Web as Corpus

Adam Kilgarriff[*]
Lexicography MasterClass Ltd. and ITRI,
University of Brighton

Gregory Grefenstette[†]
Clairvoyance Corporation

The web, teeming as it is with language data, of all manner of varieties and languages, in vast quantity and freely available, is a fabulous linguists' playground. The Special Issue explores ways in which this dream is being explored.

1 Introduction

The web is immense, free and available by mouse-click. It contains hundreds of billions of words of text and can be used for all manner of language research.

The simplest language use is spell checking. Is it *speculater* or *speculator*? Google gives 67 for the former (usefully suggesting the latter might have been intended) and 82,000 for the latter. Question answered.

Language scientists and technologists are increasingly turning to it as a source of language data, because it is so big, because it is the only available source for the type of language they are interested in, or simply because it is free and instantly available. The mode of work has increased dramatically from a standing start seven years ago with the web being used as a data source in a wide range of research activities: the papers in the Special Issue form a sample of the best of it. This introduction aims to survey the activities and explore recurring themes.

We first consider whether the web is indeed a corpus; then present a history of the theme in which we view it as a development of the empiricist turn which has brought corpora center-stage in the course of the 1990s. We briefly survey the range of web-based NLP research, then present estimates of the size of the web, for English and for other languages, and a simple method for translating phrases. Next we open the Pandora's Box of representativeness (concluding that the web is not representative of anything other than itself, but then nor are other corpora, and that more work needs doing on text types). We then introduce the papers in the Special Issue, and conclude with some thoughts on how the web could be put at the linguist's disposal rather more usefully than current search engines allow.

Is the web a corpus?

To establish whether the web is a corpus we need to find out, discover or decide what a corpus is. McEnery and Wilson (1996) say

> In principle, any collection of more than one text can be called a corpus .... But the term "corpus" when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition. The following list describes the four main characteristics of the modern corpus.

[*] Lewes Rd, Brighton, BN2 4JG, UK
[†] Suite 700, 5001 Baum Blvd, Pittsburgh, PA 15213-1854

McEnery and Wilson's list is "sampling and representativeness, finite (and usually fixed) size, machine-readable, a standard reference".

We would like to reclaim the term from the connotations. Many of the collections of texts which people use and refer to as their corpus, in a given linguistic, literary, or language-technology study, do not fit. A corpus comprising the complete published works of Jane Austen is not a sample, nor representative of anything else. Closer to home, Manning and Schütze (1999, p 120) observe:

> In Statistical NLP, one commonly receives as a corpus a certain amount of data from a certain domain of interest, without having any say in how it is constructed. In such cases, having more training data is normally more useful than any concerns of balance, and one should simply use all the text that is available.

We wish to avoid a smuggling-in of values into the criterion for corpus-hood. McEnery and Wilson (following others before them) mix the question "what is a corpus?" with "what is a good corpus (for certain kinds of linguistic study)", muddying the simple question "is corpus x good for task y?" with the semantic question, "is x a corpus at all?" The semantic question then becomes a distraction, all too likely to absorb energies which would otherwise be addressed to the practical one. In order that the semantic question may be set aside, the definition of corpus should be broad. We define a corpus simply as "a collection of texts". If that seems too broad, the one qualification we allow relates to the domains and contexts in which the word is used rather than its denotation: *a corpus is a collection of texts when considered as an object of language or literary study.*

The answer to the question "is the web a corpus?" is yes.

2 History

For chemistry or biology, the computer is merely a place to store and process information gleaned about the object of study. For linguistics the object of study itself (in one of its two primary forms, the other being acoustic) is found on computers. Text is an information object, and a computer's hard disk is as valid a place to go for its realization as the printed page or anywhere else.

The one-million word Brown corpus opened the chapter on computer-based language study in the early 1960s. Noting the singular needs of lexicography for big data, in the 1970s Sinclair and Atkins inaugurated the COBUILD project, which raised the threshold of viable corpus size from one million to, by the early 1980s, eight million words (Sinclair, 1987). Ten years on, Atkins again took the lead with the development (from 1988) of the British National Corpus (Burnard, 1995, hereafter BNC), which raised horizons tenfold once again, with its 100M words, and was in addition widely available at low cost and covered a wide spectrum of varieties of contemporary British English.[1] As in all matters Zipfian, logarithmic graph paper is required. Where corpus size is concerned, the steps of interest are 1, 10, 100 . . . , not 1, 2, 3.

Corpora crashed into computational linguistics at the 1989 ACL meeting in Vancouver: but they were large, messy, ugly objects clearly lacking in theoretical integrity in all sorts of ways, and many people were skeptical regarding their role in the discipline. Arguments raged, and it was not clear whether corpus work was an acceptable part of the field. It was only with the highly successful 1993 Special Issue of this journal on Us-

[1] Across the Atlantic, a resurgence in empiricism was led by the success of the noisy channel model in speech recognition (see Church and Mercer (1993) for references).

ing Large Corpora (Church and Mercer, 1993) that the relation between computational linguistics and corpora was consummated.

