Building Webcorpora of Academic Prose with BootCaT

George L. Dillon
University of Washington
PO Box 354330
Seattle, Washington, USA
dillon@u.washington.edu

Abstract

A procedure is described to gather corpora of academic writing from the web using BootCaT. The procedure uses terms distinctive of different registers and disciplines in COCA to locate and gather web pages containing them.

1 Introduction

This is a preliminary report of the results of a new procedure for building a webcorpus of academic writing using BootCaT seeded searches (Baroni and Bernardini, 2004). The procedure inverts the usual one of finding text-internal traits that correlate with externally defined corpora and subcorpora (Lee, 2001). Instead, we seek words and lexical bundles so distinctive of a text-type ("register") that they will induce BootCaT to download texts of that type and no other. In the initial phase, a list of search seed terms is developed to download academic texts; in the second phase, this procedure is refined to increase its resolution, developing search seeds that can bring in texts belonging to sub-types such as Science and Technology, Arts and Humanities, History and Political Science, and so on.

One might object that this is a solution to a non-problem: that all that is necessary is to limit searches to the `.edu` and `.ac.uk` domains to search for academic web texts, at least for US and UK academics. It will become clear, however, that quite a number of qualifying texts can be found elsewhere in other web domains such as `.org`, `.gov`, and even `.com`.

2 Definitions

2.1 Academic writing:

Academic writing is very much a commonsense (or "folk") category. There is considerable agreement that it has research articles and books for disciplinary audiences at its core, but how much more beyond that is in question. This study will draw on three corpus-based studies:

- Coxhead's Academic Word List (AWL; 2000) is drawn from a corpus of 3.5 million words of running text in 441 samples. It is sorted into four equally represented domains (Arts, Commerce, Law, and Science). AWL gives lists of key academic words and word families stratified by frequency in the corpus.
- Biber et al.'s (1999) reference corpus for their lists of academic lexical bundles is a 5.5 million word corpus of articles and sections from books (in equal halves) with a few textbooks for lay readers included. This they further divide into 13 disciplines. Academic is one of four main partitions (which they call 'registers')[1]
- Davies' Corpus of Contemporary American English (COCA) (n.d.). which divides contemporary English into five meta-types, with Academic as one of the five (80 million words)[2]. It

[1]The others are Conversation, Fiction, and News.
[2]The others are Spoken, Fiction, (popular) Magazines, Newspaper and Academic.

Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop, pages 26–31,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

is made up entirely of articles in specialist journals and in a few high-brow general interest periodicals (*American Scholar, Raritan Review, Literary Review, Cato Journal)*. It includes more disciplines (Religion, Fisheries, Photography, Military Science, Agriculture) and is sorted by Library of Congress top headings (A=General Works, B=Philosophy, Psychology, Religion, etc.) consolidated into eight groupings: Education, History, Geog/SocSci, Law/PolSci, Humanities, Phil/Rel, Sci/Tech, Medicine, Misc (=LC A+Z). These eight parts are searchable, so that we can determine for any expression not just whether it is distinctively academic, but what subtype or types it is distinctive of.

COCA is thus built along principles very similar to those of AWL and Biber et al.'s reference corpus, though it is 2 orders of magnitude larger than either. We would expect AWL words and Biber et al.'s distinctively academic lexical bundles also to be distinctively academic in COCA.

2.2 "Distinctive of"

Here are two lists of words and bundles that we can check for distinctiveness in COCA:

AWL	LGSWE
hypothesis	as we have seen
empirical	extent to which
constrain	has been shown
a priori	has been suggested that
ideology	interesting to note
bias	in the development of
concurrent	no significant difference
intrinsic	presence or absence of
whereas	was found that
qualitative	

Table 1: Academic Key Words (starter set)

