Corpus design for biomedical natural language processing

K. Bretonnel Cohen
Center for Computational Pharmacology
U. of Colorado School of Medicine
Aurora, Colorado
kevin.cohen@gmail.com

Lynne Fox
Denison Library
U. of Colorado Health Sciences Center
Denver, Colorado
lynne.fox@uchsc.edu

Philip V. Ogren
Center for Computational Pharmacology
U. of Colorado School of Medicine
Aurora, Colorado
philip.ogren@uchsc.edu

Lawrence Hunter
Center for Computational Pharmacology
U. of Colorado School of Medicine
Aurora, Colorado
larry.hunter@uchsc.edu

Abstract

This paper classifies six publicly available biomedical corpora according to various corpus design features and characteristics. We then present usage data for the six corpora. We show that corpora that are carefully annotated with respect to structural and linguistic characteristics and that are distributed in standard formats are more widely used than corpora that are not. These findings have implications for the design of the next generation of biomedical corpora.

1 Introduction

A small number of data sets for evaluating the performance of biomedical language processing (BLP) systems on a small number of task types have been made publicly available by their creators (Blaschke et al. 1999[1], Craven and Kumlein 1999[2], Pustejovsky et al. 2002[3], Franzén et al. 2002[4], Collier et al. 1999[5], Tanabe et al. 2005[6]). From a biological perspective, a number of these corpora (PDG, GENIA, Medstract, Yapex, GENETAG) are exceptionally well curated. From the perspective of system evaluation, a number of these corpora (Wisconsin, GENETAG) are very well designed, with large numbers of both positive and negative examples for system training and testing. Despite the positive attributes of all of these corpora, they vary widely in their external usage rates: some of them have been found very useful in the natural language processing community outside of the labs that created them, as evinced by their high rates of usage in system construction and evaluation in the years since they have been released. In contrast, others have seen little or no use in the community at large. These data sets provide us with an opportunity to evaluate the consequences of a variety of approaches to biomedical corpus construction. We examine these corpora with respect to a number of design features and other characteristics, and look for features that characterize widely used—and infrequently used—corpora. Our findings have implications for how the next generation of biomedical corpora should be constructed, and for how the existing corpora can be modified to make them more widely useful.

2 Materials and methods

Table 1 lists the publicly available biomedical corpora of which we are aware. We omit discussion here of the corpus currently in production by the University of Pennsylvania and the Children's Hospital of Philadelphia (Kulick et al. 2004), since it is not yet available in finished form. We also omit text collections from our discussion. By *text collection* we mean textual data sets that may include metadata about documents, but do not contain mark-up of the document contents. So, the OHSUMED text collec-

[1] We refer to this corpus as the Protein Design Group (PDG) corpus.

[2] We refer to this as the University of Wisconsin corpus.

[3] The Medstract corpus.

[4] The Yapex corpus.

[5] The GENIA corpus.

[6] Originally the BioCreative Task 1A data set, now known as the GENETAG corpus.

Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, pages 38–45, Detroit, June 2005.

Table 1: Name, date, genre, and size for the six corpora. Size is in words.

Name	date	genre	size
PDG	1999	Sentences	10,291
Wisconsin	1999	Sentences	1,529,731
GENIA	1999	Abstracts	432,560
MEDSTRACT	2001	Abstracts	49,138
Yapex	2002	Abstracts	45,143
GENETAG	2004	Sentences	342,574

Table 2: Low- and high-level tasks to which the six corpora are applicable. SS is sentence segmentation, T is tokenization, and POS is part-of-speech tagging. EI is entity identification, IE is information extraction, A is acronym/abbreviation definition, and C is coreference resolution.

Name	SS	T	POS	EI	IE	A	C
PDG				•	•		
Wisconsin				•	•		
GENIA	•	•	•	•			
Medstract				•		•	•
Yapex				•			
GENETAG				•			

tion (Hersh et al. 1994) and the TREC Genomics track data sets (Hersh and Bhupatiraju 2003, Hersh et al. 2004) are excluded from this work, although their utility in information retrieval is clear.

