A Corpus Factory for many languages

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, Avinesh PVS

Lexical Computing Ltd. IIIT Hyderabad, Masaryk University, IIIT Hyderabad
United Kingdom, India, Czech Republic, India
adam@lexmasterclass.com, gvsreddy@students.iiit.ac.in, xpomikal@fi.muni.cz, avinesh@students.iiit.ac.in

Abstract

For many languages there are no large, general-language corpora available. Until the web, all but the richest institutions could do little but shake their heads in dismay as corpus-building was long, slow and expensive. But with the advent of the Web it can be highly automated and thereby fast and inexpensive. We have developed a `corpus factory' where we build large corpora. In this paper we describe the method we use, and how it has worked, and how various problems were solved, for eight languages: Dutch, Hindi, Indonesian, Norwegian, Swedish, Telugu, Thai and Vietnamese. The corpora we have developed are available for use in the Sketch Engine corpus query tool.

1. Introduction

For the major world languages, large corpora are publicly available. But for most other languages, they are not. In this paper, we present a procedure to build large corpora for many languages. (By `large', we mean at least 50m words.)

Corpus collection used to be long, slow and expensive - but then came the internet: texts, in vast number, are now available by mouse-click. The prospects of web as corpus were first explored in the late 1990s by Resnik (1999) and early 2000s by Jones and Ghani (2000). Grefenstette and Nioche (2000) showed just how much data was available. Keller and Lapata (2003) established the validity of web corpora by comparing models of human response times for collocations drawn from web frequencies with models drawn from traditional-corpus frequencies, and showing that they compared well.

Sharoff (2006) has prepared web corpora, typically of around 100 million words, for ten major world languages, primarily for use in teaching translation. Scannell (2007) has gathered corpora of, in most cases less than a million words for several hundred languages.

Here we aim to collect large corpora for many languages. Our goal is to make the task of corpora collection easy with minimal or no human intervention. In this paper, we will describe how we identify and remove bottlenecks at each step. To date we have applied the method to eight languages: Dutch, Hindi, Indonesian, Norwegian, Swedish, Telugu, Thai and Vietnamese.

2. Method

Our method is as used by Sharoff (2006) and similar to Baroni and Kilgarriff (2006), Ferraresi et al. (2008). Like BootCaT, (Baroni and Bernardini, 2004) it piggybacks on the work of the commercial search engines. Search engines crawl and index the Web, identify text-rich pages and address character-encoding issues (though they do this with mixed success, as we see below). By using this work already done, usually very well, by the search engines, we save ourselves many tasks.

Steps involved in corpora collection are

1. Gather a `seed word' list of several hundred mid-frequency words of the language
2. Repeat several thousand times (until the corpus is large enough):
   - Randomly select three (typically) of these words to create a query
   - Send the query to a commercial search engine (we have used Google, Yahoo and Bing) which returns a 'search hits' page.
   - Retrieve pages identified in the search hits page. Store them.
3. `Clean' the text, to remove navigation bars, advertisements and other recurring material
4. Remove duplicates
5. Tokenise, and, where tools are available, lemmatise and part-of-speech tag
6. Load into a corpus query tool.

We discuss each step below.

2.1. Seed Word Selection

For each language, we need seed words to start the process. Sharoff used 500 common words drawn from word lists from pre-existing corpora: the BNC for English, RNC for Russian, IDS for German and Chinese Gigaword for Chinese. But for the languages we are most interested in, there are no corpora available (which is why we are building them).

Wikipedia (Wiki) is a huge knowledge resource built by collective effort with articles from many domains. The whole dataset can be downloaded. While one possibility would be to treat the Wiki for a language as a corpus, it may not be large enough, or diverse enough in text type, for many purposes (see also the evaluation section). For these reasons we prefer to use the Wiki for generating frequency lists to determine the

seed words and then use Web data obtained using these seeds as the actual corpus. Currently, Wikipedia hosts around 265 languages including all those for which we plan to build corpora so we can apply the same method across many languages, and the corpora so produced should be `comparable' -- or at least more similar to each other than if we had used a different method for gathering seed words in each case.

