Measuring Closure Properties of Patent Sublanguages

Irina P. Temnikova
Institute of ICT
Bulgarian Academy of Sciences
irina.temnikova@gmail.com

Negacy D. Hailu
Computational Bioscience Program
U. Colorado School of Medicine
negacy.hailu@ucdenver.edu

Galia Angelova
Institute of ICT
Bulgarian Academy of Sciences
galia@lml.bas.bg

K. Bretonnel Cohen
Computational Bioscience Program
U. Colorado School of Medicine
kevin.cohen@gmail.com

Abstract

Patent search is an important information retrieval problem in scientific and business research. Semantic search would be a large improvement to current technologies, but requires some insight into the language of patents. In this article we test the fit of the language of patents to the sublanguage model, focussing on closure properties. The research presented here is relevant to the topic of sublanguage identification for different domains, and to the study of the language of patents. We investigate the hypothesis that fit to the sublanguage model increases as one moves down the International Patent Classification hierarchy. The analysis employs a general English corpus and patent documents from the MAREC corpus. It is shown that patents generally fit the sublanguage model, with some variability between categories in the extent of the fit.

1 Introduction

The study presented in this article aims to contribute to two important Natural Language Processing (NLP) applications: patent search and sublanguage identification.

1.1 Patents and Patent Search

We define patents as "legal documents issued by a government that grant a set of rights of exclusivity and protection to the owner of an invention" (Alberts et al., 2011). Patent search is an important Information Retrieval (IR) problem due to the financial risks involved in accidentally breaking previously registered patent rights, and due to the complexity of the phenomenon. Patent search is carried out by a variety of users, including patent specialists, managers, researchers, attorneys, and inventors. There are multiple scenarios requiring patent search (Alberts et al., 2011), as well as multiple types of patent search tasks—state-of-the-art, novelty, patentability, infringement, freedom to operate, and due diligence (Hunt et al., 2007; Joho et al., 2010).

Different user types are prompted to adopt different and often complex search techniques, reflecting their different search aims and search tasks (Hunt et al., 2007). Search techniques include classification code search, keyword search, full-text search, forward and backward citation of related documents, inventor or author search, patent assignee search, patent family search, legal status, and cross-language search (Alberts et al., 2011). Among these, full-text search is considered to have relatively more advantages than the other types of search techniques, as it allows the user to access the full semantic contents of the patent document (Adams, 2010a). However, in its present state, full-text patent search still exhibits several shortcomings, such as poor precision and lack of disambiguation (Adams, 2010a; Adams, 2010b). Besides the increased IR field attention towards patent search (see the CLEF-IP[1], TREC-CHEM[2], NTCIR, and PaIR[3] tracks and workshops), full-text search still suffers from lack of linguistic processing, which prevents it from addressing real user needs (Adams, 2010a; Adams, 2010b).

1.2 Patents and Sublanguages

A major step forward in patent search could be achieved if patents could be indexed by semantic content. This could include indexing by semantic classes of named entities relevant to the domain of the patent, relationships between semantic classes of named entities, and the like. However, model-

[1] http://www.ifs.tuwien.ac.at/ clef-ip/index.html. Last accessed on May 16th, 2013.

[2] http://www.ir-facility.org/trec-chem

[3] http://www.ir-facility.org/pair-workshops

Proceedings of Recent Advances in Natural Language Processing, pages 659–666,
Hissar, Bulgaria, 7-13 September 2013.

ing the appropriate semantics requires an in-depth understanding of the contents and the linguistic characteristics of the genre. This is a daunting task for unrestricted patents in general, but if patents in some domain only exhibit a limited number of semantic classes and relations, it becomes a practical undertaking. One could then apply the "information retrieval as information extraction" (Moens, 2006) approach to patent search. But, do patents exhibit such semantic limitations? And how can we tell?

The notion of the *sublanguage* has a long history in natural language processing. Definitions of "sublanguage" vary, but have some commonalities. They are contrasted with the general language (e.g. English as a whole) in terms of restrictions in a number of areas. Sublanguages (Kittredge, 2003) are generally thought to be restricted to communication by a limited community of experts, in a limited range of genres, using a limited vocabulary, with limits on the possible semantic classes of arguments to predicators and possibly limited or deviant syntax. Although it is logical to think that patents and patent applications discuss a restricted technical topic, it is known that every inventor uses his/her own language (Alberts et al., 2011), and thus the applicability of the sublanguage model to patents is not a given. This paper reports three experiments on the application of natural language processing techniques to the problem of determining whether or not patents fit the sublanguage model.

