Lang Resources & Evaluation (2009) 43:209–226
DOI 10.1007/s10579-009-9081-4

The WaCky wide web: a collection of very large linguistically processed web-crawled corpora

Marco Baroni · Silvia Bernardini · Adriano Ferraresi · Eros Zanchetta

Published online: 10 February 2009

Abstract This article introduces ukWaC, deWaC and itWaC, three very large corpora of English, German, and Italian built by web crawling, and describes the methodology and tools used in their construction. The corpora contain more than a billion words each, and are thus among the largest resources for the respective languages. The paper also provides an evaluation of their suitability for linguistic research, focusing on ukWaC and itWaC. A comparison in terms of lexical coverage with existing resources for the languages of interest produces encouraging results. Qualitative evaluation of ukWaC versus the British National Corpus was also conducted, so as to highlight differences in corpus composition (text types and subject matters). The article concludes with practical information about format and availability of corpora and tools.

Keywords Annotated corpora · Corpus construction · General-purpose linguistic resources · English · German · Italian · Web as corpus · WaCky!

1 Introduction

This article introduces the WaCky corpora, a collection of very large (>1 billion words) corpora of English (*ukWaC*), German (*deWaC*) and Italian (*itWaC*). These corpora were built by web crawling, they contain basic linguistic annotation (part-of-speech tagging and lemmatization) and they aim to serve as general-purpose

M. Baroni (✉)
CIMeC, University of Trento, Rovereto, Italy
e-mail: marco.baroni@unitn.it

S. Bernardini · A. Ferraresi · E. Zanchetta
SITLeC, University of Bologna, Forlì, Italy

resources for the target languages. The German and Italian corpora are, to the best of our knowledge, the largest publicly documented language resources in the respective languages; ukWaC is among the largest, and the only English web-crawled resource with linguistic annotation. We developed the corpora between 2005 and 2007 as part of the WaCky project (*W*eb *a*s *C*orpus *k*ool *y*nitiative), an informal consortium of researchers interested in the exploration of the web as a source of linguistic data.

The main goals of the article are to make the (computational) linguistics community aware of the corpora, to describe in some detail the process by which they were created, that can be used to rapidly develop similar corpora for other languages, and to provide some preliminary evaluation of their fitness for language studies. For space reasons, we omit here a discussion of the relative benefits of web-derived versus traditional corpora. There is by now a large literature on the issue (see Sect. 2 for some references), and our group has expressed its own view in several other locations—see for example Sect. 2 of Baroni and Ueyama (2006) and Chap. 1 of Ferraresi (2007).

The article is structured as follows: Sect. 2 reviews similar web corpus collection efforts. In Sect. 3 we describe how the WaCky corpora were constructed. In Sect. 4 we look at similarities and differences in terms of lexis with respect to Italian and English reference corpora. Sect. 5 deals with issues related to format and availability of corpora and related materials and tools. Finally, Sect. 6 concludes by discussing what we consider the most pressing next steps of the WaCky initiative.

2 Related work

There is a growing literature on using the web for linguistic purposes, mostly via search engine queries or by crawling ad hoc data—see for example the papers in Kilgarriff and Grefenstette (2003), Baroni and Bernardini (2006), Hundt et al. (2007), Fairon et al. (2007) and references therein. On the other hand, we are not aware of much publicly documented work on developing large-scale, general-purpose web-derived corpora.

The first enterprise of this sort we know of is the terabyte corpus (53 billion words) built at the University of Waterloo and briefly described in Clarke et al. (2002). This corpus is based on a crawl of the web seeded with URLs from universities and other educational organizations. Perfect duplicates were discarded but, from what we gather, the retrieved pages did not undergo any further processing, not even to remove HTML or to ensure that they are in the target language (English). A similar initiative based on a crawl of the .gov domain resulted in the GOV2 corpus used for the TREC 'terabyte track' (Clarke et al. 2005).

Thelwall (2005) describes crawls of academic sites in Australia, New Zealand and the United Kingdom. The only form of post-crawl processing reported pertains to semi-automated filtering of suspicious pages such as those dynamically generated from a database. The largest corpus—the one of UK sites—contains about 1.3 billion words. The corpora are evaluated through the extraction and analysis of

wordlists. Other projects that have used web crawls in order to create frequency lists include Kornai et al. (2006), who work on Hungarian, and Emerson and O'Neil (2006) on Chinese. A particularly impressive web-derived frequency list is the Google terabyte n-gram collection, made publicly available in 2006 (Brants and Franz 2006).

Liu and Curran (2006) describe a 10 billion word English corpus crawled from seed URLs that are randomly selected from different topics in the Open Directory collection.[1] The corpus is split into sentences and tokenized, and sentence-level filtering is carried out using lexical resources (sentences with too few dictionary words or too many numbers, punctuation or other tokens are discarded). No near-duplicate removal is performed. The corpus is thoroughly evaluated in two NLP tasks, where performance of algorithms trained on the web corpus is shown to be similar or superior to that of the same algorithms trained on 2 billion words of newspaper text.

Shaoul and Westbury (2007) make available a corpus of USENET postings collected between 2005 and 2007. NNTP headers and perfect duplicates are discarded, and so are documents containing less than 500 words or more than 500,000 words, and documents that contain less than 90% words from an English dictionary. No further processing is performed. The resulting corpus contains over 13 billion words.

