217**	151	41	46	**238**
Total annotations	221	58	57	**336**	221	48	59	**328**

Table 1: Counts of numbers of papers with 1, 2 or 3 annotations in each section of each corpus.

in order by a biologist and selected if they contained mentions of the presence or absence of mRNA or protein in any organism or tissue. A total of 4,327 documents were examined of which 1,600 were selected for TE annotation. The TE corpus is comprised of the first 238 of these documents that were not used during piloting and not rejected by the annotators.

In both phases, documents were split into TRAIN, DEVTEST, and TEST sets in a ratio of approximately 64:16:20 (see Table 1). TRAIN was to be used for training machine learning models and deriving rules, DEVTEST for testing during system development, and TEST for testing the final system. The document selection methods were dictated, in part, by the requirements of the industrial partner that assisted in the annotation of the corpora. The terms used were based on the queries used for selecting documents for creating commercially viable curated databases. Furthermore, the results of document selection were used to create training and testing corpora for a document retrieval system designed to improve the document selection phase. These corpora will be released at a future date.

4 Corpus Annotation

4.1 Overview

Documents were selected for annotation as described in Section 3. The full-text papers were downloaded from PubMed or PubMedCentral either as XML, or as HTML if the XML version was not available, and then converted to an in-house XML format using LT-XML2 tools.[10] The LT-XML2 and LT-TTT2 tools were also used to tokenise and insert sentence boundaries into the text (Grover et al., 2006).

From each corpus a random selection of documents was chosen for double or triple annotation in order to allow calculation of IAA, which is used to track annotation quality and to provide a measure of the difficulty of the task. The counts of singly and multiply annotated documents in the TRAIN, TEST and DEVTEST sections for both corpora are shown in Table 1. Multiply annotated documents were left in the corpus and not reconciled to produce a single, gold standard version. It was found during piloting that reconciliation could be very time-consuming so we decided to focus our resources on obtaining a larger sample of papers.

During the annotation of the full-text papers, we did not annotate sections that did not contain any relevant information, e.g. contact details and reference sections, HTML navigational text. Moreover, materials and methods sections were not annotated on the grounds that they would be too time-consuming to annotate. The annotators marked unannotated paragraphs during the annotation so that these sections could be excluded from training and testing. Based on the sentence splitting and tokenisation performed during the pre-processing, the PPI corpus contains approximately 74.6K sentences and 2.0M tokens, and the TE corpus is made up of around 62.8K sentences and 1.9M tokens.[11]

[10] http://www.ltg.ed.ac.uk/software/xml/

Entity type	PPI	TE
CellLine	7,676	—
Complex	7,668	4,033
DevelopmentalStage	—	1,754
Disease	—	2,432
DrugCompound	11,886	16,131
ExperimentalMethod	15,311	9,803
Fragment	13,412	4,466
Fusion	4,344	1,459
GOMOP	—	4,647
Gene	—	12,059
Modification	6,706	—
mRNAcDNA	—	8,446
Mutant	4,829	1,607
Protein	88,607	60,782
Tissue	—	36,029

Table 2: Entity types and counts in each corpus. A long dash indicates that the entity was not marked in that corpus.

4.2 Description of Markables

In both corpora the markables, i.e. units of annotation, consist of named entities, normalisations, relations, properties and attributes.

Named entities are terms of interest to biologists which belong to pre-defined semantic classes. Table 2 shows the named entity types marked and their counts in each corpus. In the PPI corpus, the entities are either proteins and other related entities involved in PPI relations (Protein, Complex, Fusion, Fragment and Mutant) or attributes of PPI relations (CellLine, DrugCompound, ExperimentalMethod, Modification). Conversely, for the TE corpus, the entities are either those that can be involved in TE relations (Tissue, Protein, Complex, Fusion, Fragment, Mutant, Gene, mRNAcDNA and GOMOP) or those that can be attributes of TE relations (DevelopmentalStage, Disease, DrugCompound, ExperimentalMethod). All named entity types (except GOMOP) have intuitively obvious biological interpretations, which are made precise in the annotation guidelines. For example, the definition of DrugCompound is: "a chemical substance of known composition used to affect the function of an organism, cell or biological process". The GOMOP entity type was used in cases where the annotator felt that the author was referring to a "***G**ene **o**r **m**RNAcDNA **o**r **P**rotein*". We felt that having a single entity type to represent this kind of ambiguity would be simpler than allowing annotators to mark the same term as multiple entity types (e.g. Protein and Gene).

