# Recognizing sublanguages in scientific journal articles through closure properties

**Irina P. Temnikova**
Linguistic Modelling Laboratory
Bulgarian Academy of Sciences
irina.temnikova@gmail.com

**K. Bretonnel Cohen**
Computational Bioscience Program
University of Colorado School of Medicine
Department of Linguistics
University of Colorado at Boulder
kevin.cohen@gmail.com

## Abstract

It has long been realized that sublanguages are relevant to natural language processing and text mining. However, practical methods for recognizing or characterizing them have been lacking. This paper describes a publicly available set of tools for sublanguage recognition. Closure properties are used to assess the goodness of fit of two biomedical corpora to the sublanguage model. Scientific journal articles are compared to general English text, and it is shown that the journal articles fit the sublanguage model, while the general English text does not. A number of examples of implications of the sublanguage characteristics for natural language processing are pointed out. The software is made publicly available at [edited for anonymization].

## 1 Introduction

### 1.1 Definitions of "sublanguage"

The notion of *sublanguage* has had varied definitions, depending on the aspects of sublanguages on which the authors focused. (Grishman and Kittredge, 1986) focus on syntactic aspects of sublanguages: "...the term suggests a subsystem of language...limited in reference to a specific subject domain. In particular, each sublanguage has a distinctive grammar, which can profitably be described and used to solve specific language-processing problems" (Grishman and Kittredge, 1986).

(Kittredge, 2003) focuses on the spontaneous appearance of sublanguages in restricted domains, where the preconditions for a sublanguage to appear are the sharing of specialized knowledge about a restricted semantic domain and recurrent "situations" (e.g. scientific journal articles, or discharge summaries) in which domain experts communicate. According to (Kittredge, 2003), characteristics of a sublanguage include a restricted lexicon, relatively small number of lexical classes, restricted sentence syntax, deviant sentence syntax, restricted word co-occurrence patterns, and different frequencies of occurrence of words and syntactic patterns from the normal language.

(McDonald, 2000) focuses on the element of restriction in sublanguages—the notion that they are restricted to a specialized semantic domain, a very "focused" audience, and "stipulated content," with the effect that both word choice and syntactic style have reduced options as compared to the normal language.

The notions of restriction that recur in these definitions of "sublanguage" lead directly to (McEnery and Wilson, 2001)'s notion of using the quantification of closure properties to assess whether or not a given sample of a genre of language use fits the sublanguage model. *Closure* refers to the tendency of a genre of language towards finiteness at one or more linguistic levels. For example, a genre of language might or might not use a finite set of lexical items, or have a finite set of sentence structures. Notions of restriction suggest that a sublanguage should tend towards closure on at least some linguistic levels. To quantify closure, we can examine relationships between types and tokens in a corpus of the genre. In particular, we count the number of types that are observed as an increasing number of tokens is examined. If a genre does not exhibit closure, then the number of types will continue to rise continually as the number of tokens increases. On the other hand, closure is demonstrated when the number of types stops growing after some number of tokens has been examined.

*Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013)*, pages 72–79,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

### 1.2 Relevance of sublanguages to natural language processing

The relevance of sublanguages to natural language processing has long been recognized in a variety of fields. (Hirschman and Sager, 1982) and (Friedman, 1986) show how a sublanguage–based approach can be used for information extraction from clinical documents. (Finin, 1986) shows that sublanguage characterization can be used for the notoriously difficult problem of interpretation of nominal compounds. (Sager, 1986) asserts a number of uses for sublanguage–oriented natural language processing, including resolution of syntactic ambiguity, definition of frames for information extraction, and discourse analysis. (Sekine, 1994) describes a prototype application of sublanguages to speech recognition. (Friedman et al., 1994) uses a sublanguage grammar to extract a variety of types of structured data from clinical reports. (McDonald, 2000) points out that modern language generation systems are made effective in large part due to the fact that they are applied to specific sublanguages. (Somers, 2000) discusses the relevance of sublanguages to machine translation, pointing out that many sublanguages can make machine translation easier and some of them can make machine translation harder. (Friedman et al., 2001) uses a sublanguage grammar to extract structured data from scientific journal articles.

