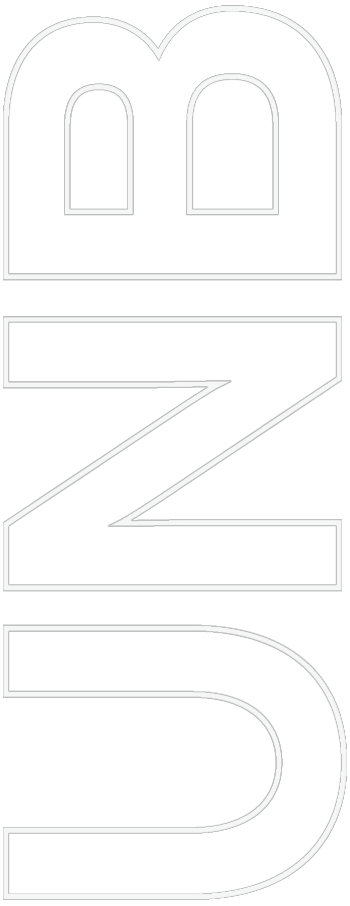
Building and evaluating web corpora representing national varieties of English

Paul Cook, Laurel J. Brinton

Language Resources and Evaluation
Online ISSN: 1574-020X
DOI: 10.1007/s10579-016-9378-z
Publisher: ACM Digital Library, Springer
Published Chapter URL: https://doi.org/10.1007/s10579-016-9378-z
Version of Record available: https://dl.acm.org/doi/abs/10.1007/s10579-016-9378-z



Language Resources and Evaluation manuscript No.
(will be inserted by the editor)

Building and evaluating web corpora representing national varieties of English

Paul Cook · **Laurel J. Brinton**

Received: date / Accepted: date

Abstract Corpora are essential resources for language studies, as well as for training statistical natural language processing systems. Although very large English corpora have been built, only relatively small corpora are available for many varieties of English. National top-level domains (e.g., `.au`, `.ca`) could be exploited to automatically build web corpora, but it is unclear whether such corpora would reflect the corresponding national varieties of English; i.e., would a web corpus built from the `.ca` domain correspond to Canadian English? In this article we build web corpora from national top-level domains corresponding to countries in which English is widely spoken. We then carry out statistical analyses of these corpora in terms of keywords, measures of corpus comparison based on the chi-square test and spelling variants, and the frequencies of words known to be marked in particular varieties of English. We find evidence that the web corpora indeed reflect the corresponding national varieties of English. We then demonstrate, through a case study on the analysis of Canadianisms, that these corpora could be valuable lexicographical resources.

Keywords Web corpora, corpus evaluation, corpus similarity, varieties of English, Canadian English

Paul Cook
Faculty of Computer Science
University of New Brunswick
P.O. Box 4400
Fredericton, N.B. E3B 5A3
E-mail: paul.cook@unb.ca

Laurel J. Brinton
Department of English
University of British Columbia
#397-1873 East Mall
Vancouver, B.C. V6T 1Z1
E-mail: brinton@mail.ubc.ca

1 Web corpora and language varieties

Corpora are essential resources for lexicographical analysis, training natural language processing (NLP) systems, and corpus-based studies of language. Other things being equal, bigger corpora are generally better. For example, large corpora are required to determine a given word's collocations, which can be very informative in lexicographical analysis (Kilgarriff et al, 2004). NLP systems trained on larger corpora will tend to perform better (Banko and Brill, 2001).

In the case of English, there are many large corpora available (e.g., Ferraresi et al, 2008). There are however, many varieties of English, including national varieties of English, such as British English and American English. For some varieties of English, large corpora are available. For example, the British National Corpus (Burnard, 2007), Corpus of Contemporary American English (Davies, 2009), and Strathy Corpus of Canadian English,[1] consist of roughly 100 million, 450 million words, and 60 million words of British, American, and Canadian English, respectively. However, for many varieties of English only much smaller corpora are available. For example, the International Corpus of English covers English in a large number of countries (including Singapore, Hong Kong, and Ireland), but the amount of material available for each country is only one million words. For other varieties of English, somewhat larger corpora are available. For example the Scottish Corpus of Texts and Speech[2] consists of roughly 4.5 million words, but this is still far smaller than the corpora available for British and American English. The availability of large corpora for further varieties of English would enable training variety-specific NLP systems, and would provide valuable resources for lexicographic analysis and linguistic study of varieties of English.

The web is a tremendous source of linguistic data, and has been widely used to automatically create a range of types of corpora. However, there has been relatively little work on building web corpora of national varieties of English. Top-level domains corresponding to countries in which English is widely spoken (e.g., `.uk`, `.au`) could be a promising source of material for building such corpora. However, it is not clear that an English web corpus built from a country's top-level domain would correspond to a corpus of that national variety of English. Although it is relatively straight-forward to build a web corpus from documents in a particular top-level domain, one key challenge is to be able to determine whether a web corpus that has been built corresponds to a particular variety of English.

Cook and Hirst (2012) built web corpora from the `.ca` and `.uk` domains and compared them to conventionally-constructed corpora of Canadian and British English. Their comparisons were based on keyword analyses (Kilgarriff, 2009), an established measure of corpus similarity (Kilgarriff, 2001), and a novel method of comparing corpora based on the relative frequencies of spelling variants (e.g., *color* and *colour*). They found preliminary evidence suggesting that web corpora from the `.ca` and `.uk` domains indeed reflect Canadian and British English, respectively.