There are parallels with web corpus work. The web is anarchic and its use is not in the familiar territory of computational linguistics. However, as students with no budget or contacts realize, it is the obvious place to obtain a corpus meeting their specifications, as companies want the research they sanction to be directly related to the language-types they need to handle (almost always available on the web), as copyright continues to constrain 'traditional' corpus development,[2] as people want to explore using more data and different text types, so web-based work will grow.

The web walked in on ACL meetings starting in 1999. Rada Mihalcea and Dan Moldovan used hit counts for carefully-constructed search engine queries to identify rank orders for word sense frequencies, as an input to a word sense disambiguation engine (Mihalcea and Moldovan, 1999). Philip Resnik showed that parallel corpora –until then a promising research avenue but largely constrained to the English-French Canadian Hansard– could be found on the web (Resnik, 1999): we can grow our own parallel corpus using the many web pages that exist in parallel in local and in major languages. We are glad to have the further development of this work (co-authored by Noah Smith) presented in this Special Issue. In the student session of ACL 2000, Rosie Jones and Rayid Ghani showed how you can build a language specific corpus using the web from a single document in that language (Jones and Ghani, 2000). In the main session Atsushi Fujii and Tetsuya Ishikawa demonstrated that descriptive, definition-like collections can be acquired from the web (Fujii and Ishikawa, 2000).

2.1 Some current themes

Since then there have been many papers, at ACL and elsewhere, and we can mention only a few. The EU MEANING project (Rigau et al., 2002) takes forward the exploration of the web as a data source for word sense disambiguation, working from the premise that within a domain, words often have just one meaning, and that domains can be identified on the web. Mihalcea and Tchklovski complement this use of web as corpus with web technology to gather manual word sense annotations on the Word Expert website.[3] Santamaría *et al.*, in this volume, discuss how to link word senses to web directory nodes, and thence to web pages.

The web is being used to address data sparseness for language modeling. In addition to Keller and Lapata (this volume) and references therein, Volk (2001) gathers lexical statistics for resolving prepositional phrase attachments, and Villasenor-Pineda et al. (2003) 'balance' their corpus using web documents.

The Information Retrieval community now has a web track as a component of their TREC evaluation initiative. The corpus for this exercise is a substantial (around 100GB) sample of the web, largely using documents in the .gov top level domain, as frozen at a given date (Hawking et al., 1999).

The web has recently been used by groups at Sheffield and Microsoft among others as a source of answers for question-answering applications, in a merge of search engine and language processing technologies (Greenwood, Roberts, and Gaizauskas, 2002; Dumais et al., 2002). AnswerBus (Zheng, 2002) will answer questions posed in English, German,

[2] Lawyers may argue that the legal issues for web corpora are no different to those around non-web corpora. However, firstly, language researchers can develop web corpora just by saving web pages on their own computer without any copying on, thereby avoiding copyright issues, and secondly, a web corpus is a very minor sub-species of the caches and indexes held by search engines and assorted other components of the infrastructure of the web: if a web corpus is infringing copyright, then it is merely doing on a small scale what search engines such as Google are doing on a colossal scale.

[3] http://teach-computers.org/word-expert.html

French, Spanish, Italian and Portuguese.

Naturally, the web is also coming into play in other areas of linguistics. Agirre et al. (2000) are exploring the automatic population of existing ontologies using the web as a source for new instances. Varantola (2000) shows how translators can use 'just-in-time' sublanguage corpora to choose correct target language terms for areas where they are not expert. Fletcher (2002) demonstrates methods for gathering and using web corpora in a language teaching context.

2.2 The 100M words of the BNC

100M words is large enough for many empirical strategies for learning about language, either for linguists and lexicographers (Baker, Fillmore, and Lowe, 1998; Kilgarriff and Rundell, 2002) or for technologies that need quantitative information about the behavior of words as input (most notably parsers (Briscoe and Carroll, 1997; Korhonen, 2000)). However for some purposes it is not large enough. This is an outcome of the Zipfian nature of word frequencies. While 100M is a vast number, and the BNC contains ample information on the dominant meanings and usage-patterns for the 10,000 words that make up the core of English, the bulk of the lexical stock occurs less than 50 times in it, which is not enough to draw statistically stable conclusions about the word. For rarer words, rare meanings of common words, and combinations of words, we frequently find no evidence at all. Researchers are obliged to look to larger data sources (Keller et al, this Special Issue; also Section 3.1 below). They find that probabilistic models of language based on very large quantities of data, even if that data is noisy, are better than ones based on estimates (using sophisticated smoothing techniques) from smaller, cleaner datasets.

Another argument is made vividly by Banko and Brill (2001). They explore the performance of a number of machine learning algorithms (on a representative disambiguation task) as the size of the training corpus grows from a million to a billion words. All the algorithms steadily improve in performance, though the question "which is best?" gets different answers for different data sizes. The moral: performance improves with data size, and getting more data will make more difference than fine-tuning algorithms.

2.3 Giving and taking

Dragomir Radev made the useful distinction between NLP 'giving' and 'taking'.[4] NLP can give to the web technologies such as summarization (for web pages or web search results); machine translation; multilingual document retrieval; question-answering and other strategies for finding not only the right document but the right part of a document; and tagging, parsing and other core technologies (to improve indexing for search engines, the viability of this being a central Information Retrieval research question for the last twenty years). 'Taking' is, simply, using the web as a source of data for any CL or NLP goal, and is the theme of this Special Issue. If we focus too closely on the giving side of the equation, we look only at short-to-medium term goals. For the longer term, for 'giving' as well as for other purposes, a deeper understanding of the linguistic nature of the web and its potential for CL/NLP is required. For that, we must take the web itself, in whatever limited way, as an object of study.

