ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA SEDE DI FORLÌ

SCUOLA SUPERIORE DI LINGUE MODERNE PER INTERPRETI E TRADUTTORI

CORSO DI LAUREA IN TRADUZIONE SETTORIALE E PER L'EDITORIA

TESI DI LAUREA

in

Traduzione dall'Inglese in Italiano I

Building a very large corpus of English obtained by Web crawling: ukWaC

CANDIDATO
Adriano Ferraresi

RELATORE
Silvia Bernardini

CORRELATORE
Marco Baroni

Anno Accademico 2006/2007

Sessione II

INDEX

INTRODUCTION

The aim of the present dissertation is to present and evaluate a new corpus resource for the English language. The corpus, called ukWaC (because it is a Web-derived Corpus constructed sampling UK sites), contains around two billion words. It was built with the intention of providing a very large and up-to-date resource that would be comparable, in terms of "balancedness" and variety of linguistic materials it contains, to traditional general-purpose corpora (in particular, the British National Corpus (BNC), a well-established standard for British English). As is the case for all corpora built with semi-automated procedures, however, the possibility to control the materials that end up in the final corpus is limited. This makes *post-hoc* evaluation a crucial task for the purpose of appraising actual corpus composition. A corpus evaluation method is therefore proposed and applied to the task of comparing ukWaC and the BNC.

Chapter 1 presents an introduction to two aspects of corpus linguistics which are central to this dissertation. On the one hand, a brief general introduction to the discipline is provided, which offers a description of the role of corpora in language studies and outlines some of the major concerns traditionally involved in the design of general-purpose corpora. On the other hand, the Chapter explores the notion of the "Web as corpus". In particular, the advantages and potential pitfalls of using Web data are taken into account, as well as the different methods through which the Web can be accessed for linguistic purposes, i.e. either as a corpus *per se*, through the use of a commercial search engine, or as a source of data that can be saved, post-processed and consulted offline. Two examples are provided of how these approaches have been applied to the actual construction of existing resources (Webcorp and WaC).

Chapter 2 discusses the reasons why ukWaC may be seen as a valid alternative to such existing Web-based resources, including its being a very large, stable and possibly balanced corpus. The procedure that was followed to collect, post-process and annotate its textual data is then explained in detail.

Chapter 3 focuses on the corpus evaluation procedure. It is argued that one way of evaluating a corpus whose composition is not known, as is the case for ukWaC, is to compare it with a benchmark. The evaluation, which in our case involves a comparison with the BNC, taken as a model of a general-purpose corpus, is therefore carried out through a comparison of different wordlists, each including all the word items belonging to the main part-of speech classes (nouns, adjectives, verbs, *-ly* adverbs and function words). The results of the analysis seem to indicate that, despite certain differences, such as the relative high proportion in ukWaC of texts related to the Web, education, and public service as well as advertising texts, and the relative low proportion of fiction and conversation, most text types and domains seem to overlap, since they do not emerge as being characteristic of either corpus. This seems to provide confirmation that the sampling strategies adopted when building our corpus were sound enough.

In Chapter 4, some directions for further work within *Web as Corpus* linguistics are outlined. First, practical improvements on ukWaC through further post-processing are envisaged. These should hopefully contribute to making this corpus a widely-used new resource for the study of English. Second, building on experience gathered in the present dissertation work it is suggested that a more thorough evaluation method for Web corpora is needed, which complements descriptive insights such as those provided here with practical usage-oriented tasks.

1

USING THE WEB AS A CORPUS: ISSUES AND APPROACHES

1.1 Introduction

With the advent and the exponential growth of the World Wide Web, an enormous amount of textual data has become available. Terabytes of information can be accessed with little effort, by simply using a computer and a modem, and, what is more, with almost no expense. As an immense, free, and easily accessible resource, it is not surprising that in recent years the WWW has attracted an increasing number of linguists, for whom the quantity of textual Web data has opened up new perspectives in language studies.

Existing resources prove sometimes inadequate for certain research questions (Kilgarriff and Grefenstette, 2003). This is the case, e.g., when less common or relatively new linguistic phenomena are the object of study, and well-established, but somewhat small (or "old"), collections of texts provide insufficient evidence for analysis. In other cases, e.g. for the study of specialized linguistic sub-domains or of minority languages, no resource exists (Scannel, 2007). In these contexts the WWW, considered as a very large repository of linguistic data, has the potential, and is indeed being exploited, to answer many research needs. The expression "Web as corpus" (or "WaC") was created to indicate these uses of the Web within language studies (Baroni and Bernardini, 2006).

In the present Chapter, the notion of the Web as a corpus is explored. In Section 1.2 a brief introduction to corpus linguistics is provided. Its aim is to offer a general description of the role of corpora in linguistic analysis, as well as to outline some of the major concerns traditionally involved in the design and construction of a linguistic corpus. In Section 1.3 attention is focused more specifically on the use of Web data in corpus building, highlighting some of the advantages and potential pitfalls that need to be taken into account when approaching the Web as a corpus. Section 1.4 attempts to describe the various

ways in which the WWW can be *used* as a corpus, i.e. either as a corpus *per se*, through the use of a commercial search engine, or as a source of data that can be saved, post-processed, and consulted offline through dedicated software. Finally, Section 1.5 focuses on the description of how such approaches translated into the construction of two different corpus resources.

1.2 A brief introduction to corpus linguistics

Corpus linguistics is a methodology for studying language whose starting point is the assumption that language is best described through an analysis of real instances of linguistic production. These can reveal patterns that could otherwise go unnoticed, even to the most acute linguist relying on his/her intuition and competence in a language (McEnery and Wilson, 2001). As a fundamentally empirical approach, corpus linguistics requires large quantities of data on which to base its observations. Corpora, i.e. collections of texts gathered according to pre-determined principles (Biber *et al*., 1998: 4), are therefore the main source of evidence in corpus linguistics.

In principle, corpora can be in printed or electronic form, but nowadays the notion of corpus is closely connected to its storage and access through computers, which allow researchers to carry out very detailed and accurate analyses of data, whose quantity is very often too large to be dealt with manually (*ibid*.). In fact, corpus linguistics studies are also more specifically linked to the analysis of data following an empirical methodology, which usually requires the use of dedicated software packages. For many corpus linguistics analyses the functions that are offered by software packages like WordSmith tools (Scott, 1996/2004) or Corpus Query Processor (Christ, 1994) are therefore vital. These include (but are not limited to) the possibility of searching for word forms, lemmas or part-of-speech tags – whether using regular expressions or not –, displaying results in KWiC (Key Word in Context) format, and sorting them according to criteria defined by the user (e.g. alphabetically, ignoring case, according to the words which precede or follow the query term, etc.). Quantitative and statistical approaches to textual data, such as frequency counts, lists of keywords, counts of the collocates of a given

word, etc. are also central to corpus linguistics, and automated functions to perform such kinds of analyses are often included in all corpus processors.

As pointed out by several authors (McEnery and Wilson, 2001; Stubbs, 1996), qualitative and quantitative approaches complement each other. While quantitative analyses are essential to demonstrating that certain patterns exist, qualitative evaluations are needed in order to provide generalisations and possible interpretations explaining why those patterns emerge. This analytical stage forms the subject of Chapter 3, where data are first compared across two corpora through a statistical method, and then classified into categories which provide meaningful explanations for the emerging patterns.

In order for qualitative analyses to correctly interpret quantitative data, it is crucial that criteria for corpus design are clear. Every emerging pattern should be explained in the light of the text types and domains that are known to be sampled in the corpus. Thus, depending on the corpus, e.g. whether it is designed to represent only certain linguistic varieties or language as a whole, it is possible to determine to what extent regularities can be generalised. Citing an example from Biber *et al*. (1998: 246):

> a corpus composed primarily of news reportage would not allow a general investigation of variation in English. Similarly, research based on a corpus containing a single type of conversation – such as conversations between teenagers – could not be generalised to conversation overall.

Hence, besides concentrating on the methods of analysis, corpus linguistics is concerned with defining criteria for corpus design. Some of these are taken into account here. The purpose is not to offer an exhaustive description, but to emphasize the most relevant points, particularly with regard to the design principles for "general language" corpora. The corpus that is presented in Chapter 2, which is the main subject of the present study, is intended to be one such corpus, and the discussion focuses accordingly on the features that should be taken into account when designing it.

Corpus size and sampling strategy are among the most important decisions that need to be made at the outset of a corpus construction task. As regards size, general language corpora are usually expected to be as large as

possible. This is so for two main reasons. Firstly, due to Zipf's law of word frequency distribution (Zipf, 1935), the possibility that rarer linguistic features are attested in a corpus increases proportionally to the largeness of the corpus itself. This means that corpora need to be very large "if they are to document as wide as possible a range of uses of as many linguistic features as possible" (Aston and Burnard, 1998: 21). Secondly, large size tends to counterbalance the relative influence that single texts can have on the results of an analysis (Biber *et al*., 1998: 249). The more numerous the texts in a corpus (and the more varied their types), the less the results are likely to reflect a language usage that is typical, e.g., of a single author, text or text genre. In this case, the issue of size is therefore intertwined with that of corpus heterogeneity and balance.

These two features rely heavily on the second design principle mentioned above, i.e. the strategy through which texts to be included in the corpus are sampled. If a corpus is to represent "general language", it should include a great variety – and, as has just been argued, a great number – of texts, possibly in such proportions so as not to introduce undue biases towards certain text genres or types. In order to avoid this, two sampling strategies can be envisaged, i.e. *proportional sampling* and *stratified sampling* (McEnery and Wilson, 2001: 77-81). If proportional sampling is chosen, text types are included in a corpus in a quantity that is proportional to the quantity of written and spoken texts that the speakers of that language[1] come into contact with during a certain period, e.g. one week. In this way, probably 90% of the corpus should be composed of spoken transcripts (Sinclair, 2005), since arguably most people spend more time speaking and listening than writing or reading. Such view of corpus balance – or, better, of theoretically justified unbalance – is challenged by Biber *et al*. (1998: 246-248), who argue that proportional sampling can be appropriate if the research question concerns, e.g. discovering "how often a person is likely to encounter a certain word in the course of a typical week" (*ibid*.: 247), but is completely inadequate to represent language

[1] This implies providing a statistical model of events of language production and reception of the population, in the same way as political polls define demographic samples on which to base their results (Biber *et al*., 1998: 247).

as a whole. They propose therefore a stratified method of corpus construction, whereby even texts that few people are likely to encounter during their life (e.g. academic prose) are included in the corpus. In their view, corpus building should ideally take into account all areas (i.e. *strata*) of language, and samples should be included from each of them. As we shall see, such distinction between proportional and stratified sampling is also central to designing Web corpora, particularly when it comes to deciding the appropriate strategies for retrieving Web texts (cf. Section 2.3.1).

Besides being concerned with defining the relative weight that text types should have in a corpus, the sampling strategy needs to define the size of samples and whether to include whole texts or only parts of them. Sinclair (2005) argues in favour of the former option, on the grounds that dismembering a text and including only parts of it is an unduly arbitrary operation, that could result in the selected part not being representative of the whole text, i.e. the distribution of its linguistic features may not correspond to that of the text taken in its entirety. However, he recognises two main drawbacks connected with this strategy, which are also acknowledged by Aston and Burnard (1998: 22). On the one hand, the size of entire texts may vary greatly, hence creating possible problems of corpus balance. On the other hand, if the corpus is to be published, it is often difficult to obtain copyright permissions to include whole texts. Both for theoretical and practical reasons, many corpora therefore include only samples of bigger texts. This is the case for the Brown (Kucera and Francis, 1967) and LOB (Johansson, 1980) corpora, which include randomly selected samples of 2000 words each, as well as for the BNC (Aston and Burnard, 1998), whose large dimensions allowed its creators to include bigger samples (40,000-50,000 words).

For Web corpora, decisions about sample size may be less stringent than for traditional corpora. Fletcher (2004b) estimates that the size of Web pages containing human-produced text usually varies between 5 and 200 kilobytes. This measure is used in his study – as well as in the present thesis – as a heuristic to decide what the "good size" of a sample is, and all the pages respecting that criterion are included in the corpus for further post-processing

(cf. Section 2.3.2.1).[2] What is most crucial is to determine whether Web pages should be included in their entirety or not. Regarding this issue, it will be argued in Section 2.3.2.2 that it is desirable to exclude from Web samples portions of text called "boilerplate", i.e. "linguistically uninteresting material repeated across the pages of a site and typically machine-generated, such as navigation information, copyright notices, advertisement, etc." (Bernardini *et al*., 2006: 20), since they provide little information about language use and tend to distort statistics about corpus composition (cf. Chapter 3).

As a conclusion to this Section, in which an attempt was made to define the methodological approach of corpus linguistics, and the most relevant design criteria for the construction of (traditional or Web) general language corpora, some applications of the discipline will be mentioned, mainly in order to suggest possible ways in which corpora can be used within language studies (for a fuller description see, e.g. McEnery and Wilson, 2001). The applications of corpora in language studies are manifold. Lexico-grammatical analyses of corpora have provided materials for grammars of English. One of the best known works is probably *A comprehensive grammar of the English language* (Quirk *et al*., 1985), which was among the first English grammars to be based on corpora. A more recent example is the *Longman grammar of spoken and written English* (Biber *et al*., 1999), whose approach is even more corpus-intensive. But corpora can also be used in many other areas of linguistics. For instance, they have been used for socio-linguistics and cultural studies (Stubbs, 1996), as well as in Natural Language Processing (Manning and Schütze, 1999). Finally – but the list could be much longer –, comparable and parallel corpora can be used in translation studies (Olohan, 2004) or in translator training (Zanettin *et al*., 2003).

[2] This does not mean that all of the pages between 5 and 200 kb contain human-produced language. The criterion of good size is a "sine qua non", but pages have to pass other filtering phases before being allowed into the final version of the corpus.

1.3 Web data: advantages and potential pitfalls

As in all tasks of corpus construction, advantages and potential pitfalls of using certain types of data instead of others (e.g. scanned in vs. manually typed texts) need to be taken into account. Both theoretical and practical reasons should be considered, such as the adequacy of data in relation to the corpus being built (e.g. does a scanned version of a newspaper article differ from its Web-published counterpart? Can they be included in a traditional or Web corpus indifferently?), and the resources that are available for corpus construction, like time, funding, people who work in the project, etc. In the present Section, advantages and pitfalls linked with using Web data are discussed.

Arguably, one of the main advantages of using Web data instead of other types of data is that texts retrieved for inclusion in the corpus are already in machine-readable form, and do not therefore need to be converted into an electronic form (unlike "traditionally published" texts). Especially if automated methods of text retrieval are used (see, e.g., Baroni and Bernardini, 2004), corpora can thus be constructed in a very short time, even by a single researcher (cf. Section 1.4.2).

Both in the field of NLP and in corpus linguistics, it is now largely acknowledged that "more data is better data". For this reason, the huge size of the Web, used as a source of linguistic data, can be seen as another advantage for corpus construction. Banko and Brill (2001) show that a simple algorithm trained on a very large corpus in a simple language disambiguation task outperforms more sophisticated algorithms created *ad hoc* for use on smaller – and "cleaner" – data sets. Clarke *et al*. (2002) demonstrate that the performance of a question answering system tends to improve with corpus size, even if it reaches an asymptote and declines slightly when the algorithm is tested on a (Web) corpus bigger than 400-500 GB. It has to be noted, however, that traditional corpora are usually much smaller than this, and that, at least for now, only the Web seems to be an adequate source for retrieving such quantity of data in a reasonable time and with reasonable effort. Keller and Lapata (2003) demonstrate that the Web, given its size, makes it possible to find bigrams (adjective-noun, noun-noun, verb-noun) that are not attested in

traditional corpora, and that counts about their frequency can be produced via a search engine with a relative degree of confidence. However, the huge size of the Web can also be exploited in more theoretically-oriented linguistic studies. Mair (2003) shows that a linguistic phenomenon like the grammaticalization of *get* as a passive in English cannot be fully investigated in the BNC, while the Web, given its size, makes such a task possible. Brekke (2000) uses the AltaVista search engine to study the grammatical constructions and textual domains in which two word items, i.e. *chaos* and *quantum*, appear. After comparing the results of queries for these words in the BNC and the Web, he concludes that the latter is a more suitable resource for such a task, both because the BNC yields fewer results[3] and because the Web includes a wider range of text domains in which the two words are attested. Brekke (2000: 243) remarks that this can also be due to the fact that only in recent times "the two test items are […] seeing increased use outside their strictly scientific domains". In this sense, the BNC, which is a synchronic corpus dating back to the early 1990's, may be seen as an insufficient resource to study recent evolutions of language.[4]

The point raised by Brekke (2000) relates to yet another important feature of Web data, i.e. their being up-to-date and constantly refreshed (Fetterly *et al.*, 2004). This constitutes an evident advantage over traditional corpus resources, "that are often subject to a certain lag between the time of production of the materials […] and the publication of the corpus itself" (Baroni and Ueyama, 2006: 32). For this reason, Web data are usually the only resource available to study recently emerged linguistic phenomena, such as the use of the suffix *–itis* in German and English words formed in non-medical domains (Lüdeling *et al.*, 2007). Moreover, Web corpora can include samples taken from "emerging text genres" (Santini, 2007) that are not attested in traditional resources, such as blogs and forums of discussion. These contain large quantities of texts, relate to a wide range of topics, and, what is perhaps

[3] In this regard, it should be noted, however, that the words were chosen precisely on the grounds that they can be considered as relatively rare.

[4] This, of course, does not imply that the BNC is not still useful for a number of purposes, ranging from historical interests to didactic applications. In fact it is also used in the present study as a benchmark corpus.

most interesting from a corpus linguist's point of view, are spontaneously produced by Web users, whose demographic characteristics (age, profession, etc.) may vary to a great extent.

Not only blogs and forums seem to deal with a great variety of topics, but also the Web in general. This often makes it the only resource available for studying specialized linguistic sub-domains, as in the field of terminology extraction. Traditional resources like the BNC contain a certain amount of specialized texts (Aston and Burnard, 1998), but, since they are not designed to represent specific technical domains, the problems connected with their use for terminological purposes may be manifold. The specialized domain under investigation may not be included in the corpus, the corpus may contain too few texts about that domain, or the texts may not be recent and able to document contemporary usages in a constantly evolving field such as that of terminology (Cabré, 1999). On the contrary, the Web contains constantly updated information, and the number of texts it contains is usually sufficient to extract relevant terms for the domain in question (Fantinuoli, 2006). Varantola (2003) also suggests that the Web can be used when the need arises to build specialized corpora in little time, as in the case of specialized translation tasks.

The last point that is going to be made is that Web data can be, and are indeed, exploited for building corpora in languages for which no well-established corpus resource exists. This is true for so-called "minority languages", such as Basque, Welsh and Hawaiian, but also for much more widespread languages, such as Italian and Japanese. One of the main problems connected with the construction of resources for these languages – this applies especially to minority languages – is that there is little chance that a corpus building project finds funding or attracts commercial enterprises (Scannel, 2007). Since Web data are freely available, and since the phenomenon of Web publishing is widespread on a global scale, the WWW seems therefore the most suitable source from which corpora for these languages can be compiled. In particular, Scannel (2007) implemented a method of corpus construction (relying on the BootCaT toolkit; Baroni and Bernardini, 2004) which, starting from a small set of training texts, allowed him and his collaborators to build

corpora for 416 languages. However, he does not provide accurate qualitative or quantitative analyses about his results. Before him, Ghani *et al*. (2003) developed a similar method for the construction of corpora for "under-resourced" languages, which required the collection of URLs through queries to a search engine and downloading and post-processing the corresponding Web pages.

After discussing the advantages offered by the Web over other types of resources, we now shift attention to the major problems that using Web data may cause. One of the most frequent pieces of criticism of Web data, in this case referred to English, is that "Web English is not representative of written or spoken English" (Thelwall, 2005: 522). Thelwall (*ibid*.) adds that the Web as a whole should not be used as a corpus,[5] and justifies his claims by affirming that the Web contains disproportionate amounts of text topics and genres (e.g. a large quantity of computer- and business-related texts, but very few fiction texts), and that Web authors cannot be considered as a representative sample of the native speakers of a language, since they tend to be young people with above average computing skills. As regards the latter point, Baroni and Ueyama (2006: 32) point out that, while observations such as those in Thelwall (2005) are founded,

> over-representation of certain groups seems a more general property of written language […]. While (almost) everybody engages in oral communication on a daily basis, only a non-random subset of a community frequently engages in written communication. If something, the Web is expanding the range of speakers who belong to this subset.