[1] `http://www.queensu.ca/strathy/corpus`

[2] `http://www.scottishcorpus.ac.uk/`

In this work we also build web corpora from top-level domains and compare them to conventially-constructed corpora of known varieties of English. This paper contributes to the evidence that English corpora constructed from national top-level domains reflect the corresponding national variety of English, and expands the work of Cook and Hirst (2012) in several important ways:

- we reproduce the findings of Cook and Hirst using entirely different web corpora, demonstrating the robustness of these findings;
- this study is itself replicable because the web corpora used are built from a publicly available web crawl, unlike the corpora used by Cook and Hirst;
- we build and analyse web corpora from a much wider range of top-level domains than the study of Cook and Hirst (although due limitations in the availability of conventionally-constructed corpora against which comparisons are made, not all the web corpora we built could be included in all analyses);
- in addition to comparing the web corpora to conventially-constructed corpora of British and Canadian English, we also compare them against a corpus of American English;
- this study considers a further measure of the extent to which a web corpus corresponds to a national variety of English based on lists of terms from dictionaries that are marked in a particular variety;
- finally, this study demonstrates that web corpora produced from top-level domains can produce useful information for lexicography, through a case study applying the web corpora built to the analysis of Canadianisms.

2 Related work

A number of approaches have been proposed to building general-purpose corpora from the web for both English and other languages.(e.g., Sharoff, 2006; Baroni et al, 2009; Kilgarriff et al, 2010) These corpora have been successful in that they have been used for both training statistical NLP systems, and as a source for evidence in lexicography (Atkins, 2010). Methods for building domain-specific corpora have also been developed (e.g., Baroni and Bernardini, 2004) and subsequently incorporated into the Sketch Engine (Baroni et al, 2006), a commonly-used tool in lexicography. Techniques for building corpora of specific genres (e.g., Dillon, 2010), and parallel corpora (e.g., Resnik and Smith, 2003) have also been proposed.

A smaller number of studies have focused on building web corpora of specific varieties or dialects (e.g., Roth, 2012; Schulz et al, 2013) or of closely related languages (e.g., Ljubešić and Klubička, 2014). This work has relied on information about the top-level domain in which a document is found, but has not focused specifically on English. Murphy and Stemle (2011) built a corpus of Hiberno-English (the variety of English spoken in Ireland) by relying on lexical markers of this variety (e.g., placenames, regionalisms, and loanwords), as opposed to top-level domains. The GLoWbE corpora contain English texts from twenty countries, collected via the web based on Google Search's region function;[3] however, the extent to which these corpora cor-

[3] `http://corpus.byu.edu/glowbe/`

respond to other English corpora from those countries does not appear to have been measured. The findings of this paper could allow us to have greater confidence in the findings of linguistic and lexicographical analyses of the GLoWbE corpora.

One important aspect of web corpus construction is analyzing and evaluating corpora that have been built to determine their composition. One common method for doing so is to examine keywords (e.g., Kilgarriff, 2009) for a corpus with respect to a reference corpus (Schäfer and Bildhauer, 2013). Empirical measures of corpus similarity have also been proposed (Kilgarriff, 2001). However, these are general notions of corpus similarity that do not necessarily reflect the extent to which two corpora correspond to the same variety of English, and could instead reflect similarities with respect to, for example, topic or genre. Cook and Hirst (2012) therefore propose a measure of corpus similarity based on the relative frequencies of spelling variants — where the spellings of words such as *realize* and *realise* are known to have preferences in some English varieties. In this paper we consider all of these methods for analyzing corpora, and in addition consider a novel measure based on the frequency of known varietal terms sourced from dictionaries.

3 Corpora

To build the corpora used in this study we start with the English portion of ClueWeb09,[4] a publicly-available web crawl containing roughly 500 million webpages crawled from January–February 2009. ClueWeb09 has subsequently been used for web corpus construction efforts (e.g., Pomikálek et al, 2012; Cook and Lui, 2012) and in a variety of shared tasks (e.g., Clarke et al, 2011). Crucially by using ClueWeb09 as the starting point for our corpus construction we enable others to recreate the corpora used in our study, to replicate our findings.

From English ClueWeb09, following Ferraresi et al (2008) we select documents of size 5K–200K bytes and MIME type text/html. We only consider documents whose URL's are in the following top-level domains which correspond to the parenthetical countries in which English is widely spoken: `.au` (Australia), `.ca` (Canada), `.hk` (Hong Kong), `.ie` (Ireland), `.in` (India), `.nz` (New Zealand), `.uk` (United Kingdom), `.us` (United States), and `.za` (South Africa). ClueWeb09 is split into a number of sections. Because there are so many documents from `.uk` in ClueWeb09, we consider only the first four sections for this top-level domain. For all other top-level domains we use all of ClueWeb09.

Following Pomikálek et al (2012), we extract textual portions of documents using justext (Pomikálek, 2011). This tool heuristically identifies and removes the boilerplate parts of a document, and was shown to perform well on the CleanEval data (Baroni et al, 2008) — a shared task focused on removing boilerplate and markup from webpages to produce cleaned documents.

The language identification for ClueWeb09 — which was done to ensure that the English portion of ClueWeb09 consists of English documents — was done using

[4] `http://www.lemurproject.org/clueweb09/`

Table 1 Number of tokens and documents in the corpora produced for each top-level domain. The corresponding country for each top-level domain is also shown.

