Corpora from the Web

Gabriela Cavaglià and Adam Kilgarriff
ITRI, University of Brighton
Email: {Gabriela.Cavaglia, Adam.Kilgarriff}@itri.brighton.ac.uk

December 19, 2000

Abstract

We investigate the potential of using the web as a huge corpus for language studies. We test the hypothesis that corpora produced by gathering web pages found when searching on a technical term are homogeneous whereas corpora produced using a general term are heterogeneous.

1 Introduction

Over the last couple of years the Web has become the corpus resource of choice for many language studies (e.g., (**?**; **?**)). It contains a huge quantity of text, of any number of varieties, for a very wide range of languages (**?**). Moreover text is available immediately, for free, and can be downloaded without concern for copyright.

Many web documents do not contain only text or do not contain text at all. Many are duplicates. Some others point at documents which do not exist any more. All these characteristics can sound like good reasons for not using the web as a linguistic resource. However they are all obstacles that can be overcome by judicious, linguistically-aware filtering of downloaded material. The prize of vast, accessible corpora will make effort spent on developing filters a good investment.

In this paper we propose a method for creating corpora built from the web (made up of web documents) using a search engine, a simple query and a downloading program. We then compare the homogeneity of corpora generated using different search terms.

When we use a search engine (Altavista, Yahoo, Google, ...) to search on a technical term, we would expect the texts that the search engine finds to be fairly homogeneous. For example, we would expect the result for the technical term as *vowel shift* to be predominantly articles and course notes concerning phonology. Instead if we use a general term - an arbitrary adjective-noun combination such as *suitable place* - we would expect the corpus generated to be much more heterogeneous, as the term will occur in texts of a very wide range of varieties. In this paper we describe an experiment designed to test the hypothesis that technical terms give rise to more homogeneous corpora than general terms.

2 Method

First, 91 technical terms and 59 general terms were identified. The technical terms were drawn from the index of a phonology book, the general terms from a novel. For each, an Altavista search was instigated. The Altavista search results were parsed and the URLs of pages containing the search term were identified. Altavista gave at least 200 hits for 43 of the technical terms and 47 of the general terms. Downloads of the first 200 pages for each of these were attempted, with most but not all pages successfully retrieved. This gave 43 technical-term corpora and 47 general-term corpora for further experimentation.

The next question was, which of these pages comprised text of the relevant human language? (in these experiments, English). To answer this, we need a working defintion of what is to count as language. At the one extreme, everything included in the Brown corpus clearly is language. At the other, images clearly are not. Problematic points in between include:

- timetables
- computer language code
- indexes
- bibiliographies
- headings

There is also HTML markup. This was clearly not English, and it was our intention from the outset to exclude the markup from the computations used to determine language-like-ness or document similarity. We would like to use our knowledge of the semantics of HTML tags to guide the identification of language segments within documents, noting, for example, that data in HTML tables is less likely to be linguistic than data in HTML paragraphs, but this is future work.

Our strategy throughout was to err towards strictness, and to aim to throw out items if we were not confident they were good instances of language, on the basis that the web is huge and one can always get more data, so there is no need to use data one is not sure of.

Where pages do not contain many words, it is harder to determine algorithmically whether they are good instances of language. We rejected all pages containing less than 2000 non-markup words.

2.1 A unigram model of language-like-ness

We used a simple unigram model to identify whether a web page was predominantly linguistic and in English.[1] We took relative frequencies of twenty high-frequency

[1] **?**) faced similar problems of identifying fragments which were Japanese text, and used a trigram language model to filter out text fragments with perplexity above a threshold. We expect both techniques would lead to similar results.

words from a reference corpus, the British National Corpus[2] (BNC). For each page, we compared the frequencies of these twenty words with their BNC frequencies, and rejected the page if the score was above a threshold. The score for document d was computed as

$$Score(d) = \sum_{i=1}^{i=20} \frac{(BNC_i - doc_{d,i})^2}{BNC_i}$$

where BNC_i is the relative frequency of the ith word in the BNC, and $doc_{d,i}$ is its relative frequency in document d.

The words used for the comparison were ideally words which were both very frequent and had a relatively stable frequency across genres. We used a BNC frequency list indicating, for each word, the variance in its frequency across a set of same-length samples from the˜Ad BNC.[3] We took the first twenty items in the list for which the variance was less than ten times the mean frequency per sample. These were:

the of and to a in it for be with on that by at not this but they from which

Items excluded because the variance was too great were:

is was I you he are his had she

Following some experimentation, the threshold was set at 0.1. Documents d for which $Score(d) > 0.1$ were rejected.

