SuperCAT: The (New and Improved) Corpus Analysis Toolkit

K. Bretonnel Cohen[1], William A. Baumgartner Jr.[1], Irina Temnikova[2]
[1]University of Colorado School of Medicine, Aurora, Colorado, USA
[2] Qatar Computing Research Institute, Doha, Qatar

Abstract

This paper reports SuperCAT, a corpus analysis toolkit. It is a radical extension of SubCAT, the Sublanguage Corpus Analysis Toolkit, from sublanguage analysis to corpus analysis in general. The idea behind SuperCAT is that representative corpora have no tendency towards closure—that is, they tend towards infinity. In contrast, non-representative corpora have a tendency towards closure—roughly, finiteness. SuperCAT focuses on general techniques for the quantitative description of the characteristics of any corpus (or other language sample), particularly concerning the characteristics of lexical distributions. Additionally, SuperCAT features a complete re-engineering of the previous SubCAT architecture.

Keywords: corpus, representativeness, sublanguage, toolkit

1. Introduction

This paper reports SuperCAT, a corpus analysis toolkit. It is a radical extension of SubCAT, the Sublanguage Corpus Analysis Toolkit (Temnikova et al., 2014), from sublanguage analysis to corpus analysis in general. SubCAT was originally developed to assess the fit of corpora to the sublanguage model. To this end, SubCAT contained Python scripts and applications for analyzing lexical, morphological, and sentence type closure (McEnery and Wilson, 2001), as well as over-represented words and syntactic deviance in corpora. The idea, taken from (McEnery and Wilson, 2001), is that sublanguages exhibit a tendency towards closure (roughly finiteness), and often syntactic deviance. The over-represented-word functionality aids in investigating the lexical/semantic characteristics of the domain represented in the sublanguage. SubCAT was tested on several different corpora, containing texts from different domains (Temnikova et al., 2013a; Temnikova and Cohen, 2013) and languages (Temnikova et al., 2013b). SuperCAT (Version 2.0 of SubCAT) represents a considerable extension of SubCAT to analyze arbitrary corpora. It quickly became evident that while SubCAT was useful for recognizing sublanguages, it also had applications to what might be thought of as the "opposite" task—recognizing when a sample of language has the characteristics of a representative sample of language. Therefore, in the current version, we focus less on sublanguage-specific issues and more on general techniques for quantitative description of the characteristics of any corpus (or other language sample). Additionally, SuperCAT features a complete re-engineering of the SubCAT architecture. This version is distributed via a Java JAR file, allowing the associated programs to be run without requiring installation of Python. A number of R scripts for graphing the results of the Java application analyses are also provided.

2. Motivation for SubCAT and SuperCAT

The original motivation for SubCAT was to provide tools for evaluating the fit of a sample of language to the sublanguage model. The sublanguage model describes language in combinations of specific domains and genres (e.g. molecular biology scientific journal articles and patents written in English, or patient records written in Bulgarian) (Temnikova et al., 2014). Examples of sublanguages that have been identified to date include immunology journal articles (Harris et al., 1989), weather reports (Kittredge, 2003), aircraft and automotive technical manuals (Rinaldi et al., 2004; Ciravegna, 1995), mathematical texts (Sojka et al., 2011), and others (McDonald, 2000; Sekine, 1994; Somers, 2000; Grishman, 2001; Mollá and Vicedo, 2007; Grishman and Kittredge, 1986; Hirschman and Sager, 1982; Finin, 1986). The ability to identify a sublanguage is important (in part) because once we know that a sample of texts represents a sublanguage, we can make use of that fact in designing natural language processing applications, e.g. machine translation, information extraction, speech recognition, natural language generation, automatic terminology recognition, and question answering (Lewis et al., 2011; Eck et al., 2004; Grishman, 2001; Kittredge, 2003; Rinaldi et al., 2004; Mollá and Vicedo, 2007; McDonald, 2000; Ciravegna, 1995; Butters and Ciravegna, 2008). (There are theoretical and knowledge representation implications, as well, but we focus on natural language processing for this audience.)

