Corpus creation using blogs: investigating [be done X] in Canadian English

Jennifer A. J. Hinnell

Simon Fraser University

jhinnell@sfu.ca

This article reports on a study using a corpus constructed from weblogs used to investigate the Canadian English construction [be done X], as in *I'm done my homework*. In examining this expression in natural usage through a corpus, the relationship between form and meaning construal that is so critical to cognitive linguists is explored. I focus here on the methodology behind the creation of the corpus. The corpus was built using the pre- and post-processor WebCorp, which, interacting with Google Blogs, allowed data gathering from blogs by national domain. However, due to parameters such as the restriction in register and context of the usage of [bdX], the frequency of other 'done' constructions that muddied searches, as well as limitations placed on WebCorp by its use of the Google Blogs API, gathering a robust, representative corpus was challenging. In particular to tease apart semantic differences inherent in Canadian [bdX] and Canadian and American [bdwX], a very fine-grained set of searches was essential. This paper demonstrates that corpus-based, quantitative observations are essential to developing accurate knowledge of the behaviour of constructions, and thus to allow robust theoretical analyses. It also demonstrates that creation of such corpora using raw text from the internet requires careful attention to ensure an adequate volume of accurate data.

Keywords: corpus development, weblogs, WebCorp, Canadian English

1. Introduction

In a cognitive linguistics (CL) framework, language is seen as being grounded in cognition and considered a 'window into the mind'. CL holds that "features of our thinking, cognitive processes and social interactions need to be brought in to the picture and correlated with their linguistic manifestations" (Fauconnier, 1999: 96). Thus in a cognitive approach *form* is important not for its internal structural properties, but because "behind form is not a thing at all but rather the human power to construct meaning" (Fauconnier & Turner, 2002: 6). Within this framework, this article reports on a study using a corpus constructed from Canadian weblogs, which was used to investigate an expression unique to Canadian English, 'be done X', as in the statement 'I'm done my homework'. In examining this expression in natural usage through a corpus, this study probes the relationship between form and meaning construal that is so critical to cognitive linguists. Within the scope of this article I focus on the methodology behind the creation of the corpus specific to weblogs written by speakers of Canadian English. The corpus was built using the pre- and post-processor WebCorp, which interacts with Google Blogs. The paper is organized as follows: in the remainder of this introduction, I give an overview of the [bdX] pattern, its usage and distribution. Section 2 first offers a review of research on using the web as a corpus (2.1), and then I describe the methodology behind this project, including the creation of a corpus from Canadian blogs (2.2). In Section 3, I describe the corpus data that was gathered, and pitfalls that were encountered that are specific to mining data from the World Wide Web. I conclude with a discussion of the

results for the [bdX] corpus and how they contribute to the theoretical analysis of the construction (Section 4).

The [bdX] pattern

To begin, I explain my discovery of the phenomena under discussion, and outline the nature and distribution of the pattern.

While traveling in the United States, whose variety of English is very similar to my own Canadian English, I asked an American colleague at a café, 'Are you done your tea?' I wanted to know if I could take his cup. Though ubiquitous in Canadian English, this formulation was entirely unacceptable to him. Upon further investigation it soon became apparent that this was a little-noted but wide difference between the two varieties of English. Generally stated, Canadians think that all speakers of English have this construction, and Americans are shocked that this is available in Canadian English. Furthermore, the conversation that ensued was a vehement discussion of the difference for Canadians between the phrases 'I am done my homework', and the variant using *with*, 'I am done with my homework', which is common to both dialects.

English contains a variety of lexemes and argument structure phrases to express that an activity has been completed. Examples (1) and (2) are standard present perfect constructions with the main verb lexemes *done* and *finished* respectively.

(1) I have finished my homework.
(2) I have done my homework.

American and Canadian usage both also contain the pattern 'be done/finished + PP', in the form 'be done with X' and 'be finished with X', as in (3) and (4)[1], where auxiliary *be* is followed by a prepositional phrase headed by *with*:

(3) I am done with my homework.
(4) I am finished with my homework.

Both varieties of English also include the gerund form *be done V-ing* [bdV-ing], as in (5) and (6):

(5) I'm done arguing about it.
(6) When the kids are done playing, it folds up for convenient storage.

In Canadian English (CE), however, speakers also have an utterance of the form 'be done X' [bdX] where X represents a direct object noun phrase, as in Examples (7)-(12):[2]

(7) I'**m done my finals** on Dec 9th, and then head home for Xmas.
(8) Martin **is done his bass tracks** and we are ready to start vocals.
(9) By the time I **am done dinner**, I don't want my side snack.
(10) My parents would extend his time block because he **was not done his homework.**
(11) So many bloggers I read are doing this. One is already **done her 50,000 words**!
(12) This will be particularly important once you**'re done the tattoo** and need to leave the shop.

This construction is widely accepted in all varieties of CE, while being highly *un*acceptable in American English (AE). The challenges encountered in this study center on the fact that while both Englishes use [bdV-ing] and [bdwX], only Canadian English allows the [bdX] variant. The construction does not appear in standard corpora; fortunately, weblogs prove a good resource, as they capture the

[1] I henceforth refer only to [bdwX], as I consider 'be done with X' and 'be finished with X' to be the same construction with possible alternation between the main verbs *done* and *finished*.

[2] These examples are all from the corpus developed for this study.

required colloquial register of Canadian English. Thus a corpus based on Canadian blogs on the World Wide Web was conducted.

The pattern described above has been attested across a wide variety of Canadian dialects in distinct geographical areas. According to interviews and qualitative surveys conducted by Yerastov (2008; 2010a; 2010b) speakers of CE in Alberta, British Columbia, Saskatchewan, Ontario, Quebec and Nova Scotia have reported the grammaticality of the [bdX] variant. Yerastov also attests that some speakers of Northeast Vermont English (NEVE) (in Orleans, Essex, Caledonia, and Lamoille counties) and some speakers in Philadelphia, Pennsylvania also accept and produce this variant.[3] In contrast, in grammatical judgment tasks conducted by Yerastov, speakers of American English from Massachusetts, upstate New York, Minnesota, Illinois, and Washington State found the construction grammatically unacceptable. The distribution of the pattern [bdX] is shown in Figure 1, where the dots indicate areas where the construction was found.