As regards the criticism about the non-representativeness of Web data, it should be noted that Thelwall (2005: 519) considers the Web in its entirety as a corpus ("The Web […] is a complete corpus, given an agreed precise definition of the Web, at a given moment in time"), thus ignoring the possibility to exploit Web data for the construction of smaller-scale, more controlled corpora (cf. Section 1.4.2; 1.4.3). In fact, if a sample of Web pages is chosen according

[5] He suggests that only specific and pre-determined sections of it (e.g. academic Web sites) should be crawled for inclusion in a corpus.

to well-defined criteria it is possible to obtain relatively balanced corpora, including a wide variety of text topics and genres (Sharoff, 2006). Moreover, as suggested by Kilgarriff and Grefenstette (2003), the issue of representativeness is far from well understood, and also "traditional" corpora may be seen as being affected by problems of non-representativeness.

Another critical issue about using Web texts is their sometimes poor linguistic quality. Especially for English, it may be frequent to find Web pages that are translations from other languages, or texts authored by speakers for whom English is not the mother tongue, as in the case of international researchers writing academic papers or their personal home-pages (Thelwall *et al.*, 2003). The fact that Web pages are typically anonymous and that the location of Web servers offers no reliable indication about the provenance of Web pages contributes to raising doubts about the texts' authoritativeness (Fletcher, 2004b). The lack of such pieces of information makes it also difficult to retrieve (and possibly encode in a corpus) meta-data about Web texts, as is instead done in traditional corpora, where most texts contain meta-information about a text's date of publication, its source, author, etc. Moreover, texts published online tend to contain typing and spelling errors (Ringsletter *et al.*, 2006), which are typically due to the relative lack of editorial control over the contents that are published online.

In addition to linguistic errors, Web pages contain significant amounts of "noise", such as automatically generated text, server logs and boilerplate. The problem of duplicate pages is also an issue that needs to be taken into account, especially when Web data are used to produce frequency counts about certain words or patterns. These problems, however, can be countered if Web pages are downloaded for inclusion in an offline corpus and subsequently post-processed (cf. Section 1.4.2, 1.4.3). In particular, many methods for boilerplate stripping (see, e.g., Marek *et al.*, 2007) and for duplicate pages detection and removal (Broder *et al.*, 1997) can be applied for the purpose of obtaining "cleaner" Web data.

Summing up, among the main advantages of using the "Web as corpus" we can mention its size, its being a source of constantly updated linguistic

materials, the variety of topics it contains, and its being the only viable resource for certain corpus construction tasks, like, e.g., for the construction of specialized corpora (Baroni and Bernardini, 2004), or corpora for minority languages (Scannell, 2007). On the negative side, Web data can pose problems when a fully controlled and noise-free linguistic resource is needed. As suggested by Baroni and Ueyama (2006: 32), however, it is ultimately a "matter of research policy, time constraints and funding to determine if, for a certain project, it is better to […] [build] a thoroughly controlled, probably relatively small corpus, or if it is better (or: the only viable solution given external constraints)" to use Web data as a source of linguistic evidence, even if this entails specific problems, that need to be fully considered, and if possible solved.

1.4 Three approaches to the "Web as Corpus"

In Section 1.3 the discussion focused on the general advantages and potential pitfalls that should be taken into account when Web data are used as a source of linguistic evidence. Most of these advantages/disadvantages apply to textual Web data in general, irrespective of the methodology that is adopted to use them for purposes of linguistic analysis.

Three approaches to the "Web as corpus" can be identified. These differ both in terms of the method through which Web data are collected, and in terms of the way in which such data can be subsequently used for linguistic analyses. In the present Section, these approaches are discussed in turn.

1. 4. 1 USING THE WEB AS A CORPUS THROUGH COMMERCIAL, NON-DEDICATED SEARCH ENGINES

One of the most widespread approaches to the Web as a linguistic corpus consists in issuing queries to a search engine, like Google, and relying on the counts of the resulting hits to estimate the frequency of the word or word-string in the language of interest. Bernardini *et al*. (2006: 10) refer to this approach as using the Web "as a corpus surrogate", since it evidently underlies the notion that the Web – or at least the large portion of the Web which is included in the

search engines' indexes – can be considered as a corpus *per se*, and that a search engine can be used as a sort of concordancer, albeit a rather rudimentary one.

Using this approach, Grefenstette (1999) demonstrated that it is possible to rely on search engines' reported results to find likely translations for noun phrases across English, German, French and Spanish. Brekke (2000) carried out a study on the frequency and the distribution across textual domains of two word items, i.e. *quantum* and *chaos* (cf. Section 1.3).

This approach, however, poses several problems. Current search engines were not developed for linguistic purposes, i.e. to make it possible to study linguistic *forms*, but to find relevant information, i.e. *contents*, in the huge and unstructured amount of data that is the Web. Thus, if one wants to use the Web as a corpus via search engines, one needs to be aware of the inherent limitations that the approach entails (for a fuller discussion, see Lüdeling, *et al.*, 2007; Kilgarriff, 2007; Thelwall, 2005).

Some of these limitations concern the low degree of flexibility allowed by search engines when they are used as a sort of concordancer. Indeed, they do not allow searches for word lemmas or part-of speech-tags, and do not support regular expressions.[6] Thus, the syntax of search engine queries is very rigid. Search engines also perform normalizations on the words that are searched for: case, dashes and apostrophes are ignored, and stemming procedures are applied (e.g. a query which includes the word "dogs" may also return results including pages containing the word "dog"). Besides the lack of flexibility and precision in the specification of the words and word combinations that can be searched, search engines do not allow to re-sort results according to user-defined criteria (e.g. according to words on either side of the query term, alphabetically, etc.), which usually makes it very difficult and time-consuming to observe recurring language patterns.

[6] Google, e.g., supports the use of the wildcard "*" in a non word-interior position, but it is not possible to specify the number of words that the wildcard "*" should stand for. Google is taken as an example since it is one of the most widely used search engines, and, to the best of my knowledge, one of the best-performing in this regard.

In order to make up for these deficiencies, linguist-oriented meta-search engines have been developed, like, e.g., WebCorp (cf. Section 1.5.1), or KWiCFinder[7] (Fletcher, 2004a). These wrap the output of traditional search engines and offer some of the basic functionalities of traditional concordancers (cf. Section 1.2).

Perhaps the most serious problem connected with the "Web as a corpus surrogate" is the fact that

> search companies, for obvious reasons, do not publish detailed information on how they gather, index and return query results, and the services they provide, being often and unpredictably updated following technological and market changes, tend to be extremely brittle. (Baroni and Ueyama, 2006: 33)

This raises a series of doubts about the methodological justification for using the Web as a source of linguistic evidence relying on search engines. First of all, search engines do not ensure that the counts they provide are accurate, since they may be extracted from only a subset of their entire index. While this makes it possible to view results more rapidly, which is an essential requirement for content-oriented search engines, the resulting counts may be distorted (Thelwall, 2005: 525). Secondly, the ranking algorithm according to which results are produced and sorted is unknown to the researcher, so that the display of the results may be biased, e.g. in favour of (paying) commercial companies (Kilgarriff, 2007). Finally, given the constant updates that search engines' indexes undergo, it is usually not possible to replicate an experiment. The problem of the non-reproducibility of experiments is a very serious one in corpus linguistics. As pointed out by Lüdeling *et al*. (2007: 11-12), both quantitative and qualitative approaches to corpus data require indeed that experiments' results can be repeated, both because their relevance "depends on the correctness and interpretability of the published numbers" and because any claims made about a certain language pattern may be "invalidated when a replication […] of the experiment brings up contradictory examples" (*ibid*.).

For these reasons, using the Web as a corpus via search engines does not seem the best solution for exploiting the potential that it offers.

[7] http://www.kwicfinder.com/

1. 4. 2 BUILDING CORPORA VIA SEARCH ENGINE QUERIES

The second way of exploiting the Web as a corpus that is taken into account consists in retrieving Web pages through search engine queries and then saving them offline to make up a corpus in the traditional sense of the term (unlike in the method presented in Section 1.4.1), which may be then post-processed. This corresponds to what Bernardini *et al*. (2006: 11) call "using the Web as a corpus shop". In this case, the Web is not used as a corpus *per se*, but as a source from which data are gathered, through manual or automated procedures, and can be exploited for the creation of either specialized or general-purpose corpora.

Varantola (2003) discusses the advantages of DIY (or "disposable") specialized corpora built in this way for the teaching of translation skills. Resnik and Smith (2003) developed an algorithm which relies on search engine queries to recognize and retrieve pairs of original and translated Web texts, which can be aligned so as to form large parallel corpora. Sharoff (2006) and Ueyama (2006) build and evaluate large reference corpora for multiple languages via automated queries to the Google search engine, and find their corpora to be relatively wide-ranging, both in terms of topics and text genres that are covered (cf. Section 3.2 for a fuller discussion).

Using the "Web as a corpus shop" has the advantage that, despite the fact that a search engine is still needed to retrieve the pages, documents are saved offline. This allows the researcher to counter some of the issues that were mentioned in Section 1.4.1. Web texts can be lemmatized and pos-tagged, and subsequently accessed via concordancing tools. Moreover, experiments can be repeated on the same data set, the search engine that is selected to collect the data.

As suggested by Baroni and Ueyama (2006: 33), however, this approach is not devoid of problems. One is that the quantity of data that it is possible to

find and download, either manually or via automated queries, is limited.[8] This appears to be a major constraint on the Web potential as a source of linguistic corpora, since building large collections of texts with this method, while possible, requires much effort and time. Moreover, the set of pages retrieved may still suffer from the problems – mentioned in Section 1.4.1 – linked to the ranking and matching algorithms of the search engine used.

1. 4. 3 CRAWLING THE WEB FOR LINGUISTIC PURPOSES

This method of approaching the "Web as corpus" consists in performing customized crawls of the Web that make it possible to collect and post-process Web data, which are then included in a potentially very large corpus.[9] The approach is radically different from the ones described in Section 1.4.1 and 1.4.2 insofar as it does not rely on commercial search engines, and therefore does not entail the drawbacks connected with their use as "intermediaries" between the researchers and the Web.

Crawls of the Web can be performed to build specialized corpora, such as collections of pages from academic Web sites (Thelwall, 2005), but the interest of the method for the purposes of the present study lies in its use to build very large general-purpose corpora. ukWaC, the general-purpose English corpus that is presented in Chapter 2, was indeed built adopting this approach. Similar corpora were built for German (Baroni and Kilgarriff, 2006) and Italian (Baroni and Ueyama, 2006), but no detailed evaluation of has been carried out at the time of this writing. A proposal to build a general-purpose corpus by Web crawling was also put forward by Rayson *et al*. (2006), who suggested that computing resources for data processing could be shared by interested researchers and institutions via a peer-to-peer network. The project, however, was never put into practice (Fletcher, 2007: 44-45).

While large corpora obtained via crawls are not affected by the inconveniences connected with the methods relying on search engines, they

[8] As regards automated queries, Google allowed users to submit automatically 1,000 queries per day. The service through which automated queries are issued (APIs), however, is no longer offered to new users.

[9] For a more thorough description of the crawling and post-processing methods, cf. Chapter 2.3.

nonetheless require that the problems linked with the use of Web data are tackled (cf. Section 1.3). The data obtained from the crawl need therefore to be post-processed, i.e. problematic pages (such as spam pages) must be eliminated, HTML code and "boilerplate" stripped off, and duplicate pages discarded. Implementing methods for carrying out these tasks requires some effort and computing skills. Furthermore, considerable computing resources are needed for managing the dozens of gigabytes of text and annotation in these corpora. These are perhaps some of the reasons why other methods of approaching the "Web as corpus" are more popular among linguists.

The advantages of performing large crawls of the Web to build linguistic corpora seem, however, to exceed the disadvantages. As pointed out by Bernardini *et al*. (2006: 13-14), a corpus obtained in such a way

> would possess both Web-derived and corpus-derived features. Like the Web, it would be very large, (relatively) up-to-date, it would contain text material from crawled Web sites and it would provide a fast Web-based interface to access the data. Like a corpus, it would be annotated (e.g., with POS and lemma information), it would allow sophisticated queries, and would be (relatively) stable.

For these reasons, the present study is guided by the assumption that the approach to the "Web as corpus" presented in this Section is the most valuable. Chapter 2 and 3, in particular, illustrate it in a more detailed way and explore its potential.

1.5 Existing "Web as corpus" resources

In Section 1.4 the main approaches to the use of the Web as a linguistic corpus were outlined, and their advantages and drawbacks were discussed, both from a theoretical and operational point of view. In the present Section the attention is focused on how these approaches have been put into practice for the actual construction of linguistic resources. Two of them are taken into account here, i.e. WebCorp (Renouf *et al*, 2007) and WaC (Fletcher, 2007). Even though the list could be much longer, these two were chosen insofar as they reflect two different approaches to the "Web as Corpus". The former, a linguistic-oriented processor and interface to commercial search engines, exemplifies the use of

the Web "as a corpus surrogate" (Bernardini *et al*., 2006: 10; cf. Section 1.4.1). The latter provides an online interface to a (general purpose) corpus built via automated queries to a search engine; i.e., it is an example of the Web "as a corpus shop" (Bernardini *et al*., 2006: 11-12; cf. Section 1.4.2). These resources, it will be argued, present both advantages and disadvantages, which are partly connected with the approaches to the "Web as corpus" they originated from.

1. 5. 1 WEBCORP

WebCorp[10] (Renouf *et al*., 2007) is a "linguist-friendly" online interface relying on search engines to retrieve occurrences of words and phrases. The tool acts as an intermediary between the search engine and the researcher, who can make use, through WebCorp, of the search and display functions that are usually integrated in a concordancer. Thus, it is possible to specify whether the search should be case sensitive, to use simple wildcards within a query, and to indicate filter words, which work as a rudimentary disambiguation method to find the desired meaning of a word, e.g. to find occurrences of the word *sole* in its "sea animal" meaning, by specifying *fish* as a filter word (example from Renouf *et al*., 2007: 54). WebCorp displays the results in a KWiC format, and the user can set parameters for the concordance span, sort the results according to the desired criteria and count collocates of the search term. The tool helps therefore overcome some of the obstacles that researchers find themselves confronted with when using search engines for linguistic purposes, i.e. the very limited query syntax and display options supported by "standard" search engines.

At the time of writing, however, the drawbacks connected with using WebCorp are manifold. Firstly, the tool suffers at times from serious problems of performance, depending, e.g., on the number of visitors using it. WebCorp does not have direct access to the search engines' indexes, so that each time a query is submitted it needs to wait for the search engine to respond, and the time varies greatly, depending on the workload that the search engine accepts

[10] http://www.webcorp.org.uk/

to handle. Data need then to be downloaded and processed according to the criteria specified by the user. This often results in very long waits before concordances can be seen. Secondly, not all kinds of queries can be handled by WebCorp. The tool exploits refined algorithms to translate the requests of its users into a format that is supported by search engines, but complex queries involving, e.g., regular expressions or part-of-speech tags are impossible for it to deal with. This is due to the fact that search engines (for obvious reasons) do not POS-tag their data, nor do they index data below the word level. Thus, as suggested by Lüdeling *et al*. (2007), WebCorp would be unsuitable if one wanted to carry out a study about the linguistic behaviour of the suffix *–itis* (cf. Section 1.3).

Apart from these practical considerations, WebCorp does not seem to tackle many of the points raised in Section 1.4.1, linked to the theoretical justification for relying on search engines' matching and ranking algorithms. The accuracy of the counts, the relevance of the results and the non-reproducibility of the experiments are therefore elements to be taken into account when turning to WebCorp for linguistic studies.

1. 5. 2 WAC

WaC[11] (Fletcher, 2007) provides an online interface to a very large corpus of English, which was built via automated queries to Microsoft's LiveSearch engine[12] and aims at reaching the size of one billion words. The corpus includes documents which were sampled randomly from all the Web domains corresponding to English speaking countries. The quantity of the samples is directly proportional to the population of the countries themselves (US, UK, Canada, Australia, Canada, Ireland and New Zealand). After retrieval, the data underwent basic post-processing, which included eliminating duplicate documents, conversion from HTML to text format, and indexing for fast retrieval of results when queries are generated. The interface supports all the

[11] http://webascorpus.org/ . It has to be noted that the acronym "WaC" is used in the present Section to refer to the proper name of the resource. As such, the meaning of the expression must not be confused with that used elsewhere in this dissertation (cf. Chapter 1).
[12] http://live.com

most important search, display and linguistic processing functions, including regular expressions, KWiC concordances, and frequency counts.

Unlike WebCorp, WaC has generally no problems of performance, thanks to a built-in corpus search engine that relies on its own indexes. This also allows it to support complex queries, which are not limited by the constraints imposed by search engines. Finally, experiments using WaC for linguistic purposes are replicable.

Despite its great potential as a very large Web-derived resource, WaC has some limitations. On the practical side, no "boilerplate stripping" procedure (cf. Section 2.3.2.2) was carried out on the data, and these are not POS-tagged.[13] Moreover, the growth of the corpus is strongly limited by the restrictions imposed on automated querying by LiveSearch. From a more theoretical point of view, it also has to be considered that, even if Web pages were saved and (partially) post-processed offline, the corpus was built via the intermediary of a search engine. The questions linked to using search engines' results as a source of Web data remain therefore untouched. Furthermore, no qualitative or quantitative evaluation of the resource was provided. These two reasons leave some doubts as to how the results drawn from WaC should be interpreted, especially if quantitative studies are carried out relying on it.

1.6 Concluding remarks

The present Section aimed at providing an introduction to corpus linguistics and exploring one of its recently emerged fields of interest, i.e. the use of Web data for linguistic purposes. In particular, some of the applications of corpora in language studies were illustrated, and the main criteria that are traditionally involved in the construction of general-purpose corpora were discussed. Attention was then shifted to the advantages and potential pitfalls of using Web data for corpus building tasks. These include, on the one hand, the huge size, timeliness and variety of topics and languages that characterize the Web, and, on the other, the supposed inadequacy of Web texts to "represent general

[13] The author, however, reports that these steps are under way (Fletcher, 2007: 51).

language" – an issue that, it was argued, is far from well-understood –, as well as the "noise" that Web-derived data usually contain. Three different approaches to the Web as a linguistic corpus were identified, two of which rely on commercial search engines for data retrieval, either providing a "linguist-friendly" query interface to them or using them to collect data that are saved off-line. WebCorp and WaC were taken as examples of how these two approaches have been exploited for the construction of language resources.

The third approach to the "Web as corpus" consists in performing large crawls of the Web, and, it was argued, presents the advantage of allowing the researcher to be in control of the corpus construction task (without the intermediary of search engines), and to collect large quantities of data, that can be subsequently post-processed and annotated for inclusion in a stable corpus. Chapter 2 explores in greater detail the advantages deriving from this approach and presents ukWaC, a very large, "balanced" corpus of English obtained by a large crawl of the Web in the `.uk` domain.

2

Building a very large general-purpose corpus of English by Web crawling

2.1 Introduction

In Chapter 1 the main approaches to the use of the Web as a linguistic corpus were outlined, and two examples of existing WaC resources were presented and discussed. It was argued that their main limitations are connected with their reliance on commercial search engines, which either impose serious constraints on the query syntax and do not make it possible to replicate experiments (in the case of WebCorp; Renouf, *et al*., 2007), or may bias the results of searches in unknown ways (both in the case of WebCorp and WaC; Fletcher, 2007).

The aim of the present Chapter is to present ukWaC, a corpus of English which was built via a crawl of the Web. In Section 2.2, the advantages deriving from such approach are outlined, and the main purposes for which ukWaC was built are discussed. It is suggested that ukWaC aims at being comparable to traditional balanced corpora, while at the same time providing a larger and more up-to-date resource. Section 2.3 describes in detail the different steps of the construction procedure.

2.2 Why building ukWaC

With its 100 million words, the BNC (Aston and Burnard, 1998) was considered at the time of its publication as a great achievement for corpus linguistics. As a large, balanced general-purpose corpus of English, the BNC is indeed still used today as a benchmark corpus for many studies involving qualitative and quantitative analyses of natural language. Despite its size and high standards of quality, however, the BNC cannot always provide sufficient evidence for analyses, especially when the research question focuses on relatively rare or recently emerged linguistic phenomena (cf. Section 1.3). For

this reason, the need for larger corpora, through which rarer linguistic features can be studied, and for more up-to-date resources, which may document recent evolutions of language, is nowadays widely felt within he linguistic community.

Different approaches to the "Web as corpus" have been envisaged to meet this need, which resulted in the construction of linguistic resources like WebCorp and WaC. In the previous Chapter it was argued that these, however, seem to be affected by several problems, mainly deriving from their reliance on commercial search engines (cf. Section 1.4.1; 1.4.2). Search engines' criteria for matching and presenting results are indeed not suitable for linguistic research, insofar as biases may be introduced in the data sets that cannot be predicted. For this reason, many questions remain open as to the suitability of these resources as benchmarks from which generalizations about language behaviour can be drawn.

ukWaC, the corpus that is presented in Section 2.3, aims at providing an alternative to such resources. Since it was built through a large crawl of the Web, its construction did not rely on search engines for retrieving data. Unlike WebCorp, it is a stable resource, and makes it possible to replicate linguistic experiments. Moreover, it is fully POS-tagged and lemmatised, so as to support very complex queries (provided, of course, it is accessed through a fully-tailored corpus search engine). In other words, ukWaC possesses all the features of a traditional corpus, by virtue of being able to support a (wide) range of analyses for research purposes.