[11] Note that all annotated versions of each paper are treated as separate documents in this calculation.

Database	Url	Prefix	PPI	TE
NCBI Taxonomy	http://www.ncbi.nlm.nih.gov/Taxonomy/	ncbitaxon:	Protein	Gene, mRNAcDNA, Protein, GOMOP
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq/	refseq:	Protein	Protein, mRNAcDNA
EntrezGene	http://www.ncbi.nlm.nih.gov/entrez/	gene:	Protein	Gene, mRNAcDNA, Protein, GOMOP
ChEBI	http://www.ebi.ac.uk/chebi/	chebi:	—	DrugCompound
MeSH	http://www.nlm.nih.gov/mesh/	mesh:	—	Tissue

Table 3: Databases used for normalisations and the entities to which they are assigned in each corpus. A long dash indicates that the database was not used in that corpus.

Corpus	Relation type	Count
PPI	PPI	11,523
PPI	FRAG	16,002
TE	TE	12,426
TE	CHILD-PARENT	4,735

Table 4: Relation types in each corpus.

When marking named entities, the annotators were permitted to nest them, but entities were not allowed to cross. For any pair of entities with a non-empty intersection, the intersection therefore had to coincide with at least one of the entities. Entities were also required to be continuous. Discontinuous coordinations such as "A and B cells" were annotated as two nesting entities "A and B cells" and "B cells", indicating that the first was discontinuous using a flag in the XML. Furthermore, annotators were able to override the tokenisation if entity boundaries and token boundaries did not coincide, by indicating the entity boundaries using character offsets. For example, in one annotated document, the term "Cdt1(193-447)" is tokenised as a single token, but the annotator decided that "Cdt1" was a Protein and "193-447" was a Fragment. The Protein was therefore marked using an end offset of -9, to indicate that the end of the Protein name was 9 characters from the end of the token, and in a similar way the Fragment had start offset 5 and end offset -1. The XML representation of the data enables retokenisation as proposed by Grover et al. (2006) to improve the original tokenisation at a later stage while preserving the entity annotation.

A number of types of entities were normalised to one or more of the standard, publicly available biomedical databases listed in Table 3. In general, for each entity term that was normalised, an ID of the appropriate database was assigned as the normalisation value with a prefix indicating the source database. If no appropriate identifier existed, the ID was left blank and only the database prefix was used as the normalised value.

Normalisation of protein, gene and mRNAcDNA entities was more complex. Two types of normalisations were added to each occurrence of such entities: *full normalisation* and *species normalisation*, where the former involves assigning RefSeq identifiers to protein and mRNAcDNA terms and EntrezGene identifiers to gene terms; and the latter involves assigning NCBI taxonomy identifiers to protein, gene and mRNAcDNA terms. The project initially aimed at providing *full normalisation* for both corpora.[12] However, *full normalisation* turned out to be too time-consuming. Given limited time and resources, only the TE corpus and the DEVTEST and TEST portions of the PPI corpus were fully normalised, while the TRAIN portion of the PPI corpus was only species-normalised. A few special cases must be considered in the normalisation annotation:

- *Species mismatch*. For the term to be normalised, there is an entry in the database (e.g. RefSeq) which matches the specific entity but the entry does not match the species of the term given the surrounding context. In this case the term was only normalised for its species (i.e. species normalisation).

- *Several host species*. The term to be normalised is discussed relative to several host species. In this case, the term was normalised multiple times and each annotated entity was assigned a unique identifier for each species mentioned. In case of more than five possible host species for the term, annotators followed the next instruction.

- *Host species not clear*. The host species of a term to be normalised cannot be determined from the text, because it is discussed in a general way rather than in relation to one or more specific species, or the text is unclear about the host species of the term. In this case, the entity was normalised as if its species was *Homo sapiens*, and the keyword "gen" (for "general") was added to any chosen identifier, e.g. "NP_004513 (gen)", and at the same time the Taxonomy identifier for Homo sapiens together with the keyword "gen" (e.g., "9606 (gen)") were entered as the species-normalisation. However, if *Homo sapiens* could not possibly be the correct host species, due to the occurrence of a general species word, such as *viral* or *bacterial*, "gen" was entered for species normalisation.