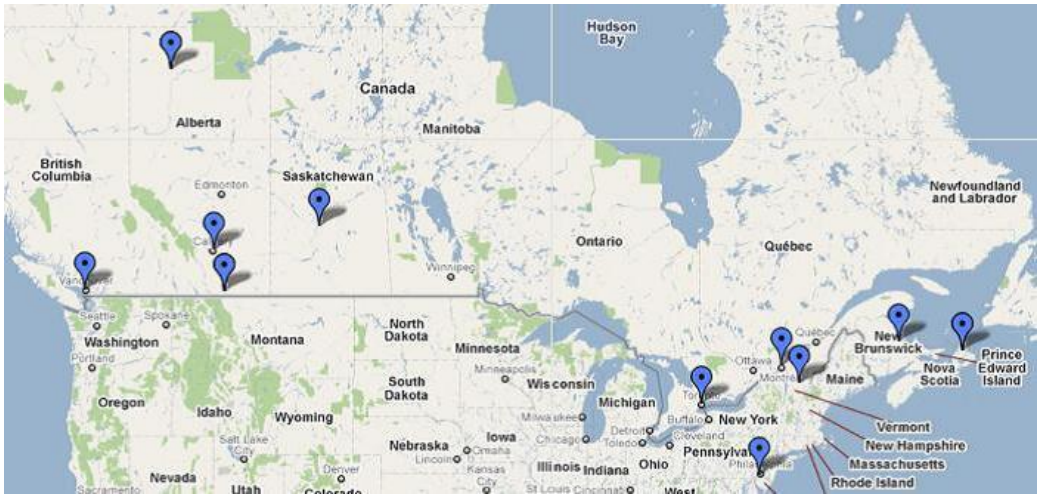

Figure 1. Distribution of [bdX] in North America (Yerastov, 2010b)

There are several factors that demand investigation in [bdX]. These are outlined in A-C below:

A. [bdX] shows syntactic variation among standard varieties of English. Generally speaking, while the phonetic markers of CE as compared to AE are noted (for example, the well-known phenomenon of Canadian raising), it is less common to have markers of a syntactic nature between these two standard varieties of English. To be sure, there are syntactic differences, but they are more usually regional, as in the case of the double modal construction in the American south, which is equally unacceptable to a speaker in New York as it is for someone from Toronto or Vancouver. In [bdX] we have a construction that is accepted in CE but not accepted at all (except for the tiny pocket noted in North Eastern Vermont and Pennsylvania) south of the border.

B. The second reason for the particular interest in this construction is the existence of two variants in CE, [bdX] and [bdwX], that have similar, but not the same, semantics. This is important in a constructionist approach, which posits a one-to-one form-meaning pairing (that is, a different form indicates a

[3] Note that not all speakers in Northeastern Vermont and Philadelphia accept this variant, in contrast to Canadian English – where it is acceptable in all regions.

different function). I argue that the two variants are not synonymous or interchangeable. Rather, [bdX] expresses *exhaustivity* or the final completion of a process, including the exhaustion of any entity referred to, and [bdwX] expresses *satiety*, or a sense of having had enough (thus *sated).* It indicates the completion of a stage in the process, but not necessarily the final stage. Crucially, [bdX] can *only* carry the exhaustive meaning. The corpus was developed to examine the following questions that arise from these preliminary observations: 1. Can the corpus study presented here demonstrate the semantic difference posited above between [bdX] and [bdwX]? 2. How does AE express this semantic contrast? 3. Is the AE [bdwX] always ambiguous? (This is especially important form a constructionist perspective: if AE does not have the contrast between [bdX] vs. [bdwX], then CE has a one-to-one form/meaning mapping, and AE has a two-to-one form/meaning mapping, which results in ambiguity).

C. Thirdly, in contrast with the [bdwX] variant, [bdX] is not entirely productive. As we will see in the description of data, it cannot be combined with just any NP, or any determiner, in any context. The corpus data show that [bdX] has very particular characteristics that need to be accounted for.

While the present study is restricted for reasons of scope to CE and AE, I conducted a cursory search to establish whether other varieties of English exhibit a [bdX] pattern. A search for 'am done my' on UK web blogs (using the same corpus-building tool as was used for the CE and AE data, WebCorp,[4] yielded the following examples:

(13) I myself <u>am done my</u> schooling and now work at a Calgary moving company.
(14) But once I <u>am done my</u> studies and I go back to Canada, how do I become
(15) I am currently on 60 mgs and <u>am done my 5th month</u> in a week.

(13) and (14) openly refer to Canada, and are thus assumed to be written by Canadians posting on UK blogs. A search for [bdX] with the definite determiner of the form 'am done the' yielded 8 instances. However, again, looking at the utterance in its context on the blog showed that the source location of 4 of those 8 blog comments was in Canada, or made reference to growing up in Canada. The results for a search of Australian blogs proved more definitive, with no instances of 'am done my', and one instance of 'am done the'. Based on the paucity of data from these searches, I believe Australian and British English can be assumed not to have the [bdX] construction.

As outlined in this introduction, the [bdX] construction provides data through which to examine the constructionist approach. In turn, the constructionist approach gives us tools to explore how to provide a unified account for the layers of meaning encoded in the construction. However, importantly, the limited scope of use of [bdX] required creativity in data mining, namely the creation of a corpus using data from weblogs written by Canadians. Thus this study offers insights into the use of corpora to examine constructions at a very specific level – i.e. in certain varieties of a language, in colloquial registers – that are under- or not at all represented in standard corpora.

