Sublanguage Corpus Analysis Toolkit: A tool for assessing the representativeness and sublanguage characteristics of corpora

Irina P. Temnikova[♡], William A. Baumgartner Jr.[♠†], Negacy D. Hailu[♠‡], Ivelina Nikolova[♣▽], Tony McEnery[♢], Adam Kilgarriff[♦], Galia Angelova[♣⊕], and K. Bretonnel Cohen[♠⊎]

[♡]Qatar Computing Research Institute, Doha, Qatar
[♠]Computational Bioscience Program, Univ. of Colorado School of Medicine, USA
[♣]IICT, Bulgarian Academy of Sciences, Sofia, Bulgaria
[♢]Department of Linguistics and English Language, Lancaster University, Lancaster, UK
[♦]Lexical Computing Ltd., Brighton, UK
[♡]itemnikova@qf.org.qa,[†]william.baumgartner@ucdenver.edu, [‡]negacy.hailu@ucdenver.edu, [▽]iva@lml.bas.bg,
[♢]a.mcenery@lancaster.ac.uk, [♦]adam@lexmasterclass.com, [⊕]galia@lml.bas.bg, [⊎]kevin.cohen@gmail.com

Abstract

Sublanguages are varieties of language that form "subsets" of the general language, typically exhibiting particular types of lexical, semantic, and other restrictions and deviance. SubCAT, the Sublanguage Corpus Analysis Toolkit, assesses the representativeness and closure properties of corpora to analyze the extent to which they are either sublanguages, or representative samples of the general language. The current version of SubCAT contains scripts and applications for assessing lexical closure, morphological closure, sentence type closure, over-represented words, and syntactic deviance. Its operation is illustrated with three case studies concerning scientific journal articles, patents, and clinical records. Materials from two language families are analyzed—English (Germanic), and Bulgarian (Slavic). The software is available at `sublanguage.sourceforge.net` under a liberal Open Source license.

Keywords: sublanguage recognition, sublanguage characterisation, corpus linguistics

1. Introduction

A fundamental early stage in working with a corpus is to analyze its properties. Such an analysis commonly includes steps such as checking to see if its contents fit Zipf's law and detecting over-represented words. These analyses can be done quite easily with simple scripts. However, there are other types of analyses that are useful but that cannot currently be accomplished without a specialized software package. This paper describes the *Sublanguage Corpus Analysis Toolkit (SubCAT)*, the first set of tools and simple format specifications for assessing the representativeness of a corpus and whether or not it is a fit to the sublanguage model.

We illustrate three use cases for SubCAT, and show that it can be applied to a wide variety of genres and to multiple languages with a variety of character encodings. The corpus of interest need only be converted to a very simple input format.

1.1. Corpora and Representativeness

Representativeness is an important, but infrequently defined, notion in corpus linguistics. We will define representativeness as the extent to which a corpus or other language sample reflects the language from which it is sampled. As such, representativeness is a continuum, ranging from large balanced samples such as the British National Corpus (BNC) to highly specialized corpora such as health care records or weather reports. Recently, McEnery and Hardie have defined it as follows: "A *representative* corpus is one that is sampled in such a way that it contains all the types of text, in the correct proportions, that are needed to make the contents of the corpus an accurate reflection of the whole of the language or variety that it samples. See also *balance*." The latter term is then defined as: "A property of a corpus (or, more properly, of a corpus sampling frame). A corpus is said to be *balanced* if the relative sizes of each of its subsections have been chosen with the aim of adequately representing the range of language that exists in the population of texts being sampled" (McEnery and Hardie, 2012). Thus, there is a relationship between balance and representativeness—in particular, balance might lead to representativeness. However, they remain separate—as a reviewer pointed out, the British National Corpus is balanced, but might no longer be considered representative of the current language.

Closure is the tendency toward finiteness in a genre or sample of language. It is exemplified, for instance, by limited vocabularies in a specialized domain. If unrestricted natural language tends toward the infinite, then we see the opposite in language samples from restricted domains. A body of work inspired by sublanguage theory and begun by McEnery and Wilson (McEnery and Wilson, 2001) has focused on studying the closure properties of language. However, no publicly available tools for language closure have been released yet. The clearest conclusion from this line of research is that corpora constructed from restricted domains exhibit closure, or a tendency towards finiteness; the corollary, which so far has not been utilized in corpus analysis, is that representative corpora do not show closure. This insight is put to practical use in SubCAT.