Top-level domain	Country	Tokens ($\times 10^6$)	Documents ($\times 10^3$)
.au	Australia	1220.5	2183.0
.ca	Canada	1301.7	2128.2
.hk	Hong Kong	69.1	134.6
.ie	Ireland	154.5	326.2
.in	India	155.4	302.5
.nz	New Zealand	294.9	552.4
.uk	United Kingdom	1979.0	3945.2
.us	United States	991.7	1522.8
.za	South Africa	245.0	412.7

an implementation of TEXTCAT (Cavnar and Trenkle, 1994).[5] TEXTCAT determines the language of a document based on the relative frequency of the byte *n*-grams it contains. Although the precision of this language identification system was reported to be 99.7% over the ten languages in ClueWeb09, Lui and Baldwin (2011) have shown that this method performs poorly when applied to documents from a different domain than the training data, as was the case for ClueWeb09, where TEXTCAT was trained on newswire and parliamentary documents, and applied to webpages. Lui and Baldwin (2011) present `langid.py`, a method for language identification that, like TEXTCAT, is based on byte *n*-grams, but improves on TEXTCAT by incorporating a Naive Bayes classifier and feature selection. Lui and Baldwin demonstrate that `langid.py` performs significantly better than TEXTCAT on a variety of text types. Furthermore Cook and Lui (2012) show the superiority of corpora built from ClueWeb09 using `langid.py` instead of TEXTCAT in an applied setting. We therefore apply `langid.py` to the selected documents from ClueWeb09 (which have already been identified as English by TEXTCAT) and eliminate any document which `langid.py` labels as non-English.

We eliminate duplicates and near-duplicates using onion (Pomikálek, 2011), which is also used by Pomikálek et al (2012) in building a corpus from ClueWeb09. We use onion with its default settings; onion makes a single pass through the corpus, and discards any paragraph for which more than 50% of its 7-grams have already been observed in the portion of the corpus processed so far.

Finally, we tokenise our corpora using the tokeniser provided by the Stanford Natural Language Processing.[6] The number of tokens and documents in the corpus produced for each top-level domain is shown in Table 1.

In addition to the web corpora described above, in the analysis in the following sections we also consider three manually-compiled English corpora of texts known to have authors from a particular counry: the Open American National Corpus (OANC),[7] the British National Corpus (BNC, Burnard, 2000), and the Strathy Corpus of Canadian English (Strathy).[8] Some of the corpora include transcripts of spoken material,

[5] `http://boston.lti.cs.cmu.edu/clueweb09/wiki/tiki-index.php?page=Language+Identification+for+ClueWeb09`

[6] `http://nlp.stanford.edu/software/tokenizer.shtml`

[7] `http://www.anc.org/OANC/`

[8] `http://www.queensu.ca/strathy/projects.html`

Table 2 Number of tokens and documents in each of the national corpora. The corresponding country for each corpus is also shown.

Corpus	Country	Tokens ($\times 10^6$)	Documents ($\times 10^3$)
BNC	United Kingdon	100.5	3.1
OANC	United States	13.4	6.4
Strathy	Canada	47.2	0.8

in addition to written texts; we restrict our analysis to the written portion of each corpus. We tokenise each of these corpora using the Stanford tools (ignoring any tokenisation provided with the corpora) to ensure that their tokenisation is consistent with the web corpora.

The number of tokens and documents in each corpus is shown in Table 2. For many of the other countries that we consider, large national corpora are not available. For example, the Australian Corpus of English (Green and Peters, 1991) consists of only one million words.[9]

4 Keywords analysis

To analyse the corpora built in the previous section we begin by considering keywords, i.e., words that are much more frequent in one corpus — the *focus corpus* — than another — the *reference corpus*. Kilgarriff (2009) presents a method for identifying keywords based on the ratio of a word's frequency per million, plus a constant, in two corpora. Kilgarriff's keywordness score ($\mathrm{KW}_{\mathrm{Kilgarriff}}$) is shown in Equation 1, where w is the target word; $fpm_x(w)$ is the frequency per million of w in corpus x; fc and rc are the focus and reference corpora, respectively; and c is a constant.

$$KW_{\mathrm{Kilgarriff}}(w) = \frac{fpm_{fc}(w) + c}{fpm_{rc}(w) + c} \tag{1}$$

Kilgarriff's approach is only suitable for the case of comparing a single focus corpus and a single reference corpus. In our case, however, we would like to compare a focus corpus against multiple reference corpora. One solution to this would be to simply concatenate the reference corpora into a single super-corpus. However, because the corpora are not of equal size, larger reference corpora will be over-represented in such a super-corpus, whereas we would like each reference corpus to have the same influence on the keyword analysis. We therefore adapt Kilgarriff's method by averaging over the frequency per million of the target word in each reference corpus, as in Equation 2 below, where rc_i is a particular reference corpus, and n is the total number of reference corpora.

[9] The Australian National Corpus (Peters, 2009, `http://www.ausnc.org.au/`) is an effort to build a larger corpus of Australian English, but is currently a collection of many corpora of diverse types — many of which are spoken — and so does not appear to be suitable for our purposes.

$$KW(w) = \frac{fpm_{fc}(w) + c}{\frac{\sum_{i=1}^{n} fpm_{rc_i}(w)}{n} + c} \quad (2)$$

We compute keywords for each web corpus relative to the other web corpora, and each national corpus relative to the other national corpora, using KW, our new notion of keywordness. We set the constant c to 100, the value recommended by Kilgarriff. We ignore case when computing keywords, and limit our analysis to keywords consisting of alphabetic characters optionally ending with a period. The top-twenty keywords for each corpus are shown in Table 3.

For the national corpora, the keywords indicate that there are domain and genre differences between the corpora, with the BNC containing more narrative (indicated by the presence of pronouns such as *she* and *him* amongst the top keywords), and the OANC containing more biomedical texts. Nevertheless, we also see evidence that the corpora contain documents related to the corresponding countries through the presence of keywords such as geographical place names (e.g. *britain*, *ontario*), demonyms (e.g., *canadian*), and prominent organisations and people (e.g., *cbc*, *clinton*, *nyt*); this trend appears to be stronger for the BNC and Strathy corpus than the OANC, which could be due to the previously-noted bias towards biomedical text in the OANC.