In the CL framework, analyses of patterns in language need to be both broad and deep, that is, they must account for *all* meaning conveyed in a language utterance. I chose to work in the cognitive and constructionist approach to

[4] http://www.webcorp.org.uk

language as it allows a unified analysis of *what* meaning is conveyed by a particular language event, and consequently examines *how* it is conveyed. I use a corpus study as a tool to examine many instances of the construction in context. Corpus work is a methodology well-aligned with the research interests of cognitive linguists. Here I rely on data collected from Canadian web blogs to inform my analysis. In the next section I examine what constitutes a corpus, and how to define the role of the World Wide Web in corpus building.

2. Corpora and the Web

The field of Cognitive linguistics favors corpus work as an increasingly important methodology, as it allows for analysis of frequency and patterns in language. In corpora "facts about the actual use of linguistic expressions such as frequencies and individual patterns that are fully compositional are recorded alongside more traditional linguistic generalizations'' (Goldberg 1995: 45). In this section I introduce the notion of corpus, and examine the advantages and shortcomings of using the World Wide Web as a corpus.

Cognitive linguists see the use of corpora as more than simply a supplementary tool to confirm intuitions; rather it is seen as a "fundamental part of theory construction. [A] schema-based approach is well-suited to the task of describing the major and minor patterns of use revealed by corpus analysis" (Barlow, 1996: 2). The growth of corpus studies in the CL framework is slowly addressing the criticism within mainstream linguistics that the Cognitive Linguistic framework has received for their "post-hoc flavor and […] lack of predictive force" reported by Gries, Hampe and Schönefeld (2005) (but cf. papers in Barlow and Kemmer 2000; Boas 2003; Diessel 2004; Goldberg et al. 2004).

2.1 What is a corpus

In the history of corpus studies there is a lasting discussion on what makes a corpus. McEnery and Wilson respond that a corpus must have the following characteristics: it must be representative, a sampling, finite in size, machine-readable and a standard reference (McEnery, 1996). However, Kilgarriff and Grefenstette (2003) revisit this argument and reframe the question as: 'what is a good corpus for linguistic task X', and define a corpus as "a collection of texts when considered as an object of language or literary study" (Kilgarriff & Grefenstette, 2003). To this I would add that a corpus is generally restricted to facilitate the object of study. For example, in the study presented here, in order to explore [bdX] at it is used in Canada, I restricted my corpus to instances of [bdX] found on Canadian web logs. The web is a corpus, but in order to render it useful, it needs to be reshaped to the purposes of a particular project.

The origins of early corpora were varied as they were built to serve different purposes. In the early 1960s, the Brown corpus was created as a resource for computer-based language study. It contains 500 samples of English-language text, totaling one million words, compiled from works published in the United States in 1961. The corpus was originally lexical, and later tagged for part of speech. In the 1970s Sinclair and Atkins developed the COBUILD (Collins Birmingham University International Language Database) project, which has yielded the Bank of English, a corpus of contemporary texts, as well as the Collins COBUILD English Dictionary, which eventually grew to 8 million words. The

COBUILD dictionary was the first to be based completely on corpus data and to give examples of real language in use from corpus data from the Bank of English corpus. Ten years later Atkins was also involved in the development of the British National Corpus, with its 100 million words capturing different varieties of British English. Corpora are no longer the domain only of computational linguists, but are used regularly by linguists in a wide variety of frameworks as a way to access actual language data. Other large English-language corpora include: the International Corpus of English (ICE), which has sub-corpora for varieties of English; the Switchboard Corpus is a corpus of spontaneous conversations collected at Texas featuring over 240 hours of recorded speech, or 3 million words of text, spoken by over 500 speakers of both sexes from every major dialect of American English; and The Corpus of Contemporary American English (COCA), 425 million word corpus collected by Brigham Young University between 1990 and 2011.

With the exponential growth of the World Wide Web since the 1980s, linguists have access to a huge amount of natural language data at their fingertips. This virtually limitless repository of actual usage data provides an incredible resource for linguists, with easy access and at no expense. Kilgarriff and Grefenstette have called it a "fabulous linguists' playground" (2003). There are many reasons for this: it is big (hundreds of billions of words); it is often the only available source for the type of language the researcher is interested in (Luedeling 2007); it is free; and it is instantly available (Kilgarriff 2003). As is the case with [bdX], the web also allows us to explore phenomena that are sparse in a standard corpus either because it belongs to a genre or register not represented in the corpus, or because it stems from a time not covered by corpus data (i.e. is too new) (Luedeling 2007: 14). In these cases the web presents an excellent source of data.

Discussions on whether the web is a corpus usually center on the question as to what a corpus represents. Indeed each corpus is only representative of the texts it uses as a source. The widely used Penn Tree Bank (PTB) is based on articles from the Wall Street Journal, and thus is representative only of the kind of language found in journal articles in that daily newspaper. Similarly the Corpus of Contemporary American English (COCA) claims to be the largest freely-available corpus of English and the "only large and balanced corpus of American English."[5] However, I would argue that the claim to being the only balanced corpus of American English is misguided: it is balanced among the genres it includes. The Switchboard Corpus is another widely used corpus, but as it is based on recorded telephone conversations, cannot claim to be representative of 'all' language any more than the PTB. The problem of representativeness is thus not restricted to the web.

