Closure Properties of Bulgarian Clinical Text

Irina P. Temnikova
Institute of ICT
Bulgarian Acad. of Sciences

Ivelina Nikolova
Institute of ICT
Bulgarian Acad. of Sciences

William A. Baumgartner Jr.
Computational Bioscience Program
U. Colorado School of Medicine

Galia Angelova
Institute of ICT
Bulgarian Academy of Sciences

K. Bretonnel Cohen
Computational Bioscience Program
U. Colorado School of Medicine

Abstract

Sublanguages are specialized genres of language associated with specific domains and document types. When sublanguages can be recognized and adequately characterized, they are useful for a variety of types of natural language processing applications. Although there are sublanguage studies related to languages other than English, all previous work on sublanguage recognition has focused on sublanguages related to general English. This paper tests whether a sublanguage detecting technique developed for English can be applied to another language. Bulgarian clinical documents are an excellent test case, because of a number of unique linguistic properties that affect their lexical and morphological characteristics. Bulgarian clinical documents were studied with respect to their closure properties and were found to fit the sublanguage model and exhibit characteristics like those noted for sublanguages related to English. It was also confirmed that the clinical sublanguage phenomenon is not a coincidental phenomenon of English, but applies to other languages as well. Implications of this fact for natural language processing are proposed.

1 Introduction and Related Work

1.1 Sublanguages

The term *sublanguage* has various definitions, depending on criteria that will be discussed in a moment. However, descriptions of the sublanguage phenomenon generally have two things in common. One is that a sublanguage is the language used to communicate in a specific genre about a specialized domain. The other is that sublanguages are restricted in some way.

Sublanguages have been described for a variety of domains, including space events (Montgomery and Glover, 1986), recipes (Kittredge, 1982), legal documents (Charrow et al., 1982), and especially for clinical documents (Hirschman and Sager, 1982; Hiz, 1982; Friedman, 1986; Dunham, 1986; Stetson et al., 2002; Friedman et al., 2002).

Concomitantly with this domain restriction, sublanguages are typically characterized as being linguistically restricted in some way. For example, Kittredge (2003) describes sublanguages as having 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.

Although sublanguage properties and sublanguage versus general language differences have been studied in various languages (e.g. (Laippala et al., 2009) and (Wermter and Hahn, 2004), for clinical language), all approaches to sublanguage recognition have been focussed on English. (We consider recognizing the existence of a sublanguage as a different task from learning the characteristics of a sublanguage; this paper is concerned with the problem of recognizing the existence of a sublanguage, although we also take preliminary steps to describe the data under investigation.) Sekine (1994) used an approach related to unsupervised learning, clustering documents and then calculating the ratio of the perplexity of the clustered documents to the perplexity of a random collection of words. Somers (1998) used weighted cumulative sums and showed that they are low in sublanguages. Stetson et al. (2002) used relative entropy and squared chi-square distance to demonstrate the existence of a sublanguage of

Proceedings of Recent Advances in Natural Language Processing, pages 667–675,
Hissar, Bulgaria, 7-13 September 2013.

cross-coverage notes. Mihaila et al. (2012) calculated distributions of a wide variety of biologically relevant semantic classes of named entities to identify and differentiate between a wide variety of scientific sublanguages in journal articles.

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

One consequence of the various types of restrictions that can be seen in various researchers' conceptions of the notion of sublanguage is that various components of the language should tend towards finiteness. That is, if we examine sufficient quantities of a sample of the language, we should observe an eventual slowing or stoppage of growth in new items in that component of the language. Take, for instance, the case of lexical items, or words. As we examine increasing numbers of tokens, we would expect the number of types to increase. If a genre of language does not fit the sublanguage model, that growth will increase indefinitely. On the other hand, if a genre of language does fit the sublanguage model, that growth will asymptote towards zero. This slowing or stoppage of growth is known as closure. If growth in the number of types stops or asymptotes, we say that closure has occurred. If it does not, then there is no closure.

An early study of closure properties (although it did not use that term) was (Grishman et al., 1984). Grishman et al. (1984) utilized a broad-coverage syntactic grammar and three English-language document collections, each of which represented a presumed sublanguage. They charted the growth in the number of syntactic productions that was used as an increasing amount of the document collections was parsed. They found that for two of the three sublanguages, both consisting of medical documents, the growth curve flattened out, indicating closure. No non-sublanguage document collection was used for comparison. A revised grammar consisting just of productions that were observed in the sublanguage document collections was then used to re-parse the document sets, and a marked increase in the speed of parsing was obtained.