The ultimate aim when building ukWaC was to provide a resource which would be comparable to the BNC. As the BNC, ukWaC is meant to be a general-purpose, balanced corpus of English. At the same time, however, it aims at providing a much larger and more up-to-date data set on which to base linguistic observations. Its size (more than two billion words), and the fact that it is derived from Web data (cf. Section 1.3 on the advantages of this approach), should thus enable linguists to find enough evidence to study rarer linguistic phenomena, and also to document recent evolutions of language.

Before moving on to the description of the steps that were followed to build the corpus, an important remark should be made. The construction of ukWaC is part of a larger project, called *WaCky*[1] (***W**eb as **C**orpus **k**ool **y**initiative*). The project is maintained by a community of linguists, who firmly believe in the potential of the Web for the construction of linguistic resources. Among the projects' achievements, the construction of two general-purpose Web-derived corpora for German (deWaC) and Italian (itWaC) should be mentioned. At the moment, work is in progress to implement a query tool available online to access the three corpora (see Baroni and Bernardini, 2006).

2.3 The construction of ukWaC

In the present Section the procedure followed to construct ukWaC is described. As was mentioned in the previous Section, the strategies presented here draw on the experience acquired while building two similar corpora for German and Italian (cf., respectively: Baroni and Kilgarriff, 2006; Baroni and Ueyama, 2006). The basics steps of the construction of ukWaC were:

- Selecting the "seed" URLs;
- Retrieving pages by crawling;
- Cleaning up the data retrieved;
- Annotating the corpus.

Each of these steps is discussed in detail.

2. 3. 1 CRAWL SEEDING AND CRAWLING

The aim in building ukWaC was to obtain a "balanced" corpus, which would ideally contain a wide range of text types and topics (cf. Section 2.2). These should include both "traditional" texts of varied nature (spanning from newspaper articles to recipes, etc.) that can also be found in electronic format on the Web, and texts which belong to typically Web-based genres, like personal pages, blogs, or postings in forums. The rationale in doing so is that

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

the corpus should include a random sample of pages that are representative of the target language, i.e. English. As pointed out by Baroni and Ciaramita (2006: 131), this is not the same as aspiring to get a random sample of Web pages, since the goal is to collect "a sample of pages that, taken together, can give a reasonably unbiased picture of a language, independently of whether they are actually representing what is out there on the Web or not" (cf. Section 1.2 and 1.3 for a discussion on sampling strategies, and on the issue of "representativeness" of Web data).

In order to pursue the goal, the strategy of mining data through a commercial search engine did not seem the best option, given the drawbacks connected with it (cf. Section 1.4.1). It was then decided to retrieve Web data by crawling (cf. Section 1.4.3) and relying on a Web-based search engine only in the first stage of corpus construction, namely that of crawl seeding (the selection of the URLs from which the crawl had to start). Previous research on the effects of seed selection upon the resulting corpus (Ueyama, 2006) suggested that automatic queries to Google which include words sampled from a traditional corpus like the BNC 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 dominated by "personal interest" pages, like blogs or bulletin boards.

Since it was desirable that both kinds of documents were included in the corpus, relevant sources were chosen from which words to be used as seeds could be sampled. The BNC was used as a first source, from which 2000 mid-frequency content words were picked, thus excluding function words, which, as suggested by Baroni and Ueyama (2006), may yield unpredictable results, since search engines usually ignore function them when submitted as part of a query. Moreover, since preliminary experiments (as reported in Baroni and Kilgarriff, 2006) demonstrated that issuing single-word queries to Google could lead to retrieval of inappropriate pages (like definitions of the word in Web-based dictionaries or pages of companies with the word in their name), the BNC sample words were paired randomly. Two other lists of bigrams were

then created, one extracted from the demographically sampled spoken section of the BNC, which should contain basic vocabulary, typical of spoken conversations, and the other from a vocabulary list for learners of English (henceforth ESL),[2] which, unlike what one might expect, contained formal or uncommon words. 20 randomly selected pairs of seeds used for the crawl are provided in Table 2.1.

BNC SEEDS	**BNC DEMOGRAPHIC SEEDS**	**ESL SEEDS**
aspects file	cooking ground	populate fist
sensitive presumably	cool police	statewide pliant
pilot consumption	general damn	reasonable frustrated
radio lots	smaller leaving	abhor colorful
johnson reduce	keen bedroom	snow visage
acceptable self	houses otherwise	attach elevator
guidance williams	thrown carrots	petal phlegmatic
yorkshire leaves	tapes double	sniff chum
session desk	chip fairly	ankle tabloid
beer scale	certain happy	lieutenant overhand
surprised raise	young given	secretarial validity
arranged eventually	beer pieces	prom overcame
dependent regulations	sink massive	deprived overhaul
gain silence	living council	ad-lib scraps
everywhere sentence	gate stuart	incompetent fanciful
ireland phase	shame shower	integral feat
ancient definition	particular poor	jargon incidentally
carefully discipline	joking bags	foible whole-wheat
bell frame	doubt prices	aerospace gender
thousands contemporary	months salad	dynamo thermos

Table 2.1. Randomly selected bigrams used as seeds for the crawl.

For each of these lists a set of URLs was obtained from the `.uk` domain by querying Google (see Table 2.2); repeated URLs were discarded and only one page per domain was kept, to ensure that the largest possible number of domains were represented. The procedure resulted in a list of 6,528 URLs, which were fed to the crawler.

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

The crawl was performed using the Heritrix[3] crawler, with a multi-threaded breadth-first crawling strategy, and was limited to pages in the `.uk` domain. This does not of course ensure that all the pages retrieved represent the British variety of English (which would be desirable, insofar as ukWaC should be comparable to the BNC). Nonetheless, the strategy was used as a simple heuristic to retrieve the largest possible number of pages which are (supposedly) published in the United Kingdom. Moreover, the crawl was restricted to pages whose URL did not end in a suffix cueing non-HTML data (`.pdf`, `.jpg`, etc.). The crawl ran for about three months, retrieving 75 GB of gzipped archives[4] (the Heritrix output format).

http://www.ilook.fsnet.co.uk/ora_sql/sql_02.htm
http://www.jubilees.co.uk/photos/45595a.html
http://www.online-betting-guide.co.uk/horse_racing.php
http://www.culture.gov.uk/alcohol_and_entertainment/ licensing_appforms.htm
http://www.nelh.shef.ac.uk/nelh/kit/msk/docs.nsf/0/ 3d01bcb0a7b09d7a80256cc400421b94?OpenDocument&Click=
http://www.derrenbrown.co.uk/news/messiah
http://www.cse.dmu.ac.uk/~cph/VR/whatisvr.html
http://www.clairecurtisthomas.labour.co.uk/ViewPage.cfm?Page=17301
http://www.jr2.ox.ac.uk/bandolier/booth/hliving/FVbreast.html
http://www.kgap.co.uk/Photo%20group%20hill.htm
http://www.woodlands-junior.kent.sch.uk/customs/questions/glossary/index.htm
http://icnorthwales.icnetwork.co.uk/news/regionalnews/
http://www.sohp.soton.ac.uk/neuro/timetable.htm
http://www.footballiq.co.uk/news/index.php?serendipity%5Baction%5D =search&serendipity%5BsearchTerm%5D=Matthew%20Spring
http://www.uk-muscle.co.uk/general-articles/14422-exercise-life-keep-fit-retirement.html
http://www.mochdrecc.freeserve.co.uk/Page143.htm
http://www.cont-ed.cam.ac.uk/BOCE/AdLib22/article2.html
http://www.bullbearings.co.uk/news.article.php?article=729653
http://www.pennardhillfarm.co.uk/
http://www.londonexternal.ac.uk/about_us/index.shtml

Table 2.2. Randomly selected URLs used as seeds for the crawl.

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

[4] It has to be highlighted, however, that the server that was used was experiencing performance problems at the time. In fact, the crawls of the aforementioned German and Italian corpora were let run for 10 days, retrieving a similar quantity of data in a much shorter time.

2. 3. 2 POST-CRAWL CLEANING

2.3.2.1 Preliminary filtering

After collecting the data from the Web, they had to be processed, so as to remove undesired noise and thus obtain a reasonably "clean" corpus. The first step consisted in identifying and discarding all sets of documents that were perfect duplicates. Experience gathered during the construction of the German and Italian corpora taught that documents that are identical before the removal of HTML code are likely to be error messages or copyright statements from the same servers; for this reason, not only the duplicates of a given document were removed, but also the document itself.[5] Subsequently, documents were discarded that were not of mime type `text/HTML`, and whose size was below 5KB or above 200KB, following an observation by Fletcher (2004b), who noted that very small documents tend to contain little human-produced text, whereas big documents are usually listings of various kinds, such as product catalogues or library indexes.

2.3.2.2 Boilerplate stripping and code removal

A crucial issue that needs to be tackled when constructing a Web-derived corpus is the presence in crawled pages of boilerplate (cf. Section 1.3). Boilerplate constitutes a serious problem for linguistic analysis of the corpus, since it may thwart attempts to analyse KWiC displays and, perhaps even more seriously, invalidate statistics and linguistic generalisations drawn from the corpus. It was therefore necessary to spot and remove as much boilerplate as possible.[6] This was done by applying a re-implementation of the algorithm of

[5] The strategy of eliminating both copies of such documents may be seen as rather arbitrary, especially because it discards texts which belong to a textual typology typical of the Web. However, it is very likely that despite the filtering procedure a considerable amount of, e.g., copyright statements remain in the corpus. This might be the case if only one copy of a document is retrieved. Thus, the strategy should be interpreted as an "operational" one, which is meant to prevent these text categories from appearing in disproportionate amounts in the corpus.

[6] As pointed out in several sources (Bernardini, *et al.*, 2006; Baroni, and Ueyama, 2006), this would not be the case if one aimed at studying the navigational structure of Web documents or its relation to the linguistic characteristics of Web pages.

the Hyppia project BTE tool,[7] which is based on the idea that content-rich sections of a page will have a low density of HTML tags, whereas boilerplate tends to be signalled by a larger amount of HTML, since it is usually characterised by special formatting, many newlines and links, etc. The main drawback of the method adopted is that it produces a corpus made up of *fragments* of Web pages. These, however, may be suitable if the aim of the collection is to provide a resource including samples of natural language, provided one is aware that complete and structured documents may not be available (cf. Section 1.2 for a discussion on sampling strategies).

After using HTML code to determine the ratio of tokens to tags for the purpose of boilerplate stripping, tags were removed.

2.3.2.3 Language and pornography filtering

Despite the crawl being in the `.uk` domain, there was no guarantee that all the pages retrieved would be in English. The strategy adopted for filtering out pages in other languages was founded on the notion that connected text should contain a high proportion of function words (Bayen, 2001), and therefore that all documents that did not meet this criterion could be discarded. The list of function words contained 151 items and included word classes like determiners, prepositions, auxiliaries and the like. Such filter should also remove pages containing too high a proportion of other undesired material, such as lists of numbers and non-linguistic characters.

Another desirable step was that of eliminating pornographic pages. This was not done for any reason of censorship or prudishness, but because they often contain long machine-generated texts, which are probably used to fool search engines. A list was therefore created of the words that are highly frequent in pornography, and all the documents that contained 3 types or 10 tokens from that list were discarded. The list was derived from the analysis of a corpus created *ad hoc* and made up of almost 200 pornographic pages; a frequency list was obtained from it and was cleaned manually, so as to remove

[7] http://www.aidanf.net/software/bte-body-text-extraction

words that, despite being very frequent in pornography, are totally "innocent" if taken in isolation (like *girls, men, young,* etc.).

The boilerplate stripping and filtering phase took almost 2 months[8] and produced a version of the corpus containing 5,690,218 documents for a total of about 19GB of uncompressed data.

2.3.2.4 Near-duplicate detection and removal

While it was relatively trivial to recognise and remove perfect duplicates from the corpus, a much more complex task was that of detecting near-duplicates, i.e. documents that share a significant portion of text but are not identical (what may differentiate them is, e.g., their header or date). In order to do this, a simplified version of the "shingling" algorithm (Broder *et al.*, 1997), implemented in perl/mysql, was adopted. The following description of the procedure is taken from Baroni and Ueyama (2006: 35), who performed the same procedure on the Italian corpus mentioned above:

> For each document, after removing all function words, we take fingerprints of a fixed number s of randomly selected n-grams (sequences of n words; we count types, not tokens – i.e., we only look at distinct n-grams, and we do not take repetitions of the same n-gram into account); then, for each pair of documents, we count the number of shared n-grams, which can be seen as an unbiased estimate of the overlap between the two documents.

If a pair of documents was found that shared more than *x n*-grams, one of the two documents was discarded. In order to avoid inconsistencies, the documents were ordered according to their ID, and only the second document of each pair was removed. The experimentations that preceded the construction of the Italian and German corpora instructed us also about the parameters that we had to set. In particular, we randomly picked 25 5-grams from each document, and looked for documents that shared as few as two of these 5-grams. If one or more documents did, they were considered as near-duplicates, and were therefore removed from the corpus (notice that, unlike in the perfect duplicate

[8] Such a long period of time was due to the aforementioned server problems. Indeed, when the machine was repaired the filtering was halfway through the process. The remaining part of the corpus was processed thereafter in less than four days.

detection phase, in this case the first copy of the document is not discarded). As pointed out by Baroni and Ueyama (*ibid.*), "this threshold might sound surprisingly low, but the chances that, after boilerplate stripping, two unrelated documents will share two sequences of five content words are very low". This phase of filtering took four days and produced a corpus made up of 2,692,645 documents, for a total size of about 12GB of uncompressed data. The decrease of the corpus with respect to the initial size of the crawled data, as can be noticed, was impressive: in this phase only, about three million documents were removed from the corpus.

2.3.2.5 Part-of-speech tagging, lemmatization and indexing

Part-of-speech tagging and lemmatization were performed using the TreeTagger.[9] This phase took about four days and resulted in a corpus which in its final version contains around two billion words, for a total size of 32 GB of uncompressed, annotated data. Figure 2.3 shows an example of the annotation procedure's output.

```
<text
id="http://www.luciesfarm.co.uk/acatalog/Dog_Cakes_and_Cookies
.html">
<s>
The     DT      the
ultimate        JJ      ultimate
birthday        NN      birthday
treat   NN      treat
for     IN      for
your    PP$     your
dog     NN      dog
.       SENT    .
</s>
<s>
A       DT      a
birthday        NN      birthday
cake    NN      cake
with    IN      with
his     PP$     his
or      CC      or
her     PP$     her
picture NN      picture
.       SENT    .
</s>
```

Figure 2.3. Example of a sentence encoded in ukWaC after the annotation procedure was carried out

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

ukWaC was then indexed, so as to make it possible to access it through a query tool in a fast and efficient way. The tool adopted was the IMS Corpus WorkBench (CWB. Christ, 1994),[10] a free indexing and retrieval toolkit. CWB is particularly suited to handle very large corpora, and supports very complex queries, such as searches for POS-tags and regular expressions. On the negative side, the tool does not index corpora larger than 450 million tokens as a single database. ukWaC had therefore to be split into various sub-corpora, which, while enabling faster retrieval of results on single portions of the corpus, makes it harder and slower to query the corpus in its entirety. In Figure 2.4 an example is provided, for merely illustrative purposes, of a complex search that it is possible to make by querying ukWaC through CWB. The search involves the use of POS tags to find the most frequent adjective-noun pairs in the first sub-portion of the corpus, which are then sorted according to their frequency:

```
UKWAC01> adjective-noun = [pos="J.*"] [pos="N.*"];

UKWAC01> count adjective-noun by lemma %cd on match..match[1];

10399   more information  [#2341889#2352287]
8979    young people  [#4204155#4213133]
7305    further information  [#1322900#1330204]
7143    last year  [#1920868#1928010]
6427    wide range  [#4151989#4158415]
6024    local authority  [#2029023#2035046]
5767    first time  [#1184215#1189981]
4881    same time  [#3414602#3419482]
4296    more detail  [#2329457#2333752]
4026    good practice  [#1438964#1442989]
3840    many people  [#2201006#2204845]
3719    high quality  [#1583731#1587449]
3221    many year  [#2219191#2222411]
3220    high level  [#1574419#1577638]
3043    long term  [#2086464#2089506]
2947    high education  [#1567710#1570656]
2935    further detail  [#1316182#1319116]
2852    last week  [#1917160#1920011]
2783    mental health  [#2261142#2263924]
```

Figure 2.4. Example of a search exploiting POS-tag annotation. The first 20 results are displayed.

[10] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

2.4 Concluding remarks

In the present Section, ukWaC was presented and its construction procedure described. It was argued that as a Web-derived, possibly "balanced", stable and annotated corpus, ukWaC may provide a valuable alternative to other existing language resources, and that, given its size and the nature of the data it contains, its construction might be seen as welcome news for researchers who are interested in studying rarer or relatively recent language phenomena.

Since semi-automated procedures were used to build it and post-process its data, however, its composition cannot be determined *a priori*. For this reason, *post-hoc* evaluation is crucial in order to assess its features and potential problems. This forms the subject of Chapter 3.

3

EVALUATING ukWaC THROUGH WORD LIST COMPARISONS

3.1 Introduction

Semi-automated methods of corpus construction allow for limited control over the contents that end up in the final corpus. A filtering phase is needed to discard documents which are deemed to constitute noise or contain uninteresting linguistic material (see Section 2.3.2), yet the actual corpus composition after this phase is still not known to the researcher. *Post-hoc* evaluation plays therefore a key role and its purposes may be manifold, from assessing what kind of documents make up the corpus (and, possibly, in what proportions), to determining the main topics and domains that are covered and examining the language that is used. As for all Web-mined corpora, the aim of the evaluation will ultimately be to ascertain the adequacy of the corpus under consideration in relation to the purpose it was built to serve. In the present case, ukWaC was built to provide a large "general-purpose" corpus of English, which would be comparable to traditional "balanced" corpora like the British National Corpus (Aston and Burnard, 1998). Since the concepts of "general language" and "balancedness" are far from well understood (for a discussion see Kilgarriff and Grefenstette, 2003; cf. Section 1.2), what can be done is therefore to assess to what extent ukWaC is similar or dissimilar to a benchmark that is widely assumed to have such features, i.e. the BNC.

The present Chapter discusses different methods for evaluating Web corpora proposed in the literature (Section 3.2) and describes in detail the one that was applied to the evaluation of ukWaC (Section 3.3). Several word lists were created for ukWaC and the BNC, each containing the word items that were identified by the TreeTagger as belonging to the main part-of speech categories. The word lists were then compared across ukWaC and the BNC via the log-likelihood association measure. Section 3.4 presents the results of the analysis, which are summarised and discussed in Section 3.5.

3.2 Related work

Despite the great interest in the Web as a source of linguistic data, limited work has been devoted so far to the qualitative analysis of Web-derived corpora. Among the researchers that have addressed this issue the names of Sharoff (2006), Ueyama and Baroni (2005) and Fletcher (2004b) can be mentioned. The former two build reference corpora for German, English, Russian (Sharoff) and Japanese (Ueyama and Baroni) using the BootCaT toolkit (Baroni and Bernardini, 2004), and then carry out an evaluation to discover how varied the collections of texts are in terms of their lexicon, and the genres and topics that are covered. In particular, Sharoff devises a statistical method to determine the number of documents that is needed to constitute an adequate sample of the whole corpus. He then randomly selects a sample and analyses it manually, in order to calculate statistics about the proportions of text genres and domains, as well as other meta-information like authorship (single, multiple, or corporate) and mode (written, transcripts of spoken language, or spontaneous communication through chats and the like). The classification of texts is carried out following a simplified version of that which was proposed by Sinclair (2003) for the European Advisory Group on Language Engineering Standards. A similar attempt to catalogue a significant random sample of texts according to their genres and domains is carried out by Ueyama and Baroni (2005). The two authors use a slightly different set of categories from Sharoff's in order to compare the composition of two Web corpora which were retrieved using the same seeds following a time interval of one year.

Fletcher (2004b), too, manually analyses his Web corpus, but his purposes are different from the other authors'. He constructs a general-reference corpus of English via automated queries to the AltaVista search engine for the 21 most frequent words in the BNC, and applies different filters to reduce various kinds of "noise" in the data retrieved, such as identical and almost identical pages. Subsequently, he skims through all the 7,038 documents that passed the first cleaning phase to detect and discard what he calls "fragmentary" texts, i.e. texts containing little or no connected prose. This

allows him to formulate a "rule of thumb" to determine the average size in bytes of a "good" Web page, if we mean by this a document that contains a reasonable amount of running text (the same rule was applied during the filtering phase of ukWaC; cf. Section 2.3.2.1).