Finally, we mention two projects that, while building resources on a smaller scale than the WaCky corpora, followed processing pipelines very similar to ours. CUCWeb (Boleda et al. 2006) is a Catalan corpus containing 166 M words from a crawl seeded with a URL list provided by a search engine. The crawl traversed the .es domain as well as retrieving pages from IP addresses assigned to networks physically located in Spain. Language filtering was applied to extract pages in Catalan, and dictionary-based heuristics were applied to exclude documents with a large proportion of non-linguistic materials. Perfect duplicates were also excluded. The resulting corpus was annotated with POS tags, morphological features and shallow syntactic information. An important feature of CUCWeb is that it is available for querying via a user-friendly web interface.[2]

Sharoff developed a collection of 'BNC-sized' corpora (around 100 M tokens) that, as of early 2008, include English, Chinese, Finnish, French, German, Italian, Japanese, Polish, Portuguese, Russian and Spanish, and that can be queried via an online interface.[3] The methodology he followed (Sharoff 2006) is similar to the one described here—indeed, many tools and ideas were developed jointly. The main differences are that Sharoff does not perform a true crawl (he retrieves and processes only the pages returned by the random Google queries, rather than using them as seed URLs), nor does he perform near-duplicate detection. Evaluation of some of these corpora is carried out in Sharoff (2006), where a comparison is made with reference points in the same languages, in terms of domain analysis and comparing wordlists similarly to what we do here.

[1] http://www.dmoz.org.

[2] http://www.catedratelefonica.upf.es/cucweb.

[3] http://corpus.leeds.ac.uk/internet.html.

This brief survey shows that we are by no means the first to build corpora by web crawling. Indeed, there are initiatives that cover more languages (Sharoff 2006), web corpora with more in-depth linguistic annotation (Boleda et al. 2006), and many larger web-derived resources (Clarke et al. 2005; Liu and Curran 2006; Shaoul and Westbury 2007). However, the WaCky collection offers a compromise between very large size (the corpora by Sharoff and Boleda et al. are about one tenth the size of the WaCky corpora) and thorough post-processing for linguistic purposes (neither Shaoul and Westbury nor the other developers of mega-corpora had corpus balance, cleaning or linguistic annotation among their priorities). Thus, the WaCky corpora are at the moment a unique resource, as far as large linguistics-oriented corpora currently go. Moreover, this article provides the most detailed and complete report we are aware of describing the pipeline used to create and evaluate a large web-derived corpus—as such, we hope it will also serve as a 'how-to' for other researchers interested in creating their own corpora.

3 Corpus construction

The procedure described in this section was carried out on a server running RH Fedora Core 3 with 4 GB RAM, Dual Xeon 4.3 GHz CPUs and about 2.5 TB hard disk space. Data about corpus size and other relevant summary statistics for each step of the creation process are reported in Table 1.

3.1 Crawl seeding and crawling

Our aim is to set up resources akin to traditional general language (or so-called 'reference') corpora, containing a wide range of text types and topics. These should include both 'pre-web' texts of a varied nature that can also be found in electronic format on the web (spanning from sermons to recipes, from technical manuals to short stories, and ideally including transcripts of spoken language as well), and texts representing web-based genres (Santini and Sharoff 2007), like personal pages,

Table 1 Size data for deWaC, itWaC and ukWaC

	deWaC	itWaC	ukWaC
Number of seed word pairs	1,653	1,000	2,000
Number of seed URLs	8,626	5,231	6,528
Raw crawl size (GB)	398	379	351
Size after document filtering (GB)	20	19	19
Number of documents after filtering (M)	4.86	4.43	5.69
Size after near-duplicate cleaning (GB)	13	10	12
Number of documents after near-duplicate cleaning (M)	1.75	1.87	2.69
Size with annotation (GB)	25.9	30.6	30
Number of tokens	1,278,177,539	1,585,620,279	1,914,150,197
Number of types	9,347,112	3,651,021	3,798,106

blogs, or postings in forums. Notice that the rationale here is for the corpus to include a sample of pages that are representative of the language of interest, rather than getting a random sample of web pages representative of the language of the web. While the latter is a legitimate object for 'web linguistics' (Kilgarriff and Grefenstette 2003), its pursuit is not among the priorities set out for the WaCky corpora.

In our approach, the first step in corpus construction consists in identifying different sets of seed URLs which ensure variety in terms of content and genre. In order to find these, random pairs of randomly selected content words in the target language are submitted to a commercial search engine through its API service.[4] We choose bigram queries because preliminary experimentation found that single word queries tend to yield potentially undesirable documents (e.g., dictionary definitions of the queried words, or the top pages of companies with the relevant word in their name), whereas combining more than two words would often retrieve pages with lists of words, rather than connected text. Content- and genre-wise, previous research on the effects of seed selection upon the resulting web corpus (Ueyama 2006) suggested that automatic queries to Google which include words sampled from traditional written sources such as newspapers and reference corpus materials tend to yield 'public sphere' documents, such as academic and journalistic texts addressing socio-political issues and the like. Queries with words sampled from a basic vocabulary list, on the contrary, tend to produce corpora featuring 'personal interest' pages, like blogs or bulletin boards. Since it is desirable that both kinds of documents are included in the corpora, relevant sources have been chosen accordingly.

For ukWaC, we construct a set of 1,000 pairs by randomly combining mid-frequency content words randomly selected from the British National Corpus (henceforth BNC; see Sect. 4.1); function words are excluded from the list, since search engines usually ignore them when they are submitted as part of a query. Two other lists of 500 random bigrams complement this, one extracted specifically from the demographically sampled spoken section of the BNC (i.e., the informal conversation component, containing basic vocabulary), and the other from a vocabulary list for foreign learners of English[5] which (however counter-intuitively) contains rather formal vocabulary, possibly required for academic study in English. Seeds for Italian and German are randomly selected among mid-frequency words in two newspaper text collections (*la Repubblica*, see Sect. 4.1, and *Süddeutsche Zeitung*), as well as from basic vocabulary lists, from which function words and particles are removed.[6] For German, 1,000 pairs are extracted from the newspaper and 653 from the basic vocabulary list. For Italian, a total of 1,000 pairs are

[4] The Google API facility was used in the construction of deWaC, itWaC and ukWaC. While this functionality is no longer offered to new users, similar ones are offered by, e.g., Microsoft Live Search and Yahoo!.