2.1.1. Extracting Wiki Corpora

For each language, a Wiki corpus is extracted from a Wiki dump of the language. A Wiki dump is a single large XML file containing all the articles of the Wikipedia. We used a slightly modified version of the Wikipedia2Text[1] tool to extract plain text (Wiki corpus) from the Wiki dump. We found that most of the Wiki articles do not have connected text but are short definitions, sets of links, or 'stubs': articles which exist for purposes of being pointed to by other articles but which have not themselves been written yet. They need filtering out. Generally they are small. Ide et al. (2002) give an estimate of minimum 2000 words as an indicator of connected text. Heuristically, we consider a Wiki file to have connected text if its word count is more than 500. We use the Wiki corpus to build a first frequency list for the language. *Table* 1 gives statistics of Wiki Corpora.

	Wiki XML dump	**Wiki plain corpus**	**Pages with >500 words**	
			MB	**Words**
Dutch	1.8 GB	2.6 GB	203 MB	30 m
Hindi	149 MB	367 MB	35 MB	2.5 m
Indonesian	475 MB	1.0 GB	58 MB	8.5 m
Norwegian	910 MB	1.6 GB	140 MB	19.1 m
Swedish	1.2 GB	2.1 GB	59 MB	9.3 m
Telugu	108 MB	337 MB	7.3 MB	0.23 m
Thai	463 MB	698 MB	93 MB	6.23 m
Vietnamese	426 MB	750 MB	78 MB	9.5 m

Table 1: Wiki Corpus Statistics

2.1.2. Building frequency lists

To get a frequency list from a Wiki Corpus, it must first be tokenised. For languages like Thai and Vietnamese where explicit word delimiters are absent, we used language-specific tools for tokenisation. For other languages we used space and other punctuation marks. Once the Wiki corpus is tokenised, term frequency and document frequency are calculated and a frequency list is built. Words are sorted in the frequency list based on document frequency.

For most languages, most search engines do not index on lemmas but on word forms. They treat different forms of the word as different words. For example the Telugu word ప్రాంతంలో ("in location") gave more Yahoo search hits than its lemma ప్రాంతం ("location"). Sharoff (2006) discusses similar findings for Russian. We used a frequency list for word forms rather than lemmas, and used word forms as seeds.

2.1.3. From frequency list to seed words

We treat the top 1000 words as the high-frequency words of the language and the next 5000 as the mid-frequency ones which we shall use as our seed words. The Wikipedias are in UTF-8 encoding and so are the seed words.

Some studies (Grefenstette and Nioche, 2000; Ghani et al., 2005) used only seed words that were unique to the target language, to avoid accidental hits for pages from other languages. Three of the eight languages in our sample (Hindi, Telugu, Thai) use their own script so, if the character encoding is correctly identified, there is no risk of accidentally getting a page for the wrong language. For other languages (Latin-script languages), we adopted different tactics.

For other languages except Vietnamese, we used a word length constraint of at least 5 characters to filter out many words which are also words in other languages: it tends to be short words which are words in multiple languages of the same script. Many words from other languages are not filtered out. However:

- We are only likely to get a page from another language if all seed terms in a query are also words from the same other language. This becomes less likely where there are multiple seeds and where many multi-language words have been filtered out
- We have a further stage of filtering for language, as a by-product for filtering for running text, using the highest-frequency words of the language (see below)
- A Vietnamese word may comprise more than one space-separated item. The lengths in characters of the space-separated items are found to be small. Word length is not a good constraint in this case. We used the constraint that a Vietnamese word should contain at least one Unicode character which is not in the ASCII range, since Vietnamese uses many diacritics.

2.2. Query Generation

Web queries are generated from the seeds using BootCaT's query generation module. It generates tuples of length n by random selection (without replacement) of n words. The tuples will not be identical nor will they be permutations of each other.

We needed to determine how to set n. Our aim is to have longer queries so that the probability of results being in the target language is high and more queries can be generated from the same seed set. At the same time, we have to make sure that the hit count is not small for most of the queries. As long as we get a hit count of more than ten for most queries (say, 90%), the query length is considered to be valid. We define the best query length as the maximum length of the query for which the hit count for most pages is more than ten. We use the following algorithm to determine the best query length for each language.