1.2. Sublanguages

According to long-accepted definitions, *sublanguages* are "subsets of general language" (Grishman and Kittredge, 1986; Kittredge, 2003), which exhibit "a systematic language-like behaviour" (Kittredge, 2003) and "arise spontaneously" in "restricted semantic domains" (Kittredge, 2003). Sublanguages are used by a community of

specialists (Kittredge, 2003) to discuss restricted semantic domain's issues in "recurrent situations". Sublanguages differ from the general language by having, among others (Grishman and Kittredge, 1986; Kittredge, 2003; McDonald, 2000):

- restricted syntax, text structure, and lexicon
- deviant syntax and lexicon (e.g. words which occur only in this variety of language)
- different frequencies of word occurrence and syntactic patterns

Some recent studies of sublanguages have been (Lippincott et al., 2011), which examined the distribution of a variety of lexical and syntactic features across a wide range of biomedical subdomains, and (Mihaila et al., 2012), which looked at the distribution of a wide variety of semantic categories across these domains. (Kilgarriff, 2001) is one study that sets out to measure the differences between different corpora.

Recognizing whether a text has been written in a sublanguage is relevant to Natural Language Processing (NLP). Knowing whether the text is written in a sublanguage can help in designing an application accordingly. A classic example of a high-performing NLP application in a restricted semantic domain is the Montreal machine translation system for weather reports, TAUM-MÉTÉO (Kittredge, 2003). Many other NLP applications have benefited from this, including information extraction, noun compound interpretation, speech recognition, natural language generation, and question answering (Hirschman and Sager, 1982; Grishman, 2001; Finin, 1986; Sekine, 1994; Somers, 2000; McDonald, 2000; Mollá and Vicedo, 2007). Awareness of the phenomenon of sublanguages is also of importance in corpus linguistics. Corpora that are intended for theoretical linguistics usage or for general-domain natural language processing both need to be representative. This requires including language samples from a variety of genres and domains. Recognizing that a type of text represents a sublanguage tells the corpus linguist that this type of text should be included in his or her sample.

1.3. Different steps of analysis: recognition and characterization

We posit two steps in the analysis of sublanguages: recognition, and characterization. *Sublanguage recognition* is the task of recognizing that a sublanguage exists in a sample. *Sublanguage characterization* is the task of describing how that sublanguage differs from the general language. The current version of SubCAT is concerned mainly with sublanguage recognition; the current state of sublanguage characterization is limited to producing a list of overrepresented words, detecting sentences with highly aberrant syntax, and measuring sentence length. A complete sublanguage characterization module is currently under production.

2. Methods

Our sublanguage recognition approach is based on a slightly modified version of the sublanguage closure detection method of McEnery and Wilson (2001). The sublanguage characterization method includes Kilgarriff's Simplemaths (Kilgarriff, 2012) and a number of scripts which measure average sentence length and the number of verbless sentences (Temnikova et al., 2013b).

2.1. Input and output files

SubCAT was designed to be very easy to use, with a minimum of format conversion required. To this end, the package requires only two file formats, as follows:

1. A file containing word – part-of-speech (POS) pairs, one pair per line, including repetitions. POS tags can be both single- and multi-word. Any tag set can be used.
2. A file containing POS tag sequences of each sentence, one sentence per line, including repetitions. Again, any tag set can be used.

The format of the input files can be seen in Table 1. Column 1 shows Input format 1 and Column 2 shows Input format 2. The examples are taken from BNC, parsed with the Machinese Connexor parser (Temnikova and Cohen, 2013), and lowercased. The corresponding words and sentence are displayed in **bold**. It has been demonstrated that tagset differences between the corpora under investigation do not affect our software's results (Temnikova and Cohen, 2013).