Once short documents and high-scoring documents had been filtered out, for some of the original search terms, there were less than twenty web pages remaining. In these cases, the entire corpus was set aside. After this process, there were corpora remaining for just seventy-four of the search terms (40 technical, 34 general). These terms and the associated corpora were the ones used used for the homogeneity experiment.

2.2 A unigram model of homogeneity

Following **?**), we expect simple word frequencies, or unigrams, to provide ample information to assess the homogeneity of a corpus. Grammatical features (as used in **?**) amongst others), type-token ratios and average sentence length are amongst the features that other researchers have focussed on, and we are happy to accept that no individual word-frequency may be as useful a feature for discriminating between text types as, for example, the type-token ratio. However, there is, we believe, more discriminating information in the frequencies of a large number of the highest-frequency words than in any short list of hand-selected features. Word frequencies have the advantage that they are easy to count, and the counts are (almost) theory-free.

For each corpus, a measure of homogeneity was calculated as follows.

First, the word frequencies for the entire corpus were calculated; this served to identify the n most frequent words, and their relative frequencies, in the corpus as a whole.

[2] `http://info.ox.ac.uk/bnc`

[3] For details and list, see `http://ftp.itri.bton.ac.uk/Ãdam.Kilgarriff/bnc-readme.html#variances`

Then, a document's similarity to the norm for the corpus was calculated as follows:

$$SimToNorm(d) = \sum_{i=1}^{i=n} \frac{(corp_i - doc_{d,i})^2}{corp_i}$$

where $corp_i$ is the relative frequency of the ith word in the corpus as a whole, and $doc_{d,i}$ is its relative frequency in the first m words of document d. We experimented with values of n —the number of high-frequency words used as features— of 100, 200 and 500, and with values of m —the size of each document— of 1000, 2000 and 5000.

Then, the corpus homogeneity was defined as (1) the mean or (2) the median of the SimToNorm scores for its component documents.

The homogeneity figures for the general-term corpora and the technical-term corpora were then compared.

3 Results

In Table 1 we present the results for 74 search terms and corpora, ordered according to homogeneity.

These results were produced using the first 2000 words of each document, with the 100 highest-frequency words as features, and using the mean (rather than the mode) as the measure of homogeneity. However we note that other permutations produced very similar results.

On each occasion, counter to our hypothesis, the results suggested that the general-term corpora were slightly more homogeneous than our technical-term corpora. Significance was examined using the Wilcoxon ranks test (also known as the Mann-Whitney U test). For some permutations, the difference was significant at the 95% or 97.5% level, for others, it was not significant at either level. For the instance shown in the table, the sum of ranks for the smaller set of general-term corpora is 1092, which gives a z-score is 1.98, so is (just) significant at a 95% level, using a two-tailed test.[4]

4 Discussion and future work

Counter to our expectations, the experiment does not confirm the hypothesis. We manually investigated some of the corpora and found high degrees of heterogeneity in the web pages wherever we looked. Amongst the web pages for *word formation*, a technical term with a low score,[5] were a philosophical article on Gadamer, a bible studies

[4] Where there are more than 20 instances in each of the two datasets, it is appropriate to use a normal approximation to the distribution of Mann-Whitney U. This is calculated as

$$z = \frac{2R - N_1(N+1)}{\sqrt{N_1 N_2 (N+1)/3}}$$

where R is the sum of ranks for the category with fewer items, N_1 is the number of items in the category with fewer items, N_2 is the number of items in the category with more items, and N is the total number of items.

[5] A low score indicates homogeneity, a high score, heterogeneity.