However, there are other well-documented (Kilgarriff & Grefenstette, 2003; Luedeling, Evert, & Baroni, 2007; Renouf, Kehoe, & Banerjee, 2007) problems inherent in using the web as a corpus. Luedeling (2007: 9) states that, in order to search a corpus, one needs:

a) a qualitative description of the items to be found that can be operationalized in the form of search conditions
b) a stable corpus (at least for the duration of data acquisition, but ideally in the long term to enable other researchers to replicate results)
c) linguistic annotation so that the search item can be located (see a)

[5] http://corpus.byu.edu/coca

d) the possibility to categorize search results according to criteria such as age/gender/genre of speaker

Many of these criteria prove problematic for the web. With respect to a) and c): while many linguists simply use frequency results from Google searches, for example, this is problematic for many reasons. Google weights its search results towards search items that appear in page titles, for example (for a more detailed discussion see Luedeling (2007)). With regards to b), the web is not stable. Users, hosts of websites, and companies, change content on the web innumerable times daily. Thus a search that is conducted at a given point in time can never be replicated. This is not desirable for linguistic research. Lastly, there is no control on the web of who the 'speaker' is, meaning that the information Luedeling lists above as desirable to categorize speakers – age, gender, etc. – cannot be ascertained.

Despite these weaknesses, however, the web does present a valuable resource due mostly to the volume of data and the aforementioned presence of language utterances not captured in other corpora. There are now tools to operationalize searches of the web that essentially provide a 'layer' between the search term and the search engine and allow the linguists more control in gathering their data. For the study conducted here, I used one such tool, WebCorp (www.webcorp.org.uk) to restrict the searches to Canadian web blogs to capture colloquial Canadian language use. While blogs are by virtue of the medium written text, their language register is more spoken than written, and therefore were well-suited to a search for colloquial Canadian speech. In the following section I outline the creation of my corpus using the WebCorp interface.

2.2 Methodology for [bdX] corpus building

Gathering the data to investigate [bdX] and related variants [bfX] and [bdwX] in AE and CE proved challenging. I searched the two large corpora of Canadian English: The Bank of Canadian English (BCE)[6] and ICE-Canada (the Canadian portion of the International Corpus of English (ICE).[7] The BCE contained 3 instances of [bdX], and ICE-Canada contained 0. Similarly, to investigate whether there were attested American instances of [bdX], and to explore the American English variant [bdwX], I undertook a search of the Corpus of Contemporary American English (COCA)[8]. The search for [bdX] in COCA yielded two results: one of these was in a clearly Canadian context, and the other was ambiguous as to the dialect of the speaker. I believe the scarcity of results in traditional corpora was due largely to the colloquial, informal register in which this construction generally is used, and to the fact that as a result it is largely spoken. The BCE is 2.4 million words, of which 100 citations were transcribed from spoken language, and the remainder is from books, periodicals such as newspapers, magazines, and journals, and websites, online data and flier texts. The ICE-Canada corpus

[6] The BCE is a database project that aims to provide a complete, representative record of Canadian English for linguistic research. Access is currently restricted to individuals for the purpose of research. I'd like to express my appreciation to Dr. Stefan Dollinger (UBC) for granting me research access to the BCE.

[7] ICE-Canada, part of the International Corpus of English (ICE), is hosted at the University of Alberta. Access is available under license by contacting Professor John Newman in the Linguistics Department of the University of Alberta. I am grateful for Dr. Newman's assistance in accessing the corpus.

[8] COCA is hosted by Brigham Young University: http://corpus.byu.edu/coca

contains 500 texts of approximately 2,000 words each for a total of approximately one million words. Of the 500 texts, 300 are spoken (phone calls, face to face conversations, classroom teaching, broadcast news, broadcast discussions, parliamentary debates, etc.) and the remaining 200 are written (e.g. student writing, exam scripts, academic writing, popular writing). Despite the higher number of texts from spoken data in ICE, [bdX] still did not appear. As we see in Section 3 [bdX] is largely restricted to exchanges regarding domestic routines or work-related tasks. This type of topic is not discussed in newscasts, for example. Initially I had thought that the telephone conversations in ICE-Canada would be a good match. However, it could be that that portion of the ICE-Canada corpus was simply too small to capture any instances, or that the subject matter and register of the calls did not lend themselves to colloquial usage about domestic routines, for example.

Faced with these difficulties in using traditional corpora I turned to the World Wide Web. WebCorp is a web-based interface to search engines that was used in this project to execute searches using of language instances on the World Wide Web. In this section I give a description of the methodology for developing the corpus using WebCorp, including the difficulties that were encountered in working with data on the Web, and a detailed sketch of the searches that were conducted. This sets the stage for Section 3 in which I describe the patterns in the data that were discovered.

WebCorp

WebCorp[9] interfaces with web search engines Google and Bing, giving the user the opportunity to specify a query in several important ways. The query itself is not limited to words, but can also include wildcards and regular expressions. The user can also restrict the search to a particular application programming interface (API), for example, deciding whether WebCorp should search using the Google search engine, the Bing search engine. Even more options are available to restrict searches, such as to only search Google News, Bing News or Google Blogs. Once the user has input the search term, defined the API, and the language of the search, WebCorp returns up to 60 hits per search, then organizes the search results in a manner similar to standard concordancers, by highlighting the keyword in context. It also returns metadata including the last update date and the URL of the source web page. This is important for replicating results, and accessing the full context of utterances that WebCorp returns. A schematic of the WebCorp architecture is presented in Figure 2 below, showing the search and analysis routine:

