Do Web Corpora from Top-Level Domains Represent National Varieties of English?

Paul Cook[1], Graeme Hirst[2]

[1]Department of Computing and Information Systems, The University of Melbourne — paulcook@unimelb.edu.au

[2]Department of Computer Science, University of Toronto — gh@cs.toronto.edu

Abstract

In this study we consider the problem of determining whether an English corpus constructed from a given national top-level domain (e.g., *.uk*, *.ca*) represents the national dialect of English of the corresponding country (e.g., British English, Canadian English). We build English corpora from two top-level domains (*.uk* and *.ca*, corresponding to the United Kingdom and Canada, respectively) that contain approximately 100M words each. We consider a previously-proposed measure of corpus similarity, and propose a new measure of corpus similarity that draws on the relative frequency of spelling variants (e.g., *color* and *colour*). Using these corpus similarity metrics we show that the Web corpus from a given top-level domain is indeed more similar to a corpus known to contain texts from authors of the corresponding country than to a corpus known to contain documents by authors from another country. These results suggest that English Web corpora from national top-level domains may indeed represent national dialects, which in turn suggests that techniques for building corpora from the Web could be used to build large dialectal language resources at little cost.

Keywords: Corpora, Web-as-corpus, comparing corpora, dialects of English.

1. Dialectal language resources

Corpora enable the study — computational or otherwise — of many aspects of language; therefore, in order to effectively study a specific dialect of a language, we require a corpus representing that dialect. Unfortunately, language resources are available for few dialects, and where they are available, they tend to be rather small by today's standards of corpus size.

Considering the case of national varieties of English, sizeable resources are available for American, British, and Canadian English (referred to as AmE, BE, and CE, respectively); the Corpus of Contemporary American English (Davies, 2009), the British National Corpus (Burnard, 2000), and the Strathy Corpus of Canadian English[1] consist of roughly 400 million, 100 million, and 40 million words of AmE, BE, and CE, respectively. Moreover, these corpora have been carefully constructed — with great manual effort and at high cost — to represent English as used in their respective countries. For many other national varieties of English, only

1 *http://www.queensu.ca/strathy/*

much smaller resources are available. For example, the International Corpus of English consists of one-million-word corpora for national or regional varieties of English, including Indian, Jamaican, and New Zealand English.[2] Given that sizeable resources are presently available for only a few English dialects, corpora for specific dialects, and methods for building them, would be a great asset to linguistic inquiry into dialects. Such resources may also have practical applications in language technology by providing dialect-specific training data for statistical natural language processing methods, enabling such methods to be tailored to dialects.

The Web has been widely used to create large corpora for many languages (e.g., Baroni *et al.*, 2009, Kilgarriff *et al.*, 2010), and corpora for specific topics (e.g., Baroni and Bernardini, 2004) and genres (e.g., Dillon, 2010). The Web has also been used to create parallel corpora (e.g., Resnik and Smith, 2003) — corpora of texts in one language and their translations into another. Crucially, these methods of corpus construction enable corpora to be created quickly and at little cost. Nevertheless, despite the impressive variety of Web corpora that have been constructed to date, there have been comparatively few efforts to create monolingual corpora representing different national or regional varieties of English. Given the cost of traditional approaches to creating such resources, methods for building corpora from the Web representing dialects are very appealing. One recent study into creating such corpora is that of Murphy and Stemle (2011), who exploit lexical markers of Hiberno-English (the variety of English spoken in Ireland) — including place-names, regionalisms, and loanwords — to identify texts containing known properties of Hiberno-English (e.g, *amn't*, a contraction of *am not*). In contrast, we focus on different varieties of English (namely, BE and CE), and do not rely on such markers in the creation of our corpora. Furthermore, we conduct a more-thorough empirical evaluation of the corpora we create.

One challenge to creating dialectal Web corpora is that it is not clear how to easily find texts written in a specific dialect on the Web. Some top-level domains (e.g., *.uk, .ca*) may contain content relevant to particular countries (e.g., the United Kingdom and Canada, respectively) but there is no guarantee that documents found in a particular top-level domain are authored by speakers from the country of that domain, and therefore may not represent that country's national dialect. Baroni *et al.* (2009) mention exactly this issue in their discussion of the ukWaC, a Web corpus created from the *.uk* domain; the ukWaC is a corpus of English, but not necessarily British English.