[5] http://wordlist.sourceforge.net/.

[6] http://www.bardito.com/language/italian_english_wordlist.html and http://homepage.bluewin.ch/cusipage/.

constructed by randomly mixing words from the newspaper and the basic vocabulary list.[7]

A maximum of 10 seed URLs are retrieved for each random seed pair query, and the retrieved URLs are collapsed in a single list. Duplicates are discarded and, to ensure maximal sparseness, only one (randomly selected) URL for each (normalized) domain name is kept. These filtered seed URLs are fed to a crawler, in random order. The crawls are limited to pages in the relevant web domains (.de/.at for German; .it for Italian; .uk for English) whose URL does not end in a suffix indicating non-HTML data (.pdf, .jpg, etc.). The English crawl is limited to the .uk domain in order to construct a relatively homogeneous resource, comparable to the BNC, and because of practical and theoretical issues arising when trying to define the country domains to crawl—e.g., including or excluding countries in which English is an official, though not a native language. Our strategy does not, of course, ensure that all the pages retrieved represent British English.

The crawls are performed using the Heritrix[8] crawler, with a multi-threaded breadth-first crawling strategy; they are stopped after 10 days of continuous running. The full seed pair and seed URL lists are available from the project page (see Sect. 5).

3.2 Post-crawl cleaning

Using information in the Heritrix logs, we only preserve documents that are of mime type text/html, and between 5 and 200 KB in size. As observed by Fletcher (2004) and confirmed by informal experimentation, very small documents tend to contain little genuine text (5 KB counts as 'very small' because of the HTML code overhead) and very large documents tend to be lists of various sorts, such as library indices, store catalogs, etc.

We also spot and remove all documents that have perfect duplicates in the collection (i.e., we do not keep *any* instance from a set of identical documents). This drastic policy derives from inspection of about 50 randomly sampled documents with perfect duplicates: most of them turn out to be of limited or no linguistic interest (e.g., warning messages, copyright statements and the like). While in this way we might also waste relevant content, our guiding principle in web-as-corpus construction is that of privileging precision over recall, given the vastness of the data source.

The contents of all the documents that pass this pre-filtering stage undergo further cleaning based on their contents. First, we need to remove code (HTML and javascript), together with the so-called 'boilerplate', i.e., following Fletcher (2004), all those parts of web documents which tend to be the same across many pages (for instance disclaimers, headers, footers, navigation bars, etc.), and which are poor in human-produced connected text. From the point of view of our target user, boilerplate

[7] The slightly different seed construction strategy used for Italian is not by design. It is an alternative course of action due to different people performing the procedure for different languages at different times.

[8] http://crawler.archive.org/.

identification is critical, since too much boilerplate will invalidate statistics collected from the corpus and impair attempts to analyze the text by looking at KWiC concordances. Boilerplate stripping is a challenging task, since, unlike HTML and javascript, boilerplate is natural language text and it is not cued by special mark-up. We adapted and re-implemented the heuristic used in the Hyppia project BTE tool,[9] which is based on the observation that the content-rich section of a page has a low HTML tag density, whereas boilerplate text tends to be accompanied by a wealth of HTML (because of special formatting, many newlines, many links, etc.). Thus, of all possible spans of text in a document, we pick the one for which the quantity $N(tokens)–N(tags)$ has the highest value. After they are used for the count, all HTML tags and javascript code and comments are removed using regular expressions. Notice in passing that this phase currently removes the links from the text, so we can no longer explore the graph structure of the web document collection.

While resource-free and efficient, the proposed boilerplate stripping method has several limits. Most importantly, it cannot extract discontinuous fragments of connected text; thus, for pages with boilerplate in the middle, depending on the tag density of this middle part, we end up either with only one of the connected text fragments, or (worse) with both, but also the boilerplate in the middle. The heuristic also has problems with the margins of the extracted section, often including some boilerplate at one end and removing some connected text at the other. Recently, more sophisticated supervised boilerplate stripping methods have been proposed as part of the 2007 CLEANEVAL competition (Fairon et al. 2007). However, the unsupervised, heuristic method we are using outperforms all the CLEANEVAL participants in the text-only task of the competition, with a score of 85.41 on average (the best competitor achieves a mean score of 84.07).[10]

Next in the pipeline, the cleaned documents are filtered based on lists of function words (124 items for German, 411 for Italian and 151 for English). Connected text is known to reliably contain a high proportion of function words (Baayen 2001), therefore documents not meeting certain minimal parameters—10 types and 30 tokens per page, with function words accounting for at least a quarter of all words—are discarded. The filter also works as a simple and effective language identifier.

Lastly, pornographic pages are identified and eliminated, since they contain long machine-generated texts, probably used to fool search engines. We create lists of words that are highly frequent in language-specific ad hoc crawls of pornography. A threshold is then set, such that documents containing at least 3 types or 10 tokens from this list are discarded.

In total, the filtering phase took about a week for each corpus.

3.3 Near-duplicate detection and removal

The next step consists in identifying near-duplicates, i.e., documents with substantial overlapping portions. There are several reasons to postpone this to after

[9] http://web.archive.org/web/*/www.smi.ucd.ie/hyppia/; our re-implementation of the Hyppia method is also available for download (see Sect. 5).