Turning to the web corpora, in all cases except `.us`, local place names (e.g., *australia*, *toronto*) and demonyms (e.g., *irish*, *african*) are common amongst the top keywords. In the case of `.us`, this top-level domain is known to contain many government-related webpages, and this is reflected in the keywords observed for this corpus (e.g., *county*, *state*, *federal*, *department*); this difference in corpus composition could be why we don't observe as many place names and demonyms amongst the keywords for `.us`. We furthermore see keywords related to the culture (e.g., *aboriginal*, *indigenous*), region (e.g., *asia*, *european*, *pacific*) and administration (e.g., *provincial*, *state*) of the corresponding countries. These findings point to a similarity between the national corpora and web corpora — despite their very different methods of construction, both types of corpus contain documents that are topically related to the corresponding country.

5 Comparisons with national corpora

A small number of studies have considered methods for measuring the similarity between corpora (e.g., Kilgarriff, 2001). If web a corpus from a given top-level domain does reflect the corresponding national variety of English, then we would expect that web corpus to be measurably more similar to a corpus known to represent that variety of English, than corpora known to represent other varieties of English. Determining whether this is the case is the focus of this section.

Kilgarriff (2001) considered a number of measures of corpus similarity in experiments on synthetically-constructed known-similarity corpora, and found a measure based on the chi-square test to give corpus similarity judgements that correlated best with the known corpus similarities. Kilgarriff's method calculates the chi-square statistic for the 500 most frequent words in the union of two corpora; this statistic is

Table 3 Top panel: top-twenty keywords for each national corpus relative to the other national corpora; bottom panel: top-twenty keywords for each web corpus relative to the other web corpora.

Corpus	Keywords
BNC	*mr*, *she*, *mrs*, *her*, *uk*, *looked*, *had*, *round*, *london*, *him*, *towards*, *sir*, *britain*, *hon.*, *eyes*, *england*, *voice*, *scotland*, *you*, *lord*
OANC	*clinton*, *cells*, *genes*, *gene*, *expression*, *cell*, *protein*, *proteins*, *sequence*, *sequences*, *percent*, *nyt*, *data*, *rna*, *dna*, *slate*, *wp*, *binding*, *mice*, *mm*
Strathy	*canada*, *canadian*, *toronto*, *ontario*, *ottawa*, *cbc*, *sub*, *quebec*, *proquest*, *id*, *document*, *vancouver*, *montreal*, *et*, *canadians*, *p.*, *iss*, *author*, *al.*, *alberta*
`.au`	*australian*, *australia*, *sydney*, *melbourne*, *nsw*, *queensland*, *posted*, *aboriginal*, *victoria*, *brisbane*, *australians*, *indigenous*, *adelaide*, *commonwealth*, *pm*, *perth*, *canberra*, *tasmania*, *victorian*, *wales*
`.ca`	*canada*, *canadian*, *ontario*, *toronto*, *alberta*, *vancouver*, *canadians*, *ottawa*, *provincial*, *province*, *quebec*, *columbia*, *manitoba*, *bc*, *faculty*, *montreal*, *saskatchewan*, *calgary*, *federal*, *aboriginal*
`.hk`	*hong*, *kong*, *chinese*, *china*, *mainland*, *ordinance*, *edit*, *software*, *internet*, *asia*, *submitted*, *vertical*, *hk*, *administration*, *asian*, *beijing*, *exchange*, *vertex*, *ip*, *chan*
`.ie`	*ireland*, *irish*, *dublin*, *cork*, *galway*, *eu*, *limerick*, *european*, *waterford*, *hotel*, *ie*, *ucd*, *clare*, *county*, *kerry*, *shannon*, *kildare*, *sligo*, *co.*, *belfast*
`.in`	*india*, *rs*, *indian*, *delhi*, *mumbai*, *shri*, *rfc*, *pradesh*, *singh*, *bangalore*, *crore*, *etc.*, *pakistan*, *temple*, *tamil*, *kerala*, *crores*, *server*, *chennai*, *various*
`.nz`	*zealand*, *nz*, *auckland*, *wellington*, *maori*, *christchurch*, *te*, *new*, *otago*, *canterbury*, *zealanders*, *pacific*, *kiwi*, *island*, *waikato*, *dunedin*, *ministry*, *bay*, *mori*, *says*
`.uk`	*uk*, *london*, *scotland*, *british*, *england*, *scottish*, *britain*, *whilst*, *wales*, *bbc*, *edinburgh*, *manchester*, *nhs*, *royal*, *holiday*, *oxford*, *advice*, *glasgow*, *charity*, *range*
`.us`	*county*, *state*, *shall*, *school*, *district*, *board*, *court*, *defendant*, *texas*, *mr.*, *federal*, *fig*, *attorney*, *license*, *invention*, *department*, *city*, *nbsp*, *pursuant*, *v.*
`.za`	*africa*, *african*, *cape*, *south*, *johannesburg*, *anc*, *mr*, *sa*, *t*, *quot*, *quot*, *durban*, *of*, *town*, *chairperson*, *africans*, *pretoria*, *lodge*, *apartheid*, *you*

then taken as the similarity between the two corpora.[10] This method was designed as a general-purpose measure of corpus similarity, and is not specifically tailored to distinguishing corpora which differ in terms of national variety.