Table 1 lists the corpora, and for each corpus, gives its release date (or the year of the corresponding publication), the genre of the contents of the corpus, and the size of the corpus[7].

The left-hand side of Table 2 lists the data sets and, for each one, indicates the lower-level general language processing problems that it could be applied to, either as a source of training data or for evaluating systems that perform these tasks. We considered here sentence segmentation, word tokenization, and part-of-speech (POS) tagging.

The right-hand side of Table 2 shows the higher-level tasks to which the various corpora can be applied. We considered here entity identification, information (relation) extraction, abbreviation/acronym definition, and coreference resolution. (Information retrieval is approached via text collections, versus corpora.) These tasks are directly related to the types of semantic annotation present in each corpus. The three EI-only corpora (GENIA, Yapex, GENETAG) are annotated with semantic classes of relevance to the molecular biology domain. In the case of the Yapex and GENETAG corpora, this annotation uses a single semantic class, roughly equivalent to the gene or gene product. In the case of the GENIA corpus, the annotation reflects a more sophisticated, if not widely used, ontology. The Medstract corpus uses multiple semantic classes, including *gene*, *protein*, *cell type*, and *molecular process*. In all of these cases, the semantic annotation was carefully curated, and in one (GENETAG) it includes alternative analyses. Two of the corpora (PDG, Wisconsin) are indicated in Table 2 as being applicable to both entity identification and information extraction tasks. From a biological perspective, the PDG corpus has exceptionally well-curated positive examples. From a linguistic perspective, it is almost unannotated. For each sentence, the entities are listed, but their locations in the text are not indicated, making them applicable to some definitions of the entity identification task but not others. The Wisconsin corpus contains both positive and negative examples. For each example, entities are listed in a normalized form, but without clear pointers to their locations in the text, making this corpus similarly difficult to apply to many definitions of the entity identification task.

The Medstract corpus is unique among these in being annotated with coreferential equivalence sets, and also with acronym expansions.

All six corpora draw on the same subject matter domain—molecular biology—but they vary widely with respect to their level of semantic restriction within that relatively broad category. One (GENIA) is restricted to the subdomain of human blood cell transcription factors. Another (Yapex) combines data from this domain with abstracts on protein binding in humans. The GENETAG corpus is considerably broader in topic, with all of PubMed/MEDLINE serving as a potential data

[7] Sizes are given in words. Published descriptions of the corpora don't generally give size in words, so this data is based on our own counts. See the web site at http://compbio.uchsc.edu/corpora for details on how we did the count for each corpus.

Table 3: External usage rates. The *systems* column gives the count of the number of systems that actually used the dataset, as opposed to publications that cited the paper but did not use the data itself. *Age* is in years as of 2005.

Name	age	systems
GENIA	6	21
GENETAG	1	8
Yapex	3	6
Medstract	4	3
Wisconsin	6	1
PDG	6	0

source. The Medstract corpus contains biomedical material not apparently related to molecular biology. The PDG corpus is drawn from a very narrow subdomain on protein-protein interactions. The Wisconsin corpus is composed of data from three separate sub-domains: protein-protein interactions, subcellular localization of proteins, and gene/disease associations.

Table 3 shows the number of systems *built outside of the lab that created the corpus* that used each of the data sets described in Tables 1 and 2. The counts in this table reflect work that actually used the datasets, versus work that cites the publication that describes the data set but doesn't actually use the data set. We assembled the data for these counts by consulting with the creators of the data sets and by doing our own literature searches[8]. If a system is described in multiple publications, we count it only once, so the number of systems is slightly smaller than the number of publications.

3 Results

Even without examining the external usage data, two points are immediately evident from Tables 1 and 2:

- Only one of the currently publicly available corpora (GENIA) is suitable for evaluating performance on basic preprocessing tasks.
- These corpora include only a very limited range of genres: only abstracts and roughly sentence-sized inputs are represented.

Examination of Table 3 makes another point immediately clear. The currently publicly available corpora fall into two groups: ones that have had a number of external applications (GENIA, GENETAG, and Yapex), and ones that have not (Medstract, Wisconsin, and PDG). We now consider a number of design features and other characteristics of these corpora that might explain these groupings[9].