Much web search engine technology has been developed with reference to language technology. The prototype for Altavista was developed in a joint project between Oxford University Press (exploring methods for corpus lexicography (Atkins, 1993)) and DEC (interested in fast access to very large databases). Language identification algorithms

[4] Remarks made in a panel discussion at the Empirical NLP Conference, Hong Kong, October 2002.

Sample phrase	BNC	WWW	WWW	WWW
	(100 M)	fall 1998	fall 2001	spring 2003
medical treatment	414	46,064	627,522	1,539,367
prostate cancer	39	40,772	518,393	1,478,366
deep breath	732	54,550	170,921	868,631
acrylic paint	30	7,208	43,181	151,525
perfect balance	38	9,735	35,494	355,538
electromagnetic radiation	39	17,297	69,286	258,186
powerful force	71	17,391	52,710	249,940
concrete pipe	10	3,360	21,477	43,267
upholstery fabric	6	3,157	8,019	82,633
vital organ	46	7,371	28,829	35,819

Table 1
Frequencies of English phrases in the BNC and on Altavista in 1998 and 2001, and on AlltheWeb in 2003. The counts for the BNC and Altavista are for individual occurrences of the phrase. The counts for AlltheWeb are page counts (the phrase may appear more than once on any page.)

(Beesley, 1988; Grefenstette, 1995), now widely used in web search engines, were developed as NLP technology. The Special Issue explores a ‘homecoming’ of web technologies, with the web now feeding one of the hands that fostered it.

3 Web size and the multilingual web

There were 56 million registered network addresses in July 1999, 125 million in January 2001, and 172 million in January 2003. A plot of this growth of the web in terms of computer hosts can easily be generated. Linguistic aspects take a little more work, and can only be estimated by sampling and extrapolation. Lawrence and Giles (1999) compared the overlap between page lists returned by different web browsers over the same set of queries and estimated that, in 1999, there were 800 million indexable web pages available. By sampling pages, and estimating an average page length of 7 to 8 kilobytes of non-markup text, they concluded that there might be 6 terabytes of text available then. In 2003, Google claims to search four times this number of web pages which raises the number of bytes of text available just through this one web server to over 20 terabytes from directly accessible web pages. At an average of ten bytes per word, a generous estimate for Latin-alphabet languages, that suggests two thousand billion words.

The web is clearly a multilingual corpus. How much of it is English? Xu (2000) estimated that 71% of the pages (453 million out of 634 million web pages indexed by the Excite engine at that time) were written in English, followed by Japanese (6.8%), German (5.1%), French (1.8%), Chinese (1.5%), Spanish (1.1%), Italian (0.9%), and Swedish (0.7%).

We have measured the counts of some English phrases according to various search engines over time and compared them with counts in the BNC, which we know has 100 million words. Table 1 shows these counts in the BNC, on Altavista in 1998 and in 2001, and then on Alltheweb in 2003. For example, the phrase *deep breath* appears 732 in the BNC. It was indexed 54,550 times by Altavista in 1998. This rose to 170,921 in 2001. And in 2003, we could find 868,631 web pages containing the contiguous words *deep breath* according to Alltheweb. The numbers found through the search engines are more than three orders of magnitude higher than the BNC counts, giving a first indication of the size of the English corpus available.

We can derive a more precise estimate of the number of words available through a

Word	Known-size-corpus relative frequency	Altavista frequency	Prediction for German-language web
oder	0.00561180	13,566,463	2,417,488,684
sind	0.00477555	11,944,284	2,501,132,644
auch	0.00581108	15,504,327	2,668,062,907
wird	0.00400690	11,286,438	2,816,750,605
nicht	0.00646585	18,294,174	2,829,353,294
eine	0.00691066	19,739,540	2,856,389,983
sich	0.00604594	17,547,518	2,902,363,900
ist	0.00886430	26,429,327	2,981,546,991
auf	0.00744444	24,852,802	3,338,438,082
und	0.02892370	101,250,806	3,500,617,348
Average			3,068,760,356

Table 2
German short words in the ECI corpus and via Altavista giving German web estimates

search engine by using the counts of function words as predictors of corpus size. Function words, such as *the*, *with*, *in*, etc., occur with a frequency that is relatively stable over many different types of texts. From a corpus of known size, we can calculate the frequency of the function words and extrapolate. In the 90-million word written-English component of the BNC *the* appears 5,776,487 times, around 7 times for every 100 words. In the American Declaration of Independence, *the* occurs 84 times. We predict that the Declaration is about $84 \times 100/7 = 1200$ words long. In fact, the text contains about 1500 words. Using the frequency of one word gives a first approximation. A better result can be obtained by using more data points.