In each corpus, two types of relations were marked (see Table 4). In the PPI corpus, relations refer to interactions between two proteins (PPI) and connect Mutants and Fragments with their parent proteins (FRAG). In the TE corpus, relations indicate when a gene or gene product is expressed in a particular tissue (TE); relations also connect Mutants and Fragments with their parent proteins (CHILD-PARENT). Annotators were permitted to mark relations between entities in the same sentence (intra-sentential) and in different sentences (inter-sentential). For the TE and PPI relations, annotators also marked "link terms" used by the authors to indicate a relation. Marked in the same way as entities, these are called InteractionWord for PPI relations and ExpressionLevelWord for TE relations.

The properties and attributes are extra pieces of information added by the annotators to both PPI and TE relations. A property is a name-value pair assigned to a relation to add extra information, for example whether a PPI is mentioned

[12] In fact, both RefSeq and EntrezGene identifiers are species-specific. When a term is "fully normalised" its host species can therefore be identified without *species normalisation*.

Name	Value	PPI	TE
IsPositive	Positive	10,718	10,243
	Negative	836	2,067
IsDirect	Direct	7,599	—
	NotDirect	3,977	—
IsProven	Proven	7,562	9,694
	Referenced	2,894	1,837
	Unspecified	1,096	736

Table 5: Property names, values and counts in each corpus. A long dash indicates that the property was not marked in this corpus.

as being direct or indirect, or whether it was experimentally proven in the paper. Both positive and negative TE and PPI relations, i.e. statements asserting that an interaction or expression did or did not occur, were also marked, with properties used to distinguish between them. The names and values for the properties were drawn from a small closed list and annotators assigned at least one value to each name, for each relation. Their counts in each corpus are listed in Table 5.

Attributes are named links between relations and other entities, e.g. to indicate the experimental method used to verify a PPI relation, or the cell line used to discover a TE relation. In the PPI corpus, all attributes, except for MethodEntity, are attached to entities. Conversely, all attributes are attached to relations in the TE corpus. Attributes are also used to link a relation to its link term and do not have to be in the same sentence as the relation. The names and counts of the attributes are listed in Tables 6 and 7.

Note that as well as being able to add multiple values for each relation property, annotators were also permitted to add multiple values for each attribute. They did this by marking extra relation entries. For example, in a sentence such as "Protein A interacts with B in the presence of Drug C but not D.", the annotators would mark two PPI relations between "A" and "B", one Positive with "C" as a DrugCompound attribute, and the other negative with "D" as a DrugCompound attribute.

4.3 The Annotation Process

Annotation was performed by a group of nine biologists, all qualified to PhD level in biology, working under the supervision of an annotation manager (also a biologist) and collaborating with a team of NLP researchers. At the beginning of the annotation of each corpus, a series of discussions between the biologists and the NLP team were held with the aim of determining a set of markables. Since the overall aim of the project was to build NLP tools for integration into a curation assistant, the markables suggested by the biologists were those which they wished the curation assistant to aid them with. The NLP team provided input as to which markables might be technically feasible and what could be reasonably accomplished within the project timescale.

A further consideration in selecting markables was how well they could be annotated in practice. Markables which could not be reliably annotated by humans would not produce good data, and as a result would be even more difficult for automated systems to extract. Using the initial list of markables, several rounds of piloting were conducted to determine the markables that could be annotated reliably. For example, four piloting iterations were conducted before commencing the annotation of the PPI corpus. As a result, it was decided to remove MutationType from the list of originally proposed entity types as this information did not occur frequently enough in the piloting documents. The piloting process also helped to produce comprehensive annotation guidelines on all markables. During the piloting phase, the same documents were annotated by two or three annotators, IAA was computed for these documents, and annotation differences were analysed. The annotators discussed points of difficulty and disagreement with the NLP team and the annotation guidelines were clarified and extended wherever necessary.

At the end of the piloting phase a final set of markables was agreed by all parties and the main body of annotation commenced. During this phase weekly annotation meetings were held to discuss the latest IAA measurements and any other issues arising from the annotation, with all the annotators in attendance plus a representative from the NLP team. IAA was measured using a sample of documents randomly selected in advance for multiple annotation. The annotation was organised so that annotators were not aware when they were assigned a document that was being annotated by someone else as well. When new annotators joined the team they went through a training phase where they annotated several documents, comparing their annotations with those created by the existing team. This was done to ensure that they were following the guidelines correctly and were consistent with the other annotators.