A sublanguage is necessarily *not* representative of the language as a whole—this is why it shows a tendency towards closure. A representative sample of the language as a whole shows the opposite tendency—that is, a tendency towards infinite growth, or what one might call "openness," on an analogy to open-class words. Thus, the same functionality in SubCAT that indicates the extent of a corpus's fit to the sublanguage model indicates equally well the opposite—its representativeness. For this reason, we have renamed the application SuperCAT, to de-emphasize the focus on sublanguages and instead emphasize its applicability to assessing representativeness (or lack thereof).

2.1. Related work on corpus analysis tools

The Sketch Engine (Kilgarriff et al., 2004) is a corpus analysis tool that takes as its input a corpus of any language and produces "word-sketches," i.e. short summaries of each word's grammatical and collocational behaviour.

LinguaStream (Widlöcher and Bilhaut, 2005; Bilhaut and

Widlöcher, 2006) uses the principal of incremental enrichment to allow visual assembly of modules for corpus analysis at various levels, from the morphological to the discursive.

The **AntConc** family of tools[1] (Anthony, 2005) comprise concordancing tools for different languages, providing additional possibilities, such as analysis of parallel corpora, vocabulary profiling, file conversion tools, etc. However, none of them is specifically designed for sublanguage detection and analysis, nor for corpus representativeness.

Upery (Bourigault, 2002) focuses on corpus analysis with respect to distributional analysis of dependency structures.

WordSmith Tools[2] is a concordancer.

GraphColl (Brezina et al., 2015) is a tool for analyzing networks of linguistic collocations.

Wmatrix[3] offers analysis of word frequencies, concordances, and collocations, and similarly to SuperCAT, it also allows the comparison of a specific domain with a larger, more general corpus.

BNCweb [4] (Hoffmann et al., 2008) is a web-interface allowing browsing of the BNC corpus. It provides functionality such as concordancing, frequency lists, and collocations.

Linguistic Inquiry and Word Count (**LIWC**) [5] is a tool which allows analyzing corpora for specific categories of words, such as those indicating different emotions, thinking styles, social concerns, and parts of speech.

WordStat is a commercial tool. It offers content analysis, authorship attribution, automatic document classification, and GIS mapping.

2.1.1. Summary of related work

Many of these tools provide functionality that is not available in SuperCAT. The primary distinguishing feature of SuperCAT is that it allows evaluation of the representativeness of a corpus, as well as of its fit to a sublanguage model.

3. Functionality

SubCAT functionality that carries over into SuperCAT includes:

- Construction of lexical closure curves
- Construction of syntactic closure curves
- Determination of over-represented words using the simplemaths algorithm (Kilgarriff, 2012; Temnikova et al., 2013b)

New functionality in SuperCAT includes:

- Pre-processing for plain text (untagged) inputs
- Rank-ordered frequency counts
- Automated Kolmogorov-Smirnov test for fit to a power law distribution
- JUnit tests and test data files for the core functionality to allow for worry-free extensions by users
- Extendable classes for representing types and tokens
- Graphing functionality
- Optional output with column headers for easier processing by statistical analysis packages

SuperCAT omits the lexical-POS (part of speech) closure calculations of SubCAT, as it is not clear that this form of closure is well-defined, and impressionistically, we never noted a difference between the lexical closure and lexical-POS closure characteristics of a corpus, regardless of whether or not it fit the sublanguage model.

Figure 1 shows an example of lexical closure curves for equivalently-sized samples of three corpora: the British National Corpus (Burnard, 1998), the CRAFT corpus (Bada et al., 2012; Verspoor et al., 2012; Cohen et al., to appear), and the GENIA corpus (Ohta et al., 2002; Kim et al., 2003a; Kim et al., 2003b). A lexical closure analysis 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—that is, a lack of closure. Sublanguages will show a tapering off in the growth of the number of types after some number of tokens has been observed—in other words, (lexical) closure. The British National Corpus is a representative corpus of English and shows rapid growth in the size of the lexicon as increasingly large numbers of tokens are observed. There is no tendency toward closure, as seen by the fact that the growth curve has not plateaued. In contrast, the CRAFT and GENIA corpora both cover very restricted domains—mouse genomics and human blood cell transcription factors, respectively. The CRAFT and GENIA corpora do not show any tendency toward closure at this sample size, either, but their lexical diversity is clearly far smaller than that of the British National Corpus, as shown by the considerably smaller values on the *y* axis (types) for a given number of tokens. The reasoning behind, and interpretation of, these graphs is discussed in more detail in (Temnikova et al., 2013a; Temnikova and Cohen, 2013; Temnikova et al., 2013b).