[10] These experiments were conducted by Jan Pomikálek, whose contribution we gratefully acknowledge.

corpus cleaning, and in particular after boilerplate stripping. Boilerplate may create both false positives (different documents that share substantial amounts of boilerplate, thus looking like near-duplicates) and false negatives (documents with nearly identical contents that differ in their boilerplate). Also, near-duplicate spotting is computationally costly and hard to parallelize, as it requires comparison of all documents in the collection; thus it is wise to reduce the number and size of documents in the collection first.

We use a simplified version of the 'shingling' algorithm (Broder et al. 1997). For each document, after removing all function words, we take fingerprints of a fixed number of randomly selected n-grams (sequences of n words, where we only consider distinct n-grams, not taking repetitions of the same n-gram into account). Then, for each pair of documents, we count the number of shared n-grams, which should provide an unbiased estimate of the overlap between the two documents (Broder et al. 1997). For pairs of documents sharing more than t n-grams, one of the two is discarded. The pairs are ordered by document ID, and, to avoid inconsistencies, the second document of each pair is always removed. Thus, if the pairs A–B, B–C and C–D are in the list, only document A is kept; however, if the list contains the pairs A–C and B–C, both A and B are kept. Devising efficient ways to identify *clusters* of near-duplicates, rather than pairs, is left to future work.

In constructing the WaCky corpora, 25 5-grams are extracted from each document (these parameters were set based on preliminary experimentation). Near-duplicates are defined as documents sharing at least two 5-grams. The threshold might sound low, yet there are very low chances that, after boilerplate stripping, two unrelated documents will share two sequences of five content words. A quick sanity check conducted on a sample of 20 pairs of documents sharing two 5-grams confirmed that they all had substantial overlapping text.

The near-duplicate detection phase took about 4 days for each corpus.

3.4 Annotation

At this point, the surviving text can be enriched with different types of annotation. In the case of the three WaCky corpora, part-of-speech tagging was performed by the TreeTagger,[11] which also provided lemmatization for the deWaC and ukWaC corpora. Lemmatization of the Italian WaCky corpus was instead performed using the Morph-it![12] lexicon. The annotation phase took about 5 days for each corpus.

In their final versions, the WaCky corpora contain between 1.2 and 1.9 billion tokens, for a total of between 10 and 13 GB of uncompressed data (25–30 GB with annotation). See Table 1 for detailed size information at the different stages (token and type counts are for words that contain nothing but alphabetic characters, apostrophes and dashes). Notice in particular the massive size reduction between the raw crawl output and the final size: in all cases, we discard over 96% of the retrieved data.

[11] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

[12] http://sslmit.unibo.it/morphit.

4 Exploring the WaCky corpora

Automated methods of corpus construction allow for limited control over the contents that end up in the final corpus. The actual corpus composition needs therefore to be investigated through post hoc evaluation methods. The ultimate test of the quality of the WaCky corpora will be how useful they are to researchers in the field. Here, we will first provide a general, mainly quantitative assessment of the overlap of itWaC and ukWaC with relatively large, widely used corpora in the respective languages (we have no comparable resource available for deWaC). We will then present a more detailed qualitative analysis of the nouns most typical of ukWaC when compared to the BNC. See Baroni and Kilgarriff (2006) and Baroni and Ueyama (2006) for further evaluation focusing specifically on deWaC and itWaC, respectively.

4.1 Overlap with reference corpora

We compare ukWaC to the BNC, a well-known, 'balanced' corpus of British English produced in the early nineties and representing a variety of written and (to a lesser extent) spoken registers.[13] For itWaC, our point of reference is the *la Repubblica* corpus,[14] a collection of 16 years of the *la Repubblica* daily. Despite its being single-source, this is widely used as an Italian reference corpus thanks to its size and the variety of newspaper contents. Applying the same filtering method used to obtain the token counts in Table 1, the BNC has 96,868,603 tokens; *la Repubblica* has 326,363,463 tokens.

Table 2 reports the number of distinct nouns, adjectives and verbs in the four corpora (here and below, we use a version of the BNC that has been re-tagged using the TreeTagger, for comparability with ukWaC; similarly, *la Repubblica* has been processed with the same tools used for itWaC). Clearly, the WaCky corpora present much more variety, both in terms of absolute number of types, and of types that occur at least 20 times. The latter (admittedly arbitrary) frequency threshold is selected following John Sinclair's (2005) claim that an experienced lexicographer will need at least 20 instances of a word to be able to build an outline description of its behaviour. Since low frequency statistics will not be of much use to NLP applications either, we take the 'Sinclair cutoff' as a rough way to estimate the number of 'useful' words in a corpus.

The greater type richness of the WaCky corpora is comforting. Yet one could hypothesize that the presence of noise of different kinds might inflate these counts, giving credit to the WaCky corpora for what is ultimately one of their limits. To provide a rough estimate of the level of noise we are likely to encounter among types above the Sinclair threshold in both the traditional and the web corpora, we randomly selected 100 nouns, 100 adjectives and 100 verbs with fq $\geq$ 20 from the lemma lists for ukWaC, the BNC, itWaC and *la Repubblica*. The resulting 600 lemmas with their tags (but no indication of the corpus of provenance) were

[13] http://www.natcorp.ox.ac.uk.

[14] http://sslmit.unibo.it/repubblica.