From the first megabyte of the German text found in the European Corpus Initiative Multilingual Corpus,[5] we extracted frequencies for function words and other short, common words. We removed from the list words that were also common words in other languages[6]. Altavista provided on their results pages, along with a page count for a query, the number of times that each query word was found on the web.[7] Table 2 shows relative frequency of the words from our known corpus, the index frequencies that Altavista gave (February 2000) and the consequent estimates of the size of the German-language web indexed by Altavista.

We set aside words which give discrepant predictions –too high or too low– as (1) Altavista does not record in its index the language a word comes from, so the count for the string *die* includes both the German and English occurrences, and (2) a word might be under- or over-represented in the training corpus or the web (consider *here* which occurs very often in "click here".) Averaging the remaining predictions gives an estimate of 3 billion words of German that could be accessed through Altavista on that day in February 2000.

This technique has been tested on controlled data (Grefenstette and Nioche, 2000) in which corpora of different languages were mixed in various proportions, and gives reliable results. Table 3 gives estimates for the number of words that were available in thirty different Latin script languages through Altavista in March 2001. English led the pack with 76 billion words, and seven further languages already had over a billion.

[5] http://www.elsnet.org/resources/eciCorpus.html

[6] These lists of short words and frequencies were initially used to create a language identifier.

[7] Altavista have recently stopped providing information about how often individual words in a query have been indexed, and now only returns a page count for the entire query.

Language	Web size	Language	Web size
Albanian	10,332,000	Catalan	203,592,000
Breton	12,705,000	Slovakian	216,595,000
Welsh	14,993,000	Polish	322,283,000
Lithuanian	35,426,000	Finnish	326,379,000
Latvian	39,679,000	Danish	346,945,000
Icelandic	53,941,000	Hungarian	457,522,000
Basque	55,340,000	Czech	520,181,000
Latin	55,943,000	Norwegian	609,934,000
Esperanto	57,154,000	Swedish	1,003,075,000
Roumanian	86,392,000	Dutch	1,063,012,000
Irish	88,283,000	Portuguese	1,333,664,000
Estonian	98,066,000	Italian	1,845,026,000
Slovenian	119,153,000	Spanish	2,658,631,000
Croatian	136,073,000	French	3,836,874,000
Malay	157,241,000	German	7,035,850,000
Turkish	187,356,000	English	76,598,718,000

Table 3
Estimates of web size in words as indexed by Altavista for various languages

From the table, we see that even 'smaller' languages such as Slovenian, Croatian, Malay and Turkish have more than one hundred million words on the web. Much of the research that has been undertaken on the BNC simply exploits its scale and could be transferred directly to these languages.

The numbers are lower bounds for a number of reasons.

- Altavista only covers a fraction of the indexable web pages available. The fraction was estimated at just 15% by Lawrence and Giles (1999).
- Altavista may be biased to North American (mainly English language) pages by the strategy it uses to crawl the web
- Altavista only indexes pages that can be directly called by a URL, and does not index text found in databases that are accessible through dialog windows on web pages, the 'hidden web'. This is vast (consider MedLine,[8] just one such database with more than 5 billion words; see also Ipeirotis, Gravano, and Sahami (2001)) and this hidden web is not considered at all in this estimate.

Repeating the procedure after an interval, the second author and Nioche showed that the proportion of non-English text to English is growing. In October 1996 there were 38 German words for every 1000 words of English indexed by Altavista. In August 1999, there were 71 and in March 2001, 92.

3.1 Finding the right translation

How can these great numbers be used for other language processing tasks? Consider the compositional French noun phrase *groupe de travail.* In the MEMODATA bilingual dictionary[9] the French word *groupe* is translated by the English words *cluster, group, grouping, concern* and *collective.* The French word *travail* translates as *work, labor* or

[8] http://www4.ncbi.nlm.nih.gov/PubMed/

[9] See http://www.elda.fr/cata/text/M0001.html. The basic multilingual lexicon produced by MEMODATA contains 30,000 entries for five languages: French, English, Italian, German, Spanish.

labor cluster	21	labour collective	428
labor grouping	28	work collective	759
labour concern	45	work concern	772
labor concern	77	labor group	3,977
work grouping	124	labour group	10,389
work cluster	279	work group	148,331
labor collective	423		

Table 4
Altavista frequencies for candidate translations of *groupe de travail*

labour. Many web search engines allow the user to search for adjacent phrases. Combining the possible translations of *groupe de travail* and submitting them to Altavista in early 2003 gave the counts in Table 4. The phrase *work group* is 15 times more frequent than any other, and is also the best translation among the tested possibilities. A set of controlled experiments of this form are described in Grefenstette (1999). A good translation was found in 87% of ambiguous cases from German to English and 86% of ambiguous cases from Spanish to English.

4 Representativeness

We know the web is big, but a common response to a plan to use the web as a corpus is "but it's not representative".

There are a great many things to be said about this. It opens up a pressing yet almost untouched practical and theoretical issue for computational linguistics and language technology.

4.1 Theory

First, 'representativeness' begs the question, 'representative of what?' Outside very narrow, specialized domains, we do not know with any precision what existing corpora might be representative of. If we wish to develop a corpus of general English, we may think it should be representative of general English, so we then need to define the population of 'general English language events' of which the corpus will be a sample. Consider the following issues.