The approach taken here is to examine the closure properties of patents. The phenomenon of closure is related to the element of restriction in sublanguages. If a genre is restricted with respect to some linguistic characteristic, then that linguistic property will tend towards finiteness. We test for this by counting the incidence of some linguistic characteristic, such as the occurrence of novel lexical items, as increasing amounts of a body of documents are observed. If the linguistic characteristic tends towards finiteness, then at some point we will see no further growth as increasing amounts of the document collection are examined. When such growth stops, *closure* is said to have occurred. In this study, we experiment with three different levels of closure, described below.

For our experiment, we follow the International Patent Classification (IPC, recently revised to IPCR), which divides all areas of technology into eight sections (A-H), each hierarchically subdivided into several levels, including classes, subclasses, groups, and sub-groups (Alberts et al., 2011). Each patent has a code assigned, which indicates its membership at each of these classification levels (e.g. *"A63B 69/02"* corresponds to *training tools for fencing*).

It may be the case that sublanguages exist at the level of patents in general, or only at the lowest levels of the hierarchy, or at some level of abstraction between the lowest levels and the general category of "patent." For this reason, we experiment with categories at multiple levels in the hierarchy.

2 Related Work in Patent Language Studies and Sublanguage Identification

Besides the interest of the IR community, not much has been done on discussing the characteristics of patent language. The existing studies have noted very complex sentences, vague definitions, presence of multiple languages in the same patent, technical concepts, inventor-specific definitions, and a high number of spelling errors (Lupu, 2011; Itoh et al., 2003; Sheremetyeva et al., 1996). There is, however, also research focussing on the linguistic aspects of patent documents. Lin and Hsieh (2004) have investigated verb-noun collocations appearing in patent claims for developing resources for teaching English for Specific Purposes, and more specifically in the legal domain. The same authors (Lin and Hsieh, 2010) later conducted a corpus-based study with the purpose of collecting the most frequent technical terms using The United States Patent and Trademark Office (USPTO) Glossary. Shinmori et al. (2003) studied the syntactic and term complexities of Japanese patent claims using the NTCIR3 patent collection (Iwayama et al., 2003), with the aim of improving readability of Japanese patent claims.

The paper most related to our work is that of Oostdijk et al. (2010), who study the language differences between the different patent domains and the genre differences between the different patent sections (title, abstract, description, and claims) for purposes of tuning a patent search engine. They use the English-language European patent documents from the MAREC400k corpus. For preprocessing, they clean the XML tags, split the texts into sentences, and parse them with the Aegir parser. On average 1000 patents containing all four text sections, from three different classes (H01L – Semiconductor devices, A61K – Medical and dental preparations, and F06G – Electric digital data processing) were compared. Genre and domain differences were measured by calculating the average sentence length, the type-token ratio and the hapax ratio. They show that there are differences between the different domains, as well as that there are more differences at section than at subdomain level.

Our approach goes beyond the work of Oostdijk et al. (2010) by testing the hypothesis that the patent categories employed in their work fit the sublanguage model. To our knowledge, no study has tested this hypothesis on patents before. In addition to that, we also calculate the average sentence length and type:token ratio for all of the examined categories.

Our research hypothesis is that all of the levels of categories fit the sublanguage model, with the lowest (more specific ones) showing more closure, and the highest (more generic) ones having characteristics closer to general English.