Each of these expressions occurs over four times more frequently in the COCA Academic register than any other and twice as frequently there as in all the others combined. Let that be the working definition of "distinctive." So for example, *the presence or absence of* occurs 3.72 times per million words in the COCA Academic subcorpus, 0.35 times per million in the Magazines, and negligibly in Spoken, Fiction, and Newspapers. It is thus 10 times more frequent in the Academic subcorpus than in Magazines and passes the 4 times or more criterion for distinctiveness. In addition, some frequent AWL words were checked for combining in bigrams, which have the potential of being even more specific for domain/genre than individual words. These indeed prove to be more distinctive than the individual words (though of course less frequent). In the first column of Table 2 are the frequencies per million words of these words, phrases, and bigrams in COCA.

words seeds	**COCA**	**words**	**bundl**	**bigram**	**wikiped**
hypothesis	74	x	145	140	23
empirical	58	x	30	189	8
constrain	5.5	x	3	10	1
a priori	7	x	3	15	1
ideology	58	x	6	12	16
bias	46	x	41	45	13
concurrent	11	x	10	20	5
intrinsic*	27	x	43	*	6
whereas	105	x	237	114	46
qualitative	38	x	22	109	3
Total (minus intrinsic)	402		498	655	122
acad_bundle					
as we have seen	6	9	x	2	0
extent to which	37	40	x	47	3
has been shown	9	11	x	30	4
has been suggested that	3	4	x	3	6
It is interesting to note	5	5	x	4	3
in the development of	19	16	x	30	10
no significant difference	8	4	x	6	0
presence or absence of	4	4	x	7	1
it was found that	5	2	x	9	2
Total	96	95		137	29
bigrams					
face validity	2	5	1	x	0
these data	18	16	98	x	1
important implications	5	4	5	x	0
basic concepts	2	4	0	x	1
theoretical framework	5	5	1	x	0
intrinsic motivation	6	*	3	x	0
these findings	32	19	75	x	1
this interpretation	6	13	4	x	3
previous work	3	3	11	x	1
indicative of	11	7	15	x	3
Total (minus intrinsic)	89	75	214		10

Table 2: Frequency/Million of Seeds in COCA, Wikipedia,[3] and the Collected Web Corpora

The next three columns give the frequencies/million words of these distinctive terms in each of the three corpora collected from the Web with words, bundles, and bigrams for seeds. (The x's replace the frequencies of the seed terms in the respective

[3] from Corpuseye's 115M word Wikipedia A, B, C corpus at http://corp.hum.sdu.dk/cqp.en.html

corpora made with them—the numbers are of course very high.) These frequencies track those of the terms in the COCA Academic subcorpus quite closely, especially the 'words' corpus, with the 'bundle' and 'bigram' corpora following the same pattern but at somewhat higher (i.e., 'richer') levels.

The Wikipedia figures are included for comparison. The low frequency of these marker words and phrases suggests that Wikipedia is not very academic in its style, which is perhaps not surprising since Wikipedia authors are not allowed to report their own research or argue positions.

Most of these putative marker terms are well represented across the eight academic domains (the "spread" is good). A word that occurs only in one domain will appear to be distinctively academic, but that is a misleading impression. *Stent,* for example, occurs only in the Medical domain in COCA (along with many other terms: *stenosis, mitral*—the list is very long). Even when the match of word and domain is not so clear cut, there are words and phrases that are found preponderantly in a discipline or group of disciplines (a "division" in university parlance) such as *the text itself* and *early modern,* both Art/Humanities terms, and similarly *of the nature of,* which scarcely occurs in Science or Medicine and only infrequently in Geography and Social Science. The next phase of this project will take up the question of increasing the resolution down to the level of a subdomain where a particular set of terms **is** distinctive.

3 Details of Constructing the Web Corpora

These three groups of seed terms were used to make BootCaT tuples (3) and to find URLs of files which contained all three terms of a tuple (with 20 tuples each) and 40 documents for each tuple. Each list of URLs was downloaded and cleaned of CSS style sheets: duplicates and dead-end (e. g., passworded) sites were removed, along with unconverted `.pdf` files. (Yahoo! rather than Google was used because Yahoo! filtered out most of the `.pdfs` and `.doc` files. BootCaT was not too successful converting `.pdf` files: a number of them seemed non-standard or corrupt).