[9] http://www.webcorp.org.uk

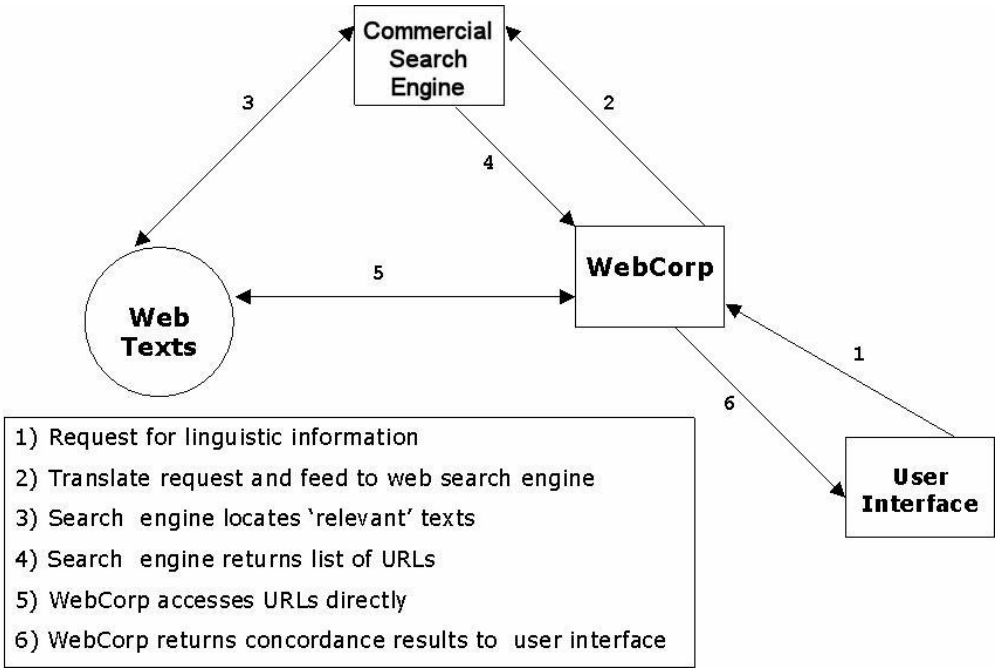

Figure 2. WebCorp system architecture (Renouf et al., 2007: 48)

However, since WebCorp accesses the web pages through the API of a particular search engine, it is limited to restrictions imposed by that search engine. For example, Google ranks the strength of the 'hits' based on its own mechanisms. While it would be more useful for linguists if WebCorp returned a random sample of results, instead the results are ordered according to Google's priorities, which are to return the pages that rank highest in popularity or topical relevance. This is not relevant to a search with a purpose of linguistic study. A second limitation is that the search is restricted at a maximum of 200 web pages (the user can set this at 10, 25, 50, 100 or 200 pages) for reasons of ensuring manageable search times. The number of returns is also capped by the Google Blogs API at 60 per search, which in the [bdX] searches proved very limiting – a point I return to in more detail below. Lastly, since part of speech is not tagged on the web, WebCorp cannot interpret part of speech tagging, for example, which is a very important feature in fixed corpora that renders the data more searchable.

The stability of a corpus developed from web data is also a challenge. Search companies are constantly updating their algorithms for search results, their databases and their interfaces. Thus experiments using tools such as WebCorp can never truly be replicated, as the control really rests with the search engine provider. To address this shortcoming, the Birmingham corpus group (the team behind WebCorp), has built its own corpus directly from the web: Linguists Search Engine (LSE). This corpus is built on web data, but is not 'live'. As a static corpus any search results using LSE are replicable. While addressing the charge of stability, the disadvantage is that in becoming a static corpus, one of the main advantages of the web – as a source of up-to-the-minute language usage data – is lost. The LSE was unfortunately not useful for the current experiment, as the corpus is not yet able to differentiate the sources of data, meaning that it was not possible to restrict the search to data that originated on Canadian web blogs. Despite the fact that the results are not replicable, using WebCorp was the best option to gather the data set required for this project. Having seen the main pitfalls of using the web, and WebCorp, to acquire data, I now turn to the factors that

made this project possible (recall that none of the established corpora contained instances of this construction at all).

The primary advantage provided by WebCorp for the current study was the ability to restrict searches to a certain genre, namely to the Google Blogs API, and to further restrict by domain name – that is to *.ca* for the CE searches – as seen in the screen capture in Figure 3 below. In the advanced settings I restricted the search to the *.ca* domain for the Canadian searches, and for American searches, changed the settings to include American blogs and newspapers.

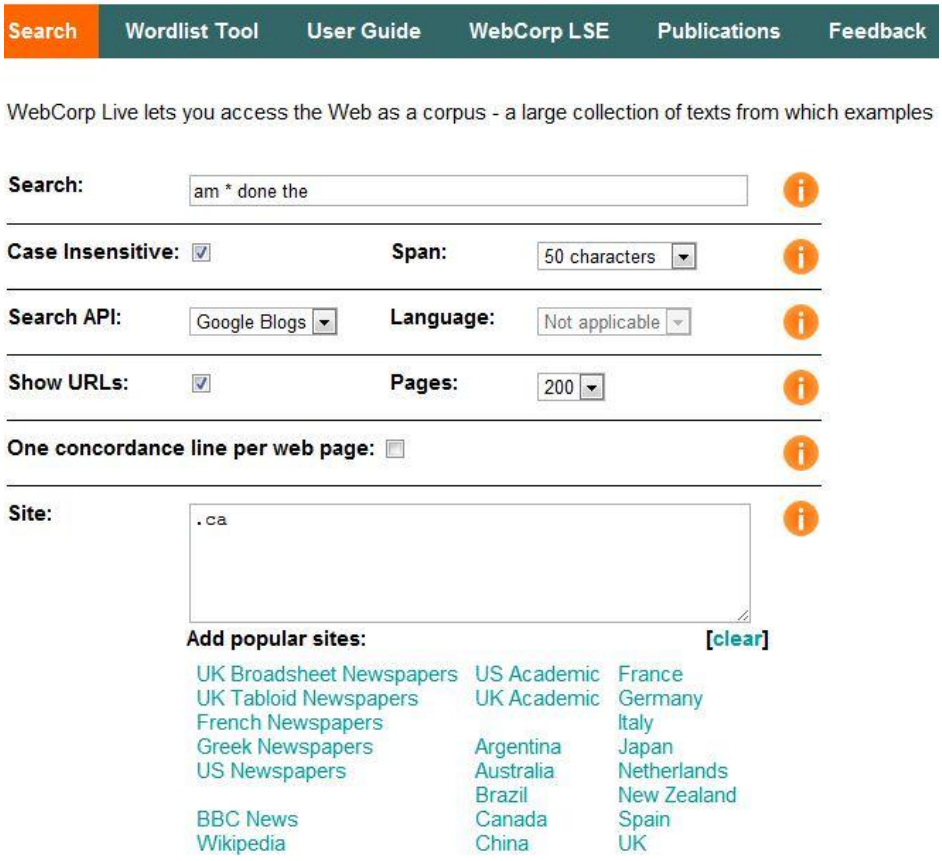

Figure 3. WebCorp screenshot (http://www.webcorp.uk.org)