Table 2 Noun, adjective and verb type counts in ukWaC/BNC and itWaC/*la Repubblica*

Corpus	Noun		Adjective		Verb	
	Total	≥20	Total	≥20	Total	≥20
ukWaC	1,528,839	115,210	538,664	42,526	182,610	15,012
BNC	167,770	21,499	76,463	9,821	23,164	5,760
itWaC	941,990	81,423	706,330	54,414	679,758	50,881
Repub	218,893	27,062	145,200	14,705	140,342	11,511

Table 3 Adjusted type counts for nouns, adjectives and verbs in ukWaC/BNC and itWaC/*la Repubblica*

Corpus	Noun		Adjective		Verb	
	% 'good'	Adjusted ≥ 20	% 'good'	Adjusted ≥ 20	% 'good'	Adjusted ≥ 20
ukWaC	80	92,168	92	39,124	82	12,310
BNC	98	21,069	99	9,723	100	5,760
itWaC	79	64,324	65	35,369	63	32,055
Repub	96	25,980	95	13,970	92	10,590

evaluated in terms of their being *actual* words or noise. For the purposes of this task we define noise as typos, garble or words from foreign language texts—e.g., the verb *woudl*, the adjective *suspensful*, the nouns *scheint* (from German texts) and *poterit* (from Latin texts). Each author rated one random fourth of the English and Italian lemmas. Table 3 shows the percentage of 'good' lemmas found in the samples, as well as adjusted counts obtained by multiplying the raw counts of types with fq ≥ 20 from Table 2 by these percentages.[15] While it is undeniable that the noise rate is larger for the web corpora (and the uneven distribution of noise across languages and parts of speech warrants further analysis in itself), even after this adjustment the web corpora remain a much richer source of types.

At this point, one might legitimately wonder if the words attested in the web corpora are the sort of words (computational) linguists and lexicographers would be typically interested in, rather than, say, web-related terms of limited general interest. To explore this issue, we look at the *overlap* between the WaCky words and those in the BNC and *la Repubblica*. We consider two measures of overlap. The *coverage* of corpus X in corpus Y (*coverage*(*Y*/*X*)) is the proportion of types that are above the Sinclair cutoff *both* in X and Y over the total number of types above the Sinclair cutoff in X. The *enrichment* of corpus X in corpus Y (*enrichment*(*Y*/*X*)) measures the proportion of words that are above the Sinclair threshold in corpus Y but *below* the threshold in corpus X, over the total number of types below the threshold in

[15] Notice that here we are estimating noise based on the type lists only. Clearly, we cannot rule out the possibility that concordances from the web corpora are also less informative due to a larger presence of duplicates. More thorough evaluations of the noise rates are needed to shed light on this issue. On the other hand, depending on the uses we envisage for the corpora, we might in fact be overestimating noise: for instance, non-standard spellings would have linguistic relevance, e.g., if one were interested in grammaticalization and language change, or in collecting statistics to train spellcheckers.

Table 4 Overlap, as measured by percentage coverage and enrichment, for ukWaC versus the BNC and itWaC versus *la Repubblica*

Corpus	Noun		Adjective		Verb	
	Coverage	Enrichment	Coverage	Enrichment	Coverage	Enrichment
ukWaC/BNC	98.4	89.2	99.1	92.3	99.7	95.9
BNC/ukWaC	18.4	0.1	22.9	0.1	38.2	0.1
itWaC/Repub	95.8	70.7	94.9	72.6	93.7	75.2
Repub/itWaC	41.8	0.1	25.6	0.8	21.2	0.8

corpus X (to avoid skewing statistics with too much noise—typos, loanwords, etc.—we only consider words that occur at least 10 times in corpus X). Coverage is a measure of the proportion of words for which we have 'enough' information in corpus X and for which we can also find enough information in corpus Y, and it is thus a (very rough) measure of the extent to which X is 'substitutable' with Y. Enrichment, on the other hand, gives an (equally rough) estimate of the proportion of words, among those attested in X, for which X does not have enough information, but Y does.[16] These statistics are reported for ukWaC/BNC and itWaC *la Repubblica* in Table 4.

The coverage data show that the near totality of BNC content words above the Sinclair cutoff are also above the cutoff in ukWaC, and a very high proportion of *la Repubblica* is covered by itWaC, indicating that the WaCky corpora include most of the vocabulary of these reference corpora. Moreover, the high enrichment values (between 90% and 95% for ukWaC/BNC, and between 70% and 75% for itWaC/*la Repubblica*) suggest that, even within the limits of general-purpose corpus vocabulary, moving to the larger web-based corpora can be very beneficial, with more solid statistics and a larger number of usage examples. Enrichment in the other direction (reference/WaCky), on the other hand, is always well below 1%.

The analysis we just presented relies on the assumption that the instances of shared vocabulary items in the WaCky corpora are (at least) comparable to those in the general corpora in terms of variety and linguistic interest. This is of course not granted. While the issue deserves a more extensive investigation, we present here a lexicography-oriented analysis of a single word (the not so randomly selected adjective *wacky*) that supports our assumption.

The word *wacky* occurs 99 times in the BNC (approximately once per million words) and 3,307 times in ukWaC (1.7 times per million words). Limiting our observations to the common nouns immediately following the adjective, we find

[16] Given two corpora C_+ and C_-, drawn from the same population of words, with C_+ N times the size of C_-, a word occurring n times in C_- should occur about Nn times in C_+. Both WaCky corpora are more than 10 times larger than their reference counterparts. Thus, if the WaCky and reference corpora were random samples from the same populations, enrichment as defined in the text should be trivially at 100% when going from reference to WaCky. However, we know that the WaCky corpora are sampling from rather different populations than the ones of the BNC and *la Repubblica*, and thus the fact that the enrichment proportion is very high is a non-trivial positive result, indicating that, despite the different sample source, the WaCky corpora also contain occurrences of the same words encountered in traditional corpora, in larger proportions than in the latter.

Table 5 *wacky* + Noun in the BNC and ukWaC

BNC	ukWaC		
3 ideas	71 world	21 stuff	11 backy
2 roles	44 ideas	21 races	10 baccy
2 photo	43 wigglers	20 things	10 fun
2 items	42 wiggler	19 idea	10 game
2 humour	28 characters	15 humour	10 inventions
2 characters	27 sense	13 games	10 names
	22 comedy	12 race	10 uses

that the BNC has only six types occurring at least twice in this position, and two of these (*roles* and *photo*) recur in the same text, thus making them suspicious as candidate collocates. As shown in Table 5, and as one would expect, ukWaC offers a much richer set of co-occurring types, each with a substantially higher number of tokens—for reasons of space we are only listing the 21 nouns co-occurring with *wacky* 10 times or more.