McEnery and Wilson (2001) first carried out a multi-faceted study of sublanguage closure properties, using two non-sublanguage document collections for comparison. Their experiment involved three document sets, one of which was suspected of fitting the sublanguage model and two of which were not. The document set that was suspected of fitting the sublanguage model consisted of a collection of IBM technical manuals. The document sets that were not suspected of fitting the sublanguage model were a collection of proceedings of the Canadian parliament known as the Hansard corpus, and a collection of works of fiction from the American Printing House for the Blind. They looked for closure on three levels: lexical closure, measured by growth in the number of word types as an increasing number of word tokens is examined; word-POS (part of speech) pair closure, where the number of different sets of combinations of a single word type with multiple POS tags is examined as an increasing number of POS-tagged words is observed; and sentence type closure, where the number of sentence types is examined as an increasing number of sentence tokens is observed.

In more recent work, Temnikova and Cohen (2013) applied similar techniques to two corpora of scientific journal articles, one from the genomics domain and one related to human blood cell transcription factors. They used the British National Corpus as the non-sublanguage comparison corpus. Scientific journal articles have been postulated to belong to a sublanguage since the seminal early work of (Harris et al., 1989). They found similar effects as in the (McEnery and Wilson, 2001) study of IBM technical manuals; lexical items and word-POS sets did not asymptote but had drastically smaller numbers than the BNC data and growth did slow considerably as the number of tokens increased. In addition, they found the type-token ratios for both of these to be consistent with the scientific journal articles fitting the sublanguage model, but not the BNC. The difference with the results of McEnery and Wilson (2001) was attributed to the fact that McEnery and Wilson (2001) probably employed a corpus of documents written in a controlled language. This factor would have restricted additionally the sublanguage corpus variety and would result in reaching closure much faster. For this reason, the significant slowing down of the growth of the specialized corpora's curves (compared with the general language corpus's), with tendency towards, but without reaching closure, was considered as a sufficient indicator of sublanguage model fit.

1.2 Relevance of Sublanguages to Natural Language Processing

The relevance of sublanguages to natural language processing is reviewed in (Temnikova and Cohen, 2013). The relevance of sublanguages to natural language processing has long been recognized in a variety of subfields. Hirschman and Sager (1982) and Friedman (1986) show how a sublanguage–based approach can be used for information extraction from clinical documents. Grishman et al. (1984) showed that a sublanguage grammar can be used to increase the speed of syntactic parsing. 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 Definition of and Prior Work on Epicrises

Since the putative sublanguage under consideration in this paper is that of Bulgarian epicrises, we define and describe them here, as well as the history of applying natural language processing techniques to them. The closest equivalents of the Bulgarian epicrises in English are discharge reports. The content of Bulgarian electronic health records is dictated by state regulatory agencies and is spelled out in Article 190 (3) of the legal agreement between the National Health Insurance Fund and the Bulgarian Medical and Dental Associations. Electronic health records must contain an *epicrisis*, or summation of the course of a medical case history. An epicrisis is typically 2-3 pages long and must contain the patient's personal details, diagnosis and comorbidities, anamnesis (personal medical history), patient status, physical examination and test findings, treatment, and recommendations. Epicrises are linguistically challenging input texts for natural language processing, for a variety of reasons. They may contain text in Latin (about 1%) and English, sometimes in the Cyrillic alphabet and sometimes in the Latin alphabet. About 3% of the text is abbreviations, both of Bulgarian and of Latin. Syntactically, the majority of the text consists of sentence fragments, rather than full sentences (Boytcheva et al., 2009).

There is some previous Natural Language Processing (NLP) work on Bulgarian epicrises which would benefit from insight into the sublanguage characteristics of Bulgarian epicrises. Boytcheva and Angelova (2009) describes a system architecture for processing Bulgarian epicrises, including a module for generating logical forms of conceptual graphs based on templates. Boytcheva et al. (2009) built a template-based system based on 106 epicrises, using it to extract structured information such as diagnoses, risk factors, and body parts. Georgiev et al. (2011) built a named entity recognizer to tag disease names in Bulgarian epicrises. Nikolova (2012) built a hybrid machine-learning-based and rule-based system to extract blood sugar levels and measures of body weight change from a collection of 2,031 sentences from 100 Bulgarian epicrises.