Figure 2 shows an example of sentence type closure curves for equivalently-sized samples of two corpora: the Bulgarian National Reference Corpus (Savkov et al., 2012), and a collection of epicrises (analytical summations of case histories, common in European medical practice) (Boytcheva and Angelova, 2009). A sentence type closure analysis 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 in SuperCAT as a sequence of part of speech tags. We note that this is arguably not a syntactic description of a sentence at all, as it is not structured. However, it is both theoretically neutral and extremely sensitive to differences in sentences. It

[1] http://www.laurenceanthony.net/software.html
[2] http://www.lexically.net/wordsmith/
[3] http://ucrel.lancs.ac.uk/wmatrix/
[4] http://corpora.lancs.ac.uk/BNCweb/
[5] http://liwc.wpengine.com

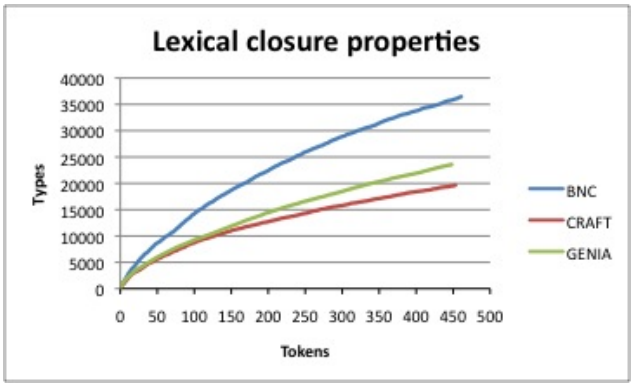

Figure 1: Lexical closure properties, comparing the British National Corpus (BNC) and two corpora of the molecular biology domain, CRAFT and GENIA. Tick-marks on the *x* axis indicate increments of 50,000 tokens. The BNC shows rapid growth in the size of the lexicon as increasingly large numbers of tokens are observed. There is no tendency toward closure, as seen by the fact that the growth curve has not plateaued. The CRAFT and GENIA corpora do not show any tendency toward closure as reflected in plateauing of the growth curve at this sample size, either, but their lexical diversity is clearly far smaller than that of the British National Corpus, as shown by the considerably smaller values on the *y* axis (types) for a given number of tokens.

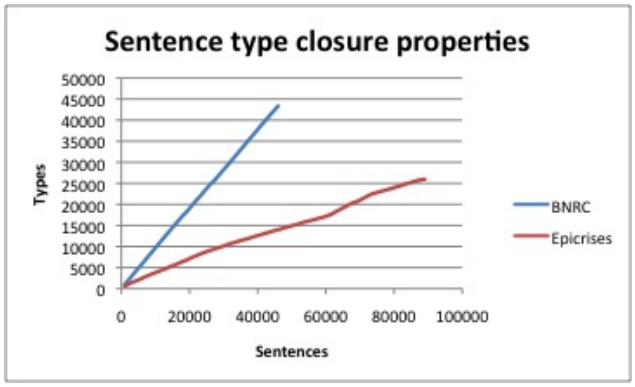

Figure 2: Sentence type closure properties in Bulgarian. *BNRC* is the Bulgarian National Reference Corpus. *Epicrises* is a collection of Bulgarian medical records. Tick-marks on the *x* axis indicate increments of 20,000 tokens. The BNRC, a representative sample of the language, shows no tendency toward closure at all, and in fact has about a 1:1 ratio of sentence types to sentence tokens. In contrast, although the epicrisis corpus also shows no tendency toward closure in terms of a growth plateau at this sample size, it has a much smaller type:token ratio, at 1:3.44.

can be seen that the Bulgarian National Reference Corpus, a representative sample of the language, shows no tendency toward closure (i.e., finiteness of sentence types) at all, and in fact has about a 1:1 ratio of sentence types to sentence tokens. In contrast, although the epicrisis corpus also shows no tendency toward closure at this sample size, it has a much smaller type:token ratio, at 1:3.44.