Algorithm 1: Best Query Length

1. set n = 1
2. generate 100 queries using n seeds per query
3. Sort queries by the number of hits they get.
4. Find hit count for 90th query (min-hits-count)
5. if min-hits-count < 10 return n-1
6. n = n +1, go to step 2

[1] `http://evanjones.ca/software/wikipedia2text.html`

	n=1	2	3	4	5	Best
Dutch	1300000	3580	74	5	-	3
Hindi	30600	86	1	-	-	2
Indonesian	29500	1150	78	9	-	3
Norwegian	49100	786	9	-	-	2
Swedish	55000	1230	33	7	-	3
Telugu	668	2	-	-	-	2
Thai	724000	1800	193	5	-	3
Vietnamese	1100000	15400	422	39	5	4

Table 2: **Query length, hit counts at 90th percentile and Best Query Length**

Best query lengths for different languages obtained from Yahoo search hits are shown in *Table* 2. We used a minimum query length of two, so did not apply the algorithm fully for Telugu.
Once query-length was established we generated around 30,000 queries for each language.

2.3. URL Collection

For each language, the top ten search hits are collected for 30,000 queries using Yahoo's or Bing's API. Recently for Swedish, Norwegian and Indonesian, we used Bing since its terms and conditions allowed us to send more queries per day. *Table* 3 gives some statistics of URL collection.
We found that Google gave more hits than Yahoo or Bing, particularly for languages that have non-ASCII characters. The reason for this may not be the difference in index size. Google normalises many non-UTF8 encoding pages to UTF8 encoding and then indexes on them whereas Yahoo and Bing do less normalisation and more often index the words in the encoding of the page itself. We verified this for Telugu. http://www.eenadu.net is a widely-used Telugu news site which uses non-UTF8 encoding. We restricted the search hits to this news site and for the unicode query చంద్రబాబు (a famous politician) we got 9930 Google search hits, 14 Yahoo hits and 10 Bing hits. We also ran the query with the original encoding. There were 0 Google hits, 2670 Yahoo hits and 1440 Bing hits. This shows that Yahoo and Bing also indexed Eenadu but did not normalise the encoding. Since we use UTF8 queries, Google would serve our purposes better for Telugu. But for licensing and usability reasons, we have used Yahoo or Bing to date. For Indian languages, to collect data in other encodings we generated queries in different encodings apart from UTF8 by converting the UTF8 seeds using encoding mappings.
While collecting the URLs, we store the query, page size and MIME type, as provided in the search engine output.

2.4. Filtering

The URLs were downloaded using unix wget. Since we already had MIME information for the URL, we downloaded only those pages whose MIME type was text/HTML. We also had page size, so we downloaded only those files above 5 KB so that the probability of connected text was greater. Files larger than 2 MB were discarded to avoid any particular domain files dominating the composition of the corpus, and also because files of this length are very often log files and other non-connected text.

The downloaded pages contain html markup and 'boilerplate' text like navigation bars, advertisements and legal disclaimers. To remove such content and extract only the connected text, we used the Body Text Extraction algorithm (BTE, Finn et al. 2001). BTE starts from the observation that Web pages typically have material at the beginning and end which is rich in boilerplate and which tends to be heavily marked up, and material in the middle, the 'body text', which is linguistic and is the material we want, and is relatively light in markup. It calculates the ratio of text to markup for different parts of the page, divides the page into three sections on the basis of this ratio, and retains only the middle one. BTE was performed on all the downloaded pages to get plain text pages.
These pages are further filtered to check for connected text. Connected text in sentences reliably contains a high proportion of function words (Baroni, 2005). If a page does not meet this criterion we discard the page. We assume that the top 500 words in the frequency list (as prepared from the Wiki corpus) include most function words. To set a threshold for the proportion of tokens to be accounted for by the top-500 words, we sorted all Wiki files according to the proportion of top-500 words in the file. We found that most of the Wiki files at the bottom (below 75-80 %) of this sorted list did not contain connected text. This is either due to bad cleaning by the Wikipedia2Text tool or because the page really did not contain connected text. The Wiki file at 70th% of the sorted list is used to set the threshold: if, in the 70th-percentile file, words from the top-500 list accounted for 65% of all words, then the threshold for the language was set at 65% and any page where less than 65% of the words were from the top-500 list was discarded.

2.5. Near Duplicate Detection

We used perl's Text::DeDuper module for near duplicate detection. This module uses the resemblance measure as proposed by Broder et al. (1997) to detect similar documents based on their text. This is a memory intensive task. N-grams (n=5) for each document are generated and similarity is measured between two documents based on the number of overlaps in their n-grams. Since main memory size is limited and can hold only a limited number of files, duplicate detection is done using a sliding window. At each iteration a fixed number of non-duplicate files, say 500, whose n-grams can fit in memory, are identified using the DeDuper module. All other files are taken one file at a time and compared with the n-grams of these non-duplicate files to identify if they are duplicates or not. This process is repeated until all files are covered. A detailed algorithm is given below. After this step, we get the final Web corpus. Sizes are given in *Table* 3.