- production and reception: is a language event an event of speaking or writing, or one of reading or hearing? Standard conversations have, for each utterance, one speaker and one hearer. A Times newspaper article has (roughly) one writer and several hundred thousand readers.

- speech and text: do speech events and written events have the same status? It seems likely that there are orders of magnitude more speech events than writing events, yet most corpus research to date has tended to focus on the more tractable task of gathering and working with text.

- background language: do muttering under one's breath or talking in one's sleep constitute speech events, and does doodling with words constitute a writing event? Or, on the reception side, does passing (and possibly subliminally reading) a roadside advertisement constitute a reading event? And what of having the radio on but not attending to it, or the conversational murmur in a restaurant?

- copying: if *I'd like to teach the world to sing*, and, like Michael Jackson or the Spice Girls, am fairly successful in this goal and they all sing my song, then does each individual singing constitute a distinct language production event?

- In the text domain, organizations such as Reuters produce news feeds which are typically adapted to the style of a particular newspaper and then re-published: is each re-publication a new writing event? (These issues, and related themes of cut-and-paste authorship, ownership and plagiarism, are explored in Wilks (2003, submitted).)

4.2 Technology

Application developers urgently need to know about what to do about sublanguages. It has often been argued that, within a sublanguage, few words are ambiguous and a limited repertoire of grammatical structures is used (Kittredge and Lehrberger, 1982). This points to sublanguage-specific application development being substantially simpler than general-language application development. However, many of the resources that developers may wish to use are general-language resources, such as, for English, WordNet, ANLT, XTag, COMLEX and the BNC. Are they relevant? Can they be used? Is it better to use a language model based on a large general-language corpus, or a relatively tiny corpus of the right kind of text? Nobody knows. There is currently no theory, no mathematical models and almost no discussion.

A related issue is that of porting an application from the sublanguage for which it was developed, to another. It should be possible to use corpora for the two sublanguages to estimate how large a task this will be, but again, our understanding is in its infancy.

4.3 Language modeling

Much work in recent years has gone into developing language models. Clearly, the statistics for different types of text will be different (Biber, 1993). This imposes a limitation on the applicability of any language model: we can only be confident that it predicts the behavior of language samples of the same text type as the training-data text type (and we can only be entirely confident if training and test samples are random samples from the same source).

When a language technology application is put to use, it will be applied to new text for which we cannot guarantee the text type characteristics. There is little work on assessing how well one language model fares, when applied to a text type which is not that of the training corpus. Two studies are Sekine (1997) and Gildea (2001), both of which show substantial variation in performance when the training corpus changes. The lack of theory of text types leaves us without a way of assessing the usefulness of language modeling work.

4.4 Language errors

Web texts are produced by a wide variety of authors. Contrary to paper-based, copy-edited published texts, web-based texts may be produced cheaply and rapidly with little concern for correctness. On Google a search for “I beleave” has 3,910 hits, and “I beleive”, 70,900 pages. The correct “I believe” appears on over 4 million pages. Table 5 present what is regarded as a common grammatical error in Spanish, comparing the frequency of such forms to the accepted forms on the web. All the “erroneous” forms exist, but much less often than the “correct” forms. The web is a dirty corpus, but expected usage is much more frequent than what might considered as noise.

pienso de que	388
pienso que	356,874
piensas de que	173
piensas que	84,896
piense de que	92
piense que	67,243
pensar de que	1,640
pensar que	661,883

Table 5
Hits for Spanish *pensar que* with and without possible 'dequeismos errors' (spurious *de* between the verb and the relative), from Alltheweb.com, March 2003. Not all items are errors, e.g. "...pienso de que manera..." *...think how...* . The correct form is always at least 500 times more common than any potentially incorrect form.

4.5 Sublanguages and general-language-corpus composition

A language can be seen as a modest core of lexis, grammar and constructions, plus a wide array of different sublanguages, as used in each of a myriad of human activities. This presents a challenge to general-language resource developers: should sublanguages be included? The three possible positions are:

- no, none should
- some, but not all, should
- yes, all should.

The problem with the first position is that, with all sublanguages removed, the residual core gives an impoverished view of language (quite apart from demarcation issues, and the problem of determining what is left). The problem with the second is that it is arbitrary. The BNC happens to include cake recipes and research papers on gastro-uterine diseases, but not car manuals or astronomy texts. The third has not, until recently, been a viable option.

4.6 Literature

To date, corpus developers have been obliged to take pragmatic decisions about the sorts of text to go into a corpus. Atkins, Clear, and Ostler (1992) describe the desiderata and criteria used for the BNC, and this stands as a good model for a general-purpose, general language corpus. The word 'representative' has tended to fall out of discussions to be replaced by the meeker 'balanced'.

The recent history of mathematically sophisticated modeling of language variation begins with Biber (1988), who identifies and quantifies the linguistic features associated with different spoken and written text types. Habert and colleagues (Folch et al., 2000; Beaudouin et al., 2001) have been developing a workstation for specifying subcorpora according to text type, using Biber-style analyses amongst others. In Kilgarriff (2001) we present a first pass at quantifying similarity between corpora and Cavaglia (2002) continues this line of work. As mentioned above, Sekine (1997) and Gildea (2001) are two papers which directly address the relation between NLP systems and text type; one further such item is (Roland et al., 2000). Buitelaar and Sacaleanu (2001) explores the relation between domain and sense disambiguation.