For the purposes of the present analysis, however, none of the methods proposed seems adequate. As pointed out by the authors themselves, the categorisations provided by Sharoff (2006) and Ueyama and Baroni (2005) rely to some extent on the researcher's subjective interpretation, which may thus vary, and are also further hindered by the lack of comprehensive and consistent schemes to classify Web pages by genre (Santini, 2005). Even if an extensive classification of the Web texts in ukWaC were carried out, the doubt would remain as to whether its results are truly comparable to those of other studies using different sets of categories to analyse the same or different corpora. As regards the method of analysis presented by Fletcher (2004b), the procedure he follows has the sole intent of discarding documents, and even though the author reports his "impressions" on the composition of the corpus in terms of topics after this "visual dash", the method is not conceived as a means to evaluate the corpus. Besides, even attempting a "visual dash" of ukWaC, with its almost two billion words, would require an unreasonable amount of work.

What the three aforementioned studies have in common is another method of corpus evaluation, namely that of analyses through comparisons of word lists. The ways in which the comparisons are carried out, however, differ. Fletcher restricts his analysis to the observation of significant differences in the frequency ranks of the most frequent word forms in his corpus and the BNC to detect those which are relatively more typical of one or the other. Baroni and Ueyama use a more refined statistical method for corpus comparison, the log-likelihood association method (Dunning, 1993), through which they investigate the most typical lexical items of the two main genre types in their corpus, i.e. *blog* and *diary*. Finally, Sharoff, whose work is the most closely related to the present study, uses the log-likelihood statistic to compare frequency lists obtained from his Web corpus and the BNC. As is suggested by his work and by other studies (Rayson and Garside, 2000), this is a fast and reliable way to

understand the major differences between a newly acquired corpus and a known benchmark corpus, and can suggest ways in which one corpus is less balanced than the other, which is also one of the main aims of the evaluation of ukWaC.

At this point, a number of issues relating to the comparison of corpora in general should be raised. In particular, we wish to challenge Rayson *et al.*'s (2004) view according to which *homogeneity* and *comparability* are important features when it comes to comparing two corpora that are viewed as *equals*, and as such should have roughly the same size. Homogeneity is defined as the presence in a corpus of the same or similar "sections" which are featured in the other corpus under consideration, and comparability as the use of the same "stratified sampling method" of corpus construction (*ibid*.: 2). It is very likely that the authors put forward such suggestions having in mind the special kind of comparison that can be carried out between the Brown (Kucera and Francis, 1967) and LOB (Johansson, 1980) corpora. As they state:

> This is the case with the Brown and LOB corpora […], since LOB was designed to be comparable to the Brown corpus, and neither corpus was designed to be homogeneous. (*ibid*.:2)

The only other kind of corpus comparison that is taken into consideration in the aforementioned study is that "of a sample corpus with a large(r) standard corpus" (*ibid*.:1), the latter being a normative corpus representative of general language. This approach may be seen as rather limited, since it only takes into account "traditional" corpora, and does not consider the instances in which comparison is used as a *post-hoc* evaluation method, i.e. when the composition of one of the corpora is not defined *a priori*, as is the case with the LOB, Brown and BNC corpora. In fact, as with all collections of texts built in (semi-) automated ways, *homogeneity* within and/or across the corpora is not a necessary condition for the comparison to be carried out, but is a feature that the corpora may turn out to have or not to have *after* the comparison is carried out. Thus, for example, we compare the BNC and ukWaC, which was built to be similar to it (irrespective of its size), *even if* it is not known in advance whether they contain the same "sections" (if by this term is meant groups of

text belonging to a similar genre or discussing similar topics), and *even if* the sampling method that was used to build them is completely different. The extent to which the two corpora can be seen as homogeneous will be an interesting datum in itself, and not the undesirable outcome of a comparison made between two non homogeneous corpora. Also, which one of the two can be seen as "more representative" of general language, and whether it makes sense to ask such a question at all, is a point that needs to be investigated empirically, and should not be taken for granted.

3. 2. 1 THE BRITISH NATIONAL CORPUS

The British National Corpus (Aston and Burnard, 1998) is a large synchronic corpus containing around 100 million words. It was published for the first time in 1994. Designed to be a balanced corpus, it is composed of written texts (90%) and spoken transcripts (10%). It is also a sample corpus, in the sense that for the most part it includes portions of texts, instead of whole texts. Each sample includes between 40,000 and 50,000 words. The written part is made up of a wide-ranging variety of texts, identified and sampled according to their domain (i.e. their subject field), time of production, and medium (e.g. book, periodical, etc.). The BNC includes therefore books (fiction, non-fiction and academic) and newspaper and magazine articles, as well as a great variety of "minor" texts, such as personal letters, brochures and reports. The spoken part was collected according to two criteria. On the one hand, spontaneous conversations were recorded, and the speakers were selected so as to constitute a significant random sample of the population, taking into account criteria such as their age, sex, social class and geographic region. On the other hand, context-governed speeches may be found, such as interviews, business and government meetings. In the intentions of its creators, the British National corpus should thus "characterize the state of contemporary British English in its various social and generic uses" (*ibid*.: 28).

3.3 Methodology

In the present Section the actual way in which the comparison was carried out is described. As mentioned in Section 3.1, the main object of comparison are word lists derived from the two corpora; the BNC is used as a benchmark corpus and the log-likelihood association measure as a statistic to analyse the differences between the word lists. Sharoff (2006), generated a single list that gave prominence to the words with the highest log-likelihood scores *in general* (the relatively most typical in either corpora). Instead, the method that is proposed here consists in creating separate lists, every one of which includes all the words that were identified by the TreeTagger as belonging to one of the main part-of-speech classes.[1] This means that lists will include, e.g., the nouns that have the highest log-likelihood score in either ukWaC or the BNC, and which are, therefore, key *nouns* (and not key *words* in general) for that corpus. While it is true that such a procedure relies heavily on the tagger's performance, it also makes it possible to carry out a more thorough analysis of the corpus than a simple keyword list would do, especially because a wider range of homogeneous word classes/language aspects can be examined in greater detail.

Hence, five pairs of lists were created for the word classes of *nouns*, *verbs*, *adjectives*, *adverbs* ending with the suffix *–ly* and *function words*,[2] with each couple including a list of the words that are, respectively, most typical of ukWaC or the BNC. For the classes of adjectives and adverbs, all the lemmas that bear the corresponding tags were extracted (see the TreeTagger Web site, [3] and Santorini, 1990 for reference to the complete tag set). The results are then lowercased and all items containing non-alphabetic characters, like word-interior hyphen, are discarded. While this procedure leads to the elimination of a considerable number of word items, even if they are well-formed, meaningful words (e.g. *bad-tempered*, *good-looking*), on the positive side it reduces noise in the lists, in particular in those pertaining to ukWaC, where one would expect

[1] In fact, function words were conflated into a single list for ease of comparison.
[2] As we shall see in greater detail in Section 3.4.4, this class corresponds to the words that are considered by the TreeTagger as grammatical (rather than content-rich) words.
[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

words containing non-alphabetic characters to be frequent (e.g. e-mail addresses, acronyms, formulae of different nature, etc.), and likely to mislead the tagger. As regards this last point, it is important to highlight that, in order to obtain truly comparable results and to minimize differences in the word lists that would be due to different tokenization and annotation procedures, the version of the BNC used (BNC World Edition) was tokenized, lemmatized and POS-tagged (by the TreeTagger) with the same scripts as ukWaC. In the final stage, each list is compared with its counterpart via the log-likelihood measure, taking the BNC as a reference corpus when calculating the key words of ukWaC, and vice versa, and then sorting the results according to their score, from the highest to the lowest. The same method is applied to the creation of the word lists of nouns and verbs, but this time word forms are used instead of lemmas, since they provide more detailed syntactic information about the words' behaviour in the corpus, such as the use of predominantly singular or plural forms for the nouns, and of present or past tense forms for the verbs.

In the next sections the results of the comparison are presented for each of the above mentioned word categories (in Appendix from 1 to 10). A more thorough analysis will be dedicated to nouns, which, it can be argued, are the most useful indicators of the composition of the corpus, mainly in terms of topics that are covered. In particular, 250 randomly selected concordances will be analysed for each of the first 100 items of the lists. For the categories of verbs, adjectives, and adverbs the number is reduced to 50 items, and to 20 for function words. The latter are certainly significant in the study of certain linguistic aspects of texts within a corpus, but they usually provide little insight into its composition, especially in terms of topics.

3.4 Results

3. 4. 1 NOUNS

3.4.1.1 Nouns most typical of ukWaC

The first word list (Appendix 1) that is analysed is that of the nouns which turn out to be the most typical of ukWaC with respect to the BNC. These, it will be remembered, are not the nouns most typical of ukWaC in absolute terms, but

only the ones that turn out to be significantly more present in ukWaC than in the BNC. At first glance, it would appear that most of these words could be categorised as being related to three macro-topics, i.e. **a)** computers and the Web (among the first ten words we find *website*, *web*, *email*, and *internet*), **b)** education (e.g. *students*, *research*) and **c)** public service (e.g. *organisations*, *nhs*, *health*). This suggests that, compared to the BNC, ukWaC contains a higher proportion of texts dealing with these topics, and may therefore be seen as "unbalanced" in this respect. However, a closer look at the contexts in which the words occur may provide a better insight both into the categories outlined above and into the composition of the corpus.

Perhaps the most prominent category is that of the words which seem to belong, broadly speaking, to the semantic fields of computers and the World Wide Web (a). In this category words are found like, e.g. *website*, *site*, *click*, *web*, *email*, *internet* (top of the list), *browser*, *software*, *server* (middle of the list), *database*, *password*, *forum* (end of the list). The relatively high frequency of these words, for each of which 250 randomly selected concordances were analysed, reveals that ukWaC contains a considerable number of texts whose topics are either issues revolving around software and hardware components for computers, or web-related issues. The category can be further split into two sub-categories. Sub-category (a.1), consists of words related to computers, and includes, e.g., the terms *pdf*, *file*, *software*, *server*, *cd*, *password*, *database*; these tend to occur in a rather limited range of text genres, which could be classified as "instruction" texts, i.e. texts which "explain how to do something" (Sharoff, 2007), like instruction manuals or online tutorials, and "discussion" texts, i.e. "texts […] aimed at discussing a state of affair" (*ibid*.), like forums in which users exchange opinions about a particular computer program or hardware component.

```
   performance can be severely impacted if either the swap
<file> or applications are on a slow drive;
   and your NDS password in the Password and Confirm
<password> boxes. If you already have another dial-up
internet connection;
   Which is why we think the release this week of
affordable <software> offering DVR-like capabilities for
web radio is significant;
```

At the lower end, PCs now account for 70 % of the total LAN <server> market. The use of servers in Japan therefore will also;

In sub-category (a.2) words are found that clearly refer to the web, like *website*, *site*, *internet*, *links*, *download*, *forum*, etc.; if the texts in which they typically occur are analysed, it becomes clear that they differ from the words in sub-category (a.1), insofar as they are distributed across a wider variety of text genres and in texts dealing with different topics. This is not too surprising if one considers that they are meta-references to the medium of communication that hosts them. Thus, for example, such words as *website* or *download* may be found not only in discussion pages about the structure of hyperlinks of the Net or problems of bandwidth, but also in promotional texts – whose communicative intention is called by Sharoff (2007) "recommendation" – introducing a firm or a web-based resource. Alongside these words, we find others that, although they are not "traditionally" or chiefly associated with the Internet or computers, are nonetheless frequently attested in ukWaC in computing- or web-related contexts. These include *access*, which often refers to "access to the Internet"; *list*, as in "mailing list"; *users*, which is frequently the subject of instruction or promotional texts regarding software programs or Web services; *format*, as in "file format"; *search*, which is frequent in help pages on how to navigate a site or discussion texts on how to surf the Web; *images*, which is featured in a number of texts about handling image files:

users say the same thing: they don't want to wait for slow <download> times." Other people did research on [computer] response times;
unified body can possibly represent the interests of both <software> publishers and software users when it comes to legal disputes over;
At the heart of our innovative degree is the belief that <software> should be imaginative and satisfy the needs of people who will be using;
It 's easier, by the way, to provide <access> as in the first examples I list above because you 're explicitly;
menus are to apply a filters to your search. Enter your <search> criteria in the text box, ie Pensions, Tax, Jobs etc, note that.

To summarise, the presence of the words belonging to category (a) among the most typical of ukWaC can be accounted for by the presence in it of

a significant number of texts which, despite the fact that they may belong to a variety of text genres, share the common topic of computers and/or the World Wide Web. It is true that such domains might be regarded as being overrepresented in ukWaC, which would mean that the objective of creating a balanced general-purpose corpus was not fully achieved. However, a number of arguments can be raised to oppose or at least modulate this view. First of all, as pointed out, among others, by Kilgarriff and Grefenstette (2003), building a "general-language" corpus does not entail the exclusion of sublanguages, as can be considered those associated with Web and computer technologies. Evidence of this is that the BNC itself contains texts belonging to technical and specialised sub-domains (Lee, 2001). Secondly, and perhaps more importantly, a corpus like ukWaC could be used to study the usage of the relatively "new" words (or the re-lexicalisation of "old" words) that are produced within the constantly changing field of new technologies, and that are unattested in traditional corpora. As an example of this, a word which has become part of everyday language like *website* does not appear at all in the BNC. On the other hand, the fact that certain words, such as *site*, occur *typically* in ukWaC in Web-related contexts, does not imply that other usages of the same words are not attested. On the contrary, *site* also appears both in its metaphorical and concrete sense, as well as in medical contexts:

```
    the market benefits of water liberalisation, seeing the
industry as a <site> for economic growth;
  the proposal in relation to other buildings within the
<site> and <site> boundaries together with the position of
buildings and highways…;
  The duodenum is the most common <site> for a peptic
ulcer to occur.
```

More problematic in terms of corpus composition is the presence of a set of words which only an analysis of the concordances can reveal. These words are not typical of any particular domain and can be found, within connected text or – much more frequently – as isolated text elements, in any text of the corpus, irrespective of its genre or topic. In other words, they represent boilerplate (cf. Section 1.2). Some examples of these words are *information*, *click*, *details*, *links*, *comments*, *contact*, *fax*, *copyright*, *feedback*. Other words

that can be classified as belonging to this class are words which only apparently belong to the category we discussed in the previous paragraph, like *download* or *file*. Such words usually appear in highly recurrent patterns such as "For further information", "Click here", "Contact details", "Download the file",[4] or in invitations to users to leave comments or feedback about a website or a service. For the purpose of evaluating the composition of the corpus, it is evident that these words and the contexts in which they appear give no hint as to the topic or genre of the text they appear in. Unless researchers are interested in studying the language of web pages, which is not our case (cf. Section 2.3.2.2), they are therefore undesirable items. Their only utility for our purposes could be that of providing inputs as to how to refine data cleaning techniques.

Other examples of problematic words in terms of corpus composition are *pm*, *aug* and *feb*. These appear in all the occurrences analysed as part of the details concerning the time ("p.m.") and date (respectively "August" and "February") when a message was posted to an online discussion forum or blog. On the one hand, it can be argued that they reveal that a significant number of texts belonging to these genres are featured in the corpus. This is a welcome finding, since it demonstrates that informal, interactive texts produced by users are included in ukWaC. However, one could argue that the ideal situation would be one in which only the user-submitted texts remain, and the repeated linguistic structures that "surround" them are eliminated by post-processing.

Finally, three other words can be mentioned that have turned out to be signals of potential problems in corpus composition, i.e. *poker*, *insurance* and *quot*. These appear the great majority of times within machine-generated texts (i.e., spam). Like "boilerplate words", such texts are uninteresting and distort statistics about corpus composition, and should therefore arguably be removed

Turning to macro-category (b), among the 100 most typical nouns of ukWaC, several seem to be related to the topic of education and universities

[4] The presence of *click* in the list of nouns, although the word is more frequently used as a verb, as the recurrent pattern "Click here" demonstrates, can be accounted for by errors of the POS-tagger. This is easily mislead by boilerplate, since, as has been argued, this usually appears within unconnected text.

(e.g. *students*, *research*, *guidance*, *training*, *learning*). The analysis of the concordance lines, for each of which the associated URL was also checked, confirmed that ukWaC contains a large proportion of texts whose "initiators" (i.e. the entities which are responsible for the publishing of contents) are universities or whose topic is education, either academic or professional. What is most remarkable is the variety of the text genres which are featured. As pointed out by Thelwall (2005), university sites may contain very different kinds of texts, whose communicative intention and register can differ significantly. To mention only a few, "traditional" texts were found, like online prospectuses for students, course outlines, and academic papers, but also "new" web-related genres like homepages of members of staff or research groups. The high frequency of these kinds of text seems also to account for the presence in the list of key nouns like *skills* (e.g. in presentation pages detailing the skills students need to acquire), *project* (as in "students' *or* research project"), *funding* and *support* (the former referring to possible sources of funding for students or scholars, the latter to financial or psychological help they might need).

```
  providing <training> in the new technologies through
both individual tuition and courses;
  All Costume Construction students will develop <skills>
in time management, resource management, budgeting and
scheduling;
  since the mid-late 1980s in Tanzania and Zimbabwe. The
<research> project covered all types of post-secondary VET
provision;
  contribution towards their tuition fees. The level of
<support> you are able to receive towards your tuition fees
and maintenance.
```

Thus, even though a certain homogeneity was found in terms of the authors of these texts, the (desirable) variety of textual genres seems to be preserved. Moreover, such important presence of universities in the role of authors/initiators can be regarded as an indication of reliability and good linguistic standards of the sections of the corpus they are featured in.

Similar points could be raised referring to category (c) of nouns, i.e. those referring to public services. The authors/initiators of the texts in which these words typically occur are non governmental organisations or departments

of the government. This can explain the high frequency of words like *services*, *organisations*, *nhs*, *health*, and others that are perhaps less easily associated with public services, like *issues*, *development* and *opportunities* (which appear frequently in "discussion" texts about politics or economic issues), *network* and *community* (which are often used to indicate groups of citizens, e.g. committed to social issues or living in the same city), and *support* and *guidance* (in texts offering help to users for all kinds of matters: HIV, finances, disabilities, children, etc.). As in the case of texts authored by universities, the variety of text genres is remarkable. As an example, the concordances of the word *nhs* revealed that only a few texts were retrieved from a National Health Service site, the rest being either newspaper articles or commentaries (e.g. in personal home pages or in NGOs' sites) about administrative or quality issues regarding the services to the patients, or the treatment of diseases. Besides newspaper articles – which, however, were not among the most represented genres –, promotional ("recommendation") texts were found, such as introductory pages of NGOs and charities describing their mission and asking for donations, as indicated by the concordances of words like *funding*; and legal or politics-related texts, as indicated by the words *article* (which is featured in ukWaC, among other contexts, in pieces of legislation) or *consultation*:

```
   that much could be learnt for first wave consumer
protection <issues> given the perception that the
introduction of the euro in first wave;
   You may require a variety of services such as advice and
<support>, or relief from caring. Your disabled relative or
friend may require;
   Rural Development and the relevant district council. The
<consultation> responses will be considered in reaching a
decision on the final.
```

The purpose of the categorisation provided in this Section was to describe and generalise certain features relating to the composition of the corpus. Thus, it was not meant to define clear-cut patterns of usage of the nouns featured in the list. It does not aim to suggest that if a word is included in one macro-category of topics, the usage of that word in ukWaC is limited to the contexts mentioned. On the contrary, there is evidence that a significant number of the most typical nouns in the corpus appear in very diversified

textual genres dealing with different topics. More macro-categories could be included to account for the presence of other words in the list. For example, *event*, *team*, and *training* appear in sports contexts; *delivery*, *experience*, and *resource* are frequent in commercial sites; *design*, *music*, *album*, and *reviews* are often featured in text related to arts. For the sake of clarity, only the most significant categories were discussed, i.e. those for which data made it possible to infer clearly emerging patterns.

3.4.1.2 Nouns most typical of the BNC

The purpose of the analysis presented in Section 3.4.1.1 was twofold. On the one hand, it was intended to reveal in what regards ukWaC turns out to be "unbalanced" compared to the BNC, and, on the other, to assess the corpus' diverseness, or lack thereof, in terms of topics and genres that are covered. In other words, its aim was to investigate the differences between the two corpora while at the same time exploring the one whose composition was not known. Since the composition of the other corpus is well known (Burnard, 2007), the analysis does not need to call into question its internal structure. It can limit itself to focusing on the features that distinguish one corpus from the other, and therefore, in our case, investigate in what regards the BNC turns out to be "unbalanced" compared to ukWaC.