At 3.4 million words, the 'single word' corpus was the largest and had the most pages; the 'bundles' corpus was intermediate in word count but had the fewest pages. The corpus made with the bigram seeds was notably shorter (2.2 million words), but it was very efficient at bringing in occurrences of seed terms from the other sets. The seed terms from all sets were used to cross-check (probe) for occurrence in the other two corpora. These results are given in Table 2 in the second and third columns. There was no overlap of files (i. e., no files in common) in the three downloaded Web corpora and only one overlap between probe term (intrinsic) and file seed (intrinsic motivation).

A further set of lexical bundles (not used as seeds) were run as probes and produced the same pattern (See Table 3). Most of these are *it*-impersonal constructions, and it is not news that academic writing cultivates the impersonal (though it does allow a little *we* plural first person to slip in); in fact, at this proof-of-concept stage, expected findings are confirmation that the collected corpora have the properties of academic writing as we know it across many distinctive lexical bundles, not just the ones used as seeds.

lexical bundles	COCA	20words40	20bun40	bigram4
it should be noted	13	14	16	19
It is possible that	16	22	49	13
it is necessary to	11	16	9	13
it is important to	39	39	30	55
as we shall see	4	1	5	1
it can be seen that	1	1	2	1
in the context of	44	49	28	70
the degree to which	19	16	15	27
of the nature of	7	13	7	12
in this paper	14	27	24	53
Total	188	198	182	264

Table 3: Further probes of the 3 Collected Web Corpora

Again, the three web corpora track COCA very closely, with the 'bigram' corpus as the most efficient.

4 Analysis of Web Corpora

4.1 Top-level domains

Table 4 shows that these corpora draw from several web top-level domains, with `.com` either first or second in rank for the three corpora. (The top four domains account for a little over 90% of the pages.)

3 Webcorpora	20word40	20bun40	bigrams4
CURRENT URL	636	464	556
tokens	3.4M	2.8M	2.23M
tokens/url	5346	6034	4011
.org/	182	245	127
.edu/	174	36	163
.com/	197	113	176
.gov/	27	32	46
.net/	35	13	18
.ca/	19	4	15
.de/	19	13	11
.ac.uk/	6	3	7
.ca/	23	5	18

Table 4: Size of Corpora and Range of Domains (Estimated Domain Counts)

The domain counts are estimated, since a `grep` search over-counts by including URLs referred to on the page as well as that of the page itself. These figures are estimated from the URLs in the "cleaned_urls_list" that is used to download the pages. Clearly the `.edu` top-level domain is not the only or even the most productive source of pages containing the search key words. If these are indeed pages of academic writing, then quite a lot of academic writing on the web would get filtered out by using an `.edu` domain filter and a great deal filtered out using `ac.uk`.

4.2 Types of sites

The 1656 downloaded pages came from a wide range of sites. 281 URLs had words such as *article, journal, paper, research, publication,* or other term that identified the contents as scholarly articles. On the other hand, there were 26 entries from `en.wikipedia.org/wiki/`; 17 pages had the word *blog* in their URL and 17 had the word *courses,* the latter being teaching sites with handouts. There were nine pages of customer reviews from `www.amazon.com/review` and 15 pages from `plato.stanford.edu/entries` which is an encyclopedia of philosophy. All of these sites might be said to be academic in a looser sense, the Amazon reviews being the most distant.

5.0 The Next Phase: Increasing Resolution

It is probably only a minority of 'academic' terms that are commonly used across the board in all disciplines (or groups of disciplines). All disciplines use *have argued*, presumably because argument is at the core of academic endeavor and because the present perfect is just right for summarizing other people's claims and your own. And similarly, all disciplines have, or agree they should have, a *theoretical framework*. But one does NOT write *I argue* in Medicine, or so COCA records, nor has the word *interpretive* any use in Medicine, though it is widely used in all the other disciplines. On the other hand, Medicine has its own distinctive bundles including *it is known that*, and *it has been shown/suggested that* (Oakey, 2002)[4].

