Introducing and evaluating ukWaC, a very large web-derived corpus of English

Adriano Ferraresi,[*] Eros Zanchetta,[*] Marco Baroni,[†] Silvia Bernardini[*]

[*] SITLeC – University of Bologna (Forlì)
Corso Diaz 64, 47100 Forlì – Italy
adriano@sslmit.unibo.it, eros@sslmit.unibo.it, silvia@sslmit.unibo.it

[†] CIMeC – University of Trento
Corso Bettini 31, 38068 Rovereto (TN) – Italy
marco.baroni@unitn.it

Abstract

In this paper we introduce ukWaC, a large corpus of English constructed by crawling the `.uk` Internet domain. The corpus contains more than 2 billion tokens and is one of the largest freely available linguistic resources for English. The paper describes the tools and methodology used in the construction of the corpus and provides a qualitative evaluation of its contents, carried out through a vocabulary-based comparison with the BNC. We conclude by giving practical information about availability and format of the corpus.

1. Introduction

This article introduces ukWaC, a very large (>2 billion words) corpus of English, and presents an evaluation of its contents. UkWaC was built by web crawling, contains basic linguistic annotation (part-of-speech tagging and lemmatization) and aims to serve as a general-purpose corpus of English, comparable in terms of document heterogeneity to traditional "balanced" resources. Since the aim was to build a corpus of British English, the crawl was limited to the `.uk` Internet domain. The corpus is, to the best of our knowledge, among the largest resources of its kind, and the only web-derived, freely available English resource with linguistic annotation. It was created in 2007 as part of the WaCky project, an informal consortium of researchers interested in the exploration of the web as a source of linguistic data.

The aims of this article are to introduce ukWaC to the community of linguistic researchers, to describe the procedure that was followed in constructing it and to provide some preliminary evaluation of its contents.

The article is structured as follows: in Section 2. we outline the corpus creation process. In Section 3. we carry out a vocabulary-based comparison between ukWaC and the British National Corpus (BNC), which sheds light on the main differences between the two corpora. Section 4. deals with issues related to format and availability of the corpus. Finally, Section 5. briefly discusses previous work on web corpora, and Section 6. hints at what we consider the most pressing next steps of the WaCky initiative.

2. Corpus construction

The procedure described in this Section was carried out on a server running Fedora Core 3 with 4 GB RAM, 2 Dual Xeon 4.3 GHz CPUs and 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 at the end of the Section.

2.1. Seed selection and crawling

Our aim was to set up a resource comparable to more traditional general language 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, blogs, or postings in forums. It should be noted that our goal here was for the corpus to be representative of the language of interest, rather than being 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.

The first step was identifying sets of seed URLs which ensured variety in terms of content and genre. In order to find these, random pairs of randomly selected content words in the target language were submitted to Google. The queries were formed by two-words tuples because preliminary experimentation found that single word queries tend to yield potentially undesirable documents (e.g., dictionary definitions of the queried words), 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. Issuing queries with words sampled from a basic vocabulary list, on the contrary, tends 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 corpus, relevant sources have been chosen accordingly.

Three sets of queries were generated: the first set (1,000 word pairs) was obtained by combining mid-frequency content words randomly selected from the BNC; function words were excluded from the list, since search engines usually ignore them when submitted as part of a query. The second list of queries (500 word pairs) was obtained by randomly combining words sampled from the spoken section of the BNC, while the third list (500 word pairs) was generated from a vocabulary list for foreign learners of English[1] which (however counter-intuitively) contains rather formal vocabulary, possibly required for academic study in English. Once the various seed words had been identified, they were paired randomly before submission to Google.

A maximum of ten seed URLs were retrieved for each random seed pair query, and the retrieved URLs were collapsed in a single list. Duplicates were discarded and, to ensure maximal sparseness, only one (randomly selected) URL for each (normalized) domain name was kept. The remaining URLs were fed to a crawler in random order. The crawl was limited to pages in the `.uk` web domain whose URL does not end in a suffix cuing non-html data (`.wav`, `.jpg`, etc.). The rationale for the choice of limiting the crawl to `.uk` pages is that our goal was to construct a relatively homogeneous resource, comparable to the BNC, and because of practical issues arising when trying to define the country domains to crawl (i.e., including or excluding countries in which English is an official, though not a native language), as well as how to treat U.S. pages (relying on the `.us` domain would provide a very skewed sample of texts). Our strategy does not, of course, ensure that all the pages retrieved represent British English.