Input format 1	Input format 2
the, det	**det;n nom sg;v pres sg3;adv;en**
body, n nom sg	**n nom sg;n nom sg;v pres sg3;en**
is, v pres sg3	**pron;v pres;prep;det;n nom sg**
seriously, adv	pron sup pl;v pres;adv
infected, en	prep;det;det;n nom sg;v pres sg3
hospital, n nom sg	pron wh;v pres;pron pers nom pl3
treatment, n nom sg	n nom sg;n nom sg
is, v pres sg3	abbr nom sg;n nom sg
needed, en	n nom sg;v pres sg3;det sg;n nom sg
some, pron	adv wh;v pres;n nom;v inf;prep;n nom pl
die, v pres	pron nom sg3;v pres sg3;a abs
of, prep	<ex> adv;v pres sg3;det sg;n nom
the, det	n nom sg;v pres sg3;ing;a abs;a abs
infection, n nom sg	pron pers nom sg1;v pres sg1;a abs

Table 1: Format of SubCAT's input files. The Connexor Machinese tag set is shown; any tag set can be used.

The user can choose to use the whole documents for the analysis or to do Brown-corpus-style sampling. A command-line switch allows the user to specify the size of samples to be extracted from individual documents, in case the user prefers Brown-corpus-style samples.

The output in all cases is a CSV file, easily imported into Excel, OpenOffice, or other data plotting programs.

2.2. Description of algorithms

The software implements four different algorithms for corpus analysis:

- Lexical closure analysis
- Type/POS closure analysis

- Sentence type analysis
- Over-represented words

Lexical closure analysis: The lexical closure analysis algorithm detects the way that vocabulary size changes as increasingly large amounts of the corpus are observed. As tokens are observed sequentially, the number of types that those tokens represent is counted. The number of types observed is output at every 1,000 tokens. General language samples will tend to show continued growth in the number of types as long as new tokens are observed – a lack of closure. Sublanguages will show a tapering off in the growth of the number of types after some number of tokens have been observed—in other words, closure.

Type/POS closure analysis: The type/POS closure analysis algorithm detects the way that the set of part of speech tags for word types changes as increasingly large amounts of the corpus are observed. As tokens are observed sequentially, the number of types that those tokens represent is counted. The number of types observed is output at every 1,000 tokens. Representative samples show increasing number of type/POS sets as more tokens are observed. In sublanguages, word types are coerced into more parts of speech, and closure is observed after some number of tokens have been observed.

Sentence type closure analysis: The sentence type closure algorithm detects the way that the size of the set of sentence types observed changes as increasingly large amounts of the corpus are observed. A sentence type is defined as a sequence of part of speech tags. We note that this is arguably not a syntactic description of a sentence at all. However, it is both theoretically neutral and extremely sensitive to differences in sentences. Additionally, this representation has been found to yield results similar to experiments using more linguistically motivated representations, as will be seen below.

Over-represented words analysis: This analysis finds words that are over-represented in the corpus under analysis as compared to some reference corpus. Note that it does not find the most frequent words – it finds words that occur more often than would be expected. The basic principle is to calculate the ratio of frequencies in the corpus under analysis to frequencies in the background or reference corpus (Kilgarriff, 2012). Rather than simple smoothing, an adjustable parameter in the range from 1 to infinity can be set to bias the analysis towards finding over-represented content words or over-represented function words (also conceivable as over-represented rare words versus over-represented common words).

Additionally, format conversion scripts are included for a variety of corpora, currently including the British National Corpus, the Bulgarian National Reference Corpus, the GENIA corpus, and the CRAFT corpus.

3. Results

SubCAT has been applied and evaluated in three different scenarios. Here we discuss the three scenarios, give an example of each of the three closure measures described above, and describe the portability and availability of the software.

3.1. Scientific journal articles

In (Temnikova and Cohen, 2013), SubCAT was used to assess the fit of two corpora of scientific journal articles from the molecular biology domain to the sublanguage model. Using the British National Corpus as a reference corpus, SubCAT showed that both molecular biology corpora were good fits to the sublanguage model, while in contrast, the British National Corpus has the characteristics of a representative corpus. Figure 1 shows the lexical closure characteristics of the three corpora. It reveals that both of the molecular biology corpora show lexical closure – growth in the number of lexical types is much slower than in the British National Corpus and asymptotes after about 50,000 lexical tokens have been examined – while the British National Corpus shows no tendency towards lexical closure.