3.1 Effect of age

We considered the very obvious hypothesis that it might be length of time that a corpus has been available that determines the amount of use to which it has been put. (Note that we use the terms "hypothesis" and "effect" in a non-statistical sense, and there is no significance-testing in the work reported here.) Tables 1 and 3 show clearly that this is not the case. The age of the PDG, Wisconsin, and GENIA data is the same, but the usage rates are considerably different—the GENIA corpus has been much more widely used. The GENETAG corpus is the newest, but has a relatively high usage rate. Usage of a corpus is determined by factors other than the length of time that it has been available.

3.2 Effect of size

We considered the hypothesis that size might be the determinant of the amount of use to which a corpus is put—perhaps smaller corpora simply do not provide enough data to be helpful in the development and validation of learning-based systems. We can

[8] In the cases of the two corpora for which we found only zero or one external usage, this search was repeated by an experienced medical librarian, and included reviewing 67 abstracts or full papers that cite Blaschke et al. (1999) and 37 that cite Craven and Kumlein (1999).

[9] Three points should be kept in mind with respect to this data. First, although the sample includes all of the corpora that we are aware of, it is small. Second, there is a variety of potential confounds related to sociological factors which we are aware of, but do not know how to quantify. One of these is the effect of association between a corpus and a shared task. This would tend to increase the usage of the corpus, and could explain the usage rates of GENIA and GENETAG, although not that of Yapex. Another is the effect of association between a corpus and an influential scientist. This might tend to increase the usage of the corpus, and could explain the usage rate of GENIA, although not that of GENETAG. Finally, there may be interactions between any of these factors, or as a reviewer pointed out, there may be a separate explanation for the usage rate of each corpus in this study. Nevertheless, the analysis of the quantifiable factors presented above clearly provides useful information about the design of successful corpora.

reject this hypothesis: the Yapex corpus is one of the smallest (a fraction of the size of the largest, and only roughly a tenth of the size of GENIA), but has achieved fairly wide usage. The Wisconsin corpus is the largest, but has a very low usage rate.

3.3 Effect of structural and linguistic annotation

We expected a priori that the corpus with the most extensive structural and linguistic annotation would have the highest usage rate. (In this context, by *structural annotation* we mean tokenization and sentence segmentation, and by *linguistic annotation* we mean POS tagging and shallow parsing.) There isn't a clear-cut answer to this.

The GENIA corpus is the only one with curated structural and POS annotation, and it has the highest usage rate. This is consistent with our initial hypothesis.

On the other hand, the Wisconsin corpus could be considered the most "deeply" linguistically annotated, since it has both POS annotation and—unique among the various corpora—shallow parsing. It nevertheless has a very low usage rate. However, the comparison is not clearcut, since both the POS tagging and the shallow parsing are fully automatic and not manually corrected. (Additionally, the shallow parsing and the tokenization on which it is based are somewhat idiosyncratic.) It *is* clear that the Yapex corpus has relatively high usage despite the fact that it is, from a linguistic perspective, very lightly annotated (it is marked up for entities only, and nothing else). To our surprise, structural and linguistic annotation do not appear to uniquely determine usage rate.

3.4 Effect of format

Annotation format has a large effect on usage. It bears repeating that these six corpora are distributed in six different formats—even the presumably simple task of populating the *Size* column in Table 1 required writing six scripts to parse the various data files. The two lowest-usage corpora are annotated in remarkably unique formats. In contrast, the three more widely used corpora are distributed in relatively more common formats. Two of them (GENIA and Yapex) are distributed in XML, and one of them (GENIA) offers a choice for POS tagging information between XML and the whitespace-separated, one-token-followed-by-tags-per-line format that is common to a number of POS taggers and parsers. The third (GENETAG) is distributed in the widely used slash-attached format (e.g. *sense/NN*).