The crawl was performed using the Heritrix[2] crawler and was stopped after 10 days. The full seed pair and seed URL lists are available from the project page (see Section 4.).

2.2. Post-crawl cleaning

Using information in the Heritrix logs, we only preserved documents that were of mime type `text/html`, and between 5 and 200KB in size. As observed by Fletcher (2004) and confirmed by informal experimentation, very small documents tend to contain little genuine text (5KB 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 identified and removed all documents that had perfect duplicates in the collection (i.e., we did not keep even one instance of a set of identical documents). Decision to apply this drastic policy followed inspection of about fifty randomly sampled documents with perfect duplicates: most of them turned 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 have wasted relevant content, the guiding principle in our web-as-corpus construction approach is that of privileging precision over recall, given the vastness of the data source.[3] The contents of all the documents that passed this prefiltering stage underwent further cleaning based on their contents. First, we had 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, navigation bars, etc.), and which are poor in human-produced connected text. From the point of view of our target user, boilerplate 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 markup. We adapted and re-implemented the heuristic used in the Hyppia project BTE tool,[4] 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, etc.). Thus, of all possible spans of text in a document, we pick the one for which the quantity $N(tokens) - N(tags)$ takes the highest value. After they are used for the count, all html tags and javascript code and comments are removed using regular expressions.

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 – see algorithms described in 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).[5]

Next in the pipeline, the cleaned documents were filtered based on a list of 151 function words. Connected text is known to reliably contain a high proportion of function words (Baayen, 2001), therefore documents not meeting certain minimal parameters – ten types and thirty tokens per page, with function words accounting for at least a quarter of all words – were discarded. The filter also works as a simple and effective language identifier.

Lastly, pornographic pages were identified and eliminated, since they tend to contain long machine-generated texts, probably used to 'trap' search engines. Lists were created of words that are highly frequent in ad-hoc crawls of pornography websites. A threshold was then set, such that documents containing at least 3 types or 10 tokens from this

[1] `http://wordlist.sourceforge.net/`

[2] `http://crawler.archive.org/`

[3] This is also the reason for excluding such documents as `.pdf` and `.doc` files from the crawl (cf. Section 2.1.). It is true that these documents may contain interesting linguistic materials, but, on the negative side, they require ad-hoc post-processing techniques, and we are not aware of reliable tools for converting them to text files. We plan to tackle this issue in future work.

[4] `http://www.smi.ucd.ie/hyppia/`

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

list were discarded.

2.3. Near-duplicate detection and removal

The next step consisted in identifying near-duplicates, i.e., documents with substantial overlapping portions. There are several reasons to postpone this to after 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 took fingerprints of a fixed number `s` of randomly selected n-grams (sequences of `n` words; we counted types, not tokens – i.e., we only looked at distinct n-grams, and we did not take repetitions of the same n-gram into account). Then, for each pair of documents, we counted the number of shared n-grams, which can be taken to 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 was discarded. The pairs were ordered by document ID, and, to avoid inconsistencies, the second document of each pair was always removed. Thus, if the pairs A-B, B-C and C-D were in the list, only document A was kept; however, if the list contained the pairs A-C and B-C, both A and B were kept. The devising of efficient ways to identify *clusters* of near-duplicates, rather than pairs, is left to future work.

In constructing ukWaC, 25 5-grams were extracted from each document, based on preliminary experimentation. Near-duplicates are defined as documents sharing as few as two of these 5-grams. This threshold might sound surprisingly 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 twenty pairs of documents sharing two 5-grams confirmed that they all had substantial overlapping text. The near-duplicate detection phase took about four days.

2.4. Annotation

At this point, the surviving text could be enriched with different types of annotation. Part-of-speech tagging and lemmatization was performed by the TreeTagger.[6] The annotation phase took about five days.

In its final, annotated version, ukWaC contains 1.9 billion tokens, for a total of 12 GB of uncompressed data (30 GB with annotation). See Table 1 for detailed size information at the different stages.

n of seed word pairs	2,000
n of seed URLs	6,528
raw crawl size	351 GB
size after document filtering	19 GB
n of documents after filtering	5.69 M
size after near-duplicate cleaning	12 GB
n of documents after near-duplicate cleaning	2.69 M
size with annotation	30 GB
n of tokens	1,914,150,197
n of types	3,798,106

Table 1: Size data for ukWaC.