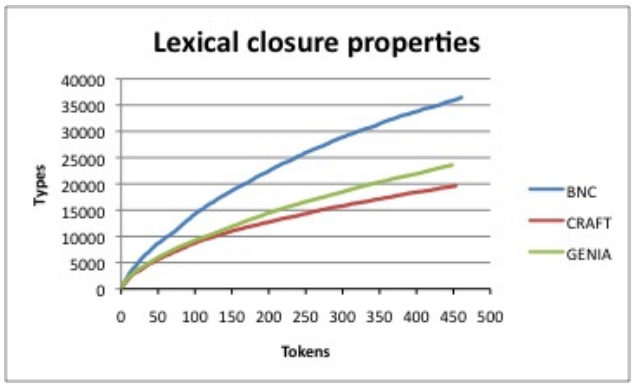

Figure 1: Lexical closure properties, comparing the British National Corpus and two corpora of the molecular biology domain, CRAFT and GENIA. Tick-marks on the *x* axis indicate increments of 50,000 tokens.

3.2. Patents

In (Temnikova et al., 2013a), SubCAT was used to assess the fit of a variety of collections of patents to the sublanguage model. Patents are hierarchically classified, with lower classifications in the hierarchy corresponding to more granularly divided domains. (Temnikova et al., 2013a) tested the hypothesis that fit to the sublanguage model increases the further in the hierarchy one descends. Figure 2 shows the type-POS closure properties of the patents and the British National Corpus. The British National Corpus shows no tendency towards closure at all. Patents at all levels of the hierarchy show clear tendencies towards closure, with greater tendency towards closure the farther down the hierarchy one descends: the patents fit the sublanguage model, and the fit increases as one descends the hierarchy; in contrast, the British National Corpus again shows the characteristics of a representative corpus. SubCAT was also used to measure average sentence length (Temnikova et al., 2013a).

3.3. Bulgarian patient records

In (Temnikova et al., 2013b), SubCAT was used to test whether a language other than English, with quite different morphological characteristics and a non-Latin script,

Word type		Lemma	
ч	hour	ч	hour
/	/	/	/
лечение	treatment	диабетна	diabetic, f. sg.
диабет	diabetes	лечение	treatment
;	;	диабет	diabetes
х	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 2: 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.

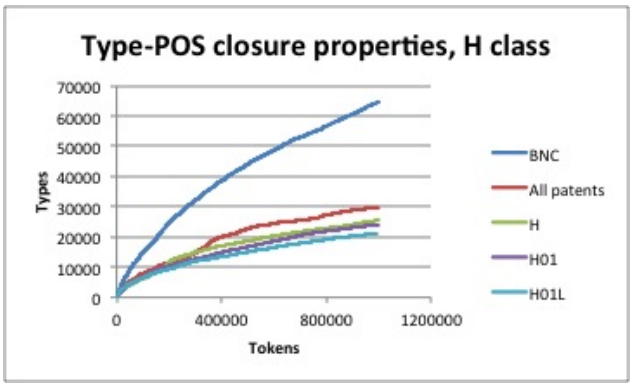

Figure 2: Type-POS closure properties of patents. *All patents* is a sample from the full collection of patents, *H* is a class within all patents, *H01* is a sub-class of the H class, and *HO1L* is a sub-class of the H01 sub-class. Tick-marks on the *x* axis indicate increments of 400,000 tokens.

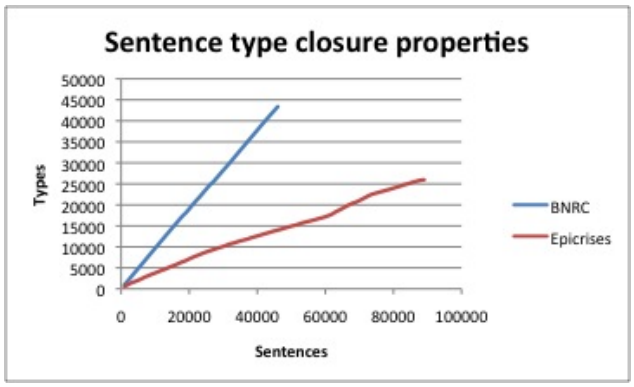

Figure 3: Sentence type closure properties in Bulgarian. *BNRC* is the Bulgarian National Reference Corpus. *Epicrises* is the collection of Bulgarian patient records. Tick-marks on the *x* axis indicate increments of 20,000 tokens.

showed similar closure properties in a restricted domain. Documents from patient health records from an endocrinology hospital were compared to the Bulgarian National Reference Corpus. It was found that the Bulgarian clinical records showed the closure properties of a sublanguage for all three metrics. In fact, this was the only study ever to demonstrate closure for sentence types; previous studies had shown lexical closure and type-POS closure, but even the experiments on a controlled language in McEnery and Wilson had not shown sentence type closure. Figure 3 shows the sentence type closure properties of the Bulgarian patient records and the Bulgarian National Reference Corpus. The Bulgarian National Reference Corpus, in contrast with the patient records, shows almost a 1:1 sentence type to sentence token ratio—there is no tendency towards closure whatsoever.