In addition to restricting domains, the pattern matching capabilities of WebCorp are more fully developed than searching Google directly. Both WebCorp and Google itself process wildcard searches. This is when the wildcard (*) stands for any word in a phrase – e.g. "the * sank" will match "the boat sank", "the ship sank", and "the ferry sank"). By searching with the wildcard (I'm * done the), I could search for modification, such *as I'm almost done the, I'm really done the, I'm half done the.* However, WebCorp also offers pattern matching where groups of characters are enclosed in square brackets and separated by the pipe (|) character. For example, while "the * sank" will match any three word phrase beginning with *the* and ending with *sank*, the pattern "the [ship | boat] sank" will only match *the ship sank* or *the boat sank*. I used these to group searches together, as in "[am|are|is] done the".

As also mentioned above, however, there are significant problems posed by relying on the internet for linguistic data. In this study, the primary downfall was that the Google Blogs API limits WebCorp to returning 60 hits per search. Thus a truly robust quantitative study based on comparing the number of hits of

one construction with the frequency of a second construction, was rendered impossible. If there were more than 60 hits per search, only 60 would show, with no indication whether this was 60 of 61, or 60 of 250, for example. This limit on search items returned proved the most frustrating and limiting feature of gathering data using WebCorp, and had significant repercussions on the data gathered. This was complicated by the fact that, because the internet is not parsed for part of speech, there were many search returns that contained uses of *done* that were not part of [bdX] or [bdwX]. In order to receive as many instances of the relevant construction as possible, I therefore had to control the search terms more than I would have liked. I had to search for full phrases 'I am done my' and 'I'm done my' for example, rather than 'done my'. For example, a search query for 'done my' returned instances of the more popular phrase 'have done my', and didn't return any instances of the [bdX] constructions within those crucial first 60 results.

The searches

Having reviewed the advantages and downfalls of using the web, we now turn to the data that was collected. Data sets were built for each of the three variants for both CE and AE, resulting in a total of six separate data sets, as shown in Table 1.

Construction	**CE: # of tokens**	**AE: # of tokens**
[bdX]	764	7
[bfX]	358	8
[bdwX]	1385	833
TOTAL	2507	848

Table 1. Data sets included in this study

The same searches were conducted on each data set. The search term for each of these six constructions was composed of 3 parts: 1. the form of the copula *be* (present and past in all persons); 2. *done* or *finished*; and 3. the determiner head of the noun phrase. I modeled my search grid on Longman's description of co-occurrence patterns of major classes of determiners and nouns in the Longman Grammar of Spoken and Written English (LGSWE) (1999:259), which lists the following major determiner types: definite and indefinite articles, personal pronouns, demonstratives and the 6 most frequent quantifiers in English, which, according to LGSWE *are all, both, each, every, many* and *some*. All of these searches were then performed a second time inserting a wildcard (*) before *done* to capture all instances of modification (*really done, half done, not done,* etc.). Past tense searches were also conducted.

In addition to the searches listed above, searches for contracted verb forms were also performed, e.g. *I'm/we're/you're/they're done* + determiner. Again these could have been collapsed into two searches: *'m done/'re done* + determiner type if it weren't for the Google Blogs API return limit of 60 instances. Of note here is the difficulty with the third person contractions *he's* and *she's*. These were not included as it was impossible to tell whether the contracted form was *has*, as in *She has done her homework* (present perfect construction) or the [bdX] form with the *be* copula, as in *She is done her homework.* An abbreviated table of the search combinations is shown in Appendix A.

It must be noted that question forms were not included in the search terms listed above to limit the scope of searches. Bare nouns were also a problem. The few attested instances of [bdX] in the BCE noted above contained a bare noun as X. Given the restrictions of using WebCorp described above, I note here that these bare nouns could only be queried on a one-off basis by entering the noun itself, for example 'am done dinner' or 'is done homework'. Because the web is not parsed, it was not possible to query bare nouns in a manner that would return all of the instances of [bdX] where X is a bare noun, and bare nouns are therefore not accounted for in this project.

A further restriction of WebCorp related to its lack of parsing is that it does not differentiate between items in a search that are at the end of one sentence and items in the same search that begin the next sentence. Thus, for example, a search for 'am done the' returned "am done. The guys…", clearly not a desired result for this construction. This again muddied the data given the limit of 60 returns. Where there were many cases of this type, it was hard to get a clear picture of how frequent the desired search result was. The same applied for searches with the indefinite article, for example, that would return "It's done a year in advance", for the search "is done a", where this is not an example of the [bdX] construction, but rather an instance the normal use of "I'm done" with a temporal adverbial phrase.

It should also be mentioned that obtaining the search results using WebCorp was the most important, but not the only step in creating a usable corpus. All relevant search items returned by WebCorp were cut and pasted individually into an excel spreadsheet as text. Subsequently a script written in python[10] was used to parse the text, consisting of the full sentence from the blog, the URL and related metadata, into three separate columns. The remainder of the data manipulation, e.g. separating the 'X' noun from the 'be done X' in order to annotate the semantic field of the noun, was done manually.

In this section I have outlined the methodology of creating the corpus, including the advantages and pitfalls of using the web, and WebCorp. Working within the restrictions of WebCorp and the Google Blogs API definitely affected the data that could be collected. The results shown here, therefore, present a representative sample of the usage of [bdX] and other patterns, but cannot be in any way considered statistically reliable. However, the results do represent a sampling of [bdX] and related constructions in the natural usage of colloquial English, and can be considered robust, if not entirely representative in a statistical sense. Having shown how the data was collected, I turn to a description of the data set in Section 3 and elucidate briefly how these are important to the theoretical analysis.