It is fairly easy to gather terms that appear to be distinctive of a certain discipline, or group of disciplines, to use them to build web corpora like the ones illustrated here, and to take frequent terms from the gathered corpora to do another search within the discipline/domain, and so to build larger and more differentiated corpora that match the COCA/Library of Congress groupings of disciplines much as has been reported here for 'Academic' writing as such. 'Distinctive' can be less stringently defined in this application: a term is distinctive in an academic subdomain when it is distinctively academic and is at least twice as frequent in the subdomain as in the Academic domain as a whole. The terms still have a strong selective effect because when used in triplets, their selective weights are as it were multiplied.

For example, the left column of Table 5 has a set of search seed terms distinctive of texts in the COCA 'Phil/Rel' subcorpus (LC: B).[5] The right column gives a set of search seeds selected from the first 100-300 most frequent words in the corpus made with the initial set of seeds. (Very frequent terms were checked as possible bigram members and the bigram used instead in the actual download of the second corpus.)

[4]Oakey's model study is based on data from the BNC. Some of his discipline-distinctive patterns scarcely occur in the much larger COCA (e. g, *it is found/concluded that*).

[5]It actually includes Philosophy, Religion, and Psychology.

Initial set of seeds	Second, derived set of seeds
this interpretation	phenomenal consciousness
incomprehensible	methodology
dialectic	scientific theories
situational	reductionist
hermeneutics	incoherent
intelligible	in this sense
materialism	in principle
	means that

Table 5: First and Second Seed Sets for Phil/Rel

The two resulting lists of URLs overlapped only slightly. By using each corpus as reference for a keyword search of the other, the terms most distinctive of each (vis-a-vis the other) were extracted. These terms fall into fairly distinct clusters: The first corpus leans toward hermeneutics/interpretation and toward marxism (via *dialectic* and *materialism*)—in short, a course in Continental philosophy (sample key words: *historical consciousness, Marx, Gadamer, bourgeois, class struggle, Scripture, exegesis*). The second has *Dennett* and *falsifiable* as keys and leans toward Anglo-American philosophy of science and of mind (other key words: *qualia, representational contents, mental states, argument for/against, sensory experience, physicalism)*. Here we begin to tap into key terms and themes of various schools of thought within and overlapping disciplines.

It is possible to determine which of the seed tuplets brought in the key phrases; i. e., the strength of particular seeds as attractors of other terms. It can also be determined when a particular web page is causing a term to spike, which happens fairly often in academic writing, since it favors frequent repetition of the key concepts of the article.

These clusters reflect dependencies (or co-locations) within texts rather than within local 'frames' of contiguous words—which is to say registers of the particular disciplines/subdisciplines. Proceeding in this way, specific lists of terms and also turns of phrase for these disciplines can be extracted. This phase of the project is nearing completion.

The power of BootCaT tuplet search to collect corpora rich in the features of academic registers is remarkable, and its potential uses are many.

References

M. Baroni and S. Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004.

D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan. Longman Grammar of Spoken and Written English. Longman, 1999.

A. Coxhead. 2000. A New Academic Word List. TESOL Quarterly, 34(2): 213-238.

_____, n.d. On-line AWL lists. http://www.victoria.ac.nz/lals/resources/academicwordlist/sublists.aspx

M. Davies. n.d. Corpus of Contemporary American English (COCA) www.americancorpus.org.

D. YW. Lee. 2001. Genres, Registers, Text Types, Do mains, and Styles: Clarifying the Concepts and Nav igating a Path Through the BNC Jungle. Language Learning & Technology 5(3): 37-72.

D. Oakey. 2004. Formulaic Language in English Academic Writing. in R. Reppen, S. M. Fitzmaurice, and Douglas Biber, eds. Using Corpora to Explore Linguistic Variation. John Benjamins. 2004.