1.4 Hypotheses

The work presented in this article is based on the closure investigation method (McEnery and Wilson, 2001; Temnikova and Cohen, 2013). Our null hypothesis is that there are no differences in the closure properties of unrestricted text and epicrises. Neither might show closure, or both might show closure. If the null hypothesis turns out not to be true, then deviations from it could logically be observed in two directions. One is that the epicrises could demonstrate closure, while the unrestricted text does not. The other is that the unrestricted text could demonstrate closure, while the epicrises do not.

2 Materials and Methods

2.1 Materials

The experiments require two bodies of data: the collection of data that is being examined for fit to the sublanguage model, and a "background" corpus consisting of material in the general (i.e. not specialized) language. The data under examination in these experiments is a collection of de-identified epicrises. The background corpus is the Bulgarian National Reference Corpus (BNRC).

2.1.1 Epicrises

The collection of epicrises was de-identified by University Specialised Hospital for Active Treatment of Endocrinology "Acad. I. Penchev". It consists of 1,000 documents in total, containing 647,498 words.

2.1.2 Bulgarian National Reference Corpus

The Bulgarian National Reference Corpus (Savkov et al., 2012) is a collection of 400,000,000 tokens of spoken and written Bulgarian, composed of 50% fiction, 30% newswire text, 10% legal text, and 10% from other genres. Following the approach of the Brown corpus to obtain a balanced, representative subset of the same size as the collection of epicrises, 8,000 words were extracted from each BNRC file until 647,498 words were reached, which is the size of the epicrises corpus.

We note that it is reasonable to question whether the size of a corpus is necessary to detect or rule out closure properties. McEnery and Wilson (2001) were successful in doing both with collections of 200,000 words—one third the size of the corpora that we are using.

2.2 Methods

2.2.1 Data Preparation

The data was processed using the pipeline described in (Savkov et al., 2012). Both document sets were split into sentences, tokenized, part-of-speech tagged, and dependency parsed. All tokens were lower-cased.

2.2.2 Measuring Lexical Closure Properties

For each document set, the number of distinct lexical types was counted as increasing numbers of tokens were encountered.

2.2.3 Type-POS Closure

It is well known that a single word type might belong to more than one part of speech. We charted the number of new type/part-of-speech sets as increasing numbers of tokens were encountered. The motivation for examining the pattern of growth here is that if a sublanguage has a restricted lexicon, then words might be coerced into more parts of speech than is the case in unconstrained language.

2.2.4 Sentence Type Closure

Following (Temnikova and Cohen, 2013), we defined sentence types as sequences of part-of-speech tags. This is a very rough approximation of syntax—arguably, it is not syntactic per se—but it increases the sensitivity of the method to diversity in sentence types and has the advantage of being theory-neutral and easily generalizable.

2.2.5 Syntactic Deviance

Sublanguages have often been claimed to have deviant syntax (e.g. (Kittredge, 2003)). In an attempt to discover deviant syntactic structures, we looked for sentences that lack verbs, as discharge letters are expected to be characterized by this type of sentence.

2.2.6 Over-Represented Lexical Items in the Epicrises

Although the primary purpose of the work reported here is to recognize the existence of a sublanguage, rather than to learn its characteristics, we performed a preliminary investigation of the contents of the epicrisis corpus, using an algorithm known as simplemaths (Kilgarriff, 2012). Simplemaths is designed to find words that are over-represented in one corpus as compared to a reference corpus. It is based on the idea of calculating frequencies of occurrences of all words in both corpora, taking the ratio of the frequency of each word in both corpora, and ranking by ratio. To avoid the problem that words of widely differing frequencies might yield the same ratio—a word that occurs 100 times in the corpus of interest and 10 times in the reference corpus produce the same ratio as a word that occurs 10,000 times in one corpus and 1,000 times in the other, but they are not equally revealing as to the domain-related contents of the corpus, since one word is quite rare and the other quite common—we add a constant value to

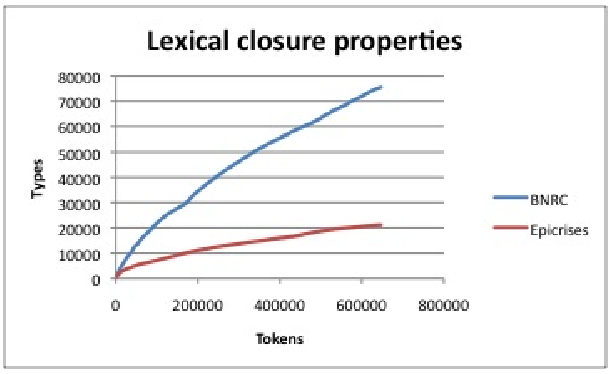

Figure 1: Lexical closure properties. Tick-marks on *x* axis indicate increments of 200,000 tokens.

all counts. This has the effect of separating out the frequency ranges of rare and common words in the corpus. (It also takes care of smoothing zero counts.) The constant number is called the "simplemaths parameter." We used the suggested value of 100 for the simplemaths parameter.