3.5 Effect of semantic annotation

The data in Table 2 and Table 3 are consistent with the hypothesis that semantic annotation predicts usage. The claim would be that corpora that are built specifically for entity identification purposes are more widely used than corpora of other types, presumably due to a combination of the importance of the entity identification task as a prerequisite to a number of other important applications (e.g. information extraction and retrieval) and the fact that it is still an unsolved problem. There may be some truth to this, but we doubt that this is the full story: there are large differences in the usage rates of the three EI corpora, suggesting that semantic annotation is not the only relevant design feature. If this analysis is in fact correct, then certainly we should see a reduction in the use of all three of these corpora once the EI problem is solved, unless their semantic annotations are extended in new directions.

3.6 Effect of semantic domain

Both the advantages and the disadvantages of restricted domains as targets for language processing systems are well known, and they seem to balance out here. The scope of the domain does not affect usage: both the low-use and higher-use groups of corpora contain at least one highly restricted domain (GENIA in the high-use group, and PDG in the low-use group) and one broader domain (GENETAG in the high-use group, and Wisconsin in the lower-use group).

4 Discussion

The data presented in this paper show clearly that external usage rates vary widely for publicly available biomedical corpora. This variability is not related to the biological relevance of the corpora—the PDG and Wisconsin corpora are clearly of high biological relevance as evinced by the number of systems that have tackled the information extraction tasks that they are meant to support. Additionally, from a biological perspective, the quality of the data in the

PDG corpus is exceptionally high. Rather, our data suggest that basic issues of distribution format and of structural and linguistic annotation seem to be the strongest predictors of how widely used a biomedical corpus will be. This means that as builders of of data sources for BLP, we can benefit from the extensive experience of the corpus linguistics world. Based on that experience, and on the data that we have presented in this paper, we offer a number of suggestions for the design of the next generation of biomedical corpora.

We also suggest that the considerable investments already made in the construction of the less-frequently-used corpora can be protected by modifying those corpora in accordance with these suggestions.

Leech (1993) and McEnery and Wilson (2001), coming from the perspective of corpus linguistics, identify a number of definitional issues and design maxims for corpus construction. Some of these are quite relevant to the current state of biomedical corpus construction. We frame the remainder of our discussion in terms of these issues and maxims.

4.1 Level of annotation

From a definitional point of view, annotation is one of the distinguishing points of a corpus, as opposed to a text collection. Perhaps the most salient characteristic of the currently publicly available corpora is that from a linguistic or language processing perspective, with the exception of GENIA and GENETAG, they are barely annotated at all. For example, although POS tagging has possibly been the sine qua non of the usable corpus since the earliest days of the modern corpus linguistic age, five of the six corpora listed in Table 2 either have no POS tagging or have only automatically generated, uncorrected POS tags. The GENIA corpus, with its carefully curated annotation of sentence segmentation, tokenization, and part-of-speech tagging, should serve as a model for future biomedical corpora in this respect. It is remarkable that with just these levels of annotation (in addition to its semantic mark-up), the GENIA corpus has been applied to a wide range of task types other than the one that it was originally designed for. Eight papers from COLING 2004 (Kim et al. 2004) used it for evaluating entity identification tasks. Yang et al. (2002) adapted a subset of the corpus for use in developing and testing a coreference resolution system. Rinaldi et al. (2004) used it to develop and test a question-answering system. Locally, it has been used in teaching computational corpus linguistics for the past two years. We do not claim that it has not required extension for some of these tasks—our claim is that it is its annotation on these structural and linguistic levels, in combination with its format, that has made these extensions practical.

4.1.1 Formatting choices and formatting standardization

A basic desideratum for a corpus is *recoverability:* it should be possible to map from the annotation to the raw text. A related principle is that it should be easy for the corpus user to extract all annotation information from the corpus, e.g. for external storage and processing: "in other words, the annotated corpus should allow the maximum flexibility for manipulation by the user" (McEnery and Wilson, p. 33). The extent to which these principles are met is a function of the annotation format. The currently available corpora are distributed in a variety of one-off formats. Working with any one of them requires learning a new format, and typically writing code to access it. At a minimum, none of the non-XML corpora meet the recoverability criterion. None[10] of these corpora are distributed in a standoff annotation format. *Standoff annotation* is the strategy of storing annotation and raw text separately (Leech 1993). Table 4 contrasts the two. Non-standoff annotation at least obscures—more frequently, destroys—important aspects of the structure of the text itself, such as which textual items are and are not immediately adjacent. Using standoff annotation, there is no information loss whatsoever. Furthermore, in the standoff annotation strategy, the original input text is immediately available in its raw form. In contrast, in the non-standoff annotation strategy, the original must be retrieved independently or recovered from the annotation (if it is recoverable at all). The standoff annotation strategy was relatively new at the time that most of the corpora in Table 1 were designed, but by now has become easy to implement, in part