For the annotation of the PPI corpus, an in-house annotation tool was developed using FilemakerPro, with data being stored in a relational database before being exported to XML for analysis by the NLP team. However, as this annotation tool did not scale well, a customised version of Callisto[13] was employed for the TE annotation project. Before the documents were presented to the annotators, they were tokenised and had sentence boundaries inserted by means of pre-processing steps implemented using the LT-XML2 and LT-TTT2 tools. The original spacing in the documents was preserved so that it could be recovered from the XML version simply by stripping off the word, sentence and paragraph elements.

All annotated documents were converted to an in-house XML format, for consumption by NLP applications. In the XML, all annotations are placed in standoff, with the normalisations included in the named entity annotation, and the properties and attributes included in the relation annotation. Listings 1, 2 and 3 show a sample of text, with its standoff entity and relation annotation. The standoff entity annotation uses word ids to refer to the start and end words of the entity, and the standoff relation annotation uses entity ids to refer to its entity pair. Note that the standoff markup for a document and its text are contained within the same file. An XML schema and format documentation will be provided with the corpus release.

```
<s><w id="A33864">Rrs1p</w>
 <w id="A33870">has</w> <w id="A33874">a</w>
 <w id="A33876">two</w><w id="A33879">-</w>
 <w id="A33880">hybrid</w>
 <w id="A33887">interaction</w>
 <w id="A33899">with</w> <w id="A33904">L5</w>
 <w id="A33906">.</w></s>
```

Listing 1: Extract from the text of an annotated document (note the original does not contain the line breaks)

[13] http://callisto.mitre.org/

Name	Entity type	Explanation	Count
ModificationBeforeEntity	Modification	Any modification applied before the interaction.	240
ModificationAfterEntity	Modification	Any modification resulting from the interaction.	1,198
DrugTreatmentEntity	DrugCompound	Any drug treatment applied to the interactors.	844
CellLineEntity	CellLine	The cell-line from which the interactor was drawn.	2,000
ExperimentalMethodEntity	ExperimentalMethod	The method used to detect the interactor.	1,197
MethodEntity	ExperimentalMethod	The method used to detect the interaction.	2,085
InteractionWordEntity	InteractionWord	The term which indicates the interaction.	11,386

Table 6: Attributes in the PPI corpus.

Name	Entity type	Explanation	Count
te_rel_ent-drug-compound	DrugCompound	Any drug compound applied.	1,549
te_rel_ent-exp-method1	ExperimentalMethod	The method used to detect the expression participants.	1,878
te_rel_ent-disease	DiseaseType	Any disease affecting the tissue.	332
te_rel_ent-dev-stage	DevelopmentalStage	The developmental stage of the tissue.	327
te_rel_ent-expr-word	ExpressionLevelWord	A term indicating the level of expression.	2,815

Table 7: Attributes in the TE corpus.

```
<ent id="e933262" norm="NP_014937" type="Protein"
 species="4932" sw="A33864" ew="A33864">Rrs1p</ent>
<ent id="e933263" norm="" type="ExperimentalMethod"
 sw="A33876" ew="A33880">two-hybrid</ent>
<ent id="e933264" norm="" type="InteractionWord"
 sw="A33887" ew="A33887">interaction</ent>
<ent id="e933265" norm="NP_015194" conf="100"
 type="Protein" species="4932" sw="A33904"
 ew="A33904">L5</ent>
```

Listing 2: Example of standoff annotation of entities

```
<relation type="ppi" id="r903106" IsProven="Proven"
 IsDirect="Direct" IsPositive="Positive">
<argument ref="e933262"></argument>
<argument ref="e933265"></argument>
<attribute name="MethodEntity" ref="e933263"/>
<attribute name="InteractionWordEntity"
 ref="e933264"/>
</relation>
```