A practical discussion of a central technical concern is Vossen (2001), who tailors a general-language resource for a domain.

Baayen (2001) presents sophisticated mathematical models for word frequency distributions and it is likely that his mixture models have potential for modeling sublanguage mixtures. His models have been developed with a specific, descriptive goal in mind and using a small number of short texts: it is unclear whether they can be usefully applied in NLP.

While the extensive literature on text classification (Manning and Schütze, 1999, pp 575-608) is certainly relevant, it most often starts from a given set of categories and cannot readily applied to the situation where the categories are not known in advance. Also, the focus is usually on content words and topics or domains, with other differences of genre or sublanguage not being examined. Exceptions focusing on genre include Kessler, Nunberg, and Schütze (1997) and Karlgren and Cutting (1994).

4.7 Representativeness: conclusion

The web is not representative of anything else. But nor are other corpora, in any well-understood sense. Picking away at the question merely exposes how primitive our understanding of the topic is, and leads inexorably to larger and altogether more interesting questions about the nature of language, and how it might be modeled.

'Text type' is an area where our understanding is, as yet, very limited. While further work is required irrespective of the web, the use of the web forces the issue. Where researchers use established corpora, such as Brown, the BNC or the Penn Treebank, researchers and readers are willing to accept the corpus name as a label for the type of text occurring in it without asking critical questions. Once we move to the web as a source of data, and our corpora have names like "April03-sample77" the issue of how the text type(s) can be characterized demands attention.

5 Introduction to Papers in this Special Issue

One use of a corpus is to extract a language model: a list of weighted words, or combinations of words that describe (i) how words are related, (ii) how they are used with each other, and (iii) how common they are in a given domain. In speech processing, language models are used to predict which word combinations are likely interpretations of a sound stream; in Information Retrieval to decide which words are useful indicators of a topic; and in Machine Translation, to identify good translation candidates.

In this volume, Celina Santamaría, Julio Gonzalo and Felisa Verdejo describe how to build sense-tagged corpora from the web by associating word meanings with web page directory nodes. The Open Directory Project (at dmoz.org) is a collaborative, volunteer project for classifying web pages into a taxonomic hierarchy. Santamaría et al. present an algorithm for attaching WordNet word senses to nodes in this same taxonomy, thus providing automatically created links between word senses and web pages. They also show how this method can be used for automatic acquisition of sense-tagged corpora, from which one could, among other things, produce language models tied to certain senses of words, or for a certain domain.

Unseen words, or word sequences –that is, words or sequences not occurring in training data– are a problem for language models. If the corpus from which the model is extracted is too small, there are many such sequences.

Taking the second author's work, as described above, as a starting point, Frank Keller and Maria Lapata examine how useful the web is as a source of frequency information for rare items: specifically, for dependency relations involving two English words such as <*fulfil* OBJECT *obligation*>. They generate pairs of common words, constructing combinations that are and are not attested in the BNC. They then compare the frequency of these combinations in a larger 325 million word corpus and on the web. They find that web frequency counts are consistent with other large corpora. They also report on a series of human-subject experiments, in which they establish that web statistics are good at predicting the intuitive plausibility of predicate-argument pairs. Other experiments show that web counts correlate reliably with counts recreated using class-based smoothing and overcome some problems of data sparseness in the BNC.

Other very large corpora are available for English, English is an exception, and the other three papers all exploit the multilinguality of the web. Andy Way and Nano Gough show how it can provide data for an Example-Based Machine Translation (Nagao, 1984) system. First, they extract 200,000 phrases from a parsed corpus. These phrases are sent to three online translation systems. Both original phrases and translations are chunked. From these pairings a set of chunk translations is extracted to be applied in a piecewise fashion to new input text. The authors use the web again at a final stage to re-rank possible translations by verifying which subsequences among the possible translations are most attested.

The two remaining papers present methods for building aligned bilingual corpora from the web. It seems plausible that this automatic construction of translation dictionaries can palliate the lack of translation resources for many language pairs. Philip Resnik was the first to recognize that it is possible to build large parallel bilingual corpora from the web. He found that one can exploit the appearance of language flags and other clues which often lead to a version of the same page in a different language[10]. Here, in this volume, Resnik and Noah Smith present their STRAND system for building bilingual corpora froms the web.

An alternative method is presented by Wessel Kraaij, Jian-Yun Nie and Michel Simard. They use the resulting parallel corpora to induce a probabilistic translation dictionary which is then embedded into a Cross Language Information Retrieval system. Various alternative embeddings are evaluated using the CLEF (Peters, 2001) multilingual information retrieval testbeds.

6 Prospects

The default means of access to the web is through a search engine such as Google. While the web search engines are dazzlingly efficient pieces of technology and excellent at the task they set themselves, for the linguist they are frustrating:

- The search engine results do not present enough instances (1000 or 5000 maximum)
- They do not present enough context for each instance (Google provides a fragment of around ten words)
- They are selected according to criteria which are, from a linguistic perspective, distorting (with uses of the search term in titles and headings going to the top of the list, and often occupying all the top slots)
- They do not allow searches to be specified according to linguistic criteria such as the citation form for a word, or word class
- The statistics are unreliable, with frequencies given for “pages containing x” varying according to search engine load and many other factors.