In this paper, we offer preliminary evidence that English Web corpora built from top-level domains do indeed represent national varieties of English. In particular, we first build corpora potentially representing national varieties of English just by focusing on text from websites in national top-level domains. The focus of this study is then to determine whether these corpora achieve their aim, or whether they are contaminated with too much text whose origin is outside the target country. For this purpose we consider previous approaches to comparing corpora (specifically Kilgarriff, 2001), and propose a new method to compare corpora based on variations in spelling — one way in which English dialects are known to vary. We show that the Web-constructed corpora exhibit properties of corpora that are known to be authored by speakers of specific national varieties of English.

2 *http://ice-corpora.net/ice/*

By creating large corpora representing national varieties of English, we open the door to further quantitative studies of dialects. For example, the methods that Peirsman *et al.* (2010) propose for identifying cross-varietal synonyms — e.g., the storage compartment in the rear of a car is typically *boot* in BE whereas *trunk* is preferred in AmE — could be applied to the newly-created corpora; the output of such a method could then be manually examined to identify previously-undocumented differences between the dialects.

2. Corpora

In the following subsections we describe the national corpora used in, and Web corpora created for, this study. We focus on BE and CE because in both cases large national corpora, and corresponding top-level domains from which to build Web corpora, are available. (For AmE, there is no top-level domain that is appropriate for constructing a general-purpose Web corpus; the *.us* domain tends to be used for specialized purposes, such as government.)

We apply the same simple tokenization method to both the national corpora (section 2.1) and the Web corpora (section 2.2), and we apply a state-of-the-art language identification method (Lui and Baldwin, 2011) to the corpora in order to retain only English documents.

2.1. National corpora

To represent BE, we use the written component of the British National Corpus (BNC), which consists of approximately 87 million words of text from a variety of genres written by British authors in the late 1900s. For CE, we use the written component of the Strathy Corpus of Canadian English, which consists of approximately 40 million words of text by Canadian authors that, like the BNC, comes from a variety of genres and the late twentieth century.

2.2. Web corpora

We construct English Web corpora from the *.uk* and *.ca* domains to test the hypothesis that corpora constructed from top-level domains represent national varieties of English. We will refer to these corpora as DotUK and DotCA respectively.

Baroni and Bernardini (2004) propose a method for Web corpus construction whereby search-engine queries are issued for random combinations of user-chosen seed words; the URLs returned by the queries are then downloaded and post-processed to produce a corpus. Similar methodologies have been used in other corpus construction efforts (e.g., Sharoff, 2006a; Kilgarriff *et al.*, 2010). One issue in this methodology is the choice of seed words, and the number of seeds to use in queries. Our present goal is only to evaluate the results of corpora created through such a process (and not to suggest improved methods for corpus construction); therefore, we follow a previously-used approach to seed word selection. A portion of the queries used by Ferraresi *et al.* (2008) are tuples of mid-frequency terms from the BNC; we form queries from similar terms. We rank the alphabetic types in the BNC with length greater than or equal to two by frequency, and select those with frequency rank 1001 to 5000 as our seed words. We use the BootCaT tools provided by Baroni and Bernardini (2004)[3] to randomly form 18,000 3-tuples from the seed words, which we use as queries. The number of tuples was chosen such that the resulting corpora would contain approximately 100M words.

3 *http://bootcat.sslmit.unibo.it*

We use the BootCaT tools to issue queries for the seed word tuples in the *.uk* and *.ca* domains, and download and post-process the retrieved URLs by removing duplicate documents and boilerplate text. (These corpora were collected in the second quarter of 2011 using the *Yahoo!* search API, which is unfortunately no longer available. Nevertheless, the BootCaT tools now support the *Bing* API, so it is still possible to create similar corpora.) To prevent any single source from biasing our corpora, we keep a maximum of three randomly selected documents per host (e.g., *www.cbc.ca*). Table 1 presents some statistics summarizing the corpora created.

	DotUK	DotCA
Queries	18K	18K
URLs	173K	179K
URLs: Duplicates removed	173K	162K
Documents	146K	134K
Documents: Duplicates removed	141K	129K
Documents: Maximum 3 documents per host	60K	37K
Words	115M	96M

Table 1: Approximate number of queries, URLs, documents, and words for the corpora, DotUK and DotCA, created from the .uk and .ca domain

3. Methods for corpus comparison

In this section, we present methods for determining the similarity of two corpora based on keywords, the chi-square test, and orthographic variants. The first of these is useful for manually examining differences and similarities between corpora, but we are also interested in methods for automatically measuring the similarity of two corpora and hence also consider the other two methods. For all methods for determining corpus similarity we consider only alphabetic types, and ignore case.