3. Evaluating ukWaC through wordlist comparisons

When corpora are built through automated procedures, as is the case for ukWaC, there is limited control over the contents that make up the final version of the corpus. Post-hoc evaluation is therefore needed to appraise actual corpus composition. Along the lines of Sharoff (2006) (cf. Section 5.), here we provide a qualitative evaluation of our web corpus based on a vocabulary comparison with the widely used BNC. A mostly quantitative evaluation of the overlap of ukWaC and the BNC in terms of lexis is presented in Baroni et al. (2008).

Separate wordlists of nouns, verbs and adjectives were created for the two corpora, which were then compared via the log-likelihood association measure.[7] This makes it possible to identify the words that are most characteristic of one corpus when compared to the other (Rayson and Garside, 2000). Since the procedure relies on the tagger's output, it should be noted that the version of the BNC used was retagged using the same tools as ukWaC, so as to minimize differences in the wordlists that would be due to different annotation procedures.

For each of the 50 words with the highest log-likelihood ratio, 250 randomly selected concordances were retrieved and analyzed. In the following Sections the results of the analysis are presented.

3.1. Nouns

The nouns most typical of ukWaC when compared to the BNC belong to three main semantic areas (see Table 2 for some examples), i.e., (a) computers and the web, (b) education, and (c) what may be called 'public sphere' issues.[8] Category (a) groups words like *website*, *email*, and *software*. If we analyse the contexts in which such words appear, it can be noticed that they are distributed across a wide variety of text types, ranging from online tutorials to promotional texts introducing, e.g. a web-based service. This

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

[7] Full lists are available from the 'download' section of the WaCky site (see Section 4.). For further details on the wordlist creation and for more detailed analysis see Ferraresi (2007).

[8] Here and below, the analysis does not take into account words typically featured in 'boilerplate' sections of web pages. An example of such words is *information*, which frequently occurs within expressions like "for more" or "for further information".

may be seen as a welcome finding, for at least two distinct reasons. First, no one-to-one correspondance is observed between a topic and a text typology (it could have been possible that, e.g., software instruction manuals emerged as a preponderant text type). Second, a corpus like ukWaC could be used to study the usage of relatively 'new' words, such as those produced within the constantly changing field of new technologies, and that are unattested in traditional corpora. As an example of this, a word like *website* does not appear at all in the BNC.

ukWaC		
Web and computers	Education	Public sphere issues
website email link software	students skills project research	services organisations nhs support
BNC		
Imaginative	Spoken	Politics and economy
eyes man door house	er cos sort mhm	government recession plaintiff party

Table 2: Examples of nouns typical of ukWaC and the BNC grouped according to their semantics.

The analysis of the concordances and associated URLs for nouns belonging to category (b) (e.g., *students*, *research*), and (c) (e.g., *organisations*, *nhs*, *support*) 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 universities – in the case of (b) – and non-governmental organizations or departments of the government – in the 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. The variety of the text genres which are featured is especially remarkable. 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-based 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[9] compared to ukWaC can also be grouped into three macro-categories (examples are provided in Table 2), i.e., (a) nouns related to the description of people or objects, (b) markers of orality (or, more precisely, typical transcriptions of such words), and (c) words related to politics, economy and public institutions. The words included in category (a) are names of body parts, like *eyes*; words used to refer to people, such as *man*, 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 *door* appears in this type of texts almost 62% 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 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". As could be expected, spoken language is less well represented in ukWaC than in the BNC, since the latter was specifically 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*, *plaintiff* and *party*. 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.

[9] As can be noticed in Table 2, some of the words taken into account here are not nouns (e.g. *er*), but rather expressions which were erroneously recognized by the tagger as nouns.