Groups of words will be analysed that show clearly emerging patterns, which are taken as indicators of the possible reasons why those words are featured in the list of the nouns most typical of the BNC. In Section 3.4.1.1 such features had to be inferred, rather rudimentarily, through analyses of the concordances and of the texts' URLs.[5] In the case of the BNC, the procedure is made much simpler by the presence of existing text classifications. In the present analysis the classification used is the one proposed by Lee (2001). Through the `/codist` function of the Corpus Query Processor (CQP; Christ,

[5] Automated methods of genre recognition for web texts are being studied (see e.g. Santini *et al.*, 2006), but it is far beyond the scope of the present study to apply them to ukWaC.

1994), it is possible to generate statistics for each of the first 100 items of the list,[6] so as to assess in what domains or genres they are most frequent.

In this regard, three points need to be raised. First, since the analysis does not aim to draw generalisations about language use, its results should not be taken to imply that the words taken into consideration are typical of one domain/genre or another. Second, macro-categories will be used. This means that only broad classes of topics and genres will be taken into account, such as fiction/imaginative vs. newspaper texts, or world affairs vs. social sciences domains. It has to be highlighted that no attempt is made at debating the theoretical justification for using such categories instead of others (on this issue, see e.g., Aston 2001). Finally, since the purpose of the analysis is to compare two corpora, and not to provide an exhaustive description of them, the results presented are not to be taken as absolutely faithful indicators of their composition. For instance, the presence of a word like "something" in the noun list – that should have been more properly tagged as a pronoun –, or "yesterday" – typically used an adverb –, suggests that the POS-tagger's performance might influence the results. It is likewise possible that using a different version of the BNC could result in different counts being produced. However, since general trends emerge which are not based on single cases, but rather on whole groups of words, the validity of the results does not seem to be undermined.

Moving on to the actual analysis of the words featured in the list (Appendix 2), three main categories can be identified, i.e. **a)** nouns related to the description of people or objects, **b)** expressions which are frequent in 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 nouns of body parts, like *eyes*, *face*, *head*, *hands*, *lips*, *arm*, *legs*, or of bodily actions, like *smile* and *breath*; words used to refer to people, such as *man*, *mother*, *woman*, *girl*, *boy*, *sir*, *husband*, *darling*, *lady*, *friend*; names of objects and places, like *door*, *house*, *bed*, *clothes*, *room*, *things*. All of these share the common characteristic of

[6] For practical reasons counts were produced for lowercase word forms only.

appearing in a clear majority of cases in texts classified as "imaginative" or "fiction/prose". As an example, *eyes* appears 74% of the times in "fiction/prose" texts; *man* appears in such kind of texts almost 41% of the times, and *room* about 47%. Other two categories of words that are found predominantly in imaginative texts are nouns indicating temporal events, such as *moment*, *night*, *tomorrow*, *morning*, and indefinite nouns and pronouns, like *something*, *nothing*, *thing*, *anything*. As we shall see, these two categories are also found in a significant number of texts belonging to the "spoken" section of the BNC.

In general, what can be inferred from the data is that, compared to ukWaC, the British National Corpus seems to contain a higher proportion of narrative fiction texts, in which we unsurprisingly find nouns related to the description of characters, objects and time. This seems to confirm that "texts aimed at recreation [such as fiction] are treated as an important category in traditional corpora" (Sharoff, 2006: 85, see also Fletcher, 2004b), 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.

The next category taken into consideration is that of expressions which are typically associated with the spoken language, including graphical transcriptions. Among the latter we find *er*, *erm*, *cos*, *mhm*, *ah*, which appear most frequently in the "spoken" sub-domain of the BNC. It is evident that these words are not nouns, but, 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 the tagset used (cf. Section 3.3).[7] Beside these words, we find others which are

[7] The presence of other words in the list can instead be explained by structural (i.e. non linguistic) differences between the two corpora. An example is represented by *ll* and *ta*: in the version of the BNC used, these forms (respectively the abbreviated form of "will", and a suffix used in verbs like "gotta") were not tokenised following the format expected by the TreeTagger, which was consequently misled by them. The word *emailinc* represents a conventional form used in the BNC to replace and hide the original email addresses present in the texts. Likewise, *speaker* and *studio* are very frequently found in transcriptions of broadcast news as conventional forms that indicate who is the speaker. For other words, like *cent*, and *pounds*, the only reason seemingly justifying their presence among the first 100 items of the

very frequently featured in the spoken section of the BNC, like *sort* (often used within the expression "sort of"), *lot* (as in "a lot of"), *bit*, and *mummy* to which the above mentioned pronouns and time indicators can be added (e.g. *something*, *nothing*, *night*, *tomorrow*). Spoken language is obviously less well represented in ukWAC than in the BNC, which was designed to contain 10% transcribed speech. This does not mean though that spoken-like genres are absent from the former, like, e.g. texts which reproduce informal, interactive, "spoken-like" language, as may be considered blogs and online forums of discussion (cf. Section 3.4.1.1).

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*, *party*, *unemployment*, *police*, *opposition*, *labour*, *court*, *state*, *republics*, and *spokesman*. All of these are mostly featured in texts that are classified as belonging to the domain "world affairs", "social sciences" or "commerce", and occur either in academic or non-academic texts, as well as in newspaper articles, e.g.:

```
  has already scored an important propaganda victory
against <government> forces, only a week after Vietnam said
it had withdrawn all its troops;
  election in which it had inflicted a massive defeat on
the <Labour> party. It was clearly not an all-party
government, yet,;
  companies controlled by Mr Cameron-Webb. Appearing in
<court> for the Corporation of Lloyd 's, Stephen Ruttle
said.
```

This may appear to be a problematic category, insofar as it seems to overlap with the group of words related to public services which, as was shown in Section 3.4.1.1, is typical of ukWaC. A possible explanation for this phenomenon could be that the texts dealing with politics and economy in ukWaC seem to be predominantly issued for "practical" purposes, such as offering guidance or promoting a certain governmental programme

---

list is different textual conventions: *cent* is very frequent in the BNC as part of the compound "per cent", which in ukWaC is more often written using the symbol "%"; the same holds true for "pounds", which is more frequent in the BNC than the symbol "£", whereas in ukWaC the proportion is reversed.

("recommendation" texts). Concordances reveal that in the BNC words like *government* or *opposition* are instead more frequently featured in texts (non-fiction books, newspaper articles, academic essays, parliamentary proceedings, etc.; cf. Lee, 2001) which comment on a given political or economic situation, and which therefore would be classified by Sharoff (2006) as "discussion" texts.

3. 4. 2 VERBS

3.4.2.1 Verbs most typical of ukWaC

Two broad categories emerge when analysing the verb forms most typical of ukWaC (see Appendix 3). The first category is that of verbs which seem to be associated with the description or the promotion of products or services.[8] In fact, verbs like *ensure*, *develop*, *offer*, *improve*, *create*, and *promote* often relate to goods or facilities that are offered either by private companies, universities or public institutions. In this respect, such class of verbs may be seen as cutting across the main domains that were identified in Section 3.4.1. The second prominent category is that of verbs which are identifiable as part of boilerplate, and includes words such as *posted*, *contact*, *updated*, and *email*. This category also includes words whose high frequency is due to systematic errors of the POS-tagger, which tagged grammatically ambiguous word forms, like *please* and *learning*, as verbs, even if concordances reveal that they are most often used as an adverb and a noun respectively, e.g.:

```
   teaching and research is best achieved through focusing
on <learning> as a process;
   There is always a risk of fire in every home so <please>
read this part carefully ; it could save your life.
```

Although such categorisation is useful to identify some types of texts that are featured in ukWaC, it cannot account for a number of the verbs in the list. Verb forms such as *need*, *required*, *allows*, or *aims* are not at first sight clearly associated with any text type or domain. In order to explain their presence in the list, it seems therefore useful to introduce a second type of categorisation.

[8] Cf. also Section 3.4.1.1, in which the presence of a considerable number of promotional texts was revealed.

Verbs will be both analysed in terms of the text types/domains they appear in (as was done in Section 3.4.1), and according to their *intrinsic* meaning.

The classification proposed by Biber *et al*. (1999: 360-378) seems particularly useful in this second respect. Such classification was applied by the authors to the most frequent verbs in the *Longman Spoken and Written corpus of English* (LSWE. *ibid*.: 24-40),[9] and was based on "seven major semantic domains" (*ibid*.: 361), corresponding to the "core meanings" of verbs. The core meaning of a verb is established on a frequency basis and represents the most typical use which is made of it. The semantic domains are as follows:

a) **activity verbs**, i.e. verbs that "denote actions and events that could be associated with choice" (*ibid*.). Examples of these verbs[10] are *use*, *provide*, and *work*;

b) **communication verbs**, i.e. "a special category of activity verbs that involve communication activities (speaking and writing)" (*ibid*.: 362). Examples are *publish* and *offer*;

c) **mental verbs**, i.e. verbs that "denote a wide range of activities and states experienced by humans; they do not involve physical action and do not necessarily entail volition" (*ibid*.). Examples are *need* and *find*;

d) **verbs of facilitation or causation**, i.e. verbs that "indicate that some person or inanimate entity brings about a new state of affairs" (*ibid*.: 363). Examples are *help*, *allow*, and *require*;

e) **verbs of simple occurrence**, i.e. verbs that "primarily report events (typically physical events) that occur apart from any volitional activity. […] They include *become*, *change*, *happen*" (*ibid*.: 364). No example of verbs belonging to this category was found in the list;

f) **verbs of existence or relationship**, which "report a state that exists between entities" (*ibid*.: 364), such as *include*, and *(be) located*;

g) **aspectual verbs**, "such as *begin*, *continue*, *finish* […] characterize the stage of progress of some […] event or activity" (*ibid*.). As was the case with

[9] The LSWE is a 40 million word corpus of British and American English, including four main text types, i.e. fiction, spoken texts, news and academic prose.

[10] The examples refer to verbs which are mentioned by Biber *et al*. (*ibid*.: 367-371) and are also featured in the list of the verbs most typical of ukWaC.

verbs of simple occurrence, no example of aspectual verbs is featured in the list.

As we shall see, some of the verbs most typical of ukWaC turn out to be among the most frequent in the LSWE, too. It has to be highlighted, however, that the categorisation provided for some verbs by Biber *et al*. does not always match the most typical use that is made of those verbs in ukWaC. As an example, *develop* is most often used in our corpus not as a verb of occurrence ("`Resistant organisms may develop in the alimentary tract`"; example from *ibid*.: 364), but rather as an activity verb (e.g. "`We have to find ways to <develop> learning software which create the same level of excitement among children`"). In such cases, verbs are classed according to evidence in ukWaC.

The approach has some evident limitations, such as the difficulty of classifying verbs whose core meaning may belong to more than one category (for a discussion, see *ibid*.: 360-361). However, it is a useful way of providing categories that are directly comparable across ukWaC and the BNC. Moreover, the results relating to the verbs of the LSWE can be used as a benchmark other than the BNC to compare ukWaC with. The final paragraphs will indeed be dedicated to a short comparison between the results obtained for ukWaC and those relating to the LSWE.

Category (a) is the most well-represented in the list (cf. Figure 3.1),[11] and includes the verb forms *use*, *provide*, *develop*, *work*, *visit*, *access*, *check*, *create*, *deliver*, *receive*, *add*, and *apply*. These occur frequently in recommendation (promotional) texts. Interestingly, as anticipated at the beginning, such texts are typical not only of advertisement materials issued by private companies, but are also found in Web pages promoting, e.g. a governmental programme, tourist destinations, university courses, or research groups' activities.

```
   powerful online assessment tool, designed to <provide>
high value computer based assessment;
```

[11] If the base form of a verb is included in the list, examples will not mention its inflected forms. However, all verb forms are taken into account when counts are produced about the distribution of verbs across semantic domains.

By joining Frank is Frank's affiliate program. You will <receive> 10 % commission for every sale you make;

The Council wants equal chances for everyone in Tameside to <work>, learn and live free from discrimination and victimisation;

These have long sandy beaches - and many places to <visit>. The climate is mild and the distances from the UK are smaller;

Aeronautics. Specifically created to perform research and to <develop> future leaders for aerospace manufacturing, civil and military aviation;

We anticipate that our findings would <provide> material for a number of papers that would be presented at academic.

Some texts are not easily classifiable as belonging to one single category, as in the case of seemingly informative texts, whose communicative intention is actually to advertise a product. A sentence like:

Future developments in hormonal treatment look to <provide> men with a contraceptive which is both highly effective and safe

published by a famous pharmaceutical company, can hardly be seen as having a merely informative function. In the same way, job vacancy announcements, which are quite frequent in ukWaC, are a mixed kind of text, partly informative – i.e. detailing the necessary skills of candidates –, and partly promotional – showing how serious and committed a firm is in recruiting its personnel, e.g.:

delivering consistent methods to establish and <develop> good working relationships with suppliers and acting as a mentor.

This corresponds to what Santini (2007: 6-8) calls “genre hybridism”, which often makes it challenging to classify web texts. For this reason, as was also pointed out in Section 3.4.1.1, the present classification of web texts according to their type or topic has to be intended as indicative, and not as a comprehensive and accurate description of the corpus composition.

Another type of texts in which activity verbs are present to a considerable extent is instruction texts. These can be either help pages or public regulations, guidelines of projects, or more traditional instruction texts, such as technical manuals for software or Web users, and recipes:

that all travellers have been immunised against polio; this <provides> protection for the individual traveller, but also, importantly;
solely in tribute to or criticism of a person or business <provided> that if: i. the Domain Name (not including the first and second level);
Selecting and deploying staff. Action 5.1 <Develop> and implement a policy to encourage vocations;
When you <visit> a web page, a copy of that page is placed in the cache;
Return soup to the saucepan. <Add> cream (if using), nutmeg, spinach and reheat.

Finally, activity verbs are frequently attested in discussion texts. Examples of this kind of texts are news and other types of articles dealing with disparate topics, such as family issues, national and international affairs, and art reviews:

have been organised because the teachers think the parents are <using> drugs. Opposite views were expressed;
into securing EU programmes that UK local authorities can <access>, so we must all make the most of this opportunity;
children are dying of AIDS. It challenges all religions to <work> together to reduce the stigma and discrimination;
As is often the case in such situations, determined artists <create> their own opportunities. The artist Algis Lankelis has curated sporadic.

Verbs belonging to category (b) and (c) are rarer in the list and seem to be less evenly distributed across text types than activity verbs. Communication verbs like *published*, *offer* and *promote* are found for the most part in promotional texts:

At Edinburgh, we <offer> a modern and innovative curriculum that provides excellent training;
Our aim is to actively <promote> responsible dog ownership and to reduce the number of stray dogs.

The same is true for the mental verbs *need*, *find* and *aims*. These are found in texts promoting, e.g. a product, or an organisation:

You can get a complete, fast, no-hassle refund. You don't even <need> to have a reason. That 's how confident I am in this material;
GuideStar UK is a registered charity that <aims> to promote the voluntary and community sector.

It can be noticed, however, that mental verbs are found in a somewhat wider range of texts, including discussion texts, such as academic papers, news articles and posts in discussion forums:

Thus, taking a firm-level perspective, this paper <aims> to question the extent to which ongoing globalisation has benefited;
But all depends on whether Member States will give their creation the resources and support it will need;
Horrified to <find> a stain of fluid under the car just betwen [sic] the radiator and front.

Verbs belonging to category (d) include *help*, *support*, *improve*, *ensure*, *required* and *allows*. As the label of the category ("verbs of facilitation or causation"; Biber *et al*., 1999: 363) seems to indicate, these verbs are frequently featured in one of the main text types that were identified in Section 3.4.1, i.e. in instruction texts such as help pages. These texts aim to facilitate the understanding of a topic, or the steps necessary to carry out a task:

Many patients with aortic regurgitation <improve> symptomatically during pregnancy;
This is precisely why our print tools <support> creating map prints at very precise, user supplied map scales;
Before the process starts we check the incoming wrought stainless steel to <ensure> it has the correct elemental composition. We use a hand-held X-ray; [12]
paper copies are acceptable . Five copies of each bid are <required> if they are in paper form. Applicants wishing to have receipt.

As in the case of activity verbs mental verbs seem to be distributed across a wide range of text types, among which recommendation texts are found, as well as a large number of discussion texts. These turn out to be mainly academic articles, and news articles, either published by organisations or online magazines and newspapers:

I am confident that these measures will <help> to increase visitor numbers to the Province and encourage local people";
These measures will <improve> NHS efficiency and staff morale and they will bring healthcare closer;
The MEMSCAP design kit <allows> users to customise the MEMS Xplorer and MEMS Pro engineering platform;

[12] In this example and the former, taken from texts issued by private companies, informative and promotional purposes seem to be intertwined.

This analysis also <allows> a confident assertion to be made about supermarket stations;
Using integers for internal storage <allows> precise equality comparisons to be done, which would not be guaranteed;
find the proper mental strategies to <help> to achieve the aim. The argument can also be turned on its head;
UK 's presidency of EU fails to <improve> consultation with the voluntary sector;
in conjunction with the revisions to HTM64, promises to <ensure> that the future supply of water in our hospitals is much safer.

The same variety of text types is attested when analysing the concordances of the verbs belonging to category (f), such as *include*, *based*, *contains*, *located*, and *designed*. These are typically used to describe (or describe and promote; cf. note 12) a product or an activity, and in discussion texts, which range from economic press articles, to academic papers, and editorials about current affairs:

umbrella that stand the test of the worst UK weather - ribs are <designed> to return back to original shape, should the umbrella be blown;
The park offers great facilities. Planet leisure <contains> a large indoor heated swimming pool, children’s play area;
The Childcare Company is a truly professionally run agency, <based> on true family values;
1.07 million in August, today's report showed. Automakers <including> General Motors Corp. have said they will cut production for the rest;
an approach to inter-operating information systems <based> upon globally defined schemas cannot work for non-centralised information;
Jews are still the favorite objects of Muslim contempt <based> on the quranic condemnation of them.

The last category that is going to be taken into account is that of boilerplate. As mentioned in the introduction of the present Section, such class includes both verbs which occur within recurrent patterns repeated across different pages, and words which were tagged incorrectly by the TreeTagger. These include *posted*, *contact*, *please*, *learning*, *top*, *posts*, *updated*, *download*, *following*, *view*, *read* and *email*. Even though some of these verbs could be included in some of the categories mentioned above, it was decided not to include them in the analysis, since their high frequency does not really indicate typical use in real, human-produced language.

If results are now compared across ukWaC and the LSWE (Biber *et al.*, 1999: 365-372), some interesting remarks can be made. A note of caution should, however, be struck on this point. Data relating to ukWaC and to the LSWE are not exactly of the same type. While Biber *et al*. (*ibid*.) take into account the verbs that are most frequent *in absolute terms* in the LSWE, data referring to ukWaC relate to the most typical verbs of ukWaC *when compared to the BNC*. Thus, if a verb form does not appear in the ukWaC list (and appears instead in Biber *et al*.' list), this does not imply that the verb is under-represented in ukWaC with respect to the LSWE. It could simply be that it is well-represented both in ukWaC and in BNC. Thus, when comparing results across ukWaC and the LSWE, it has to be reminded that the presence of a verb in both lists can be interpreted as signal that the verb is well-represented in both corpora, but the absence thereof does not necessarily indicate that the verb is under-represented in ukWaC.

Moving on to the analysis of data, it can be remarked that among the 29 verbs most typical of ukWaC,[13] 21 are indicated as also frequent in the LSWE (with a frequency of at least 20 occurrences per million words), and 16 as very frequent (occurring over 300 times per million words).[14] If attention is then focused on the text types in which such verbs typically appear in the LSWE, it can be noticed that most of them are quite evenly distributed across the four main types of texts which make up the corpus, i.e. fiction, conversation, news and academic texts. The verbs that seem most represented in a specific domain (such as *include*, *provide*, and *require*), are all associated with either news, academic texts or with both, but not with fiction and conversation.

Likewise, the distribution of the verb forms most typical of ukWaC across semantic domains (Figure 3.1) shows similar features to both the distribution of verbs in news texts and academic texts in the LSWE (cf. *ibid*.:

[13] In this case, verb lemmas are counted instead of inflected forms. This allows results to be compared, since in the cited work data are available only for lemmas. Boilerplate words are excluded from the counts.

[14] This is a positive result in terms of similarity between ukWaC and the LSWE, especially if one considers the caveat that ukWaC verbs are not the most frequent in absolute terms, but the comparatively most typical when compared to the BNC. In this regard, it would be interesting in further work to compare results taking into account absolute frequencies of verbs.

366). In these text types, activity verbs, followed by existence verbs, are the most frequent,[15] while communication and mental verbs are relatively less numerous. A trait which distinguishes ukWaC from the LSWE is the high frequency of causative verbs.