g/t	Search term	Homog	**g/t**	Search term	Homog
g	rear door	0.55	t	vowel shift	1.01
g	piercing pain	0.59	t	modal verb	1.04
t	native speaker	0.60	t	rising tone	1.04
g	soft soil	0.61	t	vowel reduction	1.07
g	dead leaves	0.61	t	pitch accent	1.08
g	medical attention	0.62	t	initial consonant	1.09
g	military parade	0.62	g	imminent arrival	1.09
g	personal effects	0.64	t	glottal stop	1.09
g	personal identification	0.65	t	surface representation	1.10
t	word formation	0.65	t	loan words	1.11
g	complete record	0.65	g	important connection	1.12
g	little patience	0.68	g	leather suitcase	1.13
t	generative phonology	0.69	t	subordinate clause	1.16
g	elevator entrance	0.72	g	strong defence	1.16
g	parked car	0.74	g	tv image	1.17
t	connected speech	0.74	t	parallel distribution	1.18
g	television program	0.74	t	regional variation	1.19
t	compensatory changes	0.74	t	local peak	1.19
g	upstairs window	0.75	t	minimal pair	1.19
t	error analysis	0.75	t	local accent	1.19
t	vocal cords	0.76	t	double articulation	1.19
t	full forms	0.76	t	immature forms	1.21
g	long corridor	0.79	t	complementary distribution	1.21
g	congested streets	0.81	t	continuous variation	1.21
t	tone language	0.81	t	stressed syllable	1.22
g	green hose	0.82	g	strange accent	1.22
g	keen interest	0.83	g	suitable place	1.22
t	citation form	0.84	t	free variation	1.26
t	tone group	0.87	t	qualitative assessment	1.33
t	formal style	0.88	g	excellent metaphor	1.33
g	basic freedom	0.90	t	consonant cluster	1.37
t	reduced forms	0.92	t	phonological process	1.45
g	excellent judgment	0.93	g	former existence	1.46
t	redundant features	0.93	g	bank officials	1.48
g	dark liquid	0.95	t	substantive evidence	1.53
t	sound change	0.95	g	excessive taxation	1.55
g	narrow space	0.97	g	essential properties	1.70

Table 1: Homogeneity scores for 74 downloaded corpora. **g** indicates the search term was ‘general’, **t** that it was technical.

piece and a page on programming alongside assorted pages on language study but these pages also formed a mixed bag, with some describing courses, one being the home page of the Lithuanian Language Club, and one a FAQ for an Estonian culture newsgroup. *Military parade*, a low scoring general term, included some tourists' pages where the military parade was an event to see, documents about films and archeological projects, and news pages. *Upstairs window* another low-scoring general term, seemed to get its low score because most of the pages were narrative; diaries, novels, stories, legends, though some were about housing (renting or maintaining), and a couple about cinema.

For technical and general terms alike, the web contains all sorts of documents, of all sorts of genres, and it is not evident whether the tendency of technical terms to occur more narrowly will ever show through the tendency, on the web, for all sorts of terms to appear in all sorts of genres.

Thus our qualitative investigation of the corpora leads us to suspect the underlying intuition will only be valid in rather more constrained ways that we had originally imagined.

There are many other considerations which may have contributed to the hypothesis not beign confirmed, and over the coming months they will be explored in greater detail. They include the following.

- Altavista's ranking of search hits returns an ordered list of search hits, with items where the search terms was highly salient at the top. Where there were many more than 200 hits, our downloading strategy retrieved only high-salience items, which will not be representative of documents containing the search term in general.

- Web pages frequently contain a mix of text and other material; to get a good text corpus, an analysis at a finer-grained level than the complete web page is required. HTML markup should be used to guide decisions about which parts of web pages are text and which are not.

- The sample sizes (eg, 2000 words or 5000 words from the beginning of the document) were too short.

- The arbitrary cutting-off of documents after 2000 or 5000 words undermined the integrity of the documents, thereby increasing the level of noise until it obscured the lingusitic effect.

- First, a classification according to genre must be undertaken; the hypthesis will only be confirmed within a genre.

- Word frequencies are not appropriate features.

- Not enough word frequency figures were used.

- The formulae used for calculating homogeneity (and language-like-ness) were inappropriate.

This was an early foray into the relation between the linguistic characterisation of corpora, and datasets obtainable by downloading. Some of the possibilities and difficulties of the interaction have been sketched, and we shall continue exploring further methods so that, in due course, language corpora of all kinds, —homogeneous, heterogenerous, of specified language varieties— can be specified according to the interests of the researcher and downloaded from the web.

References

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

Fujii, Atsushi and Tetsuya Ishikawa. 2000. Utilizing the world wide web as an encyclopedia: Extracting term descriptions from semi-structured texts. In *38th Annual Meeting of the Asociation for Computational Linguistics*, pages 488–495, Hong Kong. ACL'00.

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)*, Paris.

Kilgarriff, Adam. 2000. Comparing corpora. *International Journal of Corpus Linguistics*, in press.

Resnik, Philip. 1999. Mining the web for bilingual text. In *37th Annual Meeting of the Asociation for Computational Linguistics*, pages 527–534, Colledge Park, Maryland, USA. ACL'99.