Algorithm 2: Identify Near Duplicates

1. Sort the file names by their file sizes and store all the filenames in a list.
2. Identify first 500 non duplicate documents (traversing linearly on filenames list) using DeDuper module
3. Compare rest of the files, a file at a time, with these 500 non-duplicate documents
4. Remove any duplicate files found and store the rest of the filenames in next_filenames list
5. filenames = next_filenames
6. Continue from step 2.

In future, we expect to use methods proposed in (Pomikálek and Rychlý, 2008; Pomikálek et al., 2009)

	Unique URLs Collected	After Filtering	After Duplicate Removal	Web Corpora Size	
				MB	m Words
Dutch	97,584	22,424	19,708	739	108.6
Hindi	71,613	20,051	13,321	424	30.6
Indonesian	79,402	28,987	27,051	708	102.0
Norwegian	258,009	66,299	62,691	628	94.9
Swedish	168,511	31,683	28,842	719	114.0
Telugu	37,864	6,178	5,131	107	3.4
Thai	120,314	23,320	20,998	1200	81.8
Vietnamese	106,076	27,728	19,646	1200	149.0

Table 3: **Web Corpora Statistics**

2.6. Indian Languages

For Indian languages, we have noted that the web is relatively small given the number of speakers. We suspect this is because the dominant language of education in India is English, coupled with the confusing variety of encodings which are possible for Indian languages: most Indian web users know enough English to use the web in English, and find this easier, as they will not miss pages in the wrong encoding. (For the same reasons, web authors often choose to write in English.) As web use penetrates further, and as encodings standards are more widely adopted, we would expect this to change over the next few years.

2.7. Part-of-speech Tagging and Lemmatisation

We part-of-speech-tagged and lemmatised corpora for the languages which have open-source tools. Currently, we were able to find tools for Dutch, Vietanamese and Swedish. For other languages, we hope to either find them shortly or work with NLP groups who are developing them.

2.8. Loading into a Corpus Query Tool

The corpora were then loaded into the Sketch Engine (Kilgarriff et al., 2004), where they are accessible at http://www.sketchengine.co.uk.

3. Evaluation

What does it mean for a corpus to be good? It depends what we want to use the corpus for. The straightforward answer to the question is "if it supports us in doing what we want to do".

We anticipate that our corpora will be evaluated in this way, by a range of language researchers, over time. As they use a corpus and get to know it they will come to realise what it is good for and what it is not. We have had this experience with large English corpora, particularly the Oxford English Corpus, which has now been in use for several years and where new phases of corpus-building have been designed to address the lexicographers' criticisms of previous versions, which they had got to know very well.

But this kind of evaluation takes time: how might we do a first-pass evaluation of the corpora without waiting?

The only strategy we know of is by comparison: comparing one corpus with another, and, in particular, comparing frequency lists of the two corpora. The topic is explored in general in Kilgarriff (2001) and frequency-list-comparison methods are used for Web corpus evaluation in Baroni and Kilgarriff (2006), Sharoff (2006), Ferraresi et al. (2008). (There are also many studies using frequency list comparisons, also often called keywords analyses, to compare corpora of different text types or regional varieties, to explore the differences between the varieties. Usually word frequency lists are used, though sometimes frequencies related to word classes or grammatical constructions have been explored, notably in Biber (1988))

For each of the languages, we have two corpora available: the Web corpus and the Wiki corpus. In the case of Dutch, we also have access to a carefully-designed lexicographic corpus.

3.1. Comparing Web and Wiki corpora

The Wiki corpora were prepared as sources of seeds for the Web corpus building. But they are also corpora which may be of interest in their own right. How do they compare with the Web corpora? It is possible that they are better for some purposes: they may have a higher proportion of well-written material, as they do not include arbitrary texts in the way that the Web corpora do.

The first point to make is simply that they are far smaller, see *Table* 4.