3.2. Verbs

In Section 3.1. the nouns most characteristic of ukWaC and the BNC were grouped and analysed on the basis of the text domains and types they typically appear in. Identifying a similar relationship between textual domains and verb forms is somewhat less immediate, since verbs are often less easily associated with, e.g., a particular text topic. Alternatively, one can adopt a semantic classification based on their core meaning – here we follow that proposed by Biber et al. (1999) –, and assess whether verbs belonging to the same class show similar distributional patterns across, e.g., textual types. Another important aspect which is taken into account in the analysis of verbs is that of verb tenses, which, as we shall see, can provide further indications about the texts that characterize the two corpora.

ukWaC	
Activity verbs	Verbs of facilitation
use develop provided visit	help support improve ensure
BNC	
Activity verbs	Mental verbs
looked nodded go shrugged	know mean thought saw

Table 3: Examples of verbs typical of ukWaC and the BNC by semantic category.

A clear majority of the 50 verb forms most typical of ukWaC when compared to the BNC can be classified either as "activity verbs", i.e. verbs which "denote actions and events that could be associated with choice" (*ibid.*, p. 361), such as *use*, *provide* and *work*, or as "verbs of facilitation or causation", i.e. verbs that "indicate that some person or inanimate entity brings about a new state of affairs", such as *help* and *allow* (other examples can be found in Table 3). Taken together, the verb forms belonging to these two categories account for almost 50% of the verbs most characteristic of ukWaC. Their distribution across text types, however, seems to differ.

Activity verbs are evenly distributed across the main text types identified in Section 3.1., i.e. promotional texts – issued both by private companies and governmental departments and universities –, "discussion texts",[10] such as news articles and postings in forums, and "instruction texts", like help pages and instruction manuals. Here are some examples:

> Specifically created to perform research and to **develop** future leaders for aerospace manufacturing;

> ...children are dying of AIDS. It challenges all religions to **work** together to reduce the stigma;

> When you **visit** a web page, a copy of that page is placed in the cache.

As could be expected, verbs of causation, on the contrary, show a distinct tendency to appear in only two of these text types, i.e. instruction and promotional texts. In promotional texts, in particular, verbs of causation are used to convince readers that a certain product, service or idea can actually make a difference, as in the following sentence:

> Acas aims to **improve** organisations and working life through better employment relations.

It is interesting to notice in this respect that many texts in ukWaC are not easily classifiable as belonging to one single category. This is the case, e.g., for seemingly instructional texts, which actually also promote the product they are describing. Thus, a sentence like:

> Once again, we can **help** with any queries you may have. Products liability insurance will cover...

published on the help page of an insurance company can hardly be seen as having a merely informative function. This corresponds to what Santini (2007) calls "genre hybridism", which often makes it especially difficult to classify web text into clear-cut genre categories.

If verb tenses are taken into account, it can be noticed that most verbs in the list are in the present tense (or in their base form), and that those which could appear as past forms are, in fact, often used as past participles in passive forms. This could be due to the already noted considerable importance in ukWaC of discussion texts, which are typically concerned with current affairs, or of promotional and instruction texts, which often make use of the imperative form.

Verb forms in the BNC belong to two main semantic categories, i.e. activity verbs, like *looked* and *go*, and "mental verbs", i.e. verbs that "denote a wide range of activities and states experienced by humans, [...] do not involve physical action and do not necessarily entail volition" (Biber et al., 1999, p. 362), like "know" and "thought" (see Table 3 for other examples).

The verbs belonging to these two categories show very similar distributional patterns: verbs in the past form occur most frequently in imaginative/fiction texts, whereas present tense forms are most frequently featured in the spoken section of the corpus. As regards this point, notice that activity verbs in the BNC – which usually indicate a physical action, e.g. of a character in fiction (cf. *nodded*, *shrugged*) – seem to be less evenly distributed across text types than activity verbs in ukWaC. As an example, the past tense form *looked* appears 67% of the times in fiction texts, and *nodded* 94% of the times. As already mentioned, present tense forms – which, however, are a minority in the list, accounting for less than 15% of the total number of verbs analysed – are instead most frequent in spoken language. The verb form *go*, e.g., appears 36% of the times in spoken texts (26% of the times in fiction texts), and the mental verb *know* occurs in such texts almost 50% of the times.

[10] The terminology used to classify web texts is taken from Sharoff (2006).

Summing up, the (relatively) high frequency of activity and mental verbs in the BNC can be explained by their being frequently used within two text types, i.e. fiction and spoken texts. Moreover, when verb tenses are also taken into account, the BNC, unlike ukWaC, seems to be characterized by past-oriented (narrative) language.