3.1. Keywords

One way to compare two corpora is to examine *keywords* in those corpora, i.e., words that are much more frequent in one corpus than the other. Kilgarriff (2009) proposes a simple method to identify keywords based on the ratio of a word's frequency per million, plus a constant, in each of two corpora. In this study we compute keywords using Kilgarriff's method; we set the constant to 100, the value that he recommended.

3.2. Chi-square

Kilgarriff (2001) is one of the few studies to consider the task of automatically measuring corpus similarity. One challenge to such a study is evaluation; the similarity between pairs of corpora is generally not known, and it is difficult to obtain manual judgements of corpus similarity due to their size. Kilgarriff therefore proposes the use of known similarity corpora

to evaluate approaches to measuring corpus similarity. Specifically, Kilgarriff builds corpora consisting of varying proportions of specific sections of the BNC to create pairs of corpora of known relative similarity.

Kilgarriff considers numerous approaches to measuring corpus similarity, and concludes that a simple approach based on the chi-square statistic is most appropriate. Specifically, the chi-square statistic is calculated for the 500 most frequent words in the union of two corpora and is interpreted as the similarity between those two corpora. (The chi-square value is *not* used as the basis for a hypothesis test.)

This method of measuring corpus similarity is attractive in that it is based solely on the words in a corpus (i.e., it does not rely on any tools such as a part-of-speech tagger, which may influence the corpus similarity metric) and it taps into a very general notion of corpus similarity. Nevertheless, in this work we are concerned with the specific problem of determining whether two corpora represent the same national or regional dialect of English; it may therefore be the case that a corpus similarity measure tailored to this task is more appropriate.

3.3. Spelling variations

Different varieties of English are known to prefer different spellings of the same word. (Peters (2004), for example, notes numerous such differences.) Examples of spelling differences between BE and CE include the tendency for BE to use *aluminium* and *specialise* whereas CE prefers *aluminum* and *specialize*. Furthermore, national dialects of English can vary with respect to their *lack* of preference for a spelling variant — BE prefers *vapour* to *vapor*, whereas in CE these forms have approximately equal frequency. (These observations about BE and CE are based on the frequencies of these forms in the BNC and the Strathy Corpus.) We therefore propose a measure of corpus similarity based on such spelling differences.

Given a list of known spelling variants — i.e., a list of pairs like {*aluminium*, *aluminum*} and {*specialise*, *specialize*} — we represent a corpus as a vector where each entry *i* is the proportion of usages of either form of spelling variant *i* that corresponds to a particular form, e.g., for the spelling variant {*organise*, *organize*} we calculate

frequency(*organise*)/(frequency(*organise*)+ frequency(*organize*)).

The similarity between two corpora is then the cosine similarity of their spelling variant vectors.

Kilgarriff's (2001) method for measuring corpus similarity is based on differences in frequent words. Therefore, we expect that corpora which differ with respect to genre or topic will have different frequent words and will have low similarity according to this metric. However, in this work we are concerned with establishing whether a corpus corresponds to a given national variety of English, regardless of genre or topic. A similarity metric based on orthographic variants has the potential to pick up on differences between two corpora (with respect to orthographic variation) even when those corpora are not comparable with respect to genre or topic.

4. Classifying known-dialect corpora

Before we can consider whether the Web corpora constructed in Section 2 represent a particular national variety of English, we must establish that our methods for measuring corpus similarity are indeed able to identify national dialects.

4.1. Materials and methods

To test whether the corpus similarity metrics presented in Section 3 can identify national dialects, we consider a classification task. We randomly split the BNC and the Strathy Corpus by document into sub-corpora consisting of roughly five million words each. This size was chosen to yield sub-corpora large enough to accurately measure the frequency of common words, but small enough that the number of sub-corpora for each corpus would be sufficient; the smallest corpus — the Strathy Corpus — has eight sub-corpora.

When calculating the chi-square similarity metric, we consider only tokens that are in-vocabulary according to an English wordlist provided by Bird *et al.* (2009); we do so in an attempt to prevent any systematic non-linguistic differences between the corpora from affecting the similarity scores (e.g., the Web corpora contain many more URLs than the national corpora).