Figure 3 shows the output of SuperCAT's over-represented word analysis. It is produced by an implementation of Kilgarriff's *simplemaths* algorithm (Kilgarriff, 2012). These are not the most *frequent* words in the document collection, but rather the words that occur more often than would be expected, given the background frequencies of the words in a representative sample of the language (in this case, the Bulgarian National Reference Corpus). Figure 3 displays the top ten most over-represented lexical types, and the top ten most over-represented lemmata. The top fifty terms in each list show the heavy presence of lexical items related either to the semantics or to the unique syntax of the domain. In particular, we see heavy representation of words related to diabetes, body parts, and symptoms. The first item on the list is an abbreviation for the word *chasa* ("hours"), which occurs frequently in the epicrises to indicate the time at which one in a series of blood tests was drawn. It is essential for extraction of trends in laboratory test values. The / (forward slash) character has a variety of primarily syntactic uses in medical records, including linking systolic and diastolic blood pressures and visual acuities.

4. Elements of the new architecture

The old Python scripts of SubCAT have been reimplemented in Java in SuperCAT. The resulting functionality is distributed in a JAR file, enabling running the toolkit from the command line with easily configurable commands. This results in greater ease of use. Additional functionality making use of the graphing and statistical analysis functionality of the R language is made available through an executable R application, so installing R is not required to use SuperCAT.

All source code is included in the JAR file, as well as unit tests and some test data files.

5. Future work: Usability testing

One of the goals of the redesign of the software was to make it easier to use. The original SubCAT required pre-tagged data. The user ran a series of scripts on that data. In contrast, SuperCAT is used via a single Java .jar file. Options include giving it plain text, i.e. untagged, input. (This is in addition to the previous ability to handle pre-tagged data, which has been retained.) To investigate whether or not the new version is as easy to use as we hope, the next step in its development will be usability testing. We will request the participation of a number of groups to try out the new version and test the extent to which it is usable, in the sense of actually being applicable by third parties. Following the approach of work on reproducibility in computer science (Proebsting and Warren, 2015), the goal will be for users to be able to execute a predefined set of tasks within an hour. To eliminate the necessity for doing initial trivial data-munging (POS tagging, formatting of input files, and the like), as well as to control for the effect of data size on times, subjects will be supplied with sample data files. We will then measure:

- The percentage of users able to complete specified tasks within an hour.

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

Figure 3: The output of an over-represented words analysis of a collection of Bulgarian clinical documents. As described in the text, the top 50 terms and symbols reflect the unique syntax and semantics of the domain.

- For those users who are able to complete the specified tasks, the elapsed time.

This widely used approach to usability testing should be able to detect usability problems, as they would be indicated by (a) a low percentage of completion, and (b) high elapsed times for successful completers (Lowdermilk, 2013).

6. Acknowledgements

We gratefully acknowledge the contributions of our co-authors on the earlier SubCAT work: Galia Angelova, Negacy D. Hailu, and Ivelina Nikolova. We especially miss the input and insights of our co-author Adam Kilgarriff. KBC's work was supported by grants NIH 2R01 LM008111-09A1 NIH 2R01 to Lawrence E. Hunter, LM009254-09 NIH to Lawrence E. Hunter, 1R01MH096906-01A1 to Tal Yarkoni, and NSF IIS-1207592 to Lawrence E. Hunter and Barbara Grimpe. WAB's work was supported by NIH grant 2T15LM009451 to Lawrence E. Hunter.

7. Bibliographical References

Anthony, L. (2005). Antconc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *Professional Communication Conference, 2005. IPCC 2005. Proceedings. International*, pages 729–737. IEEE.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Jr., W. A. B., Cohen, K. B., Verspoor, K., Blake, J. A., and Hunter, L. E. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(161).

Bilhaut, F. and Widlöcher, A. (2006). Linguastream: an integrated environment for computational linguistics experimentation. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 95–98. Association for Computational Linguistics.

Bourigault, D. (2002). Upery: un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002), Nancy*, pages 75–84.

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

Brezina, V., McEnery, T., and Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2):139–173.