Although there has been extensive work on recognizing and characterizing sublanguages, little has been done on recognizing sublanguages through closure properties. The classic study is (McEnery and Wilson, 2001). McEnery and Wilson (2001) compared two corpora which were thought to be representative of the general language with one corpus which was thought to represent a sublanguage. The general language corpora were a collection of works of fiction from the American Printing House for the Blind and a collection of proceedings from the Canadian Hansard. The corpus that was thought to represent a sublanguage was a collection of IBM technical manuals. They found evidence of lexical closure and type-POS closure (described below) in the IBM technical manuals, but no evidence of closure in sentence types. Temnikova and Cohen (2013) compared a sample of general English drawn from the British National Corpus with two biomedical corpora thought to represent two distinct sublanguages and found evidence of lexical and type-POS closure in both of the biomedical corpora. Like (McEnery and Wilson, 2001), they did not observe sentence type closure in either of the sublanguage corpora. Temnikova et al. (2013) examined the closure properties of clinical documents in Bulgarian, comparing a sample from the Bulgarian National Reference Corpus, representative of the general Bulgarian language, with a corpus of Bulgarian epicrises (a document type similar to discharge summaries). They found lexical and type-POS closure, and unlike the other studies just discussed, did observe sentence type closure.

3 Materials and Methods

For consistency with the work of Oostdijk et al. (2010), we use the MAREC400k corpus. MAREC400k is a subset of the MAREC corpus[4], which is a static collection of over 19 million patent applications written in 19 languages. The patents in the MAREC collection come from four different patent authorities: the European Patent Office[5] (patents from now on called **EP**), the World Intellectual Property Organization[6] (**WP**), the United States Patent and Trademark Office (USPTO[7], patents called **US**), and the Japan Patent Office[8] (**JP**). The patents are in a normalized XML format, which splits the patent in parts. MAREC400k is a subset of 100,000 randomly collected patents from each of the four patent collections (EP, WP, US, and JP). We utilized a 77,000 US patents of MAREC400k, as this is the amount we could process in time. The US patents were chosen, as according to MAREC's statistics, only in them both the abstracts and the descriptions were written fully in English[9].

The MAREC400k documents were stripped of the XML tags, with the title, abstract, description and claims extracted and left in text format. The texts were then split into sentences and enriched with part-of-speech tags with the help of the Natural Language ToolKit (NLTK) (Bird et al., 2009).

For consistency with Oostdijk et al. (2010), we extracted 1,000,000-word subsets of the 77,000 patents, containing text from patents, classified with the **A61K** and **H01L** IPC (International Patent Classification) categories. Although Oostdijk et al. also used the F06G documents, unfortunately, there were no F06G documents in our subset, so we restricted our experiment only to the first two patent categories. 1,000,000 words samples of the categories **A61**, **H01**, **A**, **H** were also collected from patents classified with the respec-

[4] http://www.ir-facility.org/prototypes/marec

[5] http://www.epo.org. Last accessed on June 10th, 2013.

[6] http://www.wipo.int/portal/index.html.en

[7] http://www.uspto.gov

[8] http://www.jpo.go.jp

[9] http://www.ir-facility.org/prototypes/marec/statistics

tive subcategories among the 77,000 documents. Finally, a 1,000,000 words subset of **All Patents (AP)** was also collected.

In order to collect an equal distribution of words from all sub-categories of a given category, we have split the 1,000,000 words between the sub-categories and collected 2000 words from file in each sub-category, until reaching the necessary number of words. In case of sub-categories with only a few files, we copied the whole file.

This has resulted in collecting 2000 words from on average 30 files from each subcategory. This approach has been followed to collect the 1,000,000 words for All Patents (subcategories A-H), A (subcategories A01-A99), H (subcategories H01-H99), A61 (subcategories A61B-A61Q), and H01 (subcategories H01B-H01T). The 1,000,000 words for A61K and H01L have been collected by simply getting the first 2000 words from each patent, classified with these categories.

Note that the result of this sampling is that the document collections at the higher levels are not composed by addition of the document collections at the lower levels–they are distinct.

Table 1 lists the IPC categories under study, along with their topics[10].

Category	Topics
A	Human Necessities.
H	Electricity.
A61	Medical or Veterinary Science, Hygiene.
H01	Basic Electric Elements.
A61K	(Chemical) Preparations for Medical, Dental, or Toilet Purposes.
H01L	Semiconductor Devices, Electric Solid State Devices.

Table 1: IPC Categories and topics.

In this categorization, the A categories are much wider than the H categories. The A sub-categories topics include: agriculture (A01), clothes and footwear (A41 and A43), furniture (A47), and fire-fighting (A61). In contrast, the H sub-categories are restricted to only electricity-related topics. At the lowest level, while A61K groups cleaning substances and drugs, H01L includes only semiconductor devices.