### 1.3 Previous work on sublanguage recognition

Various approaches have been taken to recognizing sublanguages. We posit here two separate tasks—recognizing a sublanguage when one is present, and determining the characteristics of a sublanguage. Information-theoretic approaches have a long history. (Sekine, 1994) clustered documents and then calculated the ratio of the perplexity of the clustered documents to the perplexity of a random collection of words. (Somers, 1998) showed that texts drawn from a sublanguage corpus have low weighted cumulative sums. (Stetson et al., 2002) used relative entropy and squared chi-square distance to identify a sublanguage of cross-coverage notes. (Mihaila et al., 2012) looked at distributions of named entities to identify and differentiate between a wide variety of scientific sublanguages.

Non-information-theoretic, more heuristic methods have been used to identify sublanguages, as well. In addition to the information-theoretic measures described above, (Stetson et al., 2002) also looked at such measures as length, incidence of abbreviations, and ambiguity of abbreviations. (Friedman et al., 2002) use manual analysis to detect and characterize two biomedical sublanguages. (McEnery and Wilson, 2001) examine closure properties; their approach is so central to the topic of this paper that we will describe it in some length separately.

(McEnery and Wilson, 2001) examined the closure properties of three linguistic aspects of their material under study. As materials they used two corpora that were assumed not to meet the sublanguage model—the Canadian Hansard corpus, containing proceedings from the Canadian Parliament, and the American Printing House for the Blind corpus, made up of works of fiction. As a corpus that was suspected to meet the sublanguage model, they used a set of manuals from IBM. All three corpora differed in size, so they were sampled to match the size of the smallest corpus, meaning that all experiments were done on collections 200,000 words in size. The materials under study were evaluated for their closure properties at three linguistic levels. At the most basic level, they looked at lexical items—simple word forms. The hypothesis here was that the non-sublanguage corpora would not tend toward finiteness, i.e. would not reach closure. That is, if the number of word types found was graphed as an increasing number of tokens was examined, the resulting line would grow continually and would show no signs of asymptoting. In contrast, the sublanguage corpus would eventually reach closure, i.e. would stop growing appreciably in size as more tokens were examined.

The next level that they examined was the morphosyntactic level. In particular, they looked at the number of part-of-speech tags per lexical type. Here the intuition was that if the lexicon of the sublanguage is limited, then words might be coerced into a greater number of parts of speech. This would be manifested by a smaller overall number of unique word/part-of-speech tag combinations. Again, we would expect to see that the sublanguage corpus would have a smaller number of word/part-of-speech tag combinations, as compared to the non-sublanguage corpus. Graphing the count of word type/POS tag sets on the *y* axis and the cumulative number of tokens examined on the *x* axis, we would see slower growth and lower numbers overall.

The final level that they examined was the syntactic level. In this case, parse tree types were graphed against the number of sentences examined. The intuition here is that if the sublanguage exhibits closure properties on the syntactic level, then the growth of the line will slow and we will see lower numbers overall.

(McEnery and Wilson, 2001) found the hypotheses regarding closure to be substantiated at all levels. We will not reproduce their graphs, but will summarize their findings in terms of ratios. On the lexical level, they found type/token ratios of 1:140 for the IBM manuals (the assumed sublanguage), 1:53 for the Hansard corpus (assumed not to represent a sublanguage), and 1:17 for the American Printing House for the Blind corpus (also assumed not to represent a sublanguage). The IBM manuals consist of a much smaller number of words which are frequently repeated.

At the morphosyntactic level, they found 7,594 type/POS sets in the IBM manuals, 18,817 in the Hansard corpus, and 11,638 in the American Printing House for the Blind corpus–a much smaller number in the apparent sublanguage than in the non-sublanguage corpora. The word/part-of-speech tag averages coincided with the expected findings given these number of types. The averages were 3.19 for the IBM manuals, 2.45 for the Hansard corpus, and 2.34 for the American Printing House for the Blind corpus.