3. Results

We saw in Table 1 the six data sets that were gathered to create the corpus. The few instances (7 and 8 respectively) of [bdX] and [bfX] on AE blogs were all from the *.edu* web domain. This is the domain for educational institutions in the USA, so it is likely that these instances were posted by Canadian students studying in the United States. While the results for the first two variants listed,

[10] Many thanks to Milan Tofilofski, PhD student in Computer Science at SFU for assistance with the scripts.

[bdX] and [bfX] were predicted, the high number of [bdwX] in CE over AE was unexpected. Naturally these numbers are not necessarily indicative of overall frequency of usage in CE and AE, i.e. it can't be claimed from this corpus result that Canadians use [bdwX] more than Americans. The corpus results do, however, beg the question as to why there were 50% more instances in CE than AE. This may be due to shortcomings addressed earlier regarding the restrictions of gathering the data. However, the question also arises as to whether Canadians are more 'done' in general, than Americans, that is do they use the 'done' constructions more than alternative ways of expressing completion in the past tense. Possible directions for enquiry include: Is [bdwX] a more frequent usage in CE due to the entrenchment of other 'done' constructions such as [bdX]? Does AE use the present perfect [hdX] more often? These questions lie beyond the scope of this project but a more detailed study, especially contrasting [bdwX] and [hdX] in AE would certainly yield interesting comparisons.

The three factors that were examined across each variant were the distribution of NPs in the subject slot, determiner distribution, and the distribution of direct object noun phrase. In the presentation of the results, I compare the frequencies for [bdX] and [bfX], which pattern together for the most part, and [bdwX], across CE and AE.

Firstly, for subject distribution, the corpus results show that all four variants share a similar pattern: the first person subject 'I' is most common (~38%-47%), second person pronoun 'you' is next most common (~17-29%), and the first person plural 'we' is third most common). The high rate of first person usage is expected in a corpus of blogs, which as a genre are generally narrations in the first person; however, I believe this also relates to particular semantics of the construction, for example the strongly viewpointed nature of the construction. This is also supported by the next set of results related to determiner patterns.

The top three determiner collocations for each of the variants are shown in Table 2.

Pattern	**Top three determiners**
[bdX] – CE	the X, my X, his/her X
[bfX] – CE	the X, my X, your X
[bdwX] – CE	the X, this X, my X
[bdwX] – AE	the X, my X, this X

Table 2. Most common determiner patterns per construction

The data clearly show that *the* is dominant in all cases, while *my* is second most common in three of the four patterns. While the definite determiner and first person possessive determiner are the most frequently occurring determiner types across all constructions and variants, the remaining possessive determiners occur much more frequently with the [bdX/[bfX] construction than with the [bdwX] construction. In Table 3 below, I compare the distribution for each pattern according to the determiner classes: article (definite/indefinite), personal pronouns, demonstrative pronouns, and quantifiers.

Determiner Groupings	**[bdX] – CE**		**[bfX] – CE**		**[bdwX] – CE**		**[bdwX] - AE**	
Personal pronouns (*my, your, his/her, our, their*)	404	52.88%	194	54.19%	381	27.51%	271	32.53%
Definite/Indef. determiner (*the, a*)	168	21.99%	114	31.84%	550	39.71%	273	32.77%
Demonstrative determiner (*this X, that X, those X*)	77	10.08%	27	7.54%	227	16.39%	146	17.53%
Demonstrative pronoun (*this, that, those*)	58	7.59%	6	1.68%	149	10.76%	69	8.28%
Quantifier (*all, both, each, every, most, some*)	55	7.20%	17	4.75%	61	4.40%	69	8.28%

Table 3. Determiner distribution by determiner class for all variants

Here one can see that the most frequent determiner type for [bdX] and [bfX] is the personal pronoun, with ~52-54% of tokens containing this pattern. While one could suggest that this is due to the first person bias of a blog corpus, a look at the distribution for [bdwX] in the CE and AE quickly demonstrates that this is not the reason. [bdwX] contains a personal pronoun in the determiner position much less frequently (close to 30% for both of them).

The systematicity of the corpus searches allowed insight into patterns that may not have been noticed qualitatively. One such example, which demonstrates a highly viewpointed construal conveyed by [bdX], is the *lack* of results where the subject and possessive pronoun are not cross-mapped. That is, there are no instances of sentences such as **I'm done his homework (for him)*, which lacks reflexivity. In cursory searches for the regular present perfect, this reflexivity is widely accepted (*I've done his homework (for him)).*

Lastly, the noun phrase that had been isolated manually from the 'be done' portion of the phrase (or 'be done with' for [bdwX]]), was annotated in the corpus for semantic field, using a classification system of 14 fields. It was found that the semantic fields of [bdX] and [bfX] patterned similarly, in that four of the top five semantic fields were the same (*education, projects, work related tasks, chores,* and *leisure).* However collocational patterns differ when one sorts the corpus using exact nouns, rather than broad semantic field classes. While [bdX] prefers the nouns *shopping, year, work, class/classes,* and *exam/exams,* [bfX] prefers *degree, book/books, school/schooling, work* and *course,* in that order.

In comparing these results with the data for [bdwX] in Canadian and American English, results indicate the [bdwX] variants are more productive. They have a wider variation in semantic field, with the Canadian construction being more varied in its direct object than its American relative. The American [bdwX] variant showed very high correlation with the *education* semantic field, though this is likely skewed somewhat due to the source of data from American *.edu* blogs.

This data forms the basis for the analysis of syntactic, semantic, discourse and overall constructional properties. The full analysis lies outside the scope of this paper. Suffice it to say, for this paper, that the corpus, though limited by the aforementioned challenges of mining data from a web interface, provides rich data from which to construct a detailed and at the same time broad picture of the behavior of these constructions.