The BNC output does include a strong collocate of *wacky* found in several dictionaries, namely *ideas*. However, it provides the same (low) number of examples for the strong collocation *wacky humour* (15 occurrences in ukWaC, plus 22 occurrences of *wacky sense of humour*) and for the rather less well-established *wacky items* and *wacky photo*. The output of ukWaC requires some sifting through to bring to light interesting lexicographic regularities about standard phrases: for instance, *wacky wigglers* (toys) and *wacky races* (a famous cartoon) would be unlikely to make it into a dictionary (while at the same time being interesting information about contemporary English, or at least contemporary web English, of the sort that might be useful, e.g., for named entity extraction). But there is ample evidence about more standard general language expressions that a lexicographer could build upon. Notice that the common collocation *wacky ideas* (the most frequent phrase in the BNC, also appearing on the cover of the *Oxford Collocations Dictionary for Students of English*) occurs 44 times, scoring second after *wacky world* in the ukWaC output. The latter phrase is worth considering briefly, since it forms the core of the extended unit of meaning (Sinclair 1996) summarized in Table 6.

As Table 6 suggests, *wacky world* occurs as part of a longer expression referring to the act of *getting acquainted* with a given domain (*the wacky world of x*), and

Table 6 The *wacky world* unit of meaning

welcome to		pyramids
discover		religion
embrace		Windows
join	the wacky world of	graduate recruitment
explore		Douglas Adams
step in		Brit art
dive right into		automata
get involved in		Holliwoodland

often taking the form of an imperative utterance having the illocutionary force of an invitation.

One last observation can be made on these data. The ukWaC output contains the expressions *wacky baccy* (10) and *wacky backy* (11). The former (one occurrence in the BNC) was included in the *Oxford English Dictionary* in 2002, and described as a slang term for marijuana, while the latter is not attested in the BNC nor in the OED, though slang dictionaries on the web give it as a synonym of *wacky baccy*.[17]

Clearly, not all evidence offered by a (very large) corpus is likely to be relevant or interesting for a corpus user. Yet the observations sketched out in this section suggest that a corpus like ukWaC can provide rich, up-to-date language data on even relatively infrequent words. This evidence is in line with that provided by lexicographic resources and the BNC, but it comes on a larger scale.

4.2 Nouns in ukWaC and the BNC: a wordlist comparison

While in the previous section we focused on what WaCky corpora have in common with standard reference corpora, here we use vocabulary-based corpus comparison methods to look at the way in which they differ.

Separate lists of nouns were created for ukWaC and the BNC, applying a filtering method similar to that used to obtain the counts in Table 1 (for this task, words were also lower-cased and those containing apostrophes and dashes were discarded). These lists were then compared via the log-likelihood association measure (Dunning 1993).[18] Relying on the tagger's output, the procedure makes it possible to identify the word items tagged as nouns that are most typical of the two corpora when compared to each other (in the statistical sense that they occur in one or the other more often than one would expect by chance).

For each of the 50 words with the highest log-likelihood ratio, 250 randomly selected concordances were retrieved and analyzed.

Based on their contexts of use, the nouns that turn out to be the most typical of ukWaC when compared to the BNC belong to three main semantic domains (see Table 7 for some examples), i.e., (a) computers and the web, (b) education, and (c) what may be called 'public sphere' issues. In category (a) we find words like *website*, *link*, and *browser*. These nouns are distributed across a wide variety of text types, ranging from online tutorials to promotional texts introducing, e.g., a web-based service. Unsurprisingly, a word which has become part of everyday language like *website* does not appear at all in the BNC, which was constructed in the early nineties.

The analysis of the concordances for nouns belonging to category (b) (e.g., *students*, *research*), and (c) (e.g., *organisations*, *nhs*, *health*), for each of which the associated URL was also checked, suggests that their (relatively) high frequency can be explained by the considerable presence in ukWaC of certain entities responsible for the publishing of web contents. These are either universities—in the case of (b)—or non-governmental organizations or departments of the government—in the

[17] http://www.urbandictionary.com/define.php?term=wacky+backy.

[18] Full lists are available from the WaCky site (see Sect. 5). A more extensive analysis, that also covers adjectives, verbs and function words, is presented in Ferraresi (2007).

Table 7 Examples of nouns typical of ukWaC and the BNC by semantic domain

ukWaC			
Web and computers		Education	Public sphere issues
website	link	students	services
site	data	skills	organisations
click	download	project	nhs
web	file	research	health
email	browser	projects	support
BNC			
Imaginative		Spoken	Politics and economy
eyes	door	er	government
man	house	cos	recession
face	hair	sort	plaintiff
mother	smile	mhm	party

case of (c). Typical topics dealt with in these texts are on the one hand education and training and, on the other, public interest issues, such as assistance for citizens in need. What is most remarkable is the variety of the text genres which are featured. As pointed out by Thelwall (2005), academic sites may contain very different types of texts, whose communicative intention and register can differ substantially. We find 'traditional' texts, like online prospectuses for students and academic papers, as well as 'new' web-related genres like homepages of research groups. In the same way, the concordances of a word like *nhs* reveal that the acronym is distributed across text types as diverse as newspaper articles regarding quality issues in the services for patients and forum postings on the treatment of diseases.