At the syntactic level, they found essentially linear growth in the number of sentence types as the number of sentence tokens increased in the two non-sublanguage corpora—the ratio of sentence types to sentences in these corpora were 1:1.07 for the Hansard corpus and 1:1.02 for the American Printing House for the Blind corpus. In contrast, the growth of sentence types in the IBM manuals was not quite linear. It grew linearly to about 12,000 sentences, asymptoted between 12,000 and 16,000, and then grew essentially linearly but at a somewhat slower rate from 16,000 to 30,000 sentences. The ratio of sentence types to sentence tokens in the IBM manuals was 1:1.66—markedly higher than in the other two corpora.

### 1.4 Hypotheses tested in the paper

The null hypothesis is that there will be no difference in closure properties between the general English corpus and the two corpora of scientific journal articles that we examine. If the null hypothesis is not supported, then it might be deviated from in three ways. One is that the scientific corpora might show a greater tendency towards closure than the general English corpus. A second is that the general English corpus might show a greater tendency towards closure than the scientific corpora. A third is that there may be no relationship between the closure properties of the two scientific corpora, regardless of the closure properties of the general English corpus—one might show a tendency towards closure, and the other not.

## 2 Materials and Methods

### 2.1 Materials

The data under examination was drawn from three sources: the CRAFT corpus (Bada et al., 2012; Verspoor et al., 2012), the GENIA corpus (Kim et al., 2003), and a version of the British National Corpus (Leech et al., 1994) re-tagged with Connexor's Machinese parser (Järvinen et al., 2004). The CRAFT and GENIA corpora are composed of scientific journal articles, while the British National Corpus is a representative corpus comprising many different varieties of spoken and written English.

The CRAFT corpus is a collection of 97 full-text journal articles from the mouse genomics domain. It has been annotated for a variety of linguistic and semantic features; for the purposes of this study, the relevant ones were sentence boundaries, tokenization, and part of speech. We used the 70-document public release subset of the corpus, which comprises about 453,377 words.

The GENIA corpus is a collection of 1,999 abstracts of journal articles about human blood cell transcription factors. Like the CRAFT corpus, it has been annotated for a variety of linguistic and semantic features, again including sentence boundaries, tokenization, and part of speech. In the mid-2000's, the GENIA corpus was shown to be the most popular corpus for research in biomedical natural language processing (Cohen et al., 2005). We used version 3.02 of the corpus, containing about 448,843 words.

The experiment requires a corpus of general English for comparison. For this purpose, we used a subset of the British National Corpus. For purposes of representativeness, we followed the Brown corpus strategy of extracting the first 2,000 words from each article until a total of 453,377 words were reached (to match the size of the CRAFT corpus).

The size of the two data sets is far more than adequate for an experiment of this type—McEnery and Wilson were able to detect closure properties using corpora of only 200,000 words in their experiments.

## 2.2 Methods

### 2.2.1 Implementation details

To determine the closure properties of arbitrary corpora, we developed scripts that take a simple input format into which it should be possible to convert any annotated corpus. There are two input file types:

- A file containing one word and its corresponding part-of-speech tag per line. Part of speech tags can consist of multiple tokens, as they do in the BNC tag set, or of single tokens, as they do in most corpora. This file format is used as the input for the lexical closure script and the word type/POS tag script.
- A file containing a sequence of part of speech tags per line, one line per sentence. This file format is used as input for the sentence type closure script. We note that this is an extremely rough representation of "syntax," and arguably is actually asyntactic in that it does not represent constituent or dependency structure at all, but also point out that it has the advantage of being widely applicable and agnostic as to any particular theory of syntactic structure. It also increases the sensitivity of the method to sentence type differences, providing a stronger test of fit to the sublanguage model.

Two separate scripts then process one of these input files to determine lexical, type/POS, and sentence type closure properties. The output of every script is a comma-separated-value file suitable for importing into Excel or other applications for producing plots. The two scripts and our scripts for converting the BNC, CRAFT, and GENIA corpora into the input file formats will be made publicly available at [redacted for anonymization purposes]. To apply the scripts to a new corpus, the only necessary step is to write a script to convert from the corpus's original format to the simple format of the two input file types described above.