In addition, in (Temnikova et al., 2013b), SubCAT was used to find **over-represented words** in Bulgarian patient records and also to record the number of verbless sentences. The results showed that clinically relevant words were extractable by this methodology and that Bulgarian patient records are characterized by 66% of verbless sentences. Table 2 shows the word types and lemmata that are over-represented in Bulgarian epicrises.

3.4. Availability and portability of software

The SubCAT toolkit, as well as example files in the required formats and several corpus format conversion scripts, is available at `sublanguage.sourceforge.net`. The software has been tested on Mac OSX, Windows, and a variety of Linux operating systems.

4. Discussion and Future work

We have presented the first toolkit which allows automatic recognition of whether a corpus is written in a sublanguage or whether it is a sample of the representative language. Future extensions of SubCAT will include a sublanguage characterization module, which will provide a picture of a variety of basic characteristics of the sublanguage under investigation.

5. Acknowledgments

The authors would like to thank the three anonymous reviewers for their critiques. Chris Brew and other members of the CORPORA-L mailing list participated in a discussion of the notions of balance and representativeness that we found quite helpful in refining the final version

of the paper. Jin-Dong Kim helped us interpret the results of the experiment on scientific journal articles. The AComIn project (EC FP7 grant 316087) supported Galia Angelova, Irina Temnikova, and Ivelina Nikolova's work. Irina Temnikova's work was additionally supported by the Qatar Computing Research Institute[1].

6. References

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

Grishman, R. and Kittredge, R. (1986). *Analyzing language in restricted domains: sublanguage description and processing*. Lawrence Erlbaum Associates.

Grishman, R. (2001). Adaptive information extraction and sublanguage analysis. In *Proc. of IJCAI 2001*.

Hirschman, L. and Sager, N. (1982). Automatic information formatting of a medical sublanguage. In Kittredge, R. and Lehrberger, J., editors, *Sublanguage: studies of language in restricted semantic domains*, pages 27–80. Walter de Gruyter.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6:1–37.

Kilgarriff, A. (2012). Getting to know your corpus. In *Text, speech and dialogue*.

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

Lippincott, T., Séaghdha, D. Ó., and Korhonen, A. a. (2011). Exploring subdomain variation in biomedical language. *BMC bioinformatics*, 12(1):212.

McDonald, D. D. (2000). Natural language generation. In Dale, R., Moisl, H., and Somers, H., editors, *Handbood of Natural Language Processing*, pages 147–179. Marcel Dekker.

McEnery, T. and Hardie, A. (2012). *Corpus Linguistics*. Cambridge University Press.

McEnery, T. and Wilson, A. (2001). *Corpus Linguistics*. Edinburgh University Press, 2nd edition.

Mihaila, C., Batista-Navarro, R. T., and Ananiadou, S. (2012). Analysing entity type variation across biomedical subdomains. In *Third workshop on building and evaluating resources for biomedical text mining*, pages 1–7.

Mollá, D. and Vicedo, J. L. (2007). Question answering in restricted domains: An overview. *Computational Linguistics*, 33(1):41–61.

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

Somers, H. (2000). Machine translation. In Dale, R., Moisl, H., and Somers, H., editors, *Handbook of Natural Language Processing*, pages 329–346. Marcel Dekker.

Temnikova, I. P. and Cohen, K. B. (2013). Recognizing sublanguages in scientific journal articles through closure properties. In *Proceedings of BioNLP 2013*.

Temnikova, I. P., Hailu, N. D., Angelova, G., and Cohen, K. B. (2013a). Measuring closure properties of patent sublanguages. In *Recent Advances in Natural Language Processing*, pages 659–666.

Temnikova, I. P., Nikolova, I., Jr., W. A. B. ., Angelova, G., and Cohen, K. B. (2013b). Closure properties of Bulgarian clinical text. In *Recent Advances in Natural Language Processing*, pages 667–675.

[1]http://www.qcri.org.qa