4. Conclusion

This study has demonstrated the requirements and challenges of developing a robust corpus dedicated to investigating the [be done X] construction in Canadian English, particularly in contrast to its AE counterpart [be done with X]. The corpus was built to serve a larger study, which was a construction analysis of [bdX], including analyses of the syntactic, semantic, discourse-level and viewpointed elements of the construction. This investigation of [bdX] is the first to draw on corpus data; others have investigated the phenomenon qualitatively, but have therefore not shed light on many of the patterns of usage highlighted here. Importantly, none have discussed the difference in semantics of exhaustivity of [bdX] in contrast with [bdwX], a large part of the construction analysis.

A corpus was developed from instances of these constructions found on weblogs using the pre- and post-processor WebCorp. However, there were important factors that made the corpus creation challenging. These included the restrictions in register and context mentioned above, the frequency of other 'done' constructions that muddied searches, as well as limitations placed on WebCorp by its use of the Google Blogs API. In particular to tease apart semantic differences inherent in Canadian [bdX] and Canadian and American [bdwX], a very fine-grained set of searches was essential. The set of searches for each variant (of which there were three for each of Canadian and American English: [bdX], [bfX] (be finished X), [bdwX]) included separate searches for determiners in four classes (articles, demonstratives, possessives and quantifiers), all verb forms for present and past tense of the copula, and searches with wildcards to catch negation and modification.

To summarize the results, it was found that, as expected in a blog corpus, first person was most frequent in subject position. For all variants, the determiner of the noun phrase represented by X was most frequently *the*, though when all possessive pronouns were counted together, possessive pronouns accounted for over half of the instances of [bdX] and its co-variant [bfX]. The pattern for [bdwX] was different, in that it preferred definite articles over possessives. This was part of the support found for the conclusion that [bdX] conveys a strong viewpoint element in contrast with [bdwX]. Further evidence to support this conclusion is the lack of cross-mapping between subject pronoun and possessive determiner in the noun phrase. Lastly, the differences in semantic field that were discovered through the manual annotation of the noun phrase demonstrated the different preferences of [bdX] and [bdwX], with [bdwX] being most productive in Canadian English, and [bdX] most restricted. The AE variant of [bdwX] was in the middle.

Corpus-based, quantitative observations of this kind are essential to developing accurate knowledge of the behavior of constructions, and allows robust theoretical analyses. The Web presents a treasure-trove of data ready to be mined, and can produce exceptionally fruitful contributions to knowledge of and theoretical approaches to language. However, as shown here, it also presents a challenge for linguists in how to effectively and efficiently draw out the information that is required for analyses. If we are to continue to understand and engage with language in use, as linguists we need to work with data of all kinds, including language as it is used on the World Wide Web.

Appendix A

Search items per construction

Type of determiner		**Initial search**	**Wildcard search**
Article	the	am/are/is done the	am/are/is * done the
	a	am/are/is done a	am/are/is * done a
Demonstrative	that	am/are/is done that	am/are/is * done that
	this	am/are/is done this	am/are/is * done this
	those	am/are/is done those	am/are/is * done those
possessive	my	am done my	am* done my
	your	are done your	are * done your
	her/his	is done her/his	is * done her/his
	its	is done its	is * done its
	our	are done our	are * done our
	their	are done their	are * done their
quantifier	all	am/are/is done all	am/are/is * done all
	each	am/are/is done each	am/are/is * done each
	every	am/are/is done every	am/are/is * done every
	both	am/are/is done both	am/are/is * done both
	some	am/are/is done some	am/are/is * done some
	many	am/are/is done many	am/are/is * done many

References

Barlow, Michael. (1996). Corpora for theory and practice. *International Journal of Corpus Linguistics, 1*(1), 1-37.

Fauconnier, Gilles. (1999). Methods and generalizations. In T. Janssen & G. Redeker (Eds.), *Cognitive Linguistics: Foundations, Scope and Methodology (Cognitive Linguistics 15)* (pp. 95-128). The Hague: Mouton de Gruyter.

Fauconnier, Gilles, & Turner, Mark. (2002). The Way We Think: Conceptual Blending and the Mind's Hidden Complexities. New York: Basic Books.

Gries, S. Th., Hampe, Beate, & Schönefeld, D. (2005). Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics, 16*(4), 635-676.

Kilgarriff, Adam, & Grefenstette, Gregory. (2003). Web as corpus. *Association for Computational Linguistics*.

Luedeling, Anke, Evert, Stefan, & Baroni, Marco. (2007). Using web data for linguistic purposes. In C. Biewer, N. Nesselhauf & M. Hundt (Eds.), *Language and Computers - Studies in Practical Linguistics, Volume 59: Corpus Linguistics and the Web* (Vol. 59). Amsterdam: Rodopi.

McEnery, Tony; Wilson, Andrew. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Renouf, Antoinette, Kehoe, Andrew, & Banerjee, Jayeeta. (2007). WebCorp: an integrated system for web text search. In C. Biewer, N. Nesselhauf & M. Hundt (Eds.), *Language and Computers - Studies in Practical Linguistics, Volume 59: Corpus Linguistics and the Web* (Vol. 59). Amsterdam: Rodopi.

Yerastov, Yuri. (2008). *I am done dinner: a case of lexicalization.* Paper presented at the 2008 annual conference of the Canadian Linguistic Association.

Yerastov, Yuri. (2010a). Done, finished, and started as reflexes of the Scottish transitive be perfect in North America: their synchrony, diachrony, and current marginalisation. In R. M. Millar (Ed.), *Marginal dialects: Scotland, Ireland and beyond* (pp. 19-52). Aberdeen: Forum for Research on the Languages of Scotland and Ireland.

Yerastov, Yuri. (2010b). *I'm done dinner: When Synchrony Meets Diachrony.* Unpublished Dissertation, University of Calgary, Calgary, AB.