In order to test our hypothesis of the sublanguage model fit (McEnery and Wilson, 2001), we needed a corpus of general English. We utilized a 1,000,000-word subset of the British National Corpus (BNC) (Leech et al., 1994), syntactically parsed by the Machinese Connexor's parser (Järvinen et al., 2004).

We do not consider here the differences between the NLTK and Connexor's parser tagsets, as Temnikova and Cohen (2013) have shown that differences in the tagset granularity do not affect the sublanguage model.

4 Results

The following subsections present the results of the three experiments, starting with the H class first, as its results are more straightforward to interpret.

4.1 Lexical Closure Properties

Figure 1 shows the lexical closure properties of the H class. The lexical *types* are the different types of words, while the lexical *tokens* are the single instances of these types occurring in the text. The 'type' is not the word lemma (i.e. the token 'stops' corresponds to the type 'stops' (which may have occurred 10 times in the text, which makes 10 tokens, but 1 type) and not to 'stop').

We display the growth in types for the BNC, for all patent classes combined, for the H class with all of its subclasses, for the H01 subclass with all of its subclasses, and for the H01L subclass of H01. Note that in the figures for the H class and for the A class, the curves for the BNC and for *all patents combined* are identical.

In the H class we see the prototypical results for lexical closure in a sublanguage and lack of closure in unrestricted text. As discussed in Temnikova et al. (2013), we consider tendency towards closure, with no evident closure as a sufficient sign of the sublanguage model fit. The clear closure in McEnery and Wilson (2001) is assumed to be due to the IBM manuals presumably being written in a controlled language, which, here, is not the case.

The number of types in the BNC continues to grow rapidly even after 1,000,000 tokens have been observed—there is no closure. In contrast, the number of types for all patents combined, for the H class, the H01 subclass of H, and the H01L subclass of H01 slows down in growth after about 200,000 tokens have been observed and after 1,000,000 tokens have been observed has

[10] Information taken from http://web2.wipo.int/ipcpub.

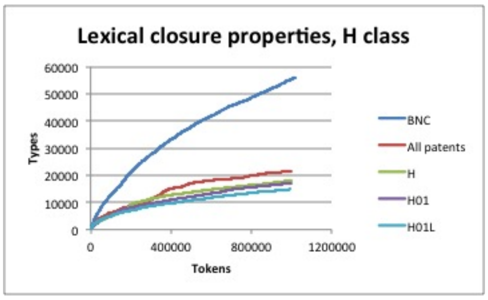

Figure 1: Lexical closure properties of the H class. Tick-marks on the *x* axis indicate increments of 400,000 tokens.

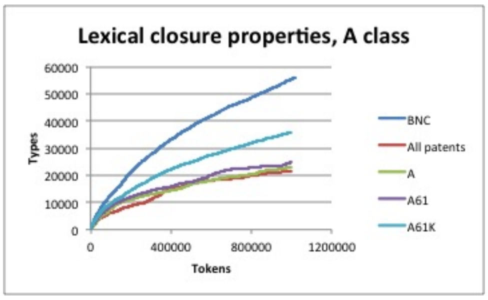

Figure 2: Lexical closure properties of the A class. Tick-marks on the *x* axis indicate increments of 400,000 tokens.

grown to a much smaller absolute number than the BNC. The evidence for closure is quite clear. In fact, closure is slightly more evident the further down the IPC hierarchy we go—looking at the ordering of the lines in Figure 1, we see that the ordering of the lines follows the descent into the hierarchy.

Figure 2 shows the lexical closure properties for the A class. Here, the picture is more complicated. Again, the BNC does not show closure. In contrast, the set of all patents, the A class, and the A61 subclass of A slow in growth after about 400,000 tokens have been observed and after 1,000,000 tokens have been observed have a much smaller absolute number of types than the BNC. However, the A61K subclass of A61 continues to exhibit rapid growth in the number of types as long as we continue to observe new tokens. After 1,000,000 tokens have been observed, the overall number of types is smaller than the BNC, but is about 1.5 times as large as the number of tokens in the classes that show closure. So, we can say that all patents, the A class, and its A61 subclass show lexical closure, but the A61K subclass does not appear to exhibit lexical closure.