3 Results

3.1 Lexical Closure

Figure 1 shows the lexical closure properties of the Bulgarian National Reference Corpus and the epicrises. As can be noted, there are drastic differences between the two. The BNRC has a much larger number of lexical types, and shows no tendency towards closure at all. In contrast, the epicrises have a much smaller number of lexical types and appear to show closure at a bit below 600,000 tokens.

The type/token ratio for lexical items in the BNRC and the epicrises is shown in Table 1. As the theory predicts, the type/token ratio of lexical items for the epicrises is much higher than that of the BNRC—more than three times higher.

Corpus	Ratio
BNRC	1:7.63
Epicrises	1:26.52

Table 1: Lexical type-to-token ratios.

3.2 Type-POS Closure

Figure 2 shows the type-POS set closure properties for the Bulgarian National Reference Corpus and the epicrises. Once again, we see drastic differences between the two. The BNRC has no tendency towards closure at all. In contrast, although the epicrises do not yet show closure, they show a clear tendency in that direction.

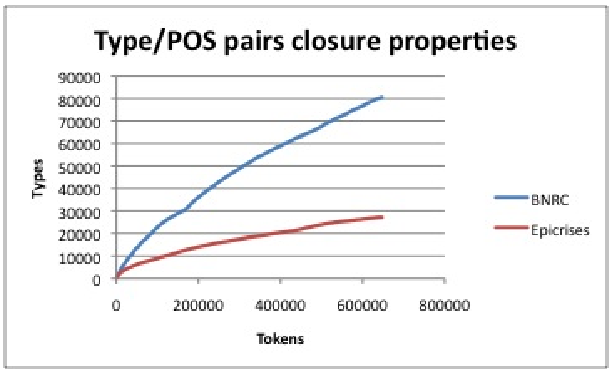

Figure 2: Type-POS closure properties. Tick-marks on *x* axis indicate increments of 200,000 tokens.

The type/token ratio for type-POS sets in the BNRC and the epicrises is shown in Table 2. Again, as the theory predicts, the type/token ratio of type-POS sets for the epicrises is much higher than that of the BNRC—more than two times higher.

Corpus	Ratio
BNRC	1:7.24
Epicrises	1:19.75

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

3.3 Sentence Type Closure

Figure 3 shows the sentence type closure properties for the Bulgarian National Reference Corpus and the epicrises. Unlike the other two graphs, where the number of tokens is the same, in the case of this graph the number of sentence tokens is different between the two corpora, since sentence length varies between them. The results are notable for a number of reasons. We see drastic differences in the growth curves for the two corpora. In the case of the BNRC, growth in sentence types almost completely matches the number of sentence tokens—sentence types are rarely repeated. In contrast, we see drastically different growth in the epicrisis sentence types—there are many more epicrisis sentence tokens, and yet far fewer sentence types overall. Sentence types are frequently repeated in the epicrises. This is an important finding—McEnery and Wilson (2001) and Temnikova and Cohen (2013) did not find find any

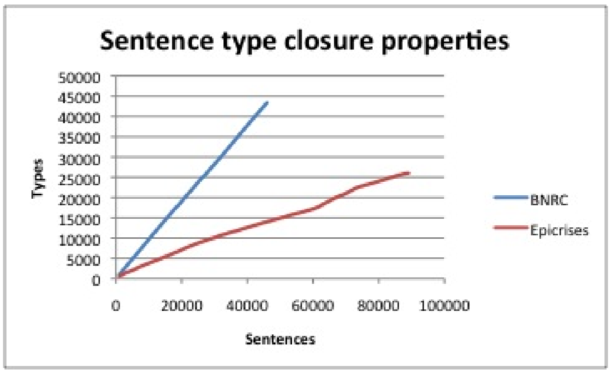

Figure 3: Sentence type closure properties. Tick-marks on *x* axis indicate increments of 20,000 tokens.

closure at the syntactic level. It is remarkable to note that this result was obtained in spite of the large number of part-of-speech tags assigned (680, due to the very complex morphology of Bulgarian). Such a large number would make the probability of any sequence of part-of-speech tags very low.