Our spelling variation-based corpus similarity metric requires a list of known English spelling variant pairs. We obtain such a list from VarCon.[4] We select those entries in VarCon that consist of exactly two alphabetic alternatives, such that both alternatives have length greater than three characters. (Some very short spelling variations in VarCon seem questionable, e.g., (*tae*, *te*).) We further restrict ourselves to those entries for which the sum of the frequency of the two alternatives is at least five in every sub-corpus of both corpora, and neither alternative is in the list of seed words used to build the Web corpora (see Section 2.2). This gives 201 pairs of spelling variants.

For each sub-corpus we compute its average similarity to the other sub-corpora from the BNC and Strathy Corpus using each corpus similarity metric (chi-square and spelling variation) separately. If a sub-corpus is found to be, on average, more similar to sub-corpora from the same corpus than to those from the other corpus, then the system has correctly classified this sub-corpus, and the system is scored as correct. We then calculate the accuracy of this classification for all sub-corpora.

4.2. Results

The accuracy using both the chi-square and spelling variation-based similarity metrics is 100%, which demonstrates that these corpus similarity metrics are able to distinguish between CE and BE corpora. It is particularly encouraging that both metrics achieve this high accuracy given that they measure corpus similarity in very different ways. Nevertheless, this very high accuracy is not entirely unexpected; Kilgarriff (2001) noted that if very different text types were chosen in the construction of his known-similarity corpora, the accuracy of his chi-square method was found to be 100%.

4.3. Keyword analysis

Before examining the Web corpora, we consider keywords for the BNC and Strathy Corpus. The top-10 keywords for these corpora — calculated using Kilgarriff's (2009) method, discussed in Section 3.1 — are presented in Table 2.

For both the BNC and Strathy corpus, many of the top-10 keywords are related to British and Canadian places or institutions (e.g., *london*, *scotland*, *britain*, and *uk* and *canada*, *canadian*,

4 *http://wordlist.sourceforge.net*

toronto, *ontario*, *ottawa, cbc*, and *quebec*) which suggests that topics related to Britain are more common in the BNC and those related to Canada are more common in the Strathy Corpus. Knowing that there are such topical differences between these corpora, we are unsurprised by the previously-observed accuracy of the chi-square corpus similarity metric. This method exploits differences in the most frequent words in the corpora, and these words are expected to differ due to the differences in topic. The BNC keywords *mrs*, *sir*, *hon*, and *mr* seem to indicate differences in genre between the corpora, which we also expect to result in differences amongst the most-frequent words of the corpora. We expected the Strathy keywords *et* and *j* to be the result of French text in this corpus. (French is an official language of Canada, and both of these types are common in French; *j* is a contracted form of the first-person pronoun *je*.) However, although we do see some French usages of *et* and *j* in the Strathy Corpus (the language identification tool we use is state-of-the-art, but it is nevertheless of course imperfect) we observe that *et* is in fact typically used in the multiword expression *et al*, while *j* appears to often be part of a name. Again, these keywords might be due to differences in genre. Finally, some of the keywords are difficult to explain. *U* is typically used as part of *U.S.*, and this usage just appears to be more frequent in the Strathy corpus, although this could possibly be due to the geographical proximity of Canada and the United States. Similarly *looked* is more frequent in the BNC.

BNC:	
	mr, uk, programme, hon, britain, sir, mrs, looked, scotland, london
Strathy:	
	canada, canadian, toronto, u, et, ontario, ottawa, cbc, quebec, j

Table 2: Top-10 keywords for the BNC and Strathy Corpus according to the method of Kilgarriff (2009).

5. Evaluating the Web corpora

Having demonstrated in the previous section that our corpus similarity metrics can distinguish between BE and CE, we now consider the focus of this paper: Do Web corpora created from the *.uk* and *.ca* domains represent BE and CE, respectively? To do so we analyze keywords and the similarity between the DotUK and DotCA corpora measured using the chi-square and spelling variation-based corpus similarity metrics.

5.1. Keywords

We begin by examining the keywords for the DotUK and DotCA Web corpora. The top-10 keywords for each corpus are presented below.

DotUK:	
	uk, london, whilst, programme, scheme, wales, scotland, england, organisation, mm
DotCA:	
	canada, canadian, ontario, toronto, ca, program, vancouver, ottawa, alberta, bc

Table 3: Top-10 keywords for the Web corpora, DotUK and DotCA, created from the .uk and .ca domains, according to the method of Kilgarriff (2009).