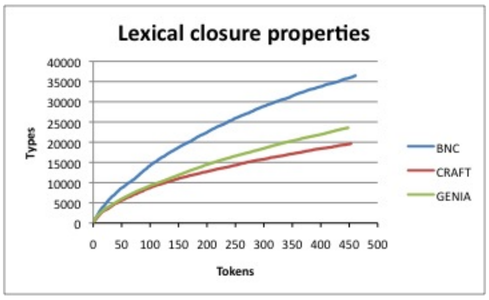

Figure 1: Lexical closure properties. Tick-marks on *x* axis indicate increments of 50,000 tokens.

### 2.2.2 Investigating closure properties

In all three cases, the number of types, whether of lexical items, lexical type/part-of-speech pair, or sentence type was counted and graphed on the *y* axis, versus the number of tokens that had been observed up to that point, which was graphed on the *x* axis. In the case of the lexical and type/POS graphs, tokens were words, and in the case of the sentence graph, "tokens" were sentences.

We then combined the lines for all three corpora and observed the total size of types, the rate of growth of the line, and whether or not there was a tendency towards asymptoting of the growth of the line, i.e. closure.

Our major deviation from the approach of (McEnery and Wilson, 2001) was that rather than parse trees, we used part-of-speech tag sequences to represent sentence types. This is suboptimal in that it is essentially asyntactic, and in that it obscures the smoothing factor of abstracting away from per-token parts of speech to larger syntactic units. However, as we point out above, it has the advantages of being widely applicable and agnostic as to any particular theory of syntactic structure, as well as more sensitive to sentence type differences.

# 3 Results

## 3.1 Lexical closure properties

Figure 1 shows the growth in number of types of lexical items as the number of tokens of lexical items increases. The British National Corpus data is in blue, the CRAFT data is in red, and the GENIA data is in green.

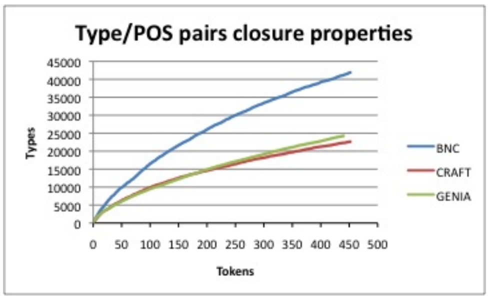

Figure 2: Type-part-of-speech tag closure properties. Tick-marks on *x* axis indicate increments of 50,000 tokens.

We note a drastic difference between the curve for the BNC and the curves for CRAFT and GENIA. The curves for CRAFT and GENIA are quite similar to each other. Overall, the curve for the BNC climbs faster and much farther, and is still climbing at a fast rate after 453,377 tokens have been examined. In contrast, the curves for CRAFT and GENIA climb more slowly, climb much less, and by the time about 50,000 tokens have been examined the rate of increase is much smaller. The increase in CRAFT and GENIA does not asymptote, as McEnery and Wilson observed for the IBM corpus. However, contrasted with the results for the BNC, there is a clear difference.

The type to token ratios for lexical items for the corpora as a whole are shown in Table 1. As the sublanguage model would predict, CRAFT and GENIA have much higher ratios than BNC.

| Corpus name | Ratio |
|---|---|
| BNC | 1: 12.650 |
| CRAFT | 1: 23.080 |
| GENIA | 1: 19.027 |

Table 1: Lexical type-to-token ratios.

### 3.2 Type/POS tag closure properties

Figure 2 shows the growth in number of type-POS tag pairs as the number of tokens of lexical item/POS tag pairs increases. The data from the different corpora corresponds to the same colors as in Figure 1.

Once again, we note a drastic difference between the curve for the BNC and the curves for CRAFT and GENIA. If anything, the differences are more pronounced here than in the case of the lexical closure graph. Again, we do not see an asymptote in the increase of the curves for CRAFT and GENIA, but there is a clear difference when contrasted with the results for the BNC.

The type-to-token sets ratios for the corpora as a whole are shown in Table 2. Again, as the sublanguage model would predict, we see much higher ratios in CRAFT and GENIA than in BNC.

| Corpus name | Ratio |
|---|---|
| BNC | 1: 10.80 |
| CRAFT | 1: 19.96 |
| GENIA | 1: 18.18 |

Table 2: Type-to-token ratios for type/POS tags.