	Wiki Corpora	Web Corpora
Dutch	30.0 m	108.6 m
Hindi	2.5 m	30.6 m
Indonesian	8.5 m	102.0 m
Norwegian	19.1 m	94.9 m
Swedish	9.3 m	114.0 m
Telugu	0.2 m	3.4 m
Thai	6.2 m	81.8 m
Vietnamese	9.5 m	149.0 m

Table 4: Sizes of Wiki and Web Corpora (in millions of words)

Another hypothesis is that the Wiki corpora are more `informational' and the Web ones more `interactional'. Biber (1988) shows how the dominant dimension of variation for English is `interactional vs informational': some kinds of language use are principally concerned with interaction between participants whereas others are principally for conveying information, and this is the principal axis along which texts are best classified for register. Biber (1995) shows how this holds across a number of languages.

Informational language is typical written, and interactional, spoken. It is usually easier to gather large quantities of informational registers, for example newspapers, official reports, academic papers and Wikipedia articles, than interactional ones, including spontaneous conversation. In general, we might expect a Web corpus to be more

Dutch				Hindi				Telugu			
Word	Web	Wiki	Ratio	Word	Web	Wiki	Ratio	Word	Web	Wiki	Ratio
ik	5786	2526	2.28	मैं	2363	360	6.55	నా	3736	603	6.18
je	4802	975	4.92	मेरा	578	90	6.39	నేను	3390	461	7.34
jezelf	96	9	10.03	तुम	827	114	7.23	నాది	44	17	2.59
kij	188	37	5.06	आप	1725	664	2.59	నన్ను	585	127	4.58
jou	102	19	5.16	आपका	192	54	3.50	మీ	2092	572	3.65
jouw	99	19	5.05	मैंने	709	65	10.76	మీరు	1756	476	3.68
jullie	367	112	3.28	मुझे	1404	122	11.50	నువ్వు	281	89	3.15
me	599	294	2.03	तू	185	50	3.65	మీకు	730	182	3.99
mezelf	41	5	6.89	तुम	827	114	7.23	నీవు	80	148	0.54
mij	768	344	2.23	तूने	23	12	1.85	నీ	465	263	1.76
Total	14221	4771	2.98	Total	8833	1645	5.36	Total	15755	3176	4.96

Thai				Vietnamese			
Word	Web	Wiki	Ratio	Word	Web	Wiki	Ratio
ผม	2935	366	8.00	anh	2255	749	3.00
ดิฉัน	133	19	7.00	bạn	1827	460	3.96
ฉัน	770	97	7.87	chị	400	36	10.91
คุณ	1722	320	5.36	em	998	199	5.00
ทาน	2390	855	2.79	mày	116	6	19.41
กระผม	21	6	3.20	tôi	4747	475	9.97
ข้าพเจ้า	434	66	6.54	tao	89	6	14.57
ตัว	2108	2070	1.01	ta	2516	675	3.72
กู	179	148	1.20	mình	2694	1487	1.81
ชั้น	431	677	0.63	mi	24	7	3.28
Total	11123	4624	2.40	Total	15666	4100	3.82

Table 5: 1st and 2nd person pronouns in Web and Wiki corpora. All figures in 'Web' and 'Wiki' columns are frequencies per million words. For Dutch and Vietnamese, counts are case-insensitive. The figure in the Ratio column is the Web:Wiki ratio.

interactional, and `traditional' and Wiki corpora more informational. The Web, particularly Web 2.0, supports interaction and informality. Ferraresi et al. (2008) explore register variation in UKWaC, a large Web corpus, comparing it with the British National Corpus, and find UKWaC to be markedly more interactional.

In our case the Wiki corpus was used, via the seed words, to generate the Web corpus. One criticism of our method would be that since we use Wikipedia texts to find seeds, we are likely to have an imbalance of informational as opposed to interactional texts in the Web corpora.

We explored the question by noting that first and second person pronouns are strong indicators of interactional language. For each pair of corpora, for each of five languages, we made a list of ten of the commonest first and second personal pronouns (for English the list would be *I me my mine you your yours we us our*) and counted their frequencies in the Web and Wiki corpora. We normalised figures to per-million and calculated the ratio, Web:Wiki, as in *Table* 5.

For forty-eight of the fifty pronouns, the ratio is greater than one, often many times greater. The ratio across all ten pronouns varies between 2.4 times more common (Thai) to over five times (Hindi). We can conclude that the Web corpora are more interactional than the Wiki corpora used to develop them.