To address this, Cook and Hirst (2012) proposed a measure of corpus similarity specifically targeted at distinguishing national varieties of English, based on the observation that national varieties of English differ in terms of their preferred spellings of some words. For example, while both BrE and CanE prefer *colour* to *color* (where the latter is more common in AmE), *realise* is more common than *realize* in BrE, but vice versa for CanE.[11] Starting from a list of *variant pairs* — pairs of words that are known to differ in spelling between varieties of English — Cook and Hirst represent a corpus as a vector for which each index corresponds to the preference in that corpus for a particular form of a variant pair, e.g., for the pair {*vapor*, *vapour*} they calculate:

[10] In this method the chi-square value is not used for statistical hypothesis testing.

[11] These observations for AmE, BrE, and CanE are based on OANC, BNC, and Strathy, respectively.

$$\frac{\text{frequency}(vapor)}{\text{frequency}(vapor) + \text{frequency}(vapour)} \tag{3}$$

The similarity of two corpora is then the cosine similarity between the vectors representing them.

Crucially, neither of these corpus similarity methods requires any corpus preprocessing other than tokenisation (i.e., no part-of-speech tagging or parsing is required, where errors introduced by systems for such tasks could influence the findings).

Cook and Hirst (2012) showed that these corpus similarity measures are able to distinguish known BrE and CanE corpora. In Section 5.1 we extend this finding to also include AmE, albeit using a slightly different experimental setup. Cook and Hirst further showed that web corpora from the `.uk` and `.ca` domains are more similar in terms of these measures to known BrE and CanE corpora, respectively. In Section 5.2 we further extend this to the `.us` domain and AmE.

5.1 Analysis of national corpora

We randomly split each of our corpora into 6M word sub-corpora by document.[12] 6M words was chosen as the largest size that would give multiple sub-corpora for our smallest corpus (OANC). For each sub-corpus, we compute the chi-square and cosine similarity with each other sub-corpus. In computing chi-square we only consider words from the NLTK (Bird et al, 2009) English word list to avoid the influence of non-linguistic tokens (which could occur in particular in the web corpora — which we consider in the following subsection — because of limitations of the corpus cleansing). To compute cosine similarity, following Cook and Hirst (2012) we use the list of known variant pairs provided by VarCon.[13] Some of the shorter entries in VarCon appear somewhat questionable as variants (e.g., (*le*,*loe*)). We therefore only consider pairs from VarCon for which both items have length at least four characters, and are entirely alphabetic. We further limit the pairs considered to those for which the sum of the frequency of both variants is greater than five in each of the sub-corpora.

The average similarities between sub-corpora from each national corpus are presented in Table 4. For both chi-square and cosine, sub-corpora from each national corpus are on average more similar to sub-corpora from that corpus than from the other corpora. We further see that for both measures, OANC is more like Strathy than BNC, and BNC is more like Strathy than OANC. This is consistent with CanE having been historically influenced by both BE and AmE (Chambers, 2008). We further considered an experiment in which we classified each sub-corpus according to the national sub-corpora it was on average most similar to; the accuracy was 100%.

[12] This is necessary because the chi-square measure of corpus similarity is only applicable to equal-size corpora.

[13] `http://wordlist.sourceforge.net`

Table 4 Left panel: Average chi-square similarities ($\times 10^5$) for sub-corpora of national corpora. Because these are chi-square similarities, smaller numbers imply higher similarity. Right panel: Average cosine similarities using the variant pairs approach for the same sub-corpora. For cosine similarities, higher numbers correspond to higher similarities.

	Chi-square			Cosine		
	BNC	OANC	Strathy	BNC	OANC	Strathy
BNC	**0.19**	1.91	1.14	**0.98**	0.66	0.83
OANC	-	**0.03**	1.19	-	**1.00**	0.87
Strathy	-	-	**0.24**	-	-	**0.99**

Table 5 Left panel: Average chi-square similarities ($\times 10^5$) for sub-corpora of web corpora and sub-corpora of national corpora. Right panel: Average cosine similarities using the variant pairs approach for the same sub-corpora. similarities.

	Chi-square			Cosine		
	BNC	OANC	Strathy	BNC	OANC	Strathy
`.ca`	1.61	1.48	**1.05**	0.82	0.94	**0.98**
`.uk`	**1.53**	1.77	1.58	**0.97**	0.74	0.83
`.us`	2.20	**1.72**	1.76	0.69	**0.99**	0.89

5.2 Analysis of web corpora

We randomly split the web corpora into 6M word sub-corpora by document, as was done for the national corpora. We then compute the chi-square and cosine similarity, using the same methodology as before, between each sub-corpus from the web corpora and each national sub-corpus. The average similarities between the sub-corpora of each web corpus and each national corpus are shown in Table 5. By looking across the rows, we see that, using each similarity measure, each web corpus is, on average, most similar to the corresponding national corpus.[14] This suggests that corpora built from national top-level domains could be representative of the corresponding national varieties of English.[15] Mirroring the findings for the national corpora, we further see that for both measures, `.us` is more like Strathy than BNC, and `.uk` is more like Strathy than OANC.

We also consider an experiment in which we classify each sub-corpus from the web corpora according to the national corpus to which it is, on average, most similar to. Results are shown in Table 6. For cosine the accuracy is, remarkably, 100% in each case; each sub-corpus from the web corpora is on average most similar to the sub-corpora from the corresponding national corpus. In the case of chi-square, the accuracy for `.uk` and `.us` are less than 100%. Nevertheless, all of these accuracies are significantly different from random classification using a one-sided binomial test

[14] By looking down the columns in Table 5 we see that each national corpus is also most similar to the corresponding web corpus for each similarity measure, except in the case of OANC for chi-square, where OANC is more similar to `.ca` than `.us`.

[15] Although we do not have a sufficiently-large national corpus for Australia to use in these experiments, we measured the similarity between sub-corpora from `.au` and the national corpora. Here we found that for chi-square, `.au` is most similar to Strathy, but for cosine it is most like BNC.