Listing 3: Example of standoff annotation of relations

Type	PPI		TE	
CellLine	81.6	(2,456)	—	
Complex	76.4	(2,243)	82.6	(886)
DevelopmentalStage	—		72.7	(357)
Disease	—		74.3	(435)
DrugCompound	76.4	(3,705)	84.9	(4,453)
ExperimentalMethod	74.0	(4,673)	76.7	(2,013)
Fragment	75.3	(3,985)	77.7	(1,179)
Fusion	78.5	(1,270)	73.9	(359)
GOMOP	—		50.2	(655)
Gene	—		77.7	(1,911)
Modification	87.6	(1,900)	—	
mRNAcDNA	—		78.1	(1,768)
Mutant	60.4	(1,008)	63.9	(310)
Protein	91.6	(32,799)	90.3	(16,329)
Tissue	—		84.1	(8,210)
All	84.9	(54,039)	83.8	(38,865)

Table 8: IAA for entities (in F_1) in each corpus. The total number of true positives is shown in brackets.

4.4 Inter-annotator Agreement

We IAA for each corpus and each markable using the multiply annotated documents. For each pair of annotations on the same document, IAA was calculated by scoring one annotator against another using precision, recall and F_1. For the PPI corpus, IAA was calculated on a total of 146 document pairs. IAA for TE corpus, having fewer triple annotations, was computed over a total of 92 document pairs. An overall corpus IAA was calculated by micro-averaging across all annotated document pairs.[14] Micro-averaging was chosen over macro-averaging, since we felt that the latter would give undue weight to documents with few or no markables. We used F_1 rather than Kappa (Cohen, 1960) to measure IAA since the latter requires comparison with a random baseline, which would not make sense for tasks such as named entity recognition and normalisation.

For named entities, IAA was calculated using precision, recall and F_1, defining two entities as equal if they had the same left and right boundaries, and the same type. The IAA figures for named entities listed in Table 8 show that annotation consistency is generally high, with important and frequently occurring entities scoring in the 80s or 90s. IAA is low for entity types which occur infrequently such as Mutant. It is particularly low for GOMOP, not only an infrequent entity but also an artificially constructed class designed to include cases of annotator uncertainty. The overall IAA is lower than that normally reported for MUC type entities, but fits with our observations that biomedical named entity annotation is more difficult.

The IAA for normalisations was only calculated when both annotators agreed on the entities. This means that the normalisation IAA only reflects agreement on normalisation annotation and is not affected by the level of agreement on the entity annotation. In addition, all entities marked as general were excluded from the IAA calculations (see Table 9). For Protein and mRNAcDNA types, only those entities that were normalised to RefSeq identifiers were included in the IAA calculations while for Gene and GOMOP entities, only those entities normalised to EntrezGene identifiers were included. The IAA was measured using F_1 where two normalisations were considered equal if both an-

[14] Micro-averaging means giving equal weight to each example, as opposed to macro-averaging which would give equal weight to each annotated document pair.

Type	PPI		TE	
DrugCompound	—		97.7	(215)
GOMOP	—		77.3	(214)
Gene	—		95.1	(1,463)
mRNAcDNA	—		88.0	(892)
Protein	88.4	(7,595)	90.0	(5,979)
Tissue	—		82.9	(6,776)
All	88.4	(7,595)	83.8	(15,785)

Table 9: IAA for normalisation (in F_1) in each corpus. The total number of true positives is shown in brackets.

Type	PPI		TE	
PPI	67.0	(2,729)	—	
TE	—		70.1	(2,078)
FRAG	84.6	(3,661)	84.0	(1,012)
All	76.1	(6,390)	74.1	(3,090)

Table 10: The IAA for relations (in F_1) in each corpus. The total number of true positives is shown in brackets. Note that FRAG relations are referred to as CHILD-PARENT in the TE corpus.

notators selected the same ID.

When calculating IAA for relations, only those relations for which both annotators agreed on the entities were included. Relation IAA was also measured using F_1, where relations are counted as equal if they connect exactly the same entity pair, and have the same type. The IAA for relations shown in Table 10 is overall lower than that for entities and normalisations, suggesting that this is a more difficult task. Since relations can span across clauses and even across sentences, the annotators need to perform a deeper analysis of the text than for entity annotations.

For properties, IAA was calculated for each name-value pair, again using precision, recall and F_1. In cases where the annotators had entered multiple relations of the same type between the same entities, these sets of equivalent relations were collapsed for the purpose of property and attribute IAA calculation. The collapsed relation was given the union of all the properties and attributes assigned to the relations in the set. This collapsing is an approximation of the annotator's intentions, but the number of occurrences of multiple equivalent relations is small so the collapsing should not have a significant effect on the IAA. The IAA for properties shown in Table 11 is generally very high, except for the IsProven-Unspecified category which was used infrequently by the annotators and suffers from being an "other" category.