The nouns most typical of the BNC compared to ukWaC can also be grouped into three macro-categories (examples are provided in Table 7), i.e., (a) nouns related to the description of people or objects, (b) expressions which are frequent in the spoken language (or, more precisely, typical transcriptions of such expressions), and (c) words related to politics, economy and public institutions. The words included in category (a) are names of body parts, like *eyes*, and *face*; words used to refer to people, such as *man* and *mother* and names of objects and places, like *door*, and *house*. All of these share the common feature of appearing in a clear majority of cases in texts classified by Lee (2001) as 'imaginative' or 'fiction/prose'. As an example, *eyes* appears 74% of the times in 'fiction/prose' texts, and *man* appears in this type of texts almost 41% of the times. In general, what can be inferred from the data is that, compared to ukWaC, the BNC seems to contain a higher proportion of narrative fiction texts, confirming that "texts aimed at recreation [such as fiction] are treated as an important category in traditional corpora" (Sharoff 2006, p. 85), whereas they are rarer in web corpora. This may be due to the nature of the web itself, since copyright restrictions often prevent published fiction texts from being freely available online.

Category (b) includes expressions which are typically associated with the spoken language, including graphical transcriptions of hesitations, backchannels and reduced forms. Among these we find *er*, *cos*, *mhm*, which appear most frequently in the spoken part of the BNC. These words are clearly not nouns. However, since the same tagging method was applied to the two corpora, it is likely that they really are more typical of the BNC, inasmuch as their relatively higher frequency cannot be accounted for by differences in tagger behavior. A noun like *sort* is also frequently featured in the spoken section of the BNC, being often found in the expression 'sort of'. Spoken language is obviously less well represented in ukWaC than in the BNC, which was designed to contain 10% transcribed speech.

The last group of words (c) which share important common traits in terms of their distribution across text genres and domains is that of words associated with politics, economy and public institutions. Examples of these nouns are *government*, *recession* and *plaintiff*. All of these are mainly featured in BNC texts that are classified as belonging to the domain 'world affairs', 'social sciences' or 'commerce', and occur both in academic and non-academic texts. As a category, this seems to overlap with the group of words related to public sphere issues which are typical of ukWaC. However, the specific vocabulary differs because the texts dealing with politics and economy in ukWaC seem to share a broad operative function, e.g. offering guidance or promoting a certain governmental program, as in the following examples:

OGC offers advice, guidance and **support**.
Local business **support** services include the recently established Sussex Business Link
... use Choice Advisers to provide practical **support** targeted at those parents most likely to need extra-help

Concordances reveal instead that in the BNC words like *government* or *recession* are more frequently featured in texts which comment on a given political or economic situation, as e.g., newspaper editorials would do, for example:

... is urging the **government** to release all remaining prisoners of conscience
Despite assurances from **government** officials that an investigation is underway
... a crucial challenge to the cornerstone of his **government**'s economic policy ...

The analysis in this section has highlighted several lexical differences between ukWaC and the BNC. Taking these as hints of differences at the level of text types and topics in the two corpora, the BNC seems to be more varied, including a comparatively higher number of instances of text types as diverse as fiction, newspaper and spoken interaction. The ukWaC corpus is characterized by a lesser degree of internal variety, at least in terms of 'typical' topics (i.e., the web, education and public sphere issues). On the other hand, it is clearly more up-to-date (as one would expect), including text types which are absent from the BNC (e.g., web-based genres), thus making a valuable candidate resource for studying contemporary English.

One last caveat. In this vocabulary-based comparison, log-likelihood scores were used to evaluate *relative* typicality in one corpus or the other. The noun *eyes*, for

instance, appears as the 4th most typical noun of the BNC, even though its absolute frequency is nearly 15 times lower than in ukWaC. Thus, the fact that a word is typical of the BNC does not imply that it is not equally well represented in ukWaC—recall that ukWaC covers a huge portion of the BNC vocabulary. Moreover, the method highlights asymmetries in the two corpora, but it conceals features that make them similar (represented by words that have a log-likelihood value close to 0). In future work, we intend to determine what kinds of text types or domains do *not* turn up as typical of either ukWaC or the BNC, and assess whether there is ground to conclude that they are similarly represented in both corpora.

5 Availability and format

The WaCky website[19] provides access to all the data and tools used in corpus construction (including lists of seed pairs and URLs) and other resources, such as unigram and bigram frequency lists. The website also provides contact information for researchers interested in obtaining the corpora.

The copyright issue remains a thorny one: there is no easy way of determining whether the content of a particular page is copyrighted, nor is it feasible to ask millions of potential copyright holders for usage permission. However, our crawler does respect the download policies imposed by website administrators (i.e. the robots.txt file), and the WaCky website contains information on how to request the removal of specific documents from our corpora. Lastly, it must be noted that we offer highly processed versions of the web pages we download, in a format unlikely to be usable by non-linguists or for non-research purposes.

Corpora are encoded in the simple 'pseudo-XML' format illustrated in Fig. 1. This format is ready for indexing with the IMS Open Corpus WorkBench,[20] the tool we use to access the WaCky corpora. All three corpora are also available for online searching via the commercial Sketch Engine.[21]

6 Conclusion: directions for further work

We are already actively using the WaCky corpora in various projects, ranging from simulations of human learning to pedagogical lexicography and terminology. In turn, these activities will give us a clearer idea of the corpora's strengths and limits.

We believe that the most pressing issue at this moment is the need to provide a free web-based interface to the corpora, that should allow user-friendly access to those without advanced technical skills (e.g., language learners), as well as support linguists in doing extensive qualitative and quantitative research with the corpora (including the possibility of saving settings and results across sessions). We are actively working in this area.

[19] http://wacky.sslmit.unibo.it.

[20] http://cwb.sourceforge.net/.