Burnard, L. (1998). *The British National Corpus*.

Butters, J. and Ciravegna, F. (2008). Using similarity metrics for terminology recognition. In *LREC*. Citeseer.

Ciravegna, F. (1995). Understanding messages in a diagnostic domain. *Information processing & management*, 31(5):687–701.

Cohen, K. B., Verspoor, K., Fort, K., Funk, C., Bada, M., Palmer, M., and Hunter, L. E. (to appear). The Colorado Richly Annotated Full Text (CRAFT) corpus: Multi-model annotation in the biomedical domain. In Nancy Ide et al., editors, *Handbook of Linguistic Annotation*. Springer.

Eck, M., Vogel, S., and Waibel, A. (2004). Improving statistical machine translation in the medical domain using the Unified Medical Language System. In *Proceedings of the 20th international conference on Computational Linguistics*, page 792. Association for Computational Linguistics.

Finin, T. W. (1986). Constraining the interpretation of nominal compounds in a limited context. In Ralph Grishman et al., 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*.

Harris, Z., Gottfried, M., Ryckman, T., Daladier, A., Mattick, P., Harris, T., and Harris, S. (1989). *The form of information in science: Analysis of an Immunology Sublanguage*. Kluwer Academic Publishers.

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

Hoffmann, S., Evert, S., Smith, N., Lee, D., and Berglund-Prytz, Y. (2008). *Corpus linguistics with BNCweb—a practical guide*, volume 6. Peter Lang.

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The sketch engine. *Information Technology*, 105:116.

Kilgarriff, A. (2012). Getting to know your corpus. In *Text, speech and dialogue*, pages 3–15. Springer.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003a). Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl. 1):180–182.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003b). Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl. 1):180–182.

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

Lewis, W. D., Munro, R., and Vogel, S. (2011). Crisis MT: Developing a cookbook for MT in crisis situations. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 501–511. Association for Computational Linguistics.

Lowdermilk, T. (2013). *User-Centered Design: A Developer's Guide to Building User-Friendly Applications*. O'Reilly Media, Inc.

McDonald, D. D. (2000). Natural language generation. In Robert Dale, et al., editors, *Handbood of Natural Language Processing*, pages 147–179. Marcel Dekker.

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

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

Ohta, T., Tateisi, Y., Kim, J.-D., Mima, H., and Tsujii, J. (2002). The GENIA corpus: an annotated corpus in molecular biology. In *Proceedings of the Human Language Technology conference*.

Proebsting, T. and Warren, A. M. (2015). Repeatability and benefaction in computer systems research.

Rinaldi, F., Hess, M., Dowdall, J., Aliod, D. M., and Schwitter, R. (2004). Question answering in terminology-rich technical domains. In *New directions in question answering*, pages 71–86.

Savkov, A., Laskova, L., Kancheva, S., Osenova, P., and Simov, K. (2012). Linguistic analysis processing line for Bulgarian. In *Proceedings of the eighth international conference on language resources and evaluation*, pages 2959–2964.

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.

Sojka, P., Líška, M., and Ružicka, M. (2011). Building corpora of technical texts. *RASLAN 2011 Recent Advances in Slavonic Natural Language Processing*, page 69.

Somers, H. (2000). Machine translation. In Robert Dale, et al., 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.

Temnikova, I. P., Baumgartner Jr, W. A., Hailu, N. D., Nikolova, I., McEnery, T., Kilgarriff, A., Angelova, G., and Cohen, K. B. (2014). Sublanguage corpus analysis toolkit: A tool for assessing the representativeness and sublanguage characteristics of corpora. In *LREC*, pages 1714–1718.

Verspoor, K., Cohen, K. B., Lanfranchi, A., Warner, C., Johnson, H. L., Roeder, C., Choi, J. D., Funk, C., Malenkiy, Y., Eckert, M., Xue, N., Jr., W. A. B., Bada, M., Palmer, M., and Hunter, L. E. (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).

Widlöcher, A. and Bilhaut, F. (2005). La plate-forme linguastream: un outil d'exploration linguistique sur corpus. In *Actes de la 12e Conférence Traitement Automatique du Langage Naturel*, pages 517–522.