For attributes, IAA was again measured using precision, recall and F_1. Two attributes were considered equivalent if they had the same type and connected the same relation and entity. Tables 12 and 13 show the IAA figures for attributes. These are quite low in some cases, and so are the total numbers of attributes assigned. Investigation of the IAA suggests that annotators often disagreed about whether to assign an attribute or not, but if they both assigned an attribute then they generally chose the same one. The entities used as attributes sometimes appeared at a distance from the relation in the text. Therefore, it is not surprising that annotators sometimes missed them, or assigned them inconsistently.

Name	Value	PPI		TE	
IsPositive	Positive	99.6	(2,553)	97.2	(1,807)
	Negative	90.1	(155)	88.9	(280)
IsDirect	Direct	86.8	(1,746)	—	
	NotDirect	61.4	(449)	—	
IsProven	Proven	87.8	(1,543)	92.8	(1,547)
	Referenced	88.6	(626)	75.3	(204)
	Unspecified	34.4	(448)	29.3	(38)
	All	87.2	(7.165)	91.2	(3,779)

Table 11: IAA for properties (in F_1) in each corpus. The total number of true positives is shown in brackets.

Name	IAA	
ModificationBeforeEntity	65.3	(31)
ModificationAfterEntity	86.7	(248)
DrugTreatmentEntity	45.4	(61)
CellLineEntity	64.0	(244)
ExperimentalMethodEntity	36.9	(94)
MethodEntity	55.4	(274)
All	59.6	(952)

Table 12: IAA of attributes (in F_1) in the PPI corpus. The total number of true positives is shown in brackets.

Name	IAA	
te_rel_ent-drug-compound	77.9	(229)
te_rel_ent-exp-method1	81.3	(261)
te_rel_ent-disease	64.0	(16)
te_rel_ent-dev-stage	57.8	(13)
All	77.2	(521)

Table 13: IAA of attributes (in F_1) in the TE corpus. The total number of true positives is shown in brackets.

5 Discussion and Conclusions

In terms of the amount of text annotated, the ITI TXM corpora are the result of one of the largest biomedical corpus annotation projects attempted to date. The two domains covered (protein-protein interactions and tissue expression) are both of crucial importance to biologists. Although there are several corpora already available with annotations of PPI, most of these only include protein annotation, and do not include the range of entities and normalisations available in the ITI TXM corpora. There are few available annotated corpora addressing tissue expression, and we are unaware of any large-scale efforts whose main focus is that domain.

Another interesting aspect of the ITI TXM corpora is the annotation of normalisations for multiple types of entity mentions, and for multiple species. This annotation was motivated by the role of the NLP system, as an assistant to curators, as it was suspected that mapping proteins, genes and other terms to standard databases occupied a significant proportion of curators' time. The annotation of multi-species normalisations was difficult in situations where it was unclear which species was being referred to for a given named entity mention. These issues were resolved by deriving a series of annotation guidelines, as detailed in Section 4.2. The annotation guidelines were also reasonably successful in ensuring annotator consistency, as evidenced by the normalisation IAA provided in Section 4.4.

During the annotation we found that the interaction between the NLP team and the biologists was essential at all

stages. In the design phase, the biologists, as the domain experts, provided insight into what information should be annotated. At the same time, the NLP team were able to explain to the biologists what their technology is capable of. However, although both parties have an insight into what can be reliably annotated, the only sure way to determine this is empirically through extensive piloting. The piloting phase not only provided experimental data on annotation agreement and timing, but also helped the NLP team and the biologists to improve their shared understanding of the annotation process and its difficulties. During the main annotation phase, it was helpful to have regular contact between the NLP and the annotation teams in order to ensure that doubts and difficulties were noted, discussed and resolved as quickly as possible. The NLP team analysed the data as it was produced by the annotators and drew their attention to any recurring sources of disagreement.