[10]The semantic annotation of the GENETAG corpus is in a standoff format, but neither the tokenization nor the POS tagging is.

Table 4: Contrasting standoff and non-standoff annotation

Raw text
MLK2 has a role in vesicle formation
Non-standoff annotation
MLK2/NN has/VBZ a/DT role/NN in/IN vesicle/NN formation/NN
Standoff annotation
<POS="NN" start=0 end=3> <POS="VBZ" start=5 end=7> <POS="DT" start=9 end=9> <POS="NN" start=11 end=14> <POS="IN" start=16 end=17> <POS="NN" start=19 end=25> <POS="NN" start=27 end=35>

due to the availability of tools such as the University of Pennsylvania's WordFreak (Morton and LaCivita 2003).

Crucially, this annotation should be based on character offsets, avoiding a priori assumptions about tokenization. See Smith et al. (2005) for an approach to refactoring a corpus to use character offsets.

4.1.2 Guidelines

The maxim of *documentation* suggests that annotation guidelines should be published. Further, basic data on who did the annotations and on their level of agreement should be available. The currently available datasets mostly lack assessments of inter-annotator agreement, utilize a small or unspecified number of annotators, and do not provide published annotation guidelines. (We note the Yang et al. (2002) coreference annotation guidelines, which are excellent, but the corresponding corpus is not publicly available.) This situation can be remedied by editors, who should insist on publication of all of these. The GENETAG corpus is notable for the detailed documentation of its annotation guidelines. We suspect that the level of detail of these guidelines contributed greatly to the success of some rule-based approaches to the EI task in the BioCreative competition, which utilized an early version of this corpus.

4.1.3 Balance and representativeness

Corpus linguists generally strive for a well-structured stratified sample of language, seeking to "balance" in their data the representation of text types, different sorts of authors, and so on. Within the semantic domain of molecular biology texts, an important dimension on which to balance is the genre or text type.

As is evident from Table 1, the extant datasets draw on a very small subset of the types of genres that are relevant to BLP: we have not done a good job yet of observing the principle of balance or representativeness. The range of genres that exist in the research (as opposed to clinical) domain alone includes abstracts, full-text articles, GeneRIFs, definitions, and books. We suggest that all of these should be included in future corpus development efforts.

Some of these genres have been shown to have distinguishing characteristics that are relevant to BLP. Abstracts and isolated sentences from them are inadequate, and also unsuited to the opportunities that are now available to us for text data mining with the recent announcement of the NIH's new policy on availability of full-text articles (NIH 2005). This policy will result in the public availability of a large and constantly growing archive of current, full-text publications. Abstracts and sentences are inadequate in that experience has shown that significant amounts of data are not found in abstracts at all, but are present only in the full texts of articles, sometimes not even in the body of the text itself, but rather in tables and figure captions (Shatkay and Feldman 2003). They are not suited to the upcoming opportunities in that it is not clear that practicing on abstracts will let us build the necessary skills for dealing with the flood of full-text articles that PubMedCentral is poised to deliver to us. Furthermore, there are other types of data—GeneRIFs and domain-specific dictionary definitions, for instance—that are fruitful sources of biological knowledge, and which may actually be easier to process automatically than abstracts. Space does not permit justifying the importance of all of these genres, but we discuss the rationale for including full text at some length due to the recent NIH announcement and due to the large body of evidence that can currently be brought to bear on the issue. A growing body of recent research makes

two points clear: full-text articles are different from abstracts, and full-text articles must be tapped if we are to build high-recall text data mining systems.