Table 6 Classification accuracy for web corpora using the chi-square and cosine similarity measures.

Corpus	Accuracy	
	Chi-square	Cosine
`.ca`	1.000	1.000
`.uk`	0.964	1.000
`.us`	0.830	1.000

($p << 10^{-50}$), which would be expected if web corpora from top-level domains did not represent national varieties of English, suggesting that this is not the case. We further examined the mis-classifications and found that all of the `.uk` and `.us` sub-corpora that are incorrectly classified are labelled as being most similar to Strathy. This is not unexpected, given the previous finding that `.uk` and `.us` are more like Strathy than OANC and BNC, respectively.

6 Comparisons with varietal wordlists

If a corpus built from a particular top-level domain indeed reflects the corresponding national variety of English, then we would expect terms particular to that variety of English to be over-represented in that corpus. For example, if a corpus built from the `.ca` domain indeed corresponds to CanE, then we would expect that corpus to contain many Canadianisms. Specifically, we would expect Canadianisms to be relatively more frequent in that corpus than in a corpus from another top-level domain corresponding to another country (e.g., `.au` or `.uk`). This intuition forms the basis for the evaluation of the corpora in this section.

In order to carry out this evaluation we require lists of words known to be particular to, or particularly common in, specific varieties of English. We obtain such lists for AmE, AuE, and CanE from dictionaries of these language varieties, as described below.

AmE The Dictionary of American Regional English (DARE, Hall, 2012) documents regional usage in the United States. We obtain DARE's headword list online from its website.[16]

AuE The Australian National Dictionary is a historical dictionary documenting terms that are particular to AuE, or more common in AuE than other English varieties; originated in Australia; or are especially important to Australia for historical reasons (AND, Ramson, 1988). We obtained the AND headword list from the online version of this dictionary.[17]

CanE The Canadian Oxford Dictionary (CanOx, Barber, 2005) is a general-purpose English dictionary, with a particular focus on CanE. We used a list of all entries from this dictionary labelled as Canadian, which was provided by lexicographers working on DCHP-2.[18]

[16] `http://dare.wisc.edu/sites/dare.wisc.edu/files/DAREindex.htm`

[17] `http://www.australiannationaldictionary.com/`

[18] `http://www.dchp.ca/`

Table 7 For the corpus from each top-level domain, the proportion (Prop.) of Australianisms, Canadianisms, and Americanisms that are more frequent in the corpus from `.au`, `.ca`, and `.us` is shown. The corresponding p values are also shown.

	Australianisms		Canadianisms		Americanisms	
	Prop.	p	Prop.	p	Prop.	p
`.au`	-	-	0.75	8.46×10^{-6}	0.60	4.9×10^{-7}
`.ca`	0.87	3.6×10^{-55}	-	-	0.58	7.7×10^{-7}
`.hk`	0.90	4.1×10^{-60}	0.91	2.24×10^{-13}	0.76	3.6×10^{-35}
`.ie`	0.91	1.8×10^{-67}	0.84	6.38×10^{-9}	0.75	6.5×10^{-33}
`.in`	0.91	1.0×10^{-66}	0.94	2.94×10^{-15}	0.77	4.7×10^{-36}
`.nz`	0.86	1.3×10^{-50}	0.83	1.12×10^{-8}	0.70	1.2×10^{-21}
`.uk`	0.79	2.6×10^{-35}	0.74	1.46×10^{-5}	0.53	3.6×10^{-2}
`.us`	0.88	5.6×10^{-56}	0.89	9.14×10^{-12}	-	-
`.za`	0.89	8.8×10^{-59}	0.86	6.88×10^{-10}	0.74	2.9×10^{-29}

In many cases it is not a lemma, but rather a lexical unit — i.e., a particular sense of a lemma — that is particular to a language variety. For example, the usage of *pot* for a small glass of beer (or cider) is an Australianism, but *pot* is also common in other varieties of English with other meanings. For this analysis we therefore focus on words which are monosemous, or words for which all of their senses are particular to the language variety of interest.

For each wordlist above (i.e., that from DARE, AND, and CanOx) we extract all single word entries consisting of only alphabetic characters and hyphens. We then eliminate any word found in one of two other general-purpose English wordlists — the English wordlist provided in NLTK (Bird et al, 2009) and Aspell (v0.60.6.1)[19] — ignoring case. This final step eliminates many words that are recorded in general-purpose dictionaries, and therefore not suitable for the analysis here. As for the evaluation in Section 5, we do not linguistically process the corpora (other than tokenisation) to avoid the possibility that the inevitable errors such processing would introduce would influence our findings.

For each of our three wordlists, we count the relative frequency of each term in each of our web corpora. For each wordlist, we then measure the proportion of terms that are more frequent in the corresponding web corpus than in each other web corpus; we do this individually for each of the other corpora. If `.au`, `.ca`, and `.us` did not represent the corresponding national varieties of English, then we would expect these proportions to be roughly equal — it would be equally likely for a given term to be more frequent in one corpus than another. We use a one-sided binomial test to determine whether the observed proportions are significantly different from this null hypothesis. Because of the large number of p values considered (3 wordlists $\times$ 8 corpora $= 24$ p values) we apply a Bonferroni correction and use $\frac{0.05}{24} \approx 0.0021$ as our threshold for significance. Results are shown in Table 7. In all but one case we observe that the corpus expected to include more usages of terms particular to a language variety does, the exception being Americanisms in `.uk`. This finding further

[19] `http://aspell.net/`

increases our confidence that our web corpora might correspond to national varieties of English, and particularly so for `.au` and `.ca`.