Because the Machinese Syntax parser was used to obtain the part-of-speech tagging for BNC and the Machinese Syntax parser's tagset is much more granular and therefore larger than the CRAFT and GENIA tag sets, both of which are adaptations of the Penn treebank tag set, we considered the hypothesis that the large size differences of the tag sets were the cause of the differences observed between BNC and the two corpora of scientific journal articles. To test this hypothesis, we manually mapped the BNC tag set to the Penn treebank tag set. The result was a new BNC list of tags, of the same number and granularity as the CRAFT/GENIA ones (35-36 tags). Using this mapping, the BNC part-of-speech tags were converted to the Penn treebank tag set and the experiment was re-run. The results show that there is almost no difference between the results from the first and the second experiments. The resulting graph is omitted for space, but examining it one can observe that the differences between the three corpora in the graph are almost the same in both graphs. The newly calculated type:tokens ratio for BNC are also illustrative. They are highly similar to the type-token ratio for the original tag set—1:10.82 with the mapped data set vs. 1:10.80 with the original, much larger tag set. This supports the original results and demonstrates that differences in tag set sizes do not interfere with the identification of sublanguages.

### 3.3 Sentence type closure properties

Figure 3 shows the growth in number of sentence types as the number of sentences increases. The data from the different corpora corresponds to the same colors as in Figure 1.

Here we see that all three corpora exhibit sim-

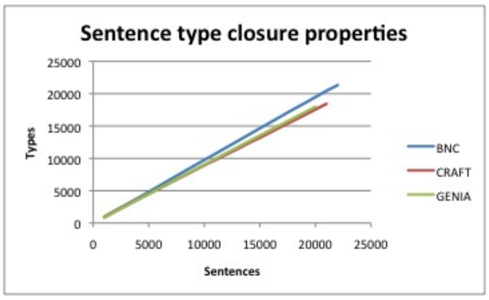

Figure 3: Sentence type closure properties. Tickmarks on *x* axis indicate increments of 5,000 sentences.

ilar curves—essentially linear, with nearly identical growth rates. This is a strong contrast with the results seen in Figures 1 and 2. We suggest some reasons for this in the Discussion section.

The ratio of sentence types to sentence tokens for the corpora as a whole are given in Table 3. As would be expected from the essentially linear growth observed with token growth for all three corpora, all three ratios are nearly 1:1.

| Corpus name | Ratio |
|---|---|
| BNC | 1: 1.03 |
| CRAFT | 1: 1.14 |
| GENIA | 1: 1.11 |

Table 3: Sentence type-to-token ratios.

## 4 Discussion and Conclusions

The most obvious conclusion of this study is that the null hypothesis can be rejected—the scientific corpora show a greater tendency towards closure than the general English corpus. Furthermore, we observe that the two scientific corpora behave quite similarly to each other at all three levels. This second observation is not necessarily a given. If we can consider for a moment the notion that there might be degrees of fit to the sublanguage model, it is clear that from a content perspective the BNC is unlimited; the CRAFT corpus is limited to mouse genomics, but not to any particular area of mouse genomics (indeed, it contains articles about development, disease, physiology, and other topics); and GENIA is more limited than CRAFT, being restricted to the topic of human blood cell transcription factors. If a technique for sublanguage detection were sufficiently precise and granular, it might be possible to show a strict ranking from BNC to CRAFT to GENIA in terms of fit to the sublanguage model (i.e., BNC showing no fit, and GENIA showing a greater fit than CRAFT since its subject matter is even more restricted). However, this does not occur—in our data, CRAFT showed a stronger tendency towards closure at the lexical level, while GENIA shows a stronger tendency towards closure at the morphosyntactic level. It is possible that the small differences at those levels are not significant, and that the two corpora show the same tendencies towards closure overall.