We believe that measuring IAA is a crucial part of any corpus annotation effort. It provides a check that the annotators are producing a reliable and consistent corpus. It also gives a measure of how difficult the task is and suggests how well an automated system can be expected to perform. We took steps to ensure that the IAA itself was reliable, by instructing annotators not to discuss papers whilst annotating them. We also did not inform annotators in advance whether they were working on a paper that was also being annotated by another person. The IAA measurements for the final set of markables shows that some proved difficult to annotate reliably, for example the GOMOP entity and some of the attributes. Annotating them was problematic in the piloting phase, and whilst we attempted to tighten up the guidelines, it was not sufficient to boost their IAA.

We hope that the two ITI TXM corpora, consisting of over 200 papers each, and with multiple types of semantic annotation, will provide a useful resource for the biomedical text-mining community when released to the academic research community later this year.

6 Acknowledgements

The ITI TXM corpora were created as part of an ITI Life Sciences Scotland (`http://www.itilifesciences.com`) research programme with Cognia EU and the University of Edinburgh. We would like to thank the annotation teams led by Elizabeth Fairley and Lynn Morrice as well as Cognia EU's software development team. Thanks are also due to Malvina Nissim, Kirsten Lillie and Henk Harkema for their help with devising the annotation guidelines and deciding on sets of markables.

7 References

Beatrice Alex, Malvina Nissim, and Claire Grover. 2006. The impact of annotation on the performance of protein tagging in biomedical text. In *Proceedings of LREC*.

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

Jean Carletta, David McKelvie, Amy Isard, Andreas Mengel, Marion Klein, and Morton Baun Møller. 2005. A generic approach to software support for linguistic annotation using XML. In Geoffrey Sampson and Diana McCarthy , editors, *Readings in Corpus Linguistics*. Continuum International.

Kevin B. Cohen, Lynne Fox, Philip V. Ogren, and Lawrence Hunter. 2005. Corpus design for biomedical natural language processing. In *Proceedings of ISMB*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

FetchProt, 2005. *The FetchProt Corpus: documentation and annotation guidelines*. Available online at: `http://fetchprot.sics.se`.

Claire Grover, Michael Matthews, and Richard Tobin. 2006. Tools to address the interdependence between tokenisation and standoff annotation. In *Proceedings of NLPXML*.

Martin Krallinger, Rainer Malik, and Alfonso Valencia. 2006. Text mining and protein annotations: the construction and use of protein description sentences. *Genome Inform*, 17(2):121–130.

Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan Mcdonald, Martha Palmer, Andrew Schein, Lyle Ungar, Scott Winters, and Pete White. 2004. Integrated annotation for biomedical information extraction. In *Proceedings of the BioLINK*.

Zhiyong Lu, Michael Bada, Philip V. Ogren, K. Bretonnel Cohen, and Lawrence Hunter. 2006. Improving biomedical corpus annotation guidelines. In *Proceedings of the Joint BioLINK and 9th Bio-Ontologies Meeting*.

Inderjeet Mani, Zhangzhi Hu, Seok Bae Jang, Ken Samuel, Matthew Krause, Jon Phillips, and Cathy H. Wu. 2005. Protein name tagging guidelines: lessons learned. *Comparative and Functional Genomics*, 6(1-2):72–76.

Tara McIntosh and James R. Curran. 2007. Challenges for extracting biomedical knowledge from full text. In *Proceedings of BioNLP*.

Claire Nedellec. 2005. Learning language in logic - genic interaction extraction challenge. In *Proceedings of the ICML Workshop on Learning Language in Logic*.

Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. 2002. GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of HLT*.

Scott Piao, Ekaterina Buyko, Yoshimasa Tsuruoka, Katrin Tomanek, Jin-Dong Kim, John McNaught, Udo Hahn, and Sophia Ananiadou. 2007. BootStrep annotation scheme - encoding information for text mining. Proceedings of the 4th Corpus Linguistics Conference.

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(1).

Parantu K. Shah, Carolina Perez-Iratxeta, Peer Bork, and Miguel A. Andrade. 2003. Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics*, 4(20).

Hagit Shatkay and Ronen Feldman. 2003. Mining the biomedical literature in the genomic era: an overview. *Journal of Computational Biology*, 10(6):821–855.

Lorraine Tanabe, Natalie Xie, Lynne H. Thom, Wayne Matten, and W. John Wilbur. 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6 Suppl 1.

GENIA Treebank, 2005. *GENIA Treebank Beta Version*. Available online at: `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/topics/Corpus/GTB.html`.

John W. Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7(1).