The type to token ratios for lexical items for all the corpora as a whole are shown in Table 2. A lower ratio means that there is more variety in the specific corpus, while higher ratios mean more repetitiveness, and thus more restriction. Besides the differences in the 'type' interpretation between us and Oostdijk et al. (2010) (they looked at lemmas, while we do not), and thus the fact that they deal with much lower numbers, our findings confirm theirs in the fact that the average values for type:token ratios for A61K are lower than for H01L. As the sublanguage model would predict, all of the patent corpora have much higher ratios (i.e. exhibit more restriction) than the BNC.

Corpus name	Ratio
BNC	1: 18.20
All Patents	1: 46.36
H	1: 55.26
H01	1: 58.50
H01L	1: 65.23
A	1: 43.23
A61	1: 40.19
A61K	1: 27.87

Table 2: Lexical type-to-token ratios.

4.2 Type-Part-Of-Speech (POS) Closure Properties

Figures 3 and 4 show the type-POS set closure properties for the H and A classes, respectively. Here, the tokens are the single instances of lexical tokens, accompanied by their part-of-speech tag (e.g. 'stops – V', 'stops – N' are two tokens).

Again, the curves for the BNC and all patents are the same in both figures. We see similar patterns to the lexical closure properties: the BNC does not even come close to reaching closure; all patents tend to closure; the H class, its subclasses, the A class, and its subclass A61 tend to closure, with the H class and its subclasses beginning to slow in growth earlier than the A class and its subclass; the A61K class, in contrast, continues to grow rapidly even after 1,000,000 tokens have been observed.

The type-to-token ratios for token-POS pairs for all the corpora as a whole are shown in Table 3. Similarly to Table 2, we see much higher ratios for all the patents corpora, than for the BNC.

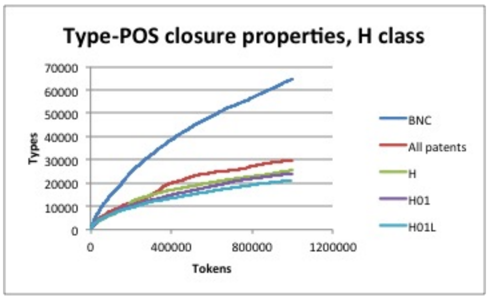

Figure 3: Type-POS closure properties of the H class. Tick-marks on the *x* axis indicate increments of 400,000 tokens.

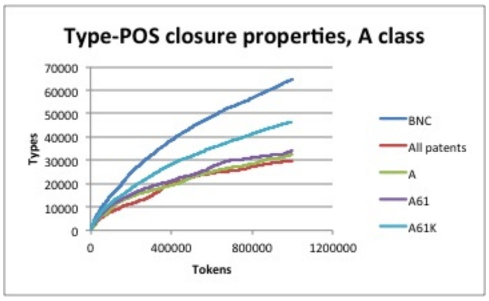

Figure 4: Type-POS closure properties of the A class. Tick-marks on the *x* axis indicate increments of 400,000 tokens.

4.3 Sentence Type Closure Properties

Figures 5 and 6 show the sentence type closure properties for the H class and the A class. Here, as in Temnikova and Cohen (2013), we define a sentence as a sequence of POS tags (every instance is a sentence token, the unique sentence is a sentence type). Again, the curves for the BNC and all patents are the same in both figures. Here we see no evidence for closure in the patents at all– the number of sentence types continues to grow rapidly even after 1,000,000 tokens have been observed.

The ratio of sentence types to sentence tokens and the average sentence lengths for the corpora as a whole are given in Table 4. As would be expected from the essentially linear growth observed in the graphics of all the corpora, all the ratios are close to 1:1. It can also be seen, that the average sentence lengths for all patents corpora are higher than the BNC, which confirms the findings of previous studies (Oostdijk et al., 2010; Shinmori et al., 2003).

Corpus name	Ratio
BNC	1: 15.46
All Patents	1: 33.36
H	1: 38.99
H01	1: 41.27
H01L	1: 46.34
A	1: 30.74
A61	1: 29.31
A61K	1: 21.41

Table 3: Type-to-token ratios for token/POS tags.