One reason that the IBM manuals in the (McEnery and Wilson, 2001) experiments showed sentence type closure but the CRAFT and GENIA corpora did not in our experiments is almost certainly related to sentence length. The average length of a sentence in the IBM manuals is 11 words, versus 24 in the Hansard corpus and 21 in the American Printing House for the Blind corpus. In this respect, the scientific corpora are much more like the Hansard and American Printing House for the Blind corpora than they are like the IBM manuals—the average length of a sentence in GENIA is 21.47 words, similar to the Hansard and American Printing House for the Blind corpora and about twice the length of sentences in the IBM manuals. Similarly, the average sentence length of the CRAFT corpus is 22.27 words (twice the average sentence length of the IBM manuals), and the average sentence length in the BNC is 20.43 words. Longer sentences imply greater chances for different sentence types.

Another reason for the tendency towards sentence type closure in the IBM manuals, which was not observed in CRAFT and GENIA, is the strong possibility that they were written in a controlled language that specifies the types of syntactic constructions that can be used in writing a manual, e.g. limiting the use of passives, etc., as well as lexical choices and limits on other options (Kuhn, under review). There is no such official controlled language for writing journal articles.

Finally, one reason that the CRAFT and GENIA corpora did not show sentence type closure while the IBM manuals did is that while McEnery and Wilson represented sentence types as parses, we represented them as sequences of part-of-speech tags. Representing sentence types as parse trees has the effect of smoothing out some variability at the leaf node level. For this reason, our representation increases the sensitivity of the method to sentence type differences, providing a stronger test of fit to the sublanguage model.

It has been suggested since Harris's classic work (Harris et al., 1989) that scientific writing forms a sublanguage. However, it is also clear from the work of (Stetson et al., 2002) and (Mihaila et al., 2012) that some putative sublanguages are a better fit to the model than others, and to date there has been no publicly available, repeatable method for assessing the fit of a set of documents to the sublanguage model. This paper presents the first such package of software and uses it to evaluate two corpora of scientific journal articles. Future work will include evaluating the effects of mapping all numbers to a fixed *NUMBER* token, which might affect the tendencies towards lexical closure; evaluating the effect of the size of tag sets on type/part-of-speech ratios, which might affect tendencies towards type/part-of-speech closure; and seeking a way to introduce more syntactic structure into the sentence type analysis without losing the generality of the current approach. We will also apply the technique to other biomedical genres, such as clinical documents. There is also an important next step to take—this work provides a means for recognizing sublanguages, but does not tackle the problem of determining their characteristics. However, despite these limitations, this paper presents a large step towards facilitating the study of sublanguages by providing a quantitative means of assessing their presence.

In analyzing the results of the study, some implications for natural language processing are apparent. Some of these are in accord with the issues for sublanguage natural language processing pointed out in the introduction. Another is that this work highlights the importance of both classic and more recent work on concept recognition for scientific journal articles (and other classes of sublanguages), such as MetaMap (Aronson, 2001; Aronson and Lang, 2010), ConceptMapper (Tanenblatt et al., 2010), and the many extant gene mention systems.

## Acknowledgments

Irina Temnikova's work on the research reported in this paper was supported by the project AComIn "Advanced Computing for Innovation", grant 316087, funded by the FP7 Capacity Programme (Research Potential of Convergence Regions). Kevin Bretonnel Cohen's work was supported by grants NIH 5R01 LM009254-07 and NIH 5R01 LM008111-08 to Lawrence E. Hunter, NIH 1R01MH096906-01A1 to Tal Yarkoni, NIH R01 LM011124 to John Pestian, and NSF IIS-1207592 to Lawrence E. Hunter and Barbara Grimpe. The authors thank Tony McEnery and Andrew Wilson for advice on dealing with the tag sets.

## References

Alan R. Aronson and Francois-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17:229–236.

A. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proc AMIA 2001*, pages 17–21.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner Jr., Kevin Bretonnel Cohen, Karin Verspoor, Judith A. Blake, and Lawrence E. Hunter. 2012. Concept annotation in the craft corpus. *BMC Bioinformatics*, 13(161).

K. B. Cohen, Lynne Fox, Philip V. Ogren, and Lawrence Hunter. 2005. Corpus design for biomedical natural language processing. In *Proceedings of the ACL-ISMB workshop on linking biological literature, ontologies and databases*, pages 38–45. Association for Computational Linguistics.