[21] http://www.sketchengine.co.uk/.

```
<text url="http://source.site.uk/thisdoc.html">
<s>
This      DT    this
is        VBZ   be
a         DT    a
sentence  NN    sentence
.         SENT  .
</s>
</text>
```

Fig. 1 Sample WaCky format

A second important line of research pertains to automated cleaning of the corpora, and to the adaptation of tools such as POS taggers and lemmatizers—that are often based on resources derived from newspaper text and other traditional sources—to web data. Moreover, corpora should be enriched with further layers of linguistic annotation. To this effect, we recently finished parsing ukWaC with a dependency parser and we are currently investigating the best way to make these data available.

Of course, further evaluation should also be conducted, including a comparison of deWaC with other German resources, and comparison of the WaCky corpora to other web-derived corpora (including Google's terabyte n-gram collection).

Finally, the method we described can be easily re-implemented to construct comparable corpora in other languages—we have recently taken the first steps towards the construction of corpora for Spanish and French, and we hope that other researchers will join us in setting up what promises to be a pool of language resources among the largest ever made available to the research community.

Acknowledgements We would like to thank the members of the World Wide WaCky community for many useful interactions, in particular: Sara Castagnoli, Tom Emerson, Stefan Evert, Bill Fletcher, Federico Gaspari, Adam Kilgarriff, Jan Pomikálek and Serge Sharoff. We would also like to thank the LREJ reviewers for useful comments.

References

Baayen, A. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Baroni, M., & Bernardini, S. (Eds.). (2006). *Wacky! Working papers on the web as corpus*. Bologna: Gedit.

Baroni, M., & Kilgarriff, A. (2006). Large linguistically-processed web corpora for multiple languages. In *Proceedings of the 11th conference of the European chapter of the association for computational linguistics*, Trento, Italy, pp. 87–90.

Baroni, M., & Ueyama, M. (2006). Building general- and special-purpose corpora by web crawling. In *Proceedings of the 13th NIJL international symposium, language corpora: Their compilation and application*, Tokyo, Japan, pp. 31–40.

Boleda, G., Bott, S., Meza, R., Castillo, C., Badia, T., & López, V. (2006). *CUCWeb: A Catalan corpus built from the web*. In Kilgarriff and Baroni (2006), pp. 19–26.

Brants, T., & Franz, A. (2006). *Web 1T 5-gram, version 1*. Philadelphia: Linguistic Data Consortium.
Broder, A., Glassman, S., Manasse, M., & Zweig, G. (1997). Syntactic clustering of the web. In *Proceedings of the sixth international world wide web conference*, Santa Clara, California, pp. 391–404.
Ciaramita, M., & Baroni, M. (2006). Measuring web corpus randomness: A progress report. In Baroni and Bernardini (2006), pp. 127–158.
Clarke, C., Cormack, G., Laszlo, M., Lynam, T., & Terra, E. (2002). The impact of corpus size on question answering performance. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval*, Tampere, Finland, pp. 369–370.
Clarke, C., Craswell, N., & Soboroff, I. (2005). The TREC terabyte retrieval track. *SIGIR Forum, 39*(1), 25.
Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61–74.
Emerson, T., & O'Neil, J. (2006). *Experience building a large corpus for Chinese lexicon construction*. In Baroni and Bernardini (2006), pp. 41–62.
Fairon, C., Naets, H., Kilgarriff, A., & de Schryver, G.-M. (Eds.). (2007). Building and exploring web corpora. In *Proceedings of the 3rd web as corpus workshop, incorporating Cleaneval*. Louvain: Presses Universitaires de Louvain.
Ferraresi, A. (2007). *Building a very large corpus of English obtained by web crawling: ukWaC*. MA Dissertation, University of Bologna. Retrieved January 28, 2008, from http://wacky.sslmit.unibo.it
Fletcher, W. (2004). *Making the web more useful as a source for linguistic corpora*. In U. Connor & T. Upton (Eds.), *Corpus linguistics in North America 2002* (pp. 191–205). Amsterdam: Rodopi.
Hundt, M., Nesselhauf, N., & Biewer, C. (Eds.). (2007). *Corpus linguistics and the web*. Amsterdam: Rodopi.
Kilgarriff, A., & Baroni, M. (Eds.). (2006). *Proceedings of the 2nd international workshop on the web as corpus*. East Stroudsburg, PA: ACL.
Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics, 29*(3), 333–347.
Kornai, A., Halácsy, P., Nagy, V., Oravecz, C., Trón, V., & Varga, D. (2006). *Web-based frequency dictionaries for medium density languages*. In Kilgarriff and Baroni (2006), pp. 1–8.
Lee, D. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology, 5*(3), 37–72.
Liu, V., & Curran, J. (2006). Web text corpus for natural language processing. In *Proceedings of the 11th conference of the European chapter of the association for computational linguistics*. Trento, Italy, pp. 233–240.
Santini, M., & Sharoff, S. (Eds.). (2007). *Proceedings of the CL 2007 colloquium: Towards a reference corpus of web genres*, Birmingham, UK.
Shaoul, C., & Westbury, C. 2007. A USENET corpus (2005–2007). Retrieved January 28, 2008, from http://www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html
Sharoff, S. (2006). *Creating general-purpose corpora using automated search engine queries*. In Baroni and Bernardini (2006), pp. 63–98.
Sinclair, J. McH. (1996). The search for units of meaning. *Textus 9*(1), 71–106.
Sinclair, J. McH. (2005). Corpus and text—Basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1–16). Oxford: Oxbow Books.
Thelwall, M. (2005). Creating and using web corpora. *International Journal of Corpus Linguistics, 10*(4), 517–541.
Ueyama, M. (2006). *Evaluation of Japanese web-based reference corpora: Effects of seed selection and time interval*. In Baroni and Bernardini (2006), pp. 99–126.