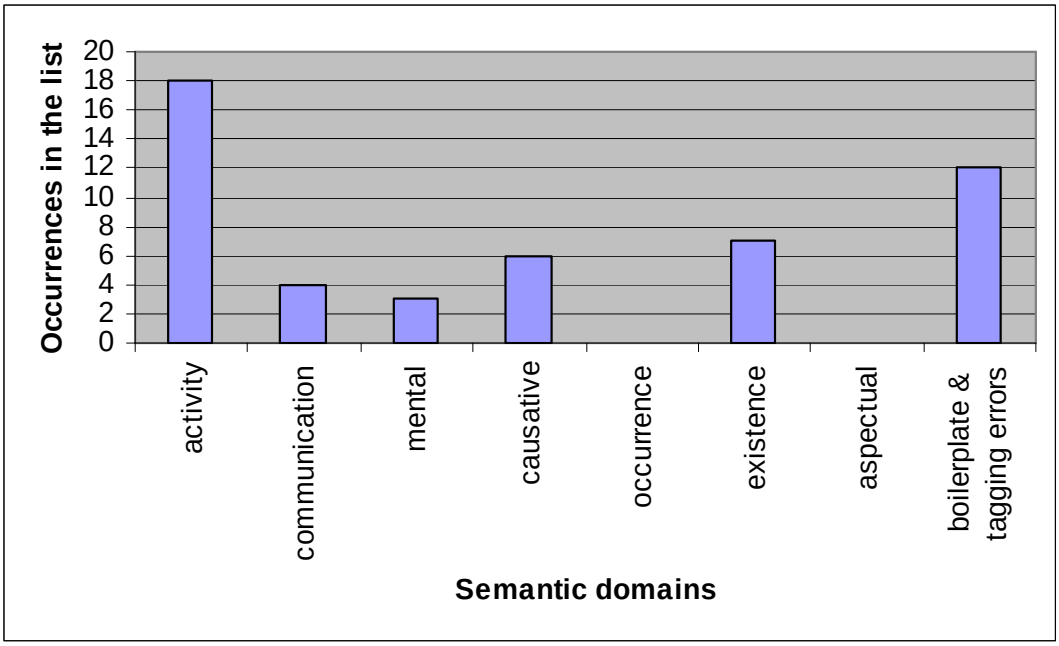

Figure 3.1. Distribution across semantic domains of the verb forms most typical of ukWaC

What can be inferred from these results is that in the continuum suggested by Biber *et al*. (*ibid*.: 25) which ranges from common, everyday language – represented by conversation –, to more specialized language – represented by academic texts –, the language of ukWaC (or, at least, the language of ukWaC which turns out to be most typical when compared to the BNC) is closer to the second pole. This could indicate, for example, that ukWaC may contain a certain amount of news and academic texts, or texts with similar linguistic features, as could be discussion pages. Like academic texts (cf. *ibid*.: 372), these usually focus on entities (either abstract, e.g. states or social issues, or concrete, e.g. children in the Third World) and describe relations among them, by using verbs of existence or relationship (cf. also category (f)). E.g.:

[15] It has to be highlighted, however, that activity verbs turn out to be the most frequent in all text types.

within the Greater London area. Such a charge would be <designed> to act as an effective incentive for operators to modify.

Like news texts, moreover, discussion pages may contain a number of communication verbs, which are frequently used to signal whose point of view is being expressed, e.g. in interviews:

This will be critical to ensure the future stability and success of our company,' Mr Moffatt <said>.

When interpreting the results, it has to be remembered that ukWaC includes a seemingly much wider range of text types than the LSWE, and that these texts may influence the results to a large extent. The presence of recommendation texts, for example, could account for the high frequency of causative verbs, which are rather infrequent in the LSWE. These verbs, which "indicate that some person or inanimate entity brings about a new state of affairs" (*ibid*.: 363), seem to be particularly used in promotional texts (cf. discussion of category (d)), whose aim is to convince readers that a certain product, service or idea can actually make a difference, e.g.:

attempts to use education to promote cultural variety and to <support> minority rights;
musical theatre as a popular entertainment genre. It will <help> you to sharpen your practical skills as a creative artist;
Entering into a relationship with Christ <allows> us to rise above whatever we were before and become someone new.

If verb tenses are taken into account, it can be noticed that most verbs are in the present tense (or in their base form), and that those which could appear as past forms are, actually, used most often as past participles in passive forms:

The candidate will also <be required> to respond to changes in learning and development;
The first year <is designed> to introduce you to the basic ideas and methods involved in the social scientific study of communications and media.

This could be due to a considerable presence of discussion texts, which are typically concerned with current affairs, or of recommendation and instruction texts, which often make use of the imperative form.

```
   We <need> to double the efficiency of the oil and gas we
<use>. We <need> to increase dramatically the energy
efficiency of our homes;
   for hard music fans, you really need to check them out
and <check out> Mike Chiplin!;
   details of work done, rates, and total contract billing.
6) <Add> VAT + the total sum due. 7) You are also perfectly
entitled to require.
```

Inversely, the relatively low frequency in ukWaC of communication verbs, which are most frequent in the LSWE in the spoken register, and past tense verb forms, which are typically used in narrative texts (*ibid*.: 456), could be a further indication of the relative absence in the Web corpus of both conversation and fiction texts.

3.4.2.2 Verbs most typical of the BNC

Some distinguishing features emerge when comparing the list of the verbs most typical of the BNC (see Appendix 3) to the list relating to ukWaC. Firstly, a considerable number of verbs which seem to be typical of narrative texts are observed. These indicate either physical actions or mental processes and seem to be connected with human beings, i.e. presumably with characters in stories (e.g. *thought*, *smiled*, *stared*, *nodded*, *walked*). Secondly, past tenses are prominently featured (e.g. *knew*, *went*, *sat*, *took*), which seems to confirm the hypothesis just mentioned. Finally, a certain number of words featured in the list seem to be connected with spoken language (e.g. *er*, *gonna*, *erm*, *fucking*).[16]

This kind of analysis, which uses verbs as indicators of the relative importance of the text types they appear in, is certainly useful. As was done in Section 3.4.1, the verbs most typical of the BNC can thus be analysed by checking their distribution across the main text domains[17] identified by Lee (2001). However, as pointed out in Section 3.4.2, when such analysis is complemented by a classification of verbs according to their inherent semantic properties, it can be more comprehensive and can account for the presence of a larger number of items in the list. For reasons of consistency, the same

[16] Of course, *er* and *erm* are not verbs and are in the list following a mistake of the POS-tagger.
[17] It has to be highlighted that in the present Section only the meta-tag `text_domain` is used, since the use of the `text_genre` tag produces too sparse results, which are difficult to interpret.

classification that was used in Section 3.4.2.1 will therefore be applied here (cf. also Biber *et al*., 1999: 360-378). This approach has two advantages. On the one hand, it provides data about the distribution of semantic *classes* of verbs (and not of *single* verbs), which are similarly represented in ukWaC and the BNC, and are thus easily comparable. On the other hand, it makes it possible to use results from the LSWE as a further benchmark for comparison. As we saw in Section 3.4.2.1, ukWaC (or, better, the features of ukWaC which turn out to be most typical when compared to the BNC) would seem to be similar to only one portion of the LSWE, i.e. the news and academic part. On the contrary, the BNC (or, better, the features of the BNC which turn out to be most typical when compared to the ukWaC) might presumably be more similar to the conversation and fiction part.

Moving on to the analysis of the semantic categories of verbs, it can be remarked that activity verbs (category (a)) are the most frequently featured in the list (cf. note 15). They include the verbs *got*, *smiled*, *go*, *nodded*, *turned*, *stared*, *come*, *shook*, *stood*, *put*, *laughed*, *glanced*, *sat*, *walked*, *shrugged*, *took*, *paused*, *leaned*, and *grinned*. Past tense forms, especially, tend to occur most frequently in imaginative texts. As an example, *looked*[18] occurs 21,782 times in imaginative texts and 10,358 times in the remaining text domains. Other verbs, like *go*, *come* and *put*, in the present tense, are most frequent in the spoken domain. Similar distributional patterns are found for communication verbs, like *say*, *tell*, *murmur* and *whisper* (category (b)). Present tense forms of these verbs (e.g. *say*, *saying*) are frequently used in spoken language, while past tense forms (e.g. *said*, *told*, *murmured*) are found more often in fiction texts, or in the domain of "world affairs", which seems to correspond to a large extent to news texts.

Mental verbs (category (c)) are the second most prominent category in the list of the verbs most typical of the BNC. They are *know*, *mean*, *think*, *felt*, *suppose*, *wanted* and *saw*. In this case, too, the present tense forms (*know*, *mean*, *think*, *suppose*) are most frequently featured in the spoken section of the BNC, while past tense forms tend to appear more often in fiction, or with a

[18] In lowercase form.

similar frequency in fiction and world affairs texts (especially the forms *knew* and *saw*).

Only one example of aspectual verbs (category (f)) is found in the list, i.e. *began*, while causative and occurrence verbs are missing.

As in the case of ukWaC a category should be added which accounts for the presence of non-verb items in the list. The high frequency of forms like *er*, *erm*, *round* (which is a transcription of the abbreviated form "*'round*", typical of spoken language), *fucking*, *ai* (which was tokenised incorrectly, and should instead be "*ain't*") can be identified as POS-tagging mistakes. In the same way, the verb forms *didn't* and *don't* were tokenised incorrectly (as *did / n't* and *do / n't*),[19] so that the occurrences of negative forms of the verb were counted as being affirmative forms. This accounts for the (erroneous) presence of *did* and *do* in the list of the verbs most typical of the BNC. These forms are labelled as "other" in Figure 3.2.

Of course, the analysis just presented is rather a sketchy one. Its aim was to identify the major textual domains across which verbs and verb classes are distributed. Only three domains were taken into account, i.e. that of spoken language events, of imaginative/fiction texts, and of world affairs/news texts. This could appear as a reductive way of interpreting the data, since many other textual domains are represented in the BNC. Our purpose, however, was not to analyse the BNC in detail, but rather to identify general features of corpus composition which distinguish it from ukWaC. Besides, it also has to be remarked that in the totality of the cases[20] the verb forms presented in Appendix 4 are most frequent, as regards the written domain, in imaginative texts; that for over 60% of these, world affairs texts represent the second most important written domain of occurrence; and that 26 % of the words occur with the absolute highest frequency in the spoken domain.

In Figure 3.2 the overall distribution across semantic domains of the verbs in the BNC list is presented. Some important differences with the distribution of verbs in ukWaC (Figure 3.1) can be observed. Firstly, activity,

[19] The error is due to the re-tokenisation procedure that was carried out when pre-processing the BNC for POS-tagging with the TreeTagger (cf. Section 3.3).
[20] Counts exclude the verbs classified as "other".

communication and mental verbs are more frequently featured in the BNC (23 occurrences vs. 18, 4 vs. 7, and 3 vs. 9, respectively). This could be due to the high frequency with which such verbs are used in the two text types – i.e. conversation and fiction – that seem the most typical of the BNC compared to ukWaC. In these text types "the typical communicative purposes" are to a large extent the same, i.e. "talking about what people have done (activity verbs), what they think or feel (mental verbs), or what they said" (Biber, *et al.*, 1999: 371). In contrast, causative, occurrence, and existence verbs seem to be much more typical of ukWaC. This datum seems to confirm what was suggested in Section 3.4.2.1, i.e. that the Web corpus contains a higher proportion of texts in which (especially) existence and occurrence verbs seem to be very frequent, i.e. academic texts (cf. *ibid.*: 366), and discussion pages.

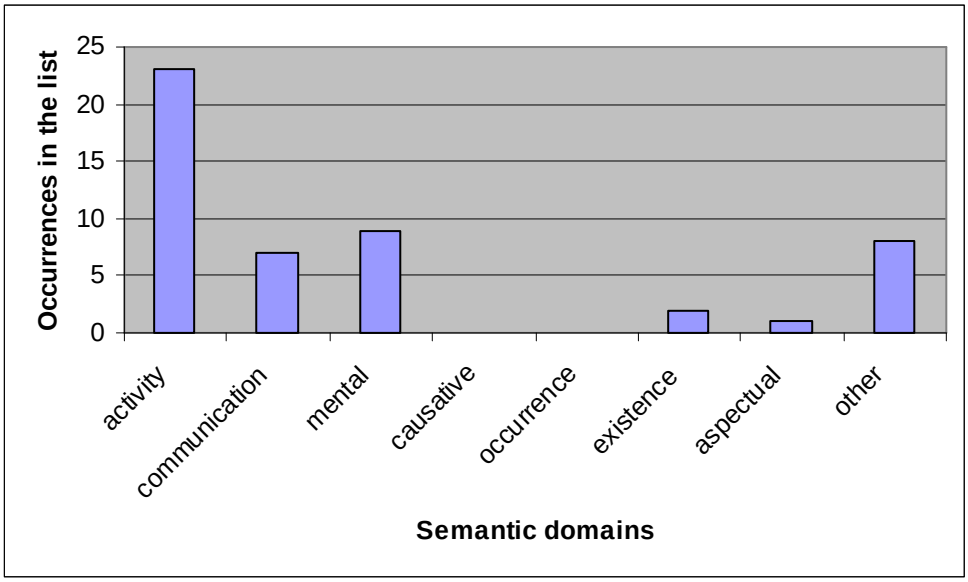

Figure 3.2. Distribution of verbs across semantic domains in the BNC.

Other interesting remarks can be made if results from the BNC are then compared to those obtained from the LSWE (*ibid.*: 373-378). The number of verb lemmas that are featured in the BNC list is 33 (for ukWaC this number was 29). Among these, 24 verbs turn out to be very frequent in both the BNC and the LSWE (with a frequency in the LSWE of at least 300 occurrences per million word), and 10 are among the verbs with the highest absolute frequency in the LSWE, maybe reflecting similar concerns in corpus design criteria.

These verbs are *say*, *get*, *go*, *know*, *think*, *see*, *come*, *take*, *want* and *mean*. Perhaps not surprisingly, all of them occur most frequently in the fiction and conversation domain.

On the whole, in the continuum of the register features mentioned in Section 3.4.2.1, which ranges from conversation to academic prose, it appears that the BNC is closer to the first pole than ukWaC.

3. 4. 3 ADJECTIVES AND *–LY* ADVERBS

3.4.3.1 Adjectives most typical of ukWaC

The analysis of the adjectives most typical of ukWaC compared to the BNC (Appendix 5) may be seen as complementary to the investigation presented in Section 3.4.1.1 In broad terms, if the analysis of nouns served as an indication of the differences between the two corpora in terms of the (comparatively) most typical topics, that of adjectives may point at differences in the way such topics are characterised. As we shall see, certain adjectives reflect the presence of certain topics, whereas others are not easily associated with any topic or domain. In both cases, analysing them means investigating further what language features, taken as indicators of the presence of certain text types or domains, distinguish ukWaC from the BNC, and thus help us to better understand the composition of the former.

As already mentioned, the correlation between certain items in the list and the topics identified in Section 3.4.1.1 is sometimes clear, as is the case for adjectives related either to the World Wide Web and new technologies (e.g. *online*, *digital*, *mobile*), or to social issues (e.g. *sustainable*, *clinical*, *affordable*, *disabled*). The presence of these adjectives among the most typical of ukWaC may be accounted for by the presence in the corpus of a considerable number of texts whose topics are connected with the Internet, or with governmental institutions and NGOs, respectively. In the majority of cases, however, no evident association between adjectives and topics emerges. This may suggest that such adjectives are distributed across a wide variety of texts, possibly dealing with different topics and belonging to different text

types. The adjectives in question can be grouped into three macro-categories, i.e. words relating to time (e.g. *current*, *ongoing*, *annual*), words indicating positive qualities (e.g. *excellent*, *fantastic*, *unique*), and what have been traditionally called "relational adjectives", that Biber *et al*. (1999: 509) define as adjectives "delimiting the referent of a noun, particularly in relation to other referents" (e.g. *key*, *additional*, *inclusive*, *specific*). Some adjectives in the list turn out to be part of boilerplate sections of texts (*more*, *full*, *top*, *related*, *registered*, *non*, *subject*, *website*, *personal*),[21] and are thus uninteresting for the purposes of the analysis.

The first category is that of the adjectives whose frequency reflects the presence of some among the topics identified in Section 3.4.1.1. Examples are *online*, *digital*, *mobile*, *electronic*, *interactive*, *audio*, and *virtual*, which can be categorised as "words which seem to belong […] to the semantic fields of computers and the World Wide Web". Like their noun "counterparts", these adjectives 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:

```
  If function calls to an object passed by value weren't
early-bound, a <virtual> call might access parts that
didn't exist;
  This prevents automating <interactive> updates, reducing
the chance of system administrators inadvertently updating;
  I've been reassuring my users that the "<mobile> phone
virus" warnings relentlessly circulating the Internet are
hoaxes;
  ensuring best practice in all aspects of our clients'
<online> systems. So if you want to get on with business
contact us.
```

In this regard, it is interesting to notice how these adjectives are not only frequent in texts whose domains are strictly related to the web or computing issues. They can also be found in presentation pages, business reports, and even in an artistic manifesto:

[21] In order to assess whether adjectives were typically part of boilerplate, some frequent collocations with these adjectives were analysed. In the present case, some of the most frequent collocations were, e.g., "more information", "full article", "back to top", "related links", "registered users", "subject to availability" "personal details". The word "non" is frequently featured in spam texts.

One of the best places for you to find a low interest tenant loan is to look <online>. Very few people realize this is the best source for getting the best deal;

A live <audio> webcast of the presentation will be available on the investor relations pages;

Today, as <digital> artisans, we now express the emancipatory potential of the information age.

This seems to confirm what was suggested in Section 3.4.1.1, that IT-related adjectives, like nouns, are presumably spread across a wide variety of texts, insofar as they represent "meta-references" to the medium of communication which hosts them. The increasing influence of IT in all fields of public and private[22] life could also be adduced as an explanation of why Web- and computing-related words are used so frequently in ukWaC, or why they are typical of the web corpus when compared to the BNC, which was published at a time when the Internet was still in its infancy.

Other words whose frequency seems to be linked to certain topics identified during the analysis of nouns are *available*, *sustainable*, *global*, *local*, *clinical*, *accessible*, *affordable* and *disabled*. These are often connected with what were called "public service" issues (cf. Section 3.4.1.1), and are typically found in texts created by departments within the government and NGOs, or in various kinds of recommendation or discussion texts, such as texts promoting a political (or humanitarian) programme, or news. The purposes of these texts are either that of attracting and persuading the general public, or debating and disclosing information that may be of interest for them. In both cases, the topics usually revolve around current affairs, and include economic issues, such as "*global* economy" and "*sustainable* growth" (which are among the most frequent collocates of the two adjectives); political concerns, like environmental issues, international relations and governmental efficiency, and equal rights and opportunities, like, e.g., accessibility of facilities and websites for disabled people, or availability of services for the less well off:

RESOLVED : That the Council use planning conditions to secure the provision of <affordable> housing and highways/environmental works in accordance with a scheme;

[22] Consider, e.g., the phenomenon of blogs and discussion forums (cf. Ueyama, 2006).

security of the entire Korean nation and <global> security, which would have a huge impact on international relation;
and thereby reduce the amount of duplicated work across <Local> Authorities;
Centre for Early Music is completely flat-floored throughout making it an <accessible> building for wheelchair users with or without an accompanied helper;
The aim of the IFI is to ensure that <disabled> people can access inclusive fitness equipment and train alongside the rest;
Building all our futures Family learning should be <available> to all families in the same way as antenatal and primary health care.

What the frequency of these adjectives – and of their noun "counterparts" – seems to point at, is that topics in ukWaC correspond to a certain extent to current themes of discussion. This, however, is also true for the BNC, in which two of the most typical adjectives compared to ukWaC are *soviet* and *cold* (cf. Appendix 6). Such datum is likely to reflect the importance that the theme of the "Soviet Union" and of the "Cold War" – which are among the most frequent collocations including the adjectives – had at the time of publication of the corpus.

Not only do the Web texts in ukWaC seem to reveal a prominent interest in current affairs, they also appear to be concerned with explicitly affirming their being contemporary. Among the most frequent adjectives of the corpus, a number of them function as references to present time, or signal a change with respect to the past, like, e.g., *new*, *current*, *innovative*, *ongoing*, and *annual*. These adjectives are often found in promotional texts, where they are used to highlight the newness of the product or service being presented. Interestingly, the authors/initiators of these promotional texts are not only private companies, as one could expect, but also universities and the government:

A high-performance platform that delivers a <new> level in small-packet performance, the IP710 exceeds 700 megabits per second;
we can boast an award-winning library, specialist research centres and <innovative> teaching facilities that are the envy of many other institutions;
access to Government monies must require high performance on race equality. The <current> legal framework under the Race Relations (Amendment) Act 2000 provides for this approach.