Timothy W. Finin. 1986. Constraining the interpretation of nominal compounds in a limited context. In Ralph Grishman and Richard Kittredge, editors, *Analyzing language in restricted domains: sublanguage description and processing*, pages 85–102. Lawrence Erlbaum Associates.

Carol Friedman, Philip O. Anderson, John H.M. Austin, James J. Cimino, and Stephen B. Johnson. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1:161–174.

Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. 2001. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(Suppl. 1):S74–S82.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35:222–235.

Carol Friedman. 1986. Automatic structuring of sublanguage information. In Ralph Grishman and Richard Kittredge, editors, *Analyzing language in* *restricted domains: sublanguage description and processing*, pages 85–102. Lawrence Erlbaum Associates.

Ralph Grishman and Richard Kittredge. 1986. *Analyzing language in restricted domains: sublanguage description and processing*. Lawrence Erlbaum Associates.

Zellig Harris, Michael Gottfried, Thomas Ryckman, Anne Daladier, Paul Mattick, T.N. Harris, and Susanna Harris. 1989. *The form of information in science: analysis of an immunology sublanguage*. Kluwer Academic Publishers.

Lynette Hirschman and Naomi Sager. 1982. Automatic information formatting of a medical sublanguage. In Richard Kittredge and John Lehrberger, editors, *Sublanguage: studies of language in restricted semantic domains*, pages 27–80. Walter de Gruyter.

Timo Järvinen, Mikko Laari, Timo Lahtinen, Sirkku Paajanen, Pirkko Paljakka, Mirkka Soininen, and Pasi Tapanainen. 2004. Robust language analysis components for practical applications. In *Robust and adaptive information processing for mobile speech interfaces: DUMAS final workshop*, pages 53–56.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl. 1):180–182.

Richard I. Kittredge. 2003. Sublanguages and controlled languages. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, pages 430–447. Oxford University Press.

Tobias Kuhn. under review. Survey and classification of controlled natural languages. *Computational Linguistics*.

G. Leech, R. Garside, and M. Bryant. 1994. The large-scale grammatical tagging of text: experience with the British National Corpus. In N. Oostdijk and P. de Haan, editors, *Corpus based research into language*.

David D. McDonald. 2000. Natural language generation. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *Handbood of Natural Language Processing*, pages 147–179. Marcel Dekker.

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

Claudiu Mihaila, Riza Theresa Batista-Navarro, and Sophia Ananiadou. 2012. Analysing entity type variation across biomedical subdomains. In *Third workshop on building and evaluating resources for biomedical text mining*, pages 1–7.

Naomi Sager. 1986. Sublanguage: linguistic phenomenon, computational tool. In Ralph Grishman and Richard Kittredge, editors, *Analyzing language in restricted domains: sublanguage description and processing*, pages 1–17. Lawrence Erlbaum Associates.

Satoshi Sekine. 1994. A new direction for sublanguage nlp. In *Proceedings of the international conference on new methods in natural language processing*, pages 123–129.

Harold Somers. 1998. An attempt to use weighted cusums to identify sublanguages. In *NeMLaP3/CoNLL98: New methods in language processing and computational natural language learning*, pages 131–139.

Harold Somers. 2000. Machine translation. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *Handbook of Natural Language Processing*, pages 329–346. Marcel Dekker.

Peter D. Stetson, Stephen B. Johnson, Matthew Scotch, and George Hripcsak. 2002. The sublanguage of cross-coverage. In *Proc. AMIA 2002 Annual Symposium*, pages 742–746.

Michael Tanenblatt, Anni Coden, and Igor Sominsky. 2010. The ConceptMapper approach to named entity recognition. In *Language Resources and Evaluation Conference*, pages 546–551.

Karin Verspoor, Kevin Bretonnel Cohen, Arrick Lanfranchi, Colin Warner, Helen L. Johnson, Christophe Roeder, Jinho D. Choi, Christopher Funk, Yuriy Malenkiy, Miriam Eckert, Nianwen Xue, William A. Baumgartner Jr., Michael Bada, Martha Palmer, and Lawrence E. Hunter. 2012. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*, 13(207).