The type/token ratio for sentence types in the BNRC and the epicrises is shown in Table 3. Once again, as the theory predicts, the type/token ratio of sentences for the epicrises is much higher than that of the BNRC—more than three times higher. The type/token ratio for the BNRC is quite close to 1:1—sentence types in unrestricted text are almost never repeated.

Corpus	Ratio
BNRC	1:1.06
Epicrises	1:3.44

Table 3: Sentence type-to-token ratios.

It is likely that the presence of repeated sentence types in the epicrises as compared to the BNRC is related to the difference in the average length of sentences in the two corpora. The average sentence length in the BNRC is 14.16 words, while the average sentence length in the epicrises is 7.40–about half the length of the average BNRC sentence. This both explains the large difference in the number of sentences seen in Figure 3 (bear in mind that the number of words in the two sets of documents is the same) and helps explain why it might be more likely for sentence types to be repeated.

3.4 Syntactic Deviance

Our preliminary attempt at characterizing syntactic deviance through counting the number of sentences with no verbs shows a strong tendency towards syntactic deviance in the epicrises as compared to the Bulgarian National Reference Corpus. In the BNRC, we noted that 11% (4,943/46,549) of the sentences were verbless (probably mostly section headers and the like). In contrast, in the epicrises, a full 66% of sentences (58,753 out of 89,331 sentences) lacked a verb, e.g. Корем - мек, неболезнен. ‘Abdomen - soft, painless.’ The epicrises show a strong tendency towards syntactic deviance, as predicted for sublanguages.

3.5 Over-Represented Lexical Items in the Epicrises

Table 4 shows lexical items that are over-represented in the epicrises. Note that these are not the most *frequent* ones, but rather the ones that occur in the document set more often than would be expected. We display just the top ten most highly over-represented lexical items, with separate lists of the over-represented word types and over-represented lemmata. Examining the top 50 terms in each list, we see heavy representation of lexical items related to diabetes, body parts, and symptoms. Even in the short list of items displayed in Table 4, almost every item is relevant to either the semantics or the syntax of the domain. ‘ч’ is an abbreviation for ‘часа’ (hours), which occurs frequently to indicate the time at which one of a series of blood levels was drawn and is essential for extracting trends in lab results. ‘/’ has a variety of uses, primarily syntactic, such as linking systolic and diastolic blood pressures. The clinical significance of the other items in the top-10 list is clear, with the exception of the semicolon ‘;’ which occurs frequently in lists of lab values and of symptoms.

4 Discussion and Conclusions

This paper has presented the first attempt to detect a sublanguage in Bulgarian.

The data demonstrate that Bulgarian clinical records fit the model, as shown by the closure properties of the lexicon, morphology, and sentence types. Unlike the previous work of McEnery and Wilson (2001) and Temnikova and Cohen (2013), sentence type closure was demonstrated for the first time.

Word type		**Lemma**	
ч	hour	ч	hour
/	/	/	/
лечение	treatment	диабетна	diabetic, f. sg.
диабет	diabetes	лечение	treatment
;	;	диабет	diabetes
x	repetition, e.g. of dosage	захарен	sugar, m. sg. adj.
мг	mg	клиника	clinic
диабетна	diabetic, f. sg.	мг	mg
тип	type	полиневропатия	polyneuropathy
полиневропатия	polyneuropathy	анамнеза	anamnesis

Table 4: Word types and lemmata that are over-represented in the epicrises. Note that these are not the most frequent word types/lemmata, but rather the ones that occur more frequently than would be expected as compared to the reference corpus.

The finding that Bulgarian clinical documents are written in a sublanguage and the logical future work on closure with respect to arguments of predicators would aid Boytcheva and Angelova (2009) and Boytcheva et al. (2009) in the discovery of additional candidates for template representations.

Our finding that epicrises seem to be written in a very restricted sublanguage would also help understand how it was possible to achieve an F-measure of 0.81 on a test collection of only ten documents and why it took almost no time to build a named entity recognizer to tag disease names in Bulgarian epicrises (Georgiev et al., 2011).

The findings described here help us understand why that was possible, when building training sets for learning to recognize other biomedical classes of named entities has been so time-consuming. By virtue of fitting the sublanguage model, the epicrises represent a smaller set of lexical items to be classified and allow for the efficacy of a smaller number of features. Finally, as mentioned in the introduction, Nikolova (2012) built a hybrid machine-learning-based and rule-based symptom to extract blood sugar levels and measures of body weight change from a collection of 2,031 sentences from 100 Bulgarian epicrises. Insight into the sublanguage properties of the input data would have helped in determining which assays would best be extracted by rule-based methods and which would best be approached through machine learning.