[10] For example, one can find Azerbaijan news feeds online at http://www.525ci.com in Azeri (written with a Turkish codeset), and on the same page are pointers to versions of the same stories in English and in Russian.

If only these constraints were removed, a search engine would be a wonderful tool for language researchers. Each of them could straightforwardly be resolved by search engine designers, but linguists are not a powerful lobby and search engine company priorities will never perfectly match our community's. This suggests a better solution: do it ourselves. Then the kinds of processing and querying would be designed explicitly to meet linguists' desiderata, without any conflict of interest or 'poor relation' role. A large numbers of possibilities open out. All those processes of linguistic enrichment which have been applied with impressive effect to smaller corpora could be applied to the web. We could parse the web. Web searches could be specified in terms of lemmas, constituents (e.g. noun phrase) and grammatical relations rather than strings. The way would be open for further anatomizing of web text types and domains. Thesauruses and lexicons could be developed directly from the web. And all for a multiplicity of languages.[11]

The web contains enormous quantities of text, in lots of languages and language types, on a vast array of topics. Our take on the web is that it is a fabulous linguists' playground. We hope the Special Issue will encourage you to come on out and play!

References

Agirre, Eneko, Olatz Ansa, Eduard Hovy, and David Martnez. 2000. Enriching very large ontologies using the WWW. In *Proceeding of the Ontology Learning Workshop of the European Conference of AI (ECAI)*, Berlin, Germany.

Atkins, Sue. 1993. Tools for computer-aided corpus lexicography: the Hector project. *Acta Linguistica Hungarica*, 41:5–72.

Atkins, Sue, Jeremy Clear, and Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing*, 7(1):1–16.

Baayen, Harald. 2001. *Word Frequency Distributions.* Kluwer, Dordrecht.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *COLING-ACL*, pages 86–90, Montreal, August.

Banko, Michele and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Toulouse.

Beaudouin, Valérie, Serge Fleury, Benoît Habert, Gabriel Illouz, Christian Licoppe, and Marie Pasquier. 2001. Typweb : décrire la toile pour mieux comprendre les parcours. In *CIUST'01, Colloque International sur les Usages et les Services des Télécommunications*, Paris, June. http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/typweb.htm.

Beesley, Kenneth R. 1988. Language identifier: A computer program for automatic natural-language identification of on-line text. In *Language at Crossroads: Proceedings of the 29th Annual Conference of the American Translators Association*, pages 47–54, Oct 12–16.

Biber, Douglas. 1988. *Variation across speech and writing.* Cambridge University Press.

Biber, Douglas. 1993. Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2):219–242.

Briscoe, Ted and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proc. Fifth Conference on Applied Natural Language Processing*, pages 356–363, Washington DC, April.

Buitelaar, Paul and Bogdan Sacaleanu. 2001. Ranking and selecting synsets by domain relevance. In *Proc. Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, NAACL*, Pittsburgh, June.

Burnard, Lou, 1995. *The BNC Reference Manual.* Oxford University Computing Service.

Cavaglia, Gabriela. 2002. Measuring corpus homogeneity using a range of measures for inter -document distance. In *Third International Conference on Language Resources and Eva luation*, pages 426–431, Las Palmas de Gran Canaria, Spain, May.

[11] The idea is developed further in Grefenstette (2001) and in Kilgarriff (2003).

Church, Kenneth W. and Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

Dumais, Susan, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. 2002. Web question answering: is more always better? In *Proc. 25th ACM SIGIR*, pages 291–298, Tampere, Finland.

Fletcher, William. 2002. Facilitating compilation and dissemination of ad-hoc web corpora. In *Teaching and Language Corpora 2002*. http://miniappolis.com/KWiCFinder/KWiCFinder.html.

Folch, Helka, Serge Heiden, Benôit Habert, Serge Fleury, Gabriel Illouz, Pierre Lafon, Julien Nioche, and Sophie Prévost. 2000. Typtex: Inductive typological text classification by multivariate statistical analysis for nlp systems tuning/evaluation. In *Second Language Resources and Evaluation Conference*, pages 141–148, Athens, Greece, May–June.

Fujii, Atsushi and Tetsuya Ishikawa. 2000. Utilizing the world wide web as an encyclopedia: Extracting term descriptions from semi-structured text. In *Proc. 38th Meeting of the ACL*, pages 488–495, Hong Kong, October.

Gildea, Daniel. 2001. Corpus variation and parser performance. In *Proc. Empirical Methods in NLP*, Pittsburgh.

Greenwood, Mark, Ian Roberts, and Robert Gaizauskas. 2002. University of sheffield trec 2002 q & a system. In E. M. Voorhees and Lori P. Buckland, editors, *The Eleventh Text REtrieval Conference (TREC-11)*, Washington. U.S. Government Printing Office. NIST Special Publication 500–XXX.

Grefenstette, Gregory. 1995. Comparing two language identification schemes. In *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data, JADT'95*, pages 263–268, Rome, Dec 11–13. www.xrce.xerox.com/competencies/content-analysis/publications/Documents/P49030/content/gg_aslib.pdf.