3.2. Comparing NLWaC and ANW

The ANW corpus is a balanced corpus of just over 100 million words compiled at the Institute for Dutch Lexicology (INL) and completed in 2004. It was built to support the lexicography for the ANW, a major new dictionary of Dutch currently in preparation. It comprises: present-day literary texts (20%), texts containing neologisms (5%), texts of various domains in the Netherlands and Flanders (32%) and newspaper texts (40%). The remainder is the 'Pluscorpus' which consists of texts, downloaded from the internet, with words that were present in an INL word list but absent in a first version of the corpus.
To compare the Dutch Web corpus (called NlWaC) with the ANW corpus, we prepared frequency lists for word forms for both corpora and found the `keywords' of each corpus in contrast to the other using the formula

$$\frac{Freq/mill\ in\ corpus1\ +\ 100}{Freq/mill\ in\ corpus2\ +\ 100}$$

(For discussion of the formula and the parameter, see Kilgarriff (2009)). We then look at the words with the highest and lowest scores.
The twenty highest-scoring (ANW) keywords and the twenty lowest-scoring (NlWaC) keywords, with English glosses and clustered by themes, are given in *Table* 6.

The classification into themes was undertaken by checking where and how the words were being used, using the Sketch Engine. The analysis shows that these two large,

ANW			NlWaC		
Theme	**Word**	**English gloss**	**Theme**	**Word**	**English gloss**
Belgian	Brussel	(city)	Religion	God	
	Belgische	Belgian		Jezus	
	Vlaamse	Flemish		Christus	
Fiction	Keek	Looked/watched		Gods	
Newspapers	vorig	previous	Web	http	
	kreek	watched/looked		Geplaatst	posted
	procent	Percent		Nl	(Web domain)
	miljoen	million		Bewerk	edited
	miljard	billion		Reacties	Replies
	frank	(Belgian) Franc		www	
	Zei	said	English	And	In book/film/song
	aldus	thus		The	titles, names etc
	Meppel	City with local newsp	History	Arbeiders	workers
	gisteren	yesterday		Dus	thus
	Foto	Photo		Macht	power
	Auteur	Author		Oorlog	war
	Van	(in names)		Volk	people
Pronouns	Hij	Him/he	Pronouns	We	we
	haar	She/her(/hair)		Ons	us
	Ze	(They/them)		Jullie	you

Table 6: Keywords in ANW and NlWaC

general corpora of Dutch have different strengths and weaknesses, and different areas that might be interpreted as over-representation or under-representation. The ANW has a much stronger representation of Flemish (the variety of Dutch spoken in Belgium). It has 20% fiction: *keek* (looked, watched) is used almost exclusively in fiction. It is 40% newspaper and newspapers talk at length about money (which also interacts with time and place: franks were the Belgian currency until 1999; also the units were small so sums in franks were often in millions or even billions). There is a particularly large chunk from the Meppel local newspaper. Most occurrences of *foto* were in ``Photo by'' or ``Photo from'' and of *auteur*, in newspaper by-lines, which might ideally have been filtered out. Daily newspapers habitually talk about what happened the day before, hence *gisteren*. *Vorig* and *aldus* (previous, thus) are fairly formal words that get used more in newspapers than elsewhere.

NlWaC has a large contingent of religious texts. It is based on Web texts, some of which could have been more rigorously cleaned to remove non-continuous-text and other non-words like URL components *www, http, nl.* The English might appear to be because we had gathered mixed-language or English pages but when we investigated, we found most of the instances of *and* and *the* were in titles and names, for example "The Good, the Bad and the Ugly", where the film was being discussed in Dutch but with the title left in English. Perhaps modern global culture, with its tendency to use English in film, book and song titles, institution names and catch phrases, is better-represented in NlWaC than in ANW. Political history is also well-represented.

Finally we note that pronouns occur in both lists: third-person ones in the ANW list, and first and second person ones in the ANW list. This confirms the hypothesis discussed above and the evidence from Ferraresi et al (2008): Web-based methods as described in this paper give us the opportunity to access more interactional language than was possible for large traditional corpora.

4. Future Work

We would like to prepare corpora for further languages. High on our priority list are Korean, Tibetan, Turkish and all the official languages of the European Union. We would like to not only extract corpora, but also estimate how large the Web is for each language.