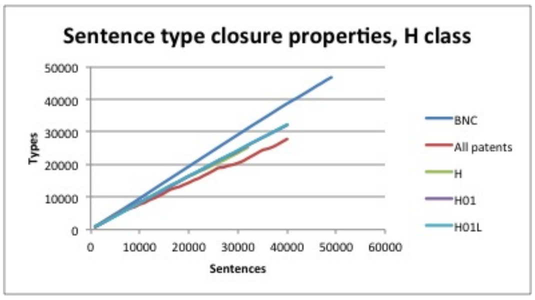

Figure 5: Sentence type closure properties of the H class. Tick-marks on *x* axis indicate increments of 50,000 tokens.

5 Discussion and Conclusions

The aim of the work reported here was to test the hypothesis that patent documents fit the sublanguage model. The motivation is that if we can detect sublanguages in any level of the patents, then there is the potential for developing methods for semantic search of patent collections.

Our most basic finding is that the patents do, in general, fit the sublanguage model. Tendency to closure at the lexical and type-POS levels were observed for all patents and for almost every class

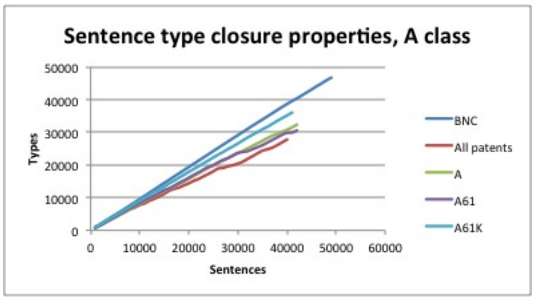

Figure 6: Sentence type closure properties of the A class. Tick-marks on *x* axis indicate increments of 50,000 tokens.

Corpus name	Ratio	Av. Sen. Length
BNC	1: 1.05	20.65
All Patents	1: 1.44	24.87
H	1: 1.27	30.95
H01	1: 1.24	24.78
H01L	1: 1.24	24.49
A	1: 1.30	23.42
A61	1: 1.37	23.53
A61K	1: 1.25	24.51

Table 4: Sentence type-to-token ratios and average sentence lengths.

and subclass that we examined, with the sole exception of A61K. Future linguistic analysis will clarify the unexpected behavior of A61K.

Sentence type closure was not observed; this result is consistent with the findings of McEnery and Wilson (2001) and Temnikova and Cohen (2013).

We examined the hypothesis that the further one descends down the IPC hierarchy, the closer the fit is to the sublanguage model. Here the results were more mixed. Descending the hierarchy of the H class, the hypothesis was supported. However, the behavior of the A class was not consistent with this hypothesis, and in fact it was unclear whether the A61K subclass fit the sublanguage model at all.

The type:token ratios showed different results for the A and the H categories. One fact that can be observed is, that in the case of H/H01/H01L the type:token ratios for lexical and token-POS pairs closures increase going down the hierarchy, as it would be expected from the increasing sublanguage specialization. The A/A61/A61K categories show the opposite: the type:token ratios are decreasing going down the hierarchy and approaching the general English values. These findings once again underline the unexpected nature of the A61K category.

The differing closure properties of the H class and the A class speak to a problem that we mulled over in the design of these experiments: is it meaningful to talk of the language of "patents" as a whole, or should we think in terms of there being many different kinds of languages of patents? The differences between the H and A class suggests that we should think of patents as representing a number of different language varieties. This raises the question of how well the language varieties line up with the IPC classification.

The size of the materials in this study allowed us to evaluate a hypothesis that has not been considered in any previous studies of the closure properties of language. McEnery and Wilson (2001) worked with samples of 200,000 words from each corpus. Temnikova and Cohen (2013) worked with samples of about 450,000 words. This study used samples of 1,000,000 words. Studies of closure properties have previously failed to consider the possibility that closure properties might be observed with small samples, but that there might be a "spikiness" to the distribution of lexical and other linguistic types that would reveal a lack of closure if larger samples were considered. The limiting factor in any study of closure properties is generally the size of the sublanguage sample; we considered here a sample more than twice the size of the previously largest sample, and still observed closure properties quite clearly. In this age of massive data sets, 1,000,000 words perhaps no longer qualifies as a "large" sample, but it is the most stringent test thus far of the ability of the sublanguage model to hold as sample size is increased beyond that of previous studies.