This work has focused on detecting the existence of sublanguages. The important next step is to develop methods for determining the characteristics of sublanguages—determining the semantic, syntactic, and other restrictions that characterize the sublanguage and reporting them to the natural language processing researcher in a utilizable way. The work here lays the groundwork for that future work, helping us to determine when a genre or domain is likely to yield results that are susceptible to such research and when such research is less likely to be fruitful.

Acknowledgments

The work of Irina Temnikova, Ivelina Nikolova and Galia Angelova presented in this article was supported by the project AComIn ”Advanced Computing for Innovation”, grant 316087, funded by the FP7 Capacity Programme (Research Potential of Convergence Regions). The work of Kevin Bretonnel Cohen and William Baumgartner was supported by the 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. We would like to specially thank Kiril Simov for the help with pre-processing the Bulgarian data.

References

Svetla Boytcheva and Galia Angelova. 2009. Towards extraction of conceptual structures from electronic health records. In *Conceptual structures: Leveraging semantic technologies*, pages 100–113.

Svetla Boytcheva, Ivelina Nikolova, and Elena Paskaleva. 2009. Context related extraction of conceptual information from electronic health records.

In *Conceptual structures for extracting natural language semantics*, pages 38–49.

Veda Charrow, Jo Ann Crandall, and Robert Charrow. 1982. Characteristics and functions of legal language. In Richard Kittredge and John Lehrberger, editors, *Sublanguage: Studies of language in restricted semantic domains*, pages 191–205. Walter de Gruyter & Company.

George Dunham. 1986. The role of syntax in the sublanguage of medical diagnostic statements. In Ralph Grishman and Richard Kittredge, editors, *Analyzing language in restricted domains: Sublanguage description and processing*, pages 175–194. Lawrence Erlbaum Associates.

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.

Georgi D. Georgiev, Valentin Zhikov, Borislav Popov, and Preslav Nakov. 2011. Building a named entity recognizer in three days: Application to disease name recognition in Bulgarian epicrises. In *Proceedings of the workshop on biomedical natural language processing, RANLP 2011*, pages 27–34.

Ralph Grishman, Ngo Thanh Nhan, Elaine Marsh, and Lynette Hirschman. 1984. Automated determination of sublanguage syntactic usage. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics*, pages 96–100. Association for Computational Linguistics.

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.

Henry Hiz. 1982. Specialized languages of biology, medicine and science and connections between them. In Richard Kittredge and John Lehrberger, editors, *Sublanguage: Studies of language in restricted semantic domains*, pages 206–212. Walter de Gruyter & Company.

Adam Kilgarriff. 2012. Getting to know your corpus. In *Text, speech and dialogue*.

Richard Kittredge. 1982. Variation and homogeneity of sublanguages. In Richard Kittredge and John Lehrberger, editors, *Sublanguage: studies of language in restricted semantic domains*, pages 107–137.

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

Veronika Laippala, Filip Ginter, Sampo Pyysalo, and Tapio Salakoski. 2009. Towards automated processing of clinical Finnish: Sublanguage analysis and a rule-based parser. *International Journal of Medical Informatics*, 78(12):e7–e12.

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.

Christine A. Montgomery and Bonnie C. Glover. 1986. A sublanguage for reporting and analysis of space events. In Ralph Grishman and Richard Kittredge, editors, *Analyzing language in restricted domains: Sublanguage description and processing*, pages 129–161. Lawrence Erlbaum Associates.

Ivelina Nikolova. 2012. Unified extraction of health condition descriptions. In *Proceedings of the NAACL HLT 2012 student research workshop*, pages 23–28.

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.

Aleksandar Savkov, Laska Laskova, Stanislava Kancheva, Petya Osenova, and Kiril Simov. 2012. Linguistic analysis processing line for Bulgarian. In *Proceedings of the eighth international conference on language resources and evaluation*, pages 2959–2964.

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.

Irina Temnikova and K. Bretonnel Cohen. 2013. Recognizing sublanguages in scientific journal articles through closure properties. In *Proceedings of BioNLP 2013*.

Joachim Wermter and Udo Hahn. 2004. Really, is medical sublanguage that different? Experimental counter-evidence from tagging medical and newspaper corpora. *Studies in health technology and informatics*, 107(Pt 1):560.