Corney et al. (2004) looked directly at the effectiveness of information extraction from full-text articles versus abstracts. They found that recall from full-text articles was more than double that from abstracts. Analyzing the relative contributions of the abstracts and the full articles, they found that more than half of the interactions that they were able to extract were found in the full text and were absent in the abstract.

Tanabe and Wilbur (2002) looked at the performance on full-text articles of an entity identification system that had originally been developed and tested using abstracts. They found different false positive rates in the Methods sections compared to other sections of full-text articles. This suggests that full-text articles, unlike abstracts, will require parsing of document structure. They also noted a range of problems related to the wider range of characters (including, e.g., superscripts and Greek letters) that occurs in full-text articles, as opposed to abstracts.

Schuemie et al. (2004) examined a set of 3902 full-text articles from *Nature Genetics* and BioMed Central, along with their abstracts. They found that about twice as many MeSH concepts were mentioned in the full-text articles as in the abstracts. They also found that full texts contained a larger number of unique gene names than did abstracts, with an average of 2.35 unique gene names in the full-text articles, but an average of only 0.61 unique gene names in the abstracts.

It seems clear that for biomedical text data mining systems to reach anything like their full potential, they will need to be able to handle full-text inputs. However, as Table 1 shows, no publicly available corpus contains full-text articles. This is a deficiency that should be remedied.

5 Conclusion

5.1 Best practices in biomedical corpus construction

We have discussed the importance of recoverability, publication of guidelines, balance and representativeness, and linguistic annotation. Corpus maintenance is also important. Bada et al. (2004) point out the role that an organized and responsive maintenance plan has played in the success of the Gene Ontology. It seems likely that the continued development and maintenance reflected in the three major releases of GENIA (Ohta et al. 2002, Kim et al. 2003) have contributed to its improved quality and continued use over the years.

5.2 A testable prediction

We have interpreted the data on the characteristics and usage rates of the various datasets discussed in this paper as suggesting that datasets that are developed in accordance with basic principles of corpus linguistics are more useful, and therefore more used, than datasets that are not.

A current project at the University of Pennsylvania and the Children's Hospital of Philadelphia (Kulick et al. 2004) is producing a corpus that follows many of these basic principles. We predict that this corpus will see wide use by groups other than the one that created it.

5.3 The next step: grounded references

The logical next step for BLP corpus construction efforts is the production of corpora in which entities and concepts are grounded with respect to external models of the world (Morgan et al. 2004).

The BioCreative Task 1B data set construction effort provides a proof-of-concept of the plausibility of building BLP corpora that are grounded with respect to external models of the world, and in particular, biological databases. These will be crucial in taking us beyond the stage of extracting information about text strings, and towards mining knowledge about known, biologically relevant entities.

6 Acknowledgements

This work was supported by NIH grant R01-LM008111. The authors gratefully acknowledge helpful discussions with Lynette Hirschman, Alex Morgan, and Kristofer Franzén, and thank Sonia Leach and Todd A. Gibson for LaTeXassistance. Christian Blaschke, Mark Craven, Lorraine Tanabe, and again Kristofer Franzén provided helpful data. We thank all of the corpus builders for their generosity in sharing their valuable resources.

References

Bada, Michael; Robert Stevens; et al. 2004. A short study on the success of the Gene Ontology. *Journal of web semantics* 1(2):235-240.

Blaschke, Christian; Miguel A. Andrade; Christos Ouzounis; and Alfonso Valencia. 1999. Automatic extraction of biological information from scientific text: protein-protein interactions. *ISMB-99*, pp. 60-67. AAAI Press.

Collier, Nigel, Hyun Seok Park, Norihiro Ogata, Yuka Tateisi, Chikashi Nobata, Takeshi Sekimizu, Hisao Imai and Jun'ichi Tsujii. 1999. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. *EACL 1999*.

Corney, David P.A.; Bernard F. Buxton; William B. Langdon; and David T. Jones. 2004. BioRAT: extracting biological information from full-length papers. *Bioinformatics* 20(17):3206-3213.