The keywords for the DotUK and DotCA Web corpora, for the most part, clearly correspond to words that are related to the United Kingdom and Canada, respectively. Like the keywords for the BNC and Strathy Corpus, the keywords for the two Web corpora include many placenames of the United Kingdom and Canada (e.g., *london* and *wales,* and *toronto* and *alberta*). We also observe spelling variants that are preferred by BE or CE (e.g., *organisation* — which has *organization* as a variant — and *programme* / *program*). From this keyword analysis it appears that the DotUK Web corpus tends to include more documents about the United Kingdom, whereas the DotCA corpus consists of more documents related to Canada. This is the same as we observed for the known BE and CE corpora, and is therefore an encouraging finding which suggests that the Web corpora might indeed represent the corresponding national varieties of English.

5.2. Chi-square and spelling variation

To compare the Web corpora (DotUK and DotCA) to the corpora of known dialect (the BNC and Strathy Corpus), we begin by randomly splitting each Web corpus by documents into sub-corpora consisting of roughly five million words each, in the same manner as in Section 4. Crucial to the experiments in this section, splitting the corpora in this way allows us to obtain independent measurements of word frequency in each sub-corpus. This allows us to form independent pairs consisting of a sub-corpus from each corpus, and enables us to use tests of significance that require such independent pairs.

We use the same list of spelling variant pairs for the spelling variant–based corpus similarity method as in Section 4. We apply the same selection criteria to these pairs as before to choose 174 of these pairs. (In this case because we are considering different corpora — we now also consider the Web corpora that weren't used in the previous experiments — the number of spelling variant pairs is different.)

To measure the similarity between corpora *a* and *b* (using either the chi-square or the spelling variant similarity metric) we pair each sub-corpus from *a* with a unique sub-corpus from *b*, and calculate the similarity of these paired sub-corpora. Because the corpora are different sizes, the number of sub-corpora available varies. As mentioned above, the smallest corpus, the Strathy Corpus, has eight sub-corpora; we simply randomly select eight sub-corpora for each corpus to get the same number of sub-corpora in each case. The similarity between corpora *a* and *b* is then the average similarity between the eight pairs of sub-corpora for *a* and *b*.

Tables 4 and 5 present the average similarity between each pair of corpora (including the DotUK-CA corpus, which will be described below) using the chi-square and spelling variant–based similarity metrics. We observe a relatively low similarity between the BNC and Strathy

Corpus — corpora which are known to reflect national dialects — for both measures. This is as expected, particularly given the results in Section 4.

Next we consider the Web corpus DotCA, created from the *.ca* domain. For both corpus similarity metrics, the pair {DotCA, Strathy} are more similar than the pair {DotCA, BNC}, showing that a corpus created from the *.ca* domain is indeed more like CE than BE in terms of these metrics. For both metrics this difference is significant according to a one-sided Wilcoxon signed rank test ($p < 0.005$). A corresponding result is observed in the case of the corpus DotUK, created from the *.uk* domain. In this case, for both similarity metrics, the similarity of the pair {DotUK, BNC} is greater than that of the pair {DotUK, Strathy}, indicating that a Web corpus from the *.uk* domain is more similar to BE than CE. These differences are also significant ($p < 0.005$). We further consider a classification experiment in which a given sub-corpus from DotUK or DotCA is classified as BE or CE depending on whether it is, on average, more similar to the known BNC or Strathy sub-corpora. The accuracy for these experiments is 100%.

We now consider whether a Web corpus built from a specific top-level domain, or from a mixture of domains, is more similar to a specific national variety of English. We build a new corpus representing a mixture of the *.uk* and *.ca* domains — referred to as DotUK-CA — by merging the DotUK and DotCA corpora. We randomly select five-million word sub-corpora from DotUK-CA as before. Tables 4 and 5 show that the pair {DotCA, Strathy} are more similar than {DotUK-CA, Strathy}; these differences are significant in both cases ($p < 0.005$) using one-sided Wilcoxon signed rank tests as before. Furthermore, {DotUK, BNC} are more similar than {DotUK-CA, BNC} using both metrics; again these differences are significant ($p < 0.05$ in this case).