3.3. Adjectives

The adjectives most typical of ukWaC can be classified as belonging to four semantic areas, i.e. (a) web-related adjectives, (b) public sphere-related adjectives, (c) time-related adjectives, and (d) emphatic adjectives conveying a positive evaluation (see Table 4 for examples). As can be noticed classes (a) and (b) correspond to two of the main text topics identified in Section 3.1., thus confirming that such topics are well represented within ukWaC.

ukWaC			
Web	Public sphere	Present time	Emphatic
online digital mobile	sustainable global disabled	new current innovative	excellent fantastic unique
BNC			
Imaginative	Politics	Present time	Sciences
pale dark afraid	political soviet conservative	last former nineteenth	gastric colonic ulcerative

Table 4: Examples of adjectives typical of ukWaC and the BNC grouped according to their semantics.

Both adjectives belonging to class (a) and (b) show distributional patterns similar to those of their noun "counterparts". Adjectives like *online* and *digital* can be found in technical instruction texts, such as tutorials and user manuals; in discussion pages, like blogs, and in promotional texts about computing-related services. Similarly, adjectives like *sustainable* and *global* typically occur in texts created by departments within the government and NGOs, or in various kinds of promotional or discussion texts, such as texts promoting a political (or humanitarian) program, or news. Topics in ukWaC thus seem to correspond to a certain extent to current themes of discussion (such as "global economy" and "sustainable growth"). This, however, is also true for the BNC, in which two of the most typical adjectives compared to ukWaC are *soviet* and *cold*. Such datum is likely to reflect the importance that such themes as the "Soviet Union" and the "Cold War" – which are among the most frequent bigrams including these adjectives – had at the time of the corpus construction.

Category (c) includes adjectives referring to present time, or signalling a change with respect to the past, like, e.g., *new* and *current*. The presence of such adjectives may be seen as also connected with the high frequency of verbs in the present tense. Taken together, these two features seem to point at the fact that the web texts in ukWaC are typically both focused on the present time and willing to signal it explicitly. This is notably true for press releases and promotional pages. In the latter type of texts, adjectives which signal a radical change with respect to the past (e.g. *innovative*) are particularly used to display how original and innovative a service or product is.

The presence of a considerable number of promotional texts is also revealed by the high frequency of adjectives which are chiefly used to indicate positive characteristics (category (d)), like *excellent*, *fantastic*, and *unique*. All of these are mainly found, e.g., in descriptions of products, services or tourist attractions, as in the following example:

```
   ...your stay in Cornwall. Fantastic views
across the ocean and countryside.
```

The adjectives most typical of the BNC when compared to ukWaC can also be classified into four main semantic areas, i.e. (a) adjectives used to describe people and objects, (b) politics-related adjectives, (c) adjectives related to past time, and (d) science-related adjectives (see examples in Table 4).

In the case of the BNC too, classes (a) and (b) correspond to two of the categories identified in Section 3.1. In category (a) we find adjectives that refer to physical characteristics of people (e.g. *pale*, *tall*), or of inanimate objects and settings in which an action takes place (e.g. *dark*, *thick*), and others that relate to people's temper (e.g. *anxious*, *angry*). As could be expected, all of these are most frequently found in imaginative texts. Adjectives belonging to category (b) include "general", hypernymic adjectives (e.g. *political*, *social*), and adjectives which designate national provenance (*soviet*, *french*) or refer to political parties (*conservative*). These are typically found in three domains, i.e. "world affairs", "social sciences" and "commerce" (Lee, 2001). As was noted in Section 3.1., this category of words seems to overlap with that of public-sphere issues identified in ukWaC. Concordances of politics-related adjectives, however, confirm that texts in which the two categories of adjectives occur differ: public sphere-related texts in ukWaC are often concerned with matter-of-fact issues (like, e.g., offering *support* to disabled people), and are mainly focused on the present (cf. Section 3.2.). Texts related to politics in the BNC, on the contrary, seem to describe events through general, abstract categories (e.g. *political*), and to report facts in the past time (cf. Section 3.2. and Section 3.1. for some examples).