Craven, Mark; and Johan Kumlein. 1999. Constructing biological knowledge bases by extracting information from text sources. *ISMB-99*, pp. 77-86. AAAI Press.

Franzén, Kristofer; Gunnar Eriksson; Fredrik Olsson; Lars Asker Per Lidén; and Joakim Cöster. 2002. Protein names and how to find them. *International Journal of Medical Informatics*, 67(1-3), pp. 49-61.

Hersh, William; Chris Buckley; TJ Leone; and David Hickam. 1994. OHSUMED: an interactive retrieval evaluation and new large test collection for research. *SIGIR94*, pp. 192-201.

Hersh, William; and Ravi Teja Bhupatiraju. 2003. TREC genomics track overview. *TREC 2003*, pp. 14-23.

Hersh et al. 2004. TREC 2004 genomics track overview. *TREC Notebook*.

Kim, Jin-Dong; Tomoko Ohta; Yuka Tateisi; and Jun'ichi Tsujii. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics* 19(Suppl. 1):i180-i182.

Kim, Jin-Dong; Tomoko Ohta; Yoshimasa Tsuruoka; and Yuka Tateisi. 2004. Introduction to the bio-entity recognition task at JNLPBA. *Proc. international joint workshop on natural language processing in biomedicine and its applications*, pp. 70-75.

Kulick, Seth; Ann Bies; Mark Liberman; Mark Mandel; Ryan McDonald; Martha Palmer; Andrew Schein; and Lyle Ungar. 2004. Integrated annotation for biomedical information extraction. *BioLink 2004*, pp. 61-68.

Leech, G. 1993. Corpus annotation schemes. *Literary and linguistic computing* 8(4):275-281.

McEnery, Tony; and Andrew Wilson. 2001. *Corpus linguistics*, 2nd edition. Edinburgh University Press.

Morgan, Alexander A.; Lynette Hirschman; Marc Colosimo; Alexander S. Yeh; and Jeff B. Colombe. 2004. Gene name identification and normalization using a model organism database. *JBMI* 37:396-410.

Morton, Thomas; and Jeremy LaCivita. 2003. WordFreak: an open tool for linguistic annotation. *HLT/NAACL 2003: demonstrations*, pp. 17-18.

NIH (National Institutes of Health). 2005. http://www.nih.gov/news/pr/feb2005/od-03.htm

Ohta, Tomoko; Yuka Tateisi; and Jin-Dong Kim. 2002. The GENIA corpus: an annotated research abstract corpus in molecular biology domain. *HLT 2002*, pp. 73-77.

Pustejovsky, James; José Castaño; R. Saurí; A. Rumshisky; J. Zhang; and W. Luo. 2002. Medstract: creating large-scale information servers for biomedical libraries. *Proc. workshop on natural language processing in the biomedical domain*, pp. 85-92. Association for Computational Linguistics.

Rinaldi, Fabio; James Dowdall; Gerold Schneider; and Andreas Persidis. 2004. Answering questions in the genomics domain. *Proc. ACL 2004 workshop on question answering in restricted domains*, pp. 46-53.

Schuemie, M.J.; M. Weeber; B.J. Schijvenaars; E.M. van Mulligen; C.C. van der Eijk; R. Jelier; B. Mons; and J.A. Kors. 2004. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics* 20(16):2597-2604.

Shatkay, Hagit; and Ronen Feldman. 2003. Mining the biomedical literature in the genomic era: an overview. *Journal of computational biology* 10(6):821-855.

Smith, Lawrence H.; Lorraine Tanabe; Thomas Rindflesch; and W. John Wilbur. 2005. MedTag: a collection of biomedical annotations. *BioLINK 2005*, this volume.

Tanabe, Lorraine; and L. John Wilbur. 2002. Tagging gene and protein names in full text articles. *Proc. ACL workshop on natural language processing in the biomedical domain*, pp. 9-13.

Tanabe, Lorraine; 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):S3.

Yang, Xiaofeng; Guodong Zhou; Jian Su; and Chew Lim Tan. Improving noun phrase coreference resolution by matching strings. 2002. *IJCNLP04*, pp. 326-333.