Grefenstette, Gregory. 1999. The www as a resource for example-based mt tasks. Invited Talk, ASLIB 'Translating and the Computer' conference, London, October.

Grefenstette, Gregory. 2001. Very large lexicons. In Walter Daelemans, Khalil Simaan, Jakub Zavrel, and Jorn Veenstra, editors, *Computational Linguistics in the Netherlands 2000. Selected Papers from the Eleventh CLIN Meeting*, Language and Computers 37, Amsterdam. Rodopi.

Grefenstette, Gregory and Julien Nioche. 2000. Estimation of english and non-english language use on the www. In *Proc. RIAO (Recherche d'Informations Assistée par Ordinateur)*, pages 237–246, Paris.

Hawking, D., E. Voorhees, N. Craswell, and P. Bailey. 1999. Overview of the TREC8 web track.

Ipeirotis, Panagiotis G., Luis Gravano, and Mehran Sahami. 2001. Probe, count, and classify: Categorizing hidden web databases. In *SIGMOD Conference*.

Jones, Rosie and Rayid Ghani. 2000. Automatically building a corpus for a minority language from the web. In *Proceedings of the Student Workshop of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 29–36.

Karlgren, Jussi and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proc. COLING-94*, pages 1071–1075, Kyoto, Japan.

Kessler, Brett, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proc. ACL and EACL*, pages 39–47, Madrid.

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

Kilgarriff, Adam. 2003. Linguistic search engine. In Kiril Simov, editor, *Shallow Processing og Large Corpora: Workshop held in association with Corpus Linguistics 2003*, Lancaster, March.

Kilgarriff, Adam and Michael Rundell. 2002. Lexical profiling software and its lexicographical applications - a case study. In *EURALEX 02*, Copenhagen, August.

Kittredge, Richard and John Lehrberger. 1982. *Sublanguage: studies of language in restricted semantic domains*. De Gruyter, Berlin.

Korhonen, Anna. 2000. Using semantically motivated estimates to help subcategorization acquisition. In *Proc. Joint Conf. on Empirical Methods in NLP and Very Large Corpora*, pages 216–223, Hong Kong, October. ACL SIGDAT.

Lawrence, Steve and C. Lee Giles. 1999. Accessibility of information on the web. *Nature*, 400:107–109.

Manning, Christopher and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

McEnery, Tony and Andrew Wilson. 1996. *Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Mihalcea, Rada and Dan Moldovan. 1999. A method for word sense disambiguation of unrestricted text. In *Proc. 37th Meeting of ACL*, pages 152–158, Maryland, June.

Nagao, Makoto. 1984. A framework of a mechanical translation between japanese and english by analogy principle. In Alick Elithorn and Ranan Banerji, editors, *Artificial and Human Intelligence*. North-Holland, Edinburgh, pages 173–180.

Peters, Carol, editor. 2001. *Cross-Language Information Retrieval and Evaluation, Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Revised Papers*, Lecture Notes in Computer Science. Springer.

Resnik, Philip. 1999. Mining the web for bilingual text. In *Proc. 37th Meeting of ACL*, pages 527–534, Maryland, June.

Rigau, German, Bernardo Magnini, Eneko Agirre, and John Carroll. 2002. Meaning: A roadmap to knowledge technologies. In *Proceedings of COLING Workshop on A Roadmap for Computational Linguistics*, Taipei, Taiwan.

Roland, Douglas, Daniel Jurafsky, Lise Menn, Susanne Gahl, Elizabeth Elder, and Chris Riddoch. 2000. Verb subcategorization frequency differences between business-news and balanced corpora: the role of verb sense. In *Proc. Workshop on Comparing Corpora, 38th ACL*, Hong Kong, October.

Sekine, Satshi. 1997. The domain dependence of parsing. In *Proc. Fifth Conference on Applied Natural Language Processing*, pages 96–102, Washington DC, April. ACL.

Sinclair, John M., editor. 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Collins, London.

Varantola, Krista. 2000. Translators and disposable corpora. In *Proc. CULT (Corpus Use and Learning to Translate)*, Bertinoro, Italy, November.

Villasenor-Pineda, L., M. Montes y Gómez, M. Pérez-Coutino, and D. Vaufreydaz. 2003. A corpus balancing method for language model construction. In *Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003)*, pages 393–401, Mexico City, February.

Volk, Martin. 2001. Exploiting the www as a corpus to resolve pp attachment ambiguities. In *Proc. Corpus Linguistics 2001*, Lancaster, UK.

Vossen, Piek. 2001. Extending, trimming and fusing wordnet for technical documents. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, June. http://www.seas.smu.edu/ rada/mwnw/papers/WNW-NAACL-105.pdf.

Wilks, Yorick. 2003, submitted. On the ownership of text. *Computers and the Humanities*.

Xu, J. L. 2000. Multilingual search on the world wide web. In *Proc. Hawaii International Conference on System Science HICSS-33*, Maui, Hawaii, January.

Zheng, Zhiping. 2002. AnswerBus question answering system. In E. M. Voorhees and Lori P. Buckland, editors, *Proceeding of HLT Human Language Technology Conference (HLT 2002)*, San Diego, CA, March 24–27.