Other typical contexts in which these adjectives are found are discussion texts, such as news and academic papers on disparate topics, like economics, politics, education and arts:

```
   In the light of the <current> Rolls-Royce situation -
and indeed the wider aerospace industry problems;
   Recent trends in housing completions, with <annual>
totals of between 1,750 and 2,160 in the last five years,
compare favourably;
   In Britain more and more mixed schools are using single-
sex classes because of <ongoing> concerns over boys'
results, which have consistently lagged behind those of
girls;
   It can lead art history to <new>, more transparent and
immediate forms of communication and co-operation.
```

Adjectives which refer to the present time may be seen as also connected with the high frequency of verbs in the present tense (cf. Section 3.4.2.1). Taken together, these two features seem to point at the fact that the texts in ukWaC are typically both focused on the present time and willing to signal it explicitly. This, as concordances reveal, is notably true for discussion texts like press releases, and recommendation texts like promotional pages. In the latter type of texts, adjectives which signal a radical change with respect to the past (e.g. *innovative*) are also used, particularly for the purpose of displaying 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, like *excellent*, *fantastic*, *unique*, *creative*, and *original*. All of these are found, e.g., in descriptions of products or tourist attractions, and in job vacancy announcements:

```
   This is of course a vintage <original> and we only have
1 available!;
   clinical supervision together with an <excellent> range
of internal and external training opportunities;
   the most beautiful space to enjoy your stay in Cornwall.
<Fantastic> views across the ocean and countryside,
contemporary en-suites.
```

“Relational adjectives”, the last category that is taken into account, are typically found in a wide range of text types. Such adjectives are used to set up conceptual relations between their referents and other referents belonging to

the same class. Examples in ukWaC of these relations are general/particular (e.g. *inclusive*, *comprehensive*, *diverse*, *multiple* VS. *appropriate*, *specific*, *dedicated*) or prominence with respect to other referents (e.g. *key*, *relevant*). Relational adjectives are often found in discussion texts, namely academic papers or essays on different topics, whose purpose is to identify and clarify abstract relations between concepts or objects (cf. also Biber *et al*., 1999: 510-511):

> In Chapter 6 we give an overview of our findings organised according to our three <key> issues - entry, retention and loss - as well as offering some conclusions;
> its treatment is a significant cost for primary care trusts. <Appropriate> nutritional and dietetic support improves patients' nutritional status.

Another typical context of use of relational adjectives is in information and instruction texts, which aim at providing readers with information or instructions on how to do something. In these texts, the relations of general/particular or prominence are used to define precisely the objects of discussion, in order to avoid any possible confusion, or signal important pieces of information. Legal texts and online tutorials represent examples of these kinds of texts:

> To satisfy the requirements of AML/CFT legislation, <additional> identity verification checks should be sought, as described in paragraphs 4.15;
> in one of two ways, at Licensee's option, subject as follows: By embedding <appropriate> provisions within Licensee's User Agreement: By obliging users to accept;
> Yes, the selling price of all goods on sale to consumers must be indicated <inclusive> of VAT, other taxes or other compulsory charges such as delivery charges;
> A brief introduction to the <relevant> standards in Web services like SOAP will help you understand.

In fact, these adjectives are also found in promotional texts, where they are often used to describe a service as being all-inclusive and suitable for all kinds of requirements, or as flexible and customized:

> A <comprehensive> hearing therapy service is provided at Saltergate, including a tinnitus clinic;
> highly experienced Translator & Interpreter used to handling subjects as <diverse> as company reports for the Financial Times through to TV documentaries;

```
  We can also offer advice on Fees , Grants & Loans and
<specific> types of funding such as that for NHS Funded
Courses.
```

3.4.3.2 Adverbs ending in *–ly* most typical of ukWaC

In order to provide a fuller description of the linguistic features which turn out to be the most typical of ukWaC, a list of adverbs ending with the suffix *–ly* was created (see Appendix 7). These are also called "derivational adverbs", since they are most often created from adjectives, with which they share many semantic properties (Quirk *et al*., 1985: 438-439). The list of adverbs, which will only be briefly analysed, reveals important common traits with that of adjectives. It includes adverbs apparently related to IT (e.g. *automatically*, *electronically*), and to social issues (e.g. *locally*, *nationally*, *globally*). A number of adverbs relate to the present time (e.g. *currently*, *recently*, *newly*). Others seem to indicate positive characteristics (e.g. *successfully*, *incredibly*, *easily*), or are forms derived from relational adjectives (e.g. *specifically*, *additionally*, *individually*). Of course, a more thorough analysis could reveal interesting patterns of usage both for these adverbs, and for others which do not seem to fit in with the present categorisation (e.g. *hopefully*, *normally*, *jointly*). However, the fact that several items featured in the list (22%) are derivative forms of the adjectives presented in Appendix 5 could be taken as an indication of the prominence in ukWaC of the semantic categories – and of the corresponding distributional patterns across typical text types – identified in the current Section.

3.4.3.3 Adjectives most typical of the BNC

As was the case for ukWaC, the adjectives most typical of the BNC can be grouped into macro-categories, only some of which reflect the presence in the corpus of a considerable number of texts associated with the topics and text types identified in Section 3.4.1.2. Among these, the most prominent category is that of the adjectives which seem to be related to the description of physical characteristics of objects and people, or of their temper; another important class

includes adjectives which refer to politics and economics. As it might be expected, however, other adjectives emerge which do not fit in with the classification that was applied to nouns. These are words related to past times, and words associated with scientific language.

In Section 3.4.1.2, a number of nouns were found that refer to people's bodies and physical actions, or to objects. An investigation of their distribution across the text domains identified by Lee (2001) showed that these words are featured for the most part in fiction texts. The same turns out to be true for several of the adjectives featured in Appendix 8, which reveal a strong correlation with those nouns. The adjectives indeed refer to physical characteristics of people (e.g. *pale*, *tall*, *thin*), or of inanimate objects and settings in which an action takes place (e.g. *dark*, *white*, *cold*, *thick*); others relate to people's temper (e.g. *sorry*, *afraid*, *angry*, *anxious*), or express an overt judgement on a situation (e.g. *right*, *terrible*, *wrong*):

```
  one side of his face, his toupee not quite straight. His
<tall> frame, in its butler's black and white, seemed to
vibrate;
  and found her in an armchair, engrossed in a <thick>,
glossy-looking book . "Something new from the library?" I
asked;
  My mind just goes on and on..." She looked bleak at the
recollection of those <dark> hours. "Well, my conscious is
clear," Miss Pinkney said archly;
  I was too anxious - far too <anxious> - and this put my
interviewers on their guard;
  the right time in the right spirit, or at least not at
the <wrong> time, in the wrong spirit, with the wrong plans
and having made the wrong preparations, with the wrong
tools.
```

Of course, these adjectives can be found in other contexts, different from fiction texts. In particular, *sorry*, *right*, and *wrong*, are also typically found in the spoken domain:[23]

```
  You will?. <Sorry> yes. I 'm I 'm really going to erm I
afraid I have experience of Who are you <sorry>? Brenda Oh
right. And you know me very well Tom.
```

[23] Other items from the list which point at the fact that spoken texts are comparatively more frequent in the BNC than in ukWaC are *er*, *erm*, *okay* and *mum*. These represent evident tagging errors (cf. also Sections 3.4.1.2 and 3.4.2.2).

Other adjectives can be identified that, although not being particularly frequent in imaginative texts, are nonetheless often featured in such text type. As an example, *black* is most frequent in the domain "world affairs", since it is often used to refer to the political and sociological issue of "black people", but is also found in fiction, where it describes, e.g. the colour of an object:

```
   though so weak as to be almost useless in practice, had
as a basis the assumption that <black> people were a part
of the community; [world affairs domain]
   keeping an eye on programme girls (most of them
certainly mature) who, in their <black> dresses and little
aprons, ushered. [imaginative domain]
```

These phenomena are not unusual, "since very common adjectives typically designate a range of meanings" (Biber *et al*., 1999: 509), and different meanings can be associated with different patterns of usage across different text types. This is the reason why, when associating words with certain text types, only classes of words are taken into account, and not single items. The fact that different words reveal similar patterns of usage makes it possible to define with some confidence what text types – associated with what words – seem to be comparatively overrepresented in the BNC or in ukWaC, which is the main purpose of the analysis. In the case under consideration, thus, the class of adjectives identified seems to point at the fact that in the BNC fiction texts are more prominent than in ukWaC.

The second category to be taken into account is that of adjectives relating to politics and economics. These include "general", hypernymic adjectives (*political*, *economic*, *social*), and adjectives which designate national provenance (*soviet*, *french*, *foreign*), political parties (*conservative*), or other forces which are usually involved in political and economic affairs (e.g. *royal*, *industrial*). All of these adjectives are typically found in three domains, which Lee (2001) calls "world affairs", "social sciences" and "commerce". As mentioned in Section 3.4.1.2, the presence of this category of adjectives may be seen as contradicting what was found about ukWaC in Section 3.4.1.1 and 3.4.3.1, i.e. that ukWaC seems to contain a larger quantity of texts revolving around politics and social issues than the BNC. However, when concordances

are analysed (only some indicative examples are given below), the apparent contradiction is clarified:

> The five were previously imprisoned from June until October 1990 for allegedly organizing a <political> party; [world affairs domain]
> General opinion is that the rate will result in increased tariffs, which under the present <economic> conditions will serve only to reduce competitiveness , " the survey said; [social science domain]
> Sales of reserves were offset by an inflow of $260m of <foreign> currency receipts from the final instalment of the sale of British Steel shares. [commerce domain]

What seems to be the case, judging both from the concordances and from the analysis of the distribution of the adjectives, is that the text types in which politics and economics are predominantly found are different across ukWaC and the BNC. In the latter, they are found in discussion texts like academic and non-academic textbooks, and newspaper articles; in the former, in addition to discussion texts, they can also be found in a considerable number of recommendation texts. Such text types, although dealing with similar topics, seem to have different features. Discussion and recommendation texts in ukWaC are often concerned with matter-of-fact issues (like, e.g., proposing solutions to improve disabled people's lives), and are mainly focused on the present. Discussion texts related to politics and economics in the BNC, on the contrary, seem to describe events through "general", abstract categories (e.g. *political*, *economic*) – which is a typical feature of essays and academic prose – and to report facts in the past time – which is typical of newspaper articles (see Biber, 1988: 191-195) –.

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, like, e.g., *last*, *long*, *nineteenth*[24] and *former*. 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

[24] The fact that these adjectives are predominantly used in a temporal sense can be confirmed by an analysis of their most frequent collocates, i.e. "last year", "last night", "last week", "last time", "last month"; "long time", "long period"; "nineteenth century".

BNC seem to adopt a retrospective, historical approach to facts, as is typical, e.g., of newspaper articles:

```
   Dr Manorani spoke at a number of Amnesty meetings <last>
autumn. The Sri Lankan government has moved to try to
counter the criticisms; [world affairs domain]
   In Czechoslovakia <former> prisoner of conscience Vaclav
Havel became President of his country; [world affairs
domain]
```

Finally, adjectives are found among those comparatively most typical of the BNC which are related to natural and applied sciences. Words like *male*, *gastric*, *colonic*, *ulcerative* and *oesophageal* are often found in academic and non-academic essays which deal with anatomy or health problems (medicine):

```
   catechin (15) to inhibit histidine decarboxylase, which
catalyses the formation of the <gastric> acid stimulator
histamine, is believed to be the basis of their antiulcer;
   Pregnancy can follow first intercourse, and can even
occur without <male> penetration;
   Hence salivary, <gastric>, pancreatic and intestinal
secretions all contribute to the large volume;
   The restricted expression of <colonic> markets is
probably the result of epigenetic alterations in the
mucosal.
```

This should not be interpreted as signalling that ukWaC does not contain texts on medicine and anatomy. In fact, a closer look at the adjectives reveals that most of them refer to the digestive system. It is therefore likely that the BNC contains a higher proportion of essays on the specific topic of human or animal digestion, rather than medicine-related texts *in general*. In this regard, such technical sub-domain may be seen as over-represented in the BNC compared to ukWaC.

3.4.3.4 Adverbs ending in –*ly* most typical of the BNC

As was done for ukWaC, a list was created for the derivational adverbs ending in –*ly* most typical of the BNC. Most of them (around 80%) are what Quirk *et al*. (1985: 482) call “manner adverbs”, i.e. adverbs which express information about how an action is performed, like, in the BNC, *suddenly*, *softly*, *quietly*, *slowly*, *ruefully*, *thoughtfully*, *warily*, etc. All of these adverbs are most typically found in fiction texts, which seems to confirm our hypothesis that the BNC contains a higher proportion of such texts than ukWaC, and that narrative

texts in general, characterised by past tense verbs, and adjectives and adverbs relating to mental states and physical actions (Biber, 1988; and Biber *et al*., 1999) are relatively less frequent in ukWaC (cf. also Section 3.4.1.2 and 3.4.2.2). A minor category that can be identified in the list is that of "epistemic stance adverbs" (Biber *et al*. 1999: 557), like *probably*, *presumably* and *reportedly*. These are used to indicate the degree of certainty with which an assertion is made, and are usually associated with texts that take into account and discuss different points of views, such as news and academic prose (Biber, 1988: 191-195). In fact, the epistemic stance adverbs most typical of the BNC are usually found in the world affairs and social sciences texts, which seems to confirm our hypothesis about the prominence of these domains, and of the text types associated with them, in the corpus.

3. 4. 4 FUNCTION WORDS

3.4.4.1 Function words most typical of ukWaC and the BNC

As mentioned in the introduction, this category is different from the rest, insofar as it is a meta-category which includes different kinds of "grammatical", instead of content-rich, words. These are subordinating and coordinating conjunctions, determiners, prepositions, modal verbs, pronouns, and all the inflected forms of the auxiliaries *be* and *have*. Of course, these words give no hint about specific topics featured in the corpus, but can nonetheless be used as indicators of the language used in it.

One of the most prominent features in the ukWaC list (Appendix 9) is the presence of first person plural pronouns and possessive adjectives (*our*, *us*), which could indicate a very strong presence of "collective" authors, as can be considered governmental departments, universities and other organisations (cf. Section 3.4.1.1). This would be consistent with Sharoff's (2006: 79-80) results, which show that what he calls "corporate authors" are significantly more represented in Web corpora than in the BNC. In the latter, according to Sharoff (*ibid*.), "single" or "multiple"[25] authors tend to prevail.

[25] The label "multiple" authors is applied to texts "created be several named co-authors" (Sharoff, 2006: 79).

The high frequency of first person plural pronouns, which was also remarked by Fletcher (2004b), is made even more noteworthy by the simultaneous presence in the list of second person pronouns (e.g. *yours*) and of present tense verb forms (e.g. *is*, *are*, *can*, *has*). All of these forms are, in fact, what Biber (1988: 105) calls signals of interactive style. Perhaps not surprisingly, this seems to point out the fact that ukWaC contains, to a large extent, texts characterised by interactive language, i.e. language which tries to build a relationship between the author(s) of the text and their intended audience (Thelwall, 2005). Another interesting datum is the presence in the list of the modal verb *will*. As also remarked by Thelwall (2005) and Fletcher (2004b), this is due to two main factors. On the hand, it is due to a high proportion in the corpus of "instruction" texts (cf. 3.3.1), and, on the other, to the fact that Web texts are more future-oriented than those in the BNC. This seems a rather interesting datum. Indeed, while in the analyses presented in Section 3.4.1 and 3.4.2 it emerged that the BNC seems to contain a higher proportion of spoken texts, the considerable presence of signals of interactivity would suggest that ukWaC texts do present some features of spoken language.

The list of the function words most typical of the BNC (Appendix 10) contains several third person pronouns and possessive adjectives, either singular (*she*, *he*, *her*, *his*, *it*, *him*, *they*, *herself*, *himself*) or plural (*they*). Moreover, a remarkable presence of past tense verb forms stands out (*had*, *was*, *were*, *could*. See also Section 3.4.2.2). According to Biber (1988: 108), these forms constitute "markers of narrative action". In narrative discourse, unlike in interactive language, person pronouns typically make reference to "referents apart from the speaker and the addressee" (*ibid*.), and past tense verbs are used to "[present] a sequential description of past events involving specific […] participants" (*ibid*.). This seems therefore to confirm our hypothesis about the abundance of narrative texts in the BNC and their relative lack in ukWaC. According to Thelwall (2005: 536), the relatively higher frequency of the first person singular pronoun in the BNC (*i*, *me*) might be another indicator of a more prominent presence of narrative (fiction) texts. The significance of this datum, however, might be limited, since the first person

pronoun "I" in ukWaC is frequently misspelled as "i", especially in user-produced texts, and that such lowercase form often misleads the TreeTagger.[26] The pronoun is therefore likely to be much more frequent in the corpus than the data reveal.

The last aspect that needs to be taken into account when analysing the ukWaC lists is that the high frequency of certain words could be partly due to their being frequently used within boilerplate sections of web-pages. In order to test this hypothesis, 15,000 occurrences of each word in the list were randomly selected, and counts were produced to check for their most frequent collocates within a span of one or two words on either side.

This procedure reveals highly recurrent patterns, which can then be evaluated in terms of their being boilerplate or not. According to the results, examples of function words whose number of occurrences could be influenced by their being part of boilerplate text are the following:[27]

- *for*: "for more information", "for further information";
- *this*: "this site", "this page", "this website";
- *can*: "can be downloaded", "can be viewed", "can be accessed", "can be contacted";
- *on*: "more *or* further information on", "click on", "password required on", "on the web" "on the site";
- *from*: "are available from", "is available from", "be available from", "be downloaded from";
- *via*: "be accessed via", "is available via", ", or via", "be contacted via", "contact us via", "are available via", "be delivered via", "via the internet", "via the web", "via email", etc.;
- *us*: "please contact us", "to contact us";
- *by*: "posted by", "Originally posted by", "published by", "Sponsored by".

It can be noticed that, out the 20 function words most typical of ukWaC compared to the BNC, 8 part of boilerplate. This a very high percentage (40%)

[26] In a randomly selected sample of the corpus consisting of 92,524,352 tokens, the form "i", which is very likely to stand for the personal pronoun "I", appears 20,946 times. In none of the occurrences is it tagged as a pronoun.

[27] All the patterns that are mentioned occur within the list of the 20 most frequent 2- or 3-grams which contain the word in question.

compared to the ratio of boilerplate in other lists, which may however be due to the fact that few items were taken into account. It is possible that if more items were considered (e.g. the 50 function words most typical of ukWaC), this ratio would get lower.

3.5 Discussion of results

In the present Chapter a method was proposed and applied to provide an evaluation of ukWaC's contents. In order to do so, different lists were created which grouped all the words belonging to each of the main part-of-speech categories, i.e. nouns, verbs, adjectives, *-ly* adverbs and function words. 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 make an example, many words were found in ukWaC that belonged to the semantic field of the Web or computing. This does not mean that nouns like *internet* are among the most frequent in the corpus in absolute terms. Rather, their frequency is comparatively higher in ukWaC than in the BNC, which is explained if one considers that the BNC was published at a time when the Web was still in its infancy. In the same way, the presence of *soviet* in the list of the adjectives most typical of the BNC should not be interpreted as a sign that the BNC is biased in absolute terms towards, e.g., newspaper articles or books about the Cold War. It simply indicates that issues revolving around Russia are more prominent in the BNC with respect to ukWaC. The second point that should be remembered is that while the method is very useful to highlight the relative "unbalances" of the two corpora, it also conceals the features that make them similar. Thus, in the analysis provided, only the differences between

ukWaC and the BNC emerged. It could be argued that a way to understand how the two corpora are similar would be to also take into account all the differences that *did not* emerge from the analysis. A (tentative) approach would therefore be, e.g., to analyse what kinds of text types or domains did not appear as typical of either ukWaC or the BNC, and assess whether there is ground to claim that they are equally represented in both corpora.

Moving on to the actual analysis of data, 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 service 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), recommendation (e.g. advertising of a traditional or Web-based service) and instruction texts (e.g. tutorials). It was argued that the presence of such words among the most typical of ukWaC is quite unsurprising, insofar as they represent meta-references to the medium of communication that hosts them. Furthermore, the fact that they are well represented in ukWaC may be seen as a welcome finding, since one of the main aims of the corpus is that of documenting recent phases of language evolution, of which the increasing importance of Web- and computing-related words could be an example. Education and public service issues are also found in a great variety of text types, ranging from "traditional" texts like academic articles and legal texts, to more recent Web-related 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. In fact, no one-to-one correspondence between a certain topic and a text type can be identified (it could have been possible, e.g., that computing-related issues were dealt with in the corpus only in online tutorials or software manuals). 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 characters' physical characteristics or emotions, and

by adverbs and 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*, *erm*) and mental verbs (e.g. *know*, *want*, *think*). The third category of texts which is considerably more present in 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 (e.g. *government*, *recession*, *political*, *economic*) and using the past tense, as is typical, e.g., of newspaper articles.