The results of this study hold out the promise of further development of semantic search for patents. However, they make it clear that this will be a broad problem, with the necessity to tackle different classes of patents separately, confirming the findings of Oostdijk et al. (2010). This study has shown that sublanguages exist in patents and that it is possible to recognize them using the techniques that we applied. Being able to recognize the presence of sublanguages in patents, the next step will be to develop techniques to characterize those sublanguages—to discover and describe *how* the patent sublanguages differ from the general language and from each other, and thence to develop methods of semantic search.

Acknowledgments

The work of Irina Temnikova and Galia Angelova 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 Negacy D. Hailu was supported by grants NIH 5R01 LM009254-07 and NIH 5R01 LM008111-08 to Lawrence E. Hunter.

Finally, we would like to thank Mihai Lupu and Allan Hanbury for providing access to the MAREC400k corpus.

References

Stephen Adams. 2010a. The text, the full text and nothing but the text: Part 1–standards for creating textual information in patent documents and general search implications. *World Patent Information*, 32(1):22–29.

Stephen Adams. 2010b. The text, the full text and nothing but the text: Part 2–the main specification, searching challenges and survey of availability. *World Patent Information*, 32(2):120–128.

Doreen Alberts, Cynthia Barcelon Yang, Denise Fobare-DePonio, Ken Koubek, Suzanne Robins, Matthew Rodgers, Edlyn Simmons, and Dominic DeMarco. 2011. Introduction to patent searching. In *Current challenges in patent information retrieval*, pages 3–43. Springer.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media.

David Hunt, Long Nguyen, and Matthew Rodgers. 2007. *Patent searching: Tools and techniques*. Wiley.

Hideo Itoh, Hiroko Mano, and Yasushi Ogawa. 2003. Term distillation in patent retrieval. In *Proceedings of the ACL-2003 workshop on patent corpus processing-Volume 20*, pages 41–45. Association for Computational Linguistics.

Makoto Iwayama, Atsushi Fujii, Noriko Kando, and Akihiko Takano. 2003. Overview of patent retrieval task at NTCIR-3. In *Proceedings of the ACL-2003 workshop on Patent corpus processing-Volume 20*, pages 24–32. Association for Computational Linguistics.

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.

Hideo Joho, Leif A. Azzopardi, and Wim Vanderbauwhede. 2010. A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. In *Proceedings of the third symposium on information interaction in context*, pages 13–24. ACM.

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

Geoffrey Leech, Roger Garside, and Michael 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*.

Darren Hsin-hung Lin and Shelley Ching-yu Hsieh. 2004. Collocation features of independent claim in US patent documents: Information retrieval from LexisNexis.

Darren Hsin-hung Lin and Shelley Ching-yu Hsieh. 2010. The specialized vocabulary of modern patent language: Semantic association in patent lexis. In *Proceedings of the 24th Pacific Asia Conference on Language, Information, and Computation (PACLIC 24)*.

Mihai Lupu. 2011. *Current challenges in patent information retrieval*, volume 29. Springer-Verlag Berlin Heidelberg.

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

Marie-Francine Moens. 2006. *Information extraction: Algorithms and prospects in a retrieval context*. Springer.

Nelleke Oostdijk, Eva D'hondt, Hans Van Halteren, and Suzan Verberne. 2010. Genre and domain in patent texts. In *Proceedings of the 3rd international workshop on Patent information retrieval*, pages 39–46. ACM.

Svetlana Sheremetyeva, Sergei Nirenburg, and Irene Nirenburg. 1996. Generating patent claims from interactive input. In *Proceedings of the Eighth International Workshop on Natural Language Generation*, pages 61–70.

Akihiro Shinmori, Manabu Okumura, Yuzo Marukawa, and Makoto Iwayama. 2003. Patent claim processing for readability. In *Proceedings of ACL 2003 Workshop on Patent Corpus Processing Workshop*.

Irina Temnikova and K. Bretonnel Cohen. 2013. Recognizing sublanguages in scientific journal articles through closure properties. In *Proceedings of the 12th Workshop on Biomedical Natural Language Processing (BioNLP 2013)*.

Irina Temnikova, Ivelina Nikolova, William A. Baumgartner Jr., Galia Angelova, and K. Bretonnel Cohen. 2013. Closure properties of Bulgarian clinical text. In *Proceedings of RANLP 2013*.