7 Applying the corpora to analyse Canadianisms

A final question to be addressed here is whether the web corpora can be used for lexicographic purposes — for example, in the compilation of dictionaries of national varieties of English — in order to establish the status of a lexical item. We take here a number of examples of known Canadianisms as test cases. The following terms are all identified as "Cdn.", i.e. terms used "exclusively in Canadian English", in (CanOx Barber, 2005). *Allophone* is a relatively new Canadianism, (first recorded in 1980)[20] modeled on the terms *anglophone* and *francophone* and denoting a speaker of a language other than English or French (in Québec). As seen in Table 8, this term is almost twice as frequent in the `.ca` corpus as in the next closest domain (`.za`). Moreover, in all of the other domains, the homophone with a linguistic meaning (a variant of a phoneme) predominates. Turning to forms for which both the meaning and form are specific to CanE, we look at another quite recent term, *heritage language*, or a language other than English or French that is spoken as one's mother tongue. This term is now used quite widely (especially in the academic literature) in a slightly modified sense, denoting a home language other than the official language(s) of the country. Despite wide dissemination, *heritage language* is still 4.3 times more frequent in the `.ca` domain than in the next closest domain (`.us`). While many Canadianisms are compounds, *humidex* is a blend of *humidity* and *index* (cf. *Canadarm* from *Canada* + *arm*). The term was first used in the mid-1960s to refer to a computed single value combining heat and humidity intended to provide a sense of how the weather feels to the average person. In the web corpora we see that it is much more frequent (11.6 times) in Canada than in the next closest domain (`.za`). When the term is used outside Canada, there is often a need to define what it means. (In the US it also seems to be a brand name of an air conditioning system.) A *pot light* in Canada is "an interior light encased in a cylindrical shall mounted recessed in a ceiling" (CanOx). This form is almost unique to Canada (no examples are found in India, Hong Kong, New Zealand, and the one example in Ireland is a mistake for "spot light"). The term *visible minority*, denoting an ethnic group which is physically distinct from the predominant group in a society, is ten times more common in Canada than in any other domain. Finally, the web corpora support the popular identification of *washroom*, for lavatory, as a Canadianism.

Thus, the use of web corpora seems to be an efficient means for undertaking lexicographic analysis. The corpora may prove a corrective to existing dictionaries or support ongoing lexicographic work. For example, *eavestrough* is identified in CanOx as North American and especially Canadian; in DARE it is shown to occur widely in the United States except in the south and south midlands. Contemporary web evidence would suggest it is now fairly rare in the United States and virtually restricted to Canada. A *statutory holiday* (known in the United Kingdom as a *bank holiday*) is

[20] Bank of Canadian English, `http://www.dchp.ca/`

Table 8 The frequency per 10 million tokens of each Canadianism discussed in Section 7 in each web corpus.

	.au	.ca	.hk	.ie	.in	.nz	.uk	.us	.za
allophone	0.503	4.284	0.000	0.911	0.065	0.000	1.140	0.335	2.010
heritage language	0.470	4.129	0.000	0.195	0.065	0.376	0.244	0.965	0.246
humidex	0.058	2.389	0.146	0.146	0.000	0.171	0.087	0.183	0.205
pot light	0.033	0.997	0.000	0.065	0.000	0.000	0.015	0.010	0.000
visible minority	0.503	55.140	0.000	0.000	0.065	0.410	0.529	0.122	0.533
washroom	5.141	70.976	15.337	15.487	5.955	7.785	9.170	3.930	7.835
eavestrough	0.008	2.250	0.000	0.000	0.000	0.000	0.081	0.010	0.410
statutory holiday	0.206	17.614	18.551	0.911	0.129	7.478	1.247	0.142	0.082
strata council	0.017	2.559	0.000	0.000	0.065	0.000	0.010	0.000	0.000

clearly a Canadianism but also widely used in New Zealand and Hong Kong. Finally, the term *strata council* is not recorded in CanOx (though it includes *strata*, a term adopted in Canada from Australia/New Zealand to denote a condominium building). It has been identified as a Canadianism in DCHP-2, and the web corpora support this identification.

One limitation of this type of analysis is that it can only be applied to items that can be easily counted in corpora, such as word forms, lemmas, or word sequences (e.g., bigrams), and not to word senses. However, in many cases, if a word sense is particularly frequent in a corpus, so will be the corresponding word form, as was the case for *allophone*, discussed above. Moreover, in future work, approaches to studying variation in word senses between language varieties (Peirsman et al, 2010) could potentially be applied to these web corpora.

8 Conclusions

Although very large English corpora have been built, for many national varieties of English, only relatively small corpora are available. Such corpora would be valuable to lexicography and corpus-based studies focused on national varieties of English, and would enable the training of statistical natural language processing systems for specific English varieties. In this paper we built web corpora from national top-level domains for countries in which English is widely spoken, and then carried out statistical analyses to determine the extent to which they reflect the corresponding national varieties of English. We found evidence from analyses based on corpus keywords, measures of corpus comparison based on the chi-square test and spelling variants, and the frequencies of words known to be marked in particular varieties of English, that English web corpora from national top-level domains indeed reflect corresponding national varieties of English. We further applied these web corpora for lexicographic analysis of Canadianisms, finding that they could indeed be a valuable resource in lexicography. In future work, we intend to investigate whether we can apply very large English web corpora built from national top-level domains for the identifica-

tion of previously undocumented varietal-specific lexical items and lexical units (e.g., previously-undocumented Canadianisms).