In this regard, it is interesting to notice that, unlike in ukWaC, the adjectives most typical of the BNC relating to time refer to the past (category (c)), like, e.g., *last*, *former*, and *nineteenth* (whose most frequent collocate is *century*). These are mainly found in two text domains, i.e. world affairs and social sciences. Their frequency in these text types may be seen as confirming that texts about politics and economics in the BNC seem to adopt a retrospective, historical approach to facts, as is typical, e.g., of academic and journal articles.

Finally, adjectives belonging to category (d) are related to natural and applied sciences. Words like *gastric*, *colonic*, and *ulcerative* are often found in academic and non-academic essays which deal with anatomy or health problems (medicine). A closer look at the adjectives reveals that several refer to the digestive system. It seems therefore likely that the BNC contains a higher proportion of essays on the specific topic of human or animal digestion than ukWaC (cf. also Kilgarriff and Grefenstette (2003)). In turn, this could be interpreted as a sign of the relative weight that even a few texts can have on a (not so small) corpus like the BNC.

3.4. Discussion

In the present Section a method was presented to provide an evaluation of ukWaC's contents. The method involved constructing different lists of nouns, verbs, and adjectives. The same procedure was carried out on the BNC, and the lists were subsequently compared across the two corpora via the log-likelihood association measure. This made it possible to find the words that are comparatively more frequent in either ukWaC or the BNC, i.e. the words that may be seen as being relatively typical of one corpus when compared to the other.

When two corpora are evaluated through word list comparisons, however, two points need to be remembered. The first is that all the words that appear in the lists should be taken as being indicators of relative typicality in one corpus or the other, and not as being absolutely typical of them. To give an example, the noun *eyes* 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. The second point is that the method is apt at highlighting strong asymmetries in the two corpora, but it conceals those 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.

Moving on to the actual data analysis, it would seem that, compared to the BNC, ukWaC contains a higher proportion of texts dealing with three domains, i.e. the Web, education, and what were called "public sphere issues". These appear in a wide range of text types. Web-related issues, in particular, are found in almost all the text types identified by Sharoff (2006), i.e. discussion (e.g. online forums of discussion about a particular software or website), promotional (e.g. advertising of a traditional or web-based service) and instruction texts (e.g. tutorials). The presence of those among the most typical of ukWaC is unsurprising, insofar as they represent meta-references to the medium of communication that hosts them, and as the BNC was published at a time when the web was still in its infancy. Education and public service issues are also found in a great variety of text types, ranging from "traditional" texts like academic articles, to more recent web-based genres, like presentation pages detailing the activity, e.g., of a research or humanitarian group. Such heterogeneity of text types is a very positive feature in terms of the internal variety of ukWaC, since no one-to-one correspondence between a certain topic and a text type can be identified. This can be interpreted as confirming the soundness of the sampling strategy adopted.

In terms of domains, the BNC features a comparatively larger presence of narrative fiction texts. These are characterised by the frequent use of nouns and adjectives referring to physical characteristics or emotions, and by verbs (in the past tense) related to human actions. Moreover, the BNC seems to contain a higher proportion of spoken texts, whose presence is signalled by a number of discourse markers (e.g. *er*) and mental verbs in the present tense (e.g. *know*, *mean*). The third category of texts typical of the BNC is that of texts which deal with political and economic issues. Such texts differ from public service texts found in ukWaC, which are characterised by a stronger focus on practical issues (e.g. offering guidance to citizens), and on the present time. Politics- and economy-related texts in the BNC, on the contrary, are more concerned with describing events through abstract categories and using the past tense, as is typical, e.g., of non-fiction prose.

Differences in temporal deixis across the two corpora prove especially noteworthy. ukWaC seems to be characterised by a stronger concern with the present time, as is demonstrated, e.g., by the use of verbs in the present tense and of adjectives which refer to the present. This may be due, among other factors, to a considerable presence of advertising texts, which also display a number of causative verbs and of adjectives conveying a positive evaluation. One of the most interesting findings in this regard was that such advertising texts are featured not only in pages selling commercial products or services, but also in pages published by universities (e.g. inviting students to enrol), and governmental departments (e.g. promoting a political program). In the BNC, on the contrary, narrative language, characterised by past tense verbs and adjectives referring to the past, is more prominent.