In a parallel stream of work focusing on English we have developed a high-accuracy, scaleable, de-duplication method (Pomikálek et al., 2009). We shall explore applying this method in the Corpus Factory.

The paper has mainly discussed the preparation of plain-text corpora. To set up the corpora for language technology and linguistic research, they should be accurately segmented, lemmatised and part-of-speech (POS) tagged; loaded into a corpus tool such as the Sketch Engine; and supplemented with a `Sketch Grammar'. Then, lexicographers and others can see `word sketches', one-page summaries of a word's grammatical and collocational behaviour. Word sketches have widely been found to be a good starting point for dictionary-writing (see eg Kilgarriff and Rundell (2002)). But for this to be realised we need the language-specific tools. For segmenters, lemmatisers and POS-taggers we have often used open-source tools, for example SWATH[2] for segmenting Thai, but for many languages they are not. In these cases we are looking out for partners with computational linguistics expertise in the language, to work together on creating the tools. We want to work with people with those skills to prepare sketch grammars.

[2]Swath: Word Segmentation Tool for Thai (`http://www.cs.cmu.edu/~paisarn/software.html`)

5. Summary

The 'corpus factory' presents a method for developing large general-language corpora which can be applied to many languages. In this paper we have described the method, and how it has worked when we have applied it to eight languages from different language families, each presenting different issues in terms of character encoding and orthography. We have produced a set of eight large corpora. We think they are high-quality resources, better for language research than any others currently in existence for at least five of the eight languages. We have evaluated the corpora, as far as we were able given the lack of other resources for comparison. The corpora are available for use in a leading corpus tool. We believe the Corpus Factory has a great deal to offer language technology and linguistic research in the years to come.

Acknowledgements

We would like to thank Diana McCarthy for proof checking the paper, Carole Tiberius for her help on Dutch, GiaoÂ ChiÂ Le Thi for hers on Vietnamese, John Hartmann for his on Thai and Phuong Le-Hong for providing the Vietnamese POS tagger.

6. References

Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proc. LREC*, pages 1313--1316.

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proc. EACL*, pages 87--90.

Marco Baroni. 2005. Distributions in text. In Anke Lüdeling and Merja Kytö, editors, *Corpus linguistics: An international handbook"*. Mouton de Gruyter, Berlin.

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

Douglas Biber. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the web. *Computer Networks*, 29(8-13):1157--1166.

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. 2008. Introducing and evaluating "ukwac", a very large web-derived corpus of English. In *Proc. WAC4 Workshop at LREC*, Marrakech, Morocco.

Rayid Ghani, Rosie Jones, and Dunja Mladenic. 2005. Building minority language corpora by learning to generate web search queries. *Knowledge and Information Systems*, 7(1):56--83.

Gregory Grefenstette and Julien Nioche. 2000. Estimation of english and non-english language use on the www. In *Proc. RIAO*, pages 237--246.

Nancy Ide, Randi Reppen, and Keith Suderman. 2002. The American National Corpus: More than the web can provide. In *Proc. LREC*, pages 839--844, Las Palmas.

Rosie Jones and Rayid Ghani. 2000. Automatically building a corpus for a minority language from the web. In *Proc. ACL Student Workshop*, pages 29--36.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational. Linguistics.*, 29(3):459--484.

Adam Kilgarriff and Michael Rundell. 2002. Lexical profiling software and its lexicographic applications : a case study. In *Proc. EURALEX*, pages 807--818, Copenhagen.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The sketch engine. In *Proc. EURALEX*, pages 105--116, Lorient, France.

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

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proc. Corpus Linguistics*, Liverpool.

Jan Pomikálek and Pavel Rychlý. 2008. Detecting coderivative documents in large text collections. In *Proc. LREC*, Marrakech, Morocco.

Jan. Pomikálek, Pavel Rychlý, and Adam Kilgarriff. 2009. Scaling to billion-plus word corpora. In *Advances in Computational Linguistics: Special Issue of Research in Computing Science*, volume 41, Mexico City.

Philip Resnik. 1999. Mining the web for bilingual text. In *Proc. ACL*, pages 527--534.

Kevin P. Scannell. 2007. The crubadan project: Corpus building for under-resourced languages. In *Proc. WAC-3: Building and Exploring Web Corpora*, Louvain-la-Neuve, Belgium.

Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. In *WaCky! Working papers on the Web as Corpus*. Gedit.