References

Atkins BTS (2010) The DANTE Database: Its contribution to English lexical research, and in particular to complementing the FrameNet data. In: de Schryver GM (ed) A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks, Menha Publishers, Kampala, Uganda

Banko M, Brill E (2001) Scaling to very very large corpora for natural language disambiguation. In: Proceedings of 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001), Toulouse, France, pp 26–33

Barber K (ed) (2005) Canadian Oxford Dictionary, 2nd edn. Oxford University Press

Baroni M, Bernardini S (2004) BootCaT: Bootstrapping corpora and terms from the Web. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal

Baroni M, Kilgarriff, Pomikálek J, Rychlý P (2006) WebBootCaT: a web tool for instant corpora. In: Proceedings XII EURALEX International Congress (EURALEX 2006), Torino, Italy, pp 123–131

Baroni M, Chantree F, Kilgarriff A, Sharoff S (2008) Cleaneval: A competition for cleaning Web pages. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, pp 638–643

Baroni M, Bernardini S, Ferraresi A, Zanchetta E (2009) The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. Language Resources and Evaluation 43(3):209–226

Bird S, Loper E, Klein E (2009) Natural Language Processing with Python. O'Reilly Media Inc., Sebastopol, CA

Burnard L (2000) The British National Corpus Users Reference Guide. Oxford University Computing Services

Burnard L (2007) Reference guide for the British National Corpus (XML Edition). Oxford University Computing Services

Cavnar WB, Trenkle JM (1994) N-gram-based text categorization. In: Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94), Las Vegas, US, pp 161–175

Chambers JK (2008) The tangled garden: Relics and vestiges in Canadian English. Anglistik 19(2):7–21, Special issue: Focus on Canadian English

Clarke CLA, Craswell N, Soboroff I, Voorhees EM (2011) Overview of the TREC 2011 Web Track. In: Proceedings of the Twentieth Text REtrieval Conference (TREC 2011), NIST Special Publication: SP 500-295.

Cook P, Hirst G (2012) Do Web corpora from top-level domains represent national varieties of English? In: Actes des 11es Journées internationales d'Analyse statistique des Données Textuelles / Proceedings of the 11th International Conference on Textual Data Statistical Analysis, Liège, Belgium, pp 281–293

Cook P, Lui M (2012) `langid.py` for better language modelling. In: Proceedings of the Australasian Language Technology Association Workshop 2012 (ALTA 2012), Dunedin, New Zealand, pp 107–112

Davies M (2009) The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. International Journal of Corpus Linguistics 14(2):159–190

Dillon G (2010) Building webcorpora of academic prose with BootCaT. In: Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop, Los Angeles, pp 26–31

Ferraresi A, Zanchetta E, Baroni M, Bernardini S (2008) Introducing and evaluating ukWaC, a very large web-derived corpus of English. In: Proceedings of the 4th Web as Corpus Workshop: Can we beat Google, Marrakech, Morocco, pp 47–54

Green E, Peters P (1991) The Australian corpus project and Australian English. International Computer Archive of Modern English 15:37–53

Hall JH (ed) (2012) Dictionary of American Regional English, Volume V: SI–Z. The Belknap Press of Harvard University Press

Kilgarriff A (2001) Comparing corpora. International Journal of Corpus Linguistics 6(1):97–133

Kilgarriff A (2009) Simple maths for keywords. In: Proceedings of the Corpus Linguistics Conference, Liverpool, UK

Kilgarriff A, Rychly P, Smrz P, Tugwell D (2004) The Sketch Engine. In: Proceedings of Euralex, Lorient, France, pp 105–116

Kilgarriff A, Reddy S, Pomikálek J, PVS A (2010) A corpus factory for many languages. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta, pp 904–910

Ljubešić N, Klubička F (2014) bs,hr,srwac - web corpora of bosnian, croatian and serbian. In: Proceedings of the 9th Web as Corpus Workshop (WaC-9), Gothenburg, Sweden, pp 29–35

Lui M, Baldwin T (2011) Cross-domain feature selection for language identification. In: Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP 2011), Chiang Mai, Thailand, pp 553–561

Murphy B, Stemle E (2011) PaddyWaC: A minimally-supervised Web-corpus of Hiberno-English. In: Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties, Edinburgh, Scotland, pp 22–29

Peirsman Y, Geeraerts D, Speelman D (2010) The automatic identification of lexical variation between language varieties. Natural Language Engineering 16(4):469–491

Peters P (2009) The architecture of a multipurpose Australian national corpus. In: Selected Proceedings of the 2008 HCSNet Workshop on Designing an Australian National Corpus, Sommerville, MA, pp 1–9

Pomikálek J (2011) Removing boilerplate and duplicate content from web corpora. PhD thesis, Masaryk University

Pomikálek J, Jakubíček M, Rychlý P (2012) Building a 70 billion word corpus of English from ClueWeb. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, pp 502–506

Ramson WS (ed) (1988) The Australian National Dictionary: A dictionary of Australianisms on historical principles. Oxford University Press

Resnik P, Smith NA (2003) The web as a parallel corpus. Computational Linguistics 29(3):349–380

Roth T (2012) Using web corpora for the recognition of regional variation in standard German collocations. In: Proceedings of the seventh Web as Corpus Workshop (WAC7), Lyon, France, pp 31–38

Schäfer R, Bildhauer F (2013) Web Corpus Construction. Morgan and Claypool, San Rafael, CA

Schulz S, Lyding V, Nicolas L (2013) STirWaC - Compiling a diverse corpus based on texts from the web for south Tyrolean German. In: Proceedings of the 8th Web as Corpus Workshop (WAC-8), Lancaster, UK, pp 37–45

Sharoff S (2006) Creating general-purpose corpora using automated search engine queries. In: Baroni M, Bernardini S (eds) Wacky! Working papers on the Web as Corpus, GEDIT, Bologna, Italy, pp 63–98