	BNC	**Strathy**	**DotCA**	**DotUK**	**DotUK-CA**
BNC	0.00	1.33	1.66	1.22	1.34
Strathy	-	0.00	1.10	1.45	1.19
DotCA	-	-	0.00	0.40	0.18
DotUK	-	-	-	0.00	0.18
DotUK-CA	-	-	-	-	0.00

Table 4: Average similarity using the chi-square corpus similarity metric between each pair of corpora. The numbers are chi-square values multiplied by a factor of 10^5; lower numbers imply greater similarity.

	BNC	Strathy	DotCA	DotUK	DOTUK-CA
BNC	1.00	0.53	0.48	0.93	0.81
Strathy	-	1.00	0.89	0.56	0.78
DotCA	-	-	1.00	0.58	0.81
DotUK	-	-	-	1.00	0.88
DotUK-CA	-	-	-	-	1.00

Table 5: Average similarity using the spelling variant–based corpus similarity metric between each pair of corpora. The numbers are cosines, so higher numbers imply greater similarity.

6. American English

We now want to test which of a number of corpora known to represent a national dialect of English our Web corpora are most similar to. For this preliminary experiment we use the Corpus of Contemporary American English (COCA, Davies, 2009), a corpus of approximately 425M words of American English from a variety of text types from 1990 to the present. We are unable to access the full text of this corpus — preventing us from splitting the corpus into 5M word samples as in our previous experiments — and instead must make do with word frequency information provided by Davies (2011). Furthermore, Kilgarriff (2001) notes that the chi-square corpus similarity measure is not suited to corpora of different sizes. Therefore we will only consider the spelling variant–based corpus similarity metric in this analysis. Specifically, we consider whether the DotUK and DotCA Web corpora are more similar to the BNC, COCA, or Strathy Corpus, corpora known to represent BE, AmE, and CE, respectively.

We apply the same criteria as in the previous experiments for selecting candidate spelling variant pairs; 1300 pairs meet this criteria. (In this case, because we compare full corpora — as opposed to five-million-word sub-corpora — the number of spelling variant pairs meeting the selection criteria is much higher.)

The similarity of the pair {DotUK, BNC} is 0.93, much higher than that of {DotUK, Strathy}, which is 0.52, and {DotUK, COCA}, which is 0.37. This finding suggests that a corpus from the *.uk* domain is more like BE than CE or AmE. In the case of DotCA, the similarity of {DotCA, Strathy} at 0.90 is quite close to that of {DotCA, COCA} at 0.88, while the pair {DotCA, BNC} are much less similar, 0.46. The two major varieties of English are often claimed to be BE (or English English) and North American English (e.g., Trudgill and Hannah, 2008). It is possible that a corpus from the *.ca* domain does not specifically represent CE, but rather represents North American English; the finding above that a corpus from the *.ca* domain has comparable similarity to both a known Canadian and known American corpus is consistent with this. However, this finding could also indicate that the *.ca* domain contains a substantial number of documents by American authors.

7. Discussion

The findings in Section 4 suggest that corpora created from the *.uk* and *.ca* domains are more similar to BE and CE than CE and BE, respectively, at least with respect to the frequency of common words and orthographic variation, the properties captured by our corpus similarity metrics. However, differences amongst varieties of English have also been noted at the lexico-syntactic and syntactic levels (e.g., Trudgill, 1984; Hargraves, 2003). The measurement of such properties could also be incorporated into a method for determining whether a corpus corresponds to a particular variety of English. However, Kilgarriff (2001) notes that any corpus comparison methodology that relies on the output of tools such as part-of-speech taggers or parsers will be influenced by the performance of those tools, and furthermore that the performance of such tools can vary from corpus to corpus. Therefore we choose to focus solely on wordforms in the present study, and leave consideration of other types of dialectal differences to future work.

One issue related to, but not considered in, the present study is whether the Web corpora created are comparable; i.e., do the corpora consist of texts from a similar mixture of topics and genres? This is not the focus of the present paper; here we are concerned with establishing whether corpora created from particular top-level domains represent corresponding national varieties of English. Nevertheless, if we are to make use of such corpora in dialectal studies it is important that they be comparable. For instance, Biber (1988) finds American texts to be more colloquial than British texts. However, such an analysis must be carried out on corpora which are balanced with respect to genre (which indeed Biber does). If the corpora consisted of very different text types — i.e., a corpus of formal American texts and informal British texts — the differences noted by Biber would not be seen. Unfortunately, with the notable exception of Kilgarriff (2001), little work has been done on the issue of measuring corpus similarity. Nevertheless, it is still possible to assess corpus similarity manually by classifying samples of documents according to topic and genre, as by Sharoff (2006b). In the present study we ensured that the Web corpora used were created following the same methodology; i.e., we issued the same queries, at approximately the same time, in the *.uk* and *.ca* domains. Because search engines are quite reliable at returning documents relevant to some query, it could reasonably be the case that these Web corpora consist of similar topics. However, the issue of whether comparable monolingual corpora can be automatically created from the Web is on its own a very interesting research question, and is left for future work.