4. Availability

UkWaC is available for download from the website of the Wacky initiative,[11] which also contains other data, such as frequency lists, seeds (words, tuples and URLs) as well as the lists used for the comparisons in Section 3. The customized tools used for corpus construction (duplicate detection, boilerplate stripping, etc.) are also available for download from the website. The corpus is available in two formats, as a plain text file (with no morphological annotation) and as a POS-tagged file encoded in a shallow XML format. This format is ready for indexing with the IMS Open Corpus Workbench (CWB),[12] a popular corpus processing tool. UkWaC is also available via the commercial "Sketch Engine".[13]

5. Related work

There is by now a large and 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). On the other hand, we are not aware of much publicly documented work on developing large-scale, general-purpose web-derived corpora.

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

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

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

The work most closely related to ours is that presented in Sharoff (2006). The author 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.[14] The methodology 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 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 corpora in the same languages, in terms of domain analysis and comparing wordlists, similarly to what we did here. For a more systematic literature review, however, we invite the reader to refer to Baroni et al. (2008).

6. Further work

UkWaC is already being actively used in several projects, including simulations of human learning, lexical semantics and langage teaching. We hope that this article will encourage other researchers to adopt ukWaC as a research tool, and that these activities will give us a clearer idea of the corpus’ strengths and limits.

We believe that the most pressing issue at this moment is the need to provide free access to the corpus, both through a web service that allows scripting access to remote corpora (to support linguists in doing extensive qualitative and quantitative research with the corpora) and via a web user interface that should allow user-friendly access to those without advanced technical skills (e.g., language learners, teachers and professionals). We are actively working in these areas.

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.

7. Acknowledgements

We would like to thank the members of the WaCky community for many useful interactions, in particular: S. Castagnoli, T. Emerson, S. Evert, W. Fletcher, F. Gaspari, A. Kilgarriff, J. Pomikálek and S. Sharoff. The WaCky project is partly funded by the University of Bologna through the LiMiNe funding scheme.

8. References

A. Baayen. 2001. *Word frequency distributions*. Kluwer, Dordrecht.

M. Baroni and S. Bernardini, editors. 2006. *Wacky! Working papers on the Web as Corpus*, Bologna. Gedit.

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2008. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. submitted.

D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *Longman grammar of spoken and written English*. Harlow, London.

A. Broder, S. Glassman, M. Manasse, and G. Zweig. 1997. Syntactic clustering of the web. In *Proceedings of the Sixth International World Wide Web Conference*, pages 391–404, Santa Clara, California.

C. Fairon, H. Naets, A. Kilgarriff, and G.-M. de Schryver, editors. 2007. *Building and exploring web corpora – Proceedings of the 3rd Web as Corpus Workshop, incorporating Cleaneval*. Presses Universitaires de Louvain, Louvain.

A. Ferraresi. 2007. Building a very large corpus of English obtained by web crawling: ukWaC. Master’s thesis, University of Bologna. Retrieved January 28, 2008 from `http://wacky.sslmit.unibo.it`.

W. Fletcher. 2004. Making the web more useful as a source for linguistic corpora. In U. Connor and T. Upton, editors, *Corpus Linguistics in North America 2002*, Amsterdam. Rodopi.

M. Hundt, N. Nesselhauf, and C. Biewer, editors. 2007. *Corpus linguistics and the web*. Rodopi, Amsterdam.

A. Kilgarriff and G. Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–347.

D. Lee. 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.

P. Rayson and R. Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of Workshop on Comparing Corpora of ACL 2000*, pages 1–6, Hong Kong, China.

M. Santini and S. Sharoff, editors. 2007. *Proceedings of the CL 2007 Colloquium: Towards a Reference Corpus of Web Genres*, Birmingham, UK.

M. Santini. 2007. Characterizing genres of web pages: genre hybridism and individualization. In *Proceedings of the 40th Hawaii International Conference on System Sciences, poster session*, pages 1–10, Waikoloa, Hawaii.

S. Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. In M. Baroni and S. Bernardini, editors, *Wacky! Working papers on the Web as Corpus*, pages 63–98, Bologna. Gedit.

M. Thelwall. 2005. Creating and using web corpora. *International Journal of Corpus Linguistics*, 10(4):517–541.

M. Ueyama. 2006. Evaluation of japanese web-based reference corpora: Effects of seed selection and time interval. In M. Baroni and S. Bernardini, editors, *Wacky! Working papers on the Web as Corpus*, pages 99–126, Bologna. Gedit.

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