Although the present study focused on national varieties of English, the corpus construction methods used may equally well apply to creating corpora of national dialects of other languages, for example, French as used in France and Quebec, and Portuguese in Portugal and Brazil.

Because of the evidence found in this study that the DotCA Web corpus created from the *.ca* domain corresponds to CE, we are using this corpus in a study of CE. In particular we are analyzing this corpus in an effort to identify previously undocumented Canadianisms — words or word senses particular to CE.

Acknowledgements

We are grateful to Marco Lui for his comments and feedback on our research. This work was financially supported by the University of Melbourne, Mitacs, the Natural Sciences and Engineering Research Council of Canada, and the University of Toronto.

References

Atwell, Eric; Arshad, Junaid; Lai, Chien-Ming; Nim, Lan; Rezapour Asheghi, Noushin; Wang, Josiah; and Washtell, Justin (2007). Which English dominates the World Wide Web, British or American? Proceedings of the Corpus Linguistics Conference (CL 2007). Birmingham, UK.

Baroni, Marco and Bernardini, Silvia (2004). BootCaT: Bootstrapping corpora and terms from the Web. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004).

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

Biber, Douglas (1988). Variation Across Speech and Writing. Cambridge University Press.

Bird, Steven; Loper, Edward; and Klein, Ewan (2009). Natural Language Processing with Python. O'Reilly Media Inc., Sebastopol, CA.

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

Davies, Mark (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.

Davies, Mark (2011). Word frequency data from the Corpus of Contemporary American English (COCA). Downloaded from http://www.wordfrequency.info on 21 November 2011.

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

Ferraresi, Adriano; Zanchetta, Eros; Baroni, Marco; and Bernardini, Silvia (2008). Introducing and evaluating ukWaC, a very large Web-derived corpus of English. Proceedings of the 4th Web as Corpus Workshop: Can we Beat Google?, pages 47–54. Marrakech, Morocco.

Hargraves, Orin (2003). Mighty Fine Words and Smashing Expressions: Making Sense of Transatlantic English. Oxford University Press.

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

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

Kilgarriff, Adam; Reddy, Siva; Pomikálek, Jan; and Avinesh PVS (2010). A corpus factory for many languages. Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010), pages 904–910. Valletta, Malta.

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

Lui, Marco and Baldwin, Timothy (2011). Cross-domain feature selection for language identification. Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP 2011), pages 553–561. Chiang Mai, Thailand.

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

Peters, Pam (2004). The Cambridge Guide to English Usage. Cambridge University Press.

Resnik, Philip and Smith, Noah A. (2003). The Web as a parallel corpus. Computational Linguistics, 29(3):349–380.

Sharoff, Serge (2006a). Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, editors, Wacky! Working Papers on the Web as Corpus, pages 63–98. GEDIT, Bologna, Italy.

Sharoff, Serge (2006b). Open-source corpora: Using the net to fish for linguistic data. Corpus Linguistics, 11(4):435–462.

Trudgill, Peter (1984). Standard English in England. In Peter Trudgill, editor, Language in the British Isles, pages 32–44. Cambridge University Press.

Trudgill, Peter and Hannah, Jean (2008). International English: A Guide to Varieties of Standard English, 5th edition. Hodder Education.

JADT 2012

Actes des 11es Journées internationales d'Analyse statistique des Données Textuelles

Proceedings of the 11th International Conference on Textual Data Statistical Analysis

Anne Dister, Dominique Longrée, Gérald Purnelle (éds)

Liège
13-15 juin 2012 / June 13-15, 2012

LASLA - SESLA
Université de Liège - Facultés Universitaires Saint-Louis Bruxelles

Université de Liège / Facultés Universitaires Saint-Louis - Bruxelles

JADT 2012

11es Journées internationales d'analyse statistique des données textuelles