Some major differences can also be found between the kind of language that turns out to be typical of each of the two corpora. 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 and adverbs which refer to the present (e.g. *current*, *recently*); moreover, interactive style seems to be prominent (use of the present tense and of first and second person pronouns). This may be due, among other factors, to a considerable presence of recommendation (advertising) texts. These are signalled in particular by the presence of a number of empathic adjectives (e.g. *excellent*, *fantastic*, *unique*), and of causative verbs (cf. Section 3.4.2.1). One 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 programme). The BNC, on the contrary, features narrative language more prominently, which is characterised by past tense verbs, adjectives and adverbs referring to the past and third person pronouns.

Besides making it possible to identify some of the main differences between ukWaC and the BNC, through which insights were provided on the composition of the Web corpus, the analysis led to the discovery of a number of problematic words, that were either part of boilerplate or frequently featured in spam sites. Their presence among the most typical words of ukWaC,

however, should not be seen as a problem of the corpus *per se*, but as an indication that better post-processing techniques are needed. Moreover, the fact that boilerplate accounts for only a minority of the words featured in the lists is an encouraging result. This would seem to confirm that ukWaC, while containing a certain amount of noise, may be considered as a valuable resource to study naturally occurring, human-produced text.

4

CONCLUSIONS

4.1 Concluding remarks

In the present dissertation a new corpus resource for the English language, i.e. ukWaC, was presented and evaluated. The ultimate aim of its construction was to obtain a very large Web-derived corpus, which would be comparable to the BNC - along very general lines - in terms of balance and variety of textual materials contained (i.e. a "general-purpose" corpus). Thus, some aspects of corpus composition were evaluated by assessing what differences emerge when ukWaC is compared to the BNC, which is widely assumed to be a model for general-purpose corpora of British English.

The corpus is central to corpus linguistics, an approach to language study whose main purpose is to analyse language as it is produced in authentic settings, and whose methodology involves quantitative and qualitative appraisals of large quantities of data. In particular, attention was focused on the main criteria that need to be taken into account when designing a general-purpose linguistic corpus, i.e. its size and sampling strategy . The aim should be that of including as large (and balanced) a quantity of text types and domains as possible.

It was argued that the Web is a very valid source from which linguistic data can be retrieved, thanks mainly to its immense size, the ease with which it makes it possible to find textual materials, its timeliness, and the variety of topics and languages it contains. Despite the inevitable pitfalls connected with using Web data, including their supposed "non-representativeness" with respect to the general language (Thelwall, 2005), and the noise they contain (e.g. duplicate pages, boilerplate, etc.), it was shown that an increasing number of researchers are now turning to the Web to find evidence for their linguistic studies.

Three different approaches to the "Web as corpus" (WaC) were then discussed. One consists in approaching Web data via commercial search engines. This, however, poses major problems in terms of the possibility to

make complex queries, of the accuracy and unbiasedness of the results, and of the reproducibility of the experiments. The second approach consists in relying on search engines to retrieve documents, and then downloading and post-processing data for inclusion in a stable corpus. Although this method makes it possible to replicate linguistic experiments and to provide a fully independent interface to the corpus, it does not solve the problems linked to the matching and ranking algorithms of the search engines. Finally, Web data can be retrieved via customised crawls of the Web. In this way, very large quantities of data can be collected and subsequently post-processed without the intermediary of search engines.

The latter is the approach that was chosen to build ukWaC. This, along with deWaC and itWaC (similar corpora of German and Italian), was built with the intention of providing a valid alternative to other currently available WaC resources. It was also suggested that as a very large, stable and possibly balanced Web-derived corpus, ukWaC is meant to meet a variety of research needs, including the need for a larger and more up-to-date resource than the BNC, which, despite its high quality standards, proves inadequate when rarer or recently emerged linguistic phenomena are taken into account. The procedure that was followed to collect and post-process the textual data of ukWaC was then explained in detail.

When semi-automated procedures of corpus construction and post-processing are used, as is the case for ukWaC, the possibility to control the materials that end up in the final corpus are limited. *Post-hoc* evaluation plays therefore a key role in determining actual corpus composition. For this reason, an evaluation method was proposed and applied to ukWaC that involved a comparison with the BNC, used as a benchmark corpus. Word lists of nouns, verbs, adjectives, *-ly* adverbs and function words were created for the two corpora, and then compared via the log-likelihood association measure. This made it possible to discover the words that are relatively most typical of either ukWaC or the BNC. Such words were thus taken as indicators of the possible "unbalances" that might characterise the two corpora when compared to each other.

The analysis indicated that ukWaC, when compared to the BNC, seems to contain a higher proportion of texts related to the Web, to education (namely universities), and public service. A great variety of text types is found in ukWaC, ranging from "traditional" texts (e.g. legal texts, instruction manuals, discussion articles, etc.), to Web-based emerging genres (e.g. blogs, forums of discussion), which are (inevitably) not attested in the BNC. The latter corpus features, instead, a comparatively larger quantity of narrative texts, politics- or economy-related articles and spoken texts. It should be noted, however, that the language of ukWaC is not devoid of spoken-language features. On the contrary, while the BNC seems to be characterised by a more narrative, past-oriented language, ukWaC's (comparatively) prominent linguistic features point at a considerable use of interactive, present-oriented language.

An important point that should be remembered is that the evaluation method proposed gives prominence only to the differences between the two corpora, and conceals the features that make them similar. It was suggested that a possible way to assess how similar ukWaC and the BNC are would be to take into account the differences that *do not* emerge from the analysis. In this respect, many text types and topics do not turn up as being typical of either corpora, which may suggest that the two of them are rather similar. This would arguably advise in favour of considering ukWaC as a general corpus of British English.

4.2 Further work

4. 2. 1 IMPROVING ON UKWAC

During the analysis provided in Chapter 3, a number of words were identified as being problematic in terms of corpus composition. These words turn out to be among those comparatively most frequent in ukWaC not because they are frequently used within connected, human-produced text (i.e. the kind of language that corpus linguistic studies are interested in), but because they belong to typical phrases used within Web pages, and as such may be repeated across different texts or even within a single text (e.g. *click*). These sequences

are typically part of what are called “boilerplate” sections of a Web page, which include navigational bars, headers, footers, and legal disclaimers.

The main drawbacks connected with a considerable presence of boilerplate in a Web corpus derive from the fact that boilerplate tends to distort statistics about corpus composition, and clutter concordance lines with uninteresting linguistic materials. Boilerplate detection and removal take therefore centre-stage in the post-processing of Web corpora. For this reason, a competition was organised recently, within which researchers and students from all over the world were invited to propose methods for Web data cleaning (CLEANEVAL; see Fairon *et al*., 2007). Future versions of ukWaC will take advantage of the techniques proposed within CLEANEVAL to eliminate boilerplate. Further improvements of ukWaC will consist in discarding all the texts that were identified as being machine-generated (i.e. spam).

It would also be interesting to apply the methods devised, e.g., by Sharoff (2007) or Santini (2006), to automatically classify Web pages into domains and genres. This would make it possible to make up, at least partially, for the lack of meta-information about the texts in the corpus, which only contain an indication of the URL they were retrieved from.

4. 2. 2 EXTENDING THE ANALYSIS

Apart from practical improvements on ukWaC, a more extensive analysis of the corpus is planned. As pointed out in Section 3.5, the method of analysis adopted in the present study did not make it possible to analyse the corpus as a whole, but only to highlight the main differences the corpus shows when compared to a benchmark corpus that is considered as balanced, i.e. the BNC.

In order to draw confident generalisations about language when using a general-purpose corpus, however, it is crucial that its composition is known, so that every result can be interpreted in the light of the text types and domains that are known to be included in it (cf. Section 1.2). For this reason, a method of analysis which makes it possible to evaluate ukWaC in its entirety should be devised. One possibility is to apply the multi-factor analysis proposed by Biber (1988). Such method, starting from a set of pre-defined linguistic features

(identified automatically), isolates several textual characteristics (that Biber calls “dimensions”; *ibid*.: 3-5), which can in turn be interpreted functionally as being characteristic of certain text types/genres. If this method is to be applied to ukWaC, however, it would be necessary to adapt the interpretative stage, so that it can account for newly emerging Web genres.

Another possibility would be to test the adequacy of ukWaC in a practical task, be it lexicographic, didactic, translational or other. For instance, within lexicography one could assess whether the corpus provides sufficient evidence to study all the possible meanings and usages of a set of randomly selected words, including neologisms and technical terms, and to provide adequate usage examples. Alternatively, materials on which to base a didactic unit could be sought in the corpus, or the latter could be used for reference purposes within a technical translation task. While such usage-oriented tasks would not offer clear indications about the composition of the corpus, they would nonetheless provide evidence as to whether ukWaC meets the purpose(s) for which it was built, i.e. to provide a comprehensive, updated and balanced resource for the study of the English language.

APPENDICES

Appendix 1

Appendix 1. Nouns most typical of ukWaC.

WORD FORM	NUMBER OF OCCURRENCES IN ukWaC	TOTAL NUMBER OF OCCURRENCES	LOG-LIKELIHOOD RATIO
information	2261755	2297546	59934.2396051694
website	657662	657662	59894.955542486
site	1169916	1179550	54716.9539809311
click	490665	491248	39285.5504093514
web	457860	458408	36628.6821961126
email	360197	360227	32355.9638299861
internet	355804	355860	31648.4996382928
students	967188	981644	27403.9639588661
page	719978	729332	23695.0118088708
details	768889	780321	21911.4786317893
skills	603025	612043	17064.3772661061
project	728814	741545	16863.7627977574
research	1001771	1023441	16144.5347691292
access	594498	603861	15802.1174384727
services	961784	982915	15019.1929238583
issues	640786	652592	13689.3057891703
links	350307	354055	13669.6983051799
service	1035457	1060091	13575.7361508639
link	357065	361033	13514.9041825772
data	746327	762477	12013.6243449224
comments	331738	335800	11542.129950665
contact	466815	474700	11318.5695752867
pm	184680	185744	10325.72644871
organisations	340879	345801	10036.3833975744
nhs	242474	244931	9861.77861431632
pages	306282	310408	9713.15910779338
pdf	107040	107061	9472.42820337902
health	738666	756471	9380.63751471236
projects	339987	345355	9004.86781388727
sites	348456	354087	8960.27395116628
download	99718	99732	8886.51009187379
advice	480956	490676	8804.74430817927
poker	114352	114550	8725.93007876096
range	754549	773440	8712.57068647623
websites	90861	90861	8269.55612048484
file	296735	301236	8253.24156390385
funding	260538	264161	7991.77595374114

text	394058	401584	7945.59203559882
delivery	248784	252320	7452.7772737125
events	469288	479539	7408.64418764165
fax	139148	140064	7326.22827886617
article	314511	319990	7300.47689130072
insurance	326309	332156	7268.64256395547
copyright	135917	136783	7267.1849214668
list	493625	504834	7184.09407927399
browser	81813	81840	7127.17279156158
application	453257	463179	7124.3340909269
users	337763	344142	6934.83964623752
support	781922	803409	6911.44096488654
format	187491	189649	6888.99986195592
software	372421	379912	6874.15507444464
info	90136	90301	6820.34449790992
search	314397	320240	6622.20821496356
design	465314	476048	6540.82525817426
address	266227	270713	6474.19223131736
staff	744483	765088	6431.4659207079
event	444053	454247	6309.32212614846
quality	591296	606544	6305.08614908653
server	129632	130674	6140.35129274494
development	924340	951990	6068.4974016596
images	221569	224990	6031.59676640767
consultation	187625	190146	5964.82529709995
guidance	204423	207408	5941.80954965743
experience	689803	708938	5914.95640645693
team	650660	668399	5894.38598665471
network	304374	310379	5810.58294300455
content	251812	256259	5730.15817437246
aug	86675	86986	5699.44462337569
resource	172594	174841	5663.50531991167
training	637754	655261	5654.18059266297
student	332235	339240	5622.22582731035
articles	187123	189780	5611.04715599883
opportunities	285434	290998	5561.7813337666
use	999796	1031112	5425.05479739021
files	182451	185092	5355.45638695262
community	605026	621676	5321.5006693301
requirements	289139	295067	5155.64284231593
learning	165432	167712	5122.0363277663
forum	105097	105902	5116.67832029424
review	307207	313797	5027.29583818266
cd	94066	94697	4907.38833710731
feedback	121014	122258	4868.2235376358
program	214758	218573	4844.64371372641

reviews	112402	113456	4830.21189510597
guide	211327	215056	4814.26671348192
password	78694	79052	4811.78315936395
album	152443	154523	4769.74127427146
feb	71722	71973	4748.25296542309
author	223736	227876	4744.15288533968
options	203031	206577	4694.57966156685
document	249212	254169	4690.2557648439
database	183385	186401	4607.08365612947
photos	97575	98364	4605.60752714569
quot	50312	50313	4561.41919994234
music	494019	507404	4560.78197209197
user	262540	268007	4548.28552390658
products	399548	409609	4541.20572364094
activities	436966	448342	4535.93904790693
card	255885	261209	4439.83649036049
history	640936	659853	4426.83544374735

Appendix 2

Appendix 2. Nouns most typical of the BNC

WORD FORM	NUMBER OF OCCURRENCES IN BNC	TOTAL NUMBER OF OCCURRENCES	LOG-LIKELIHOOD RATIO
ll	52315	54961	304834.310268148
er	46750	58306	234128.909410577
erm	37966	41336	213373.175921072
cos	12235	26677	40688.2859013132
cent	38168	276333	37399.2493143035
eyes	27356	178382	31195.327599933
man	56318	574993	29227.0663226336
yesterday	17923	106863	23021.7531669776
face	25270	193401	22688.8280901589
sort	22945	165691	22563.0319390887
mother	22036	172069	19147.2197567072
hon	9324	36694	18942.8436849022
woman	21093	166973	17916.6688857722
door	23050	192408	17873.9420003967
head	32282	336365	15975.2917023956
something	50077	616518	15910.5535116302
one	54472	705238	14743.7044373002
nothing	32191	350598	14317.810518957
father	19871	173969	14129.40065009
men	36514	420305	14067.0165103313
girl	13953	102371	13387.5769108364
voice	18701	164782	13137.7519282648
hair	13198	96091	12813.5194515618
mrs	18653	166349	12808.6322579677
round	22073	217382	12393.8147053581
women	36041	448685	11060.272851303
thing	33733	412651	10921.211775439
moment	20772	212304	10742.1724785917
government	55007	778545	10737.4657516138
forty	6378	30766	10522.0295937883
night	33735	421956	10201.1236741296
pounds	9921	71237	9836.91216558252
recession	3760	11429	9626.10130477714
way	94675	1536165	9568.11122432276
smile	6408	34005	9490.69591725221
hand	31596	400527	9156.56845447322
anything	27420	332339	9123.69471130946
boy	11081	89474	9100.28631779478
somebody	6947	41424	8920.11146053863
plaintiff	2978	7675	8715.33909579588

pound	6147	34558	8496.69827327647
mhm	1509	1663	8382.00955228208
hands	17428	184477	8305.71398425543
lips	4604	21209	7980.82469340956
wife	16398	172361	7945.93365987312
sir	4745	22894	7825.80648499145
party	27081	343794	7810.14358849744
house	33529	453202	7802.39408727094
fact	36519	507267	7671.64416139602
arm	8419	64454	7551.97864620501
husband	10498	91757	7485.82862839761
unemployment	6391	42636	7055.61666348544
tomorrow	8684	70668	7036.7310727229
ta	1955	3859	6994.9280946627
bed	14600	154595	6951.69685413182
mouth	8704	71647	6914.46168489401
police	23954	304622	6869.22241038534
speaker	7075	52156	6737.23805906714
morning	19596	241642	6189.23206843765
gentleman	4695	28094	6001.09524576948
chairman	8037	69215	5889.40663533988
kind	22564	295619	5860.43881757723
fingers	5373	36234	5841.31820214609
arms	9786	93819	5815.46881083123
relations	10105	98407	5810.02385266477
opposition	8732	80555	5611.82172996625
mind	20451	264258	5564.93281826984
labour	10897	112523	5502.93892056072
court	17353	214866	5411.67988889249
state	27870	395751	5363.31668757142
silence	5007	34329	5315.43712638464
bit	26371	371218	5255.61991988298
feet	13284	152761	5127.2554671748
money	36048	549961	5044.79443819481
darling	2111	7047	4986.78024543436
ah	1867	5460	4935.73383570732
shoulders	3891	23641	4876.08366002205
lot	27343	402575	4512.50610820848
trouble	8755	89860	4482.6964405784
sense	20774	287560	4420.32274509962
figure	13181	159702	4388.26029259223
clothes	6815	63411	4305.24228352837
emailinc	690	690	4295.50232721723
hers	1802	6017	4255.87407727879
room	27561	412133	4240.07839125334
things	40882	661132	4203.09650177643

back	13665	170289	4177.35504740677
studio	7233	70801	4112.97293825254
republics	1378	3516	4064.52333436225
breath	4757	37859	4002.83887711288
legs	6064	55341	3980.89214332703
mummy	1696	5704	3980.22784907992
pattern	8898	97826	3862.45402637412
point	35768	577005	3729.09240939899
nobody	5829	54021	3711.77128865359
spokesman	3890	29043	3631.2330440055
time	151722	2923323	3624.62965861029
lady	5458	49849	3577.58133948284
friend	14468	192947	3531.23516730313
side	31706	506074	3508.91773650427

Appendix 3

Appendix 3. Verbs most typical of the ukWaC.

WORD FORM	NUMBER OF OCCURRENCES IN ukWaC	TOTAL NUMBER OF OCCURRENCES	LOG-LIKELIHOOD RATIO
posted	534768	535475	51261.0768333588
including	1116904	1140868	29436.3748739009
contact	353613	356145	23071.7161811501
using	1013203	1037586	22124.6249412108
provide	931186	953339	20744.5623718796
include	710380	725500	18919.0967278736
please	421006	426611	18851.4478554954
use	1098542	1129118	17901.3001675439
provides	474399	482708	16271.7581531183
learning	402868	409227	15444.7091006374
help	938575	965286	14514.604498505
includes	400992	407715	14435.0385654743
based	708409	726869	13280.6299660802
ensure	478120	488276	12760.577892573
published	488407	498961	12704.953917657
top	147921	148364	12570.2530265927
posts	122935	123076	11933.3609671864
need	1082558	1117993	11426.2518585411
working	806807	831237	10666.2912063344
offers	278727	283206	10488.3086227801
develop	382153	390684	9447.50135027795
updated	139577	140530	9289.01900817703
offer	393322	402376	9250.27936435895
support	364925	373017	9117.19346360343
download	79706	79724	8386.53757269647
following	752493	776827	8237.15127990531
visit	243854	248207	8175.58375247381
view	149399	150915	8117.92693616643
providing	325521	332789	8043.5956452458
access	109667	110361	7526.43753353342
developing	280228	286285	7281.13620555086
required	517851	533249	7172.79324339706
find	1092932	1133764	7050.37745924932
improve	279129	285276	7047.16834841216
create	321994	329992	6618.30784212354
provided	538853	555677	6538.80573038529
located	160440	162912	6294.55885325156
allows	208161	212263	6167.07772457391
deliver	146349	148438	6166.52692494628
work	702613	726901	6162.15003127026

check	221287	225867	6127.18931255969
receive	295641	303026	6011.72532778989
contains	217854	222393	5974.89749843412
add	313231	321358	5914.78394589046
apply	306492	314382	5879.02827548157
read	591730	611646	5691.81870932316
designed	352909	362833	5599.98464423125
aims	132703	134639	5481.93470644071
email	50552	50555	5419.19802588881
promote	164613	167730	5135.31859366784

Appendix 4

Appendix 4. Verbs most typical of the BNC.

WORD FORM	NUMBER OF OCCURRENCES IN BNC	TOTAL NUMBER OF OCCURRENCES	LOG-LIKELIHOOD RATIO
said	195305	1582874	118618.801010009
er	13664	14186	76070.9793957405
got	90064	655766	67105.4095963126
did	135863	1260768	60316.8627835983
know	118611	1197958	41861.8388674854
mean	39542	268442	33263.5803395626
gonna	12245	32900	30922.4031310588
looked	32254	214271	28130.8809552921
thought	45237	373021	26285.6383139871
do	270143	3803754	23865.7134880833
think	88592	988840	22742.7145324624
erm	3734	3769	21598.3230549632
smiled	6889	15738	19958.6247408547
knew	23971	167151	19239.9840162432
say	66581	747218	16746.6946160258
seemed	22096	162183	16167.9230774774
went	45792	467549	15590.1370788443
told	35397	338575	14450.5637069818
nodded	4599	9781	14122.0406233664
felt	26062	231424	12807.4241438032
go	85152	1091091	12472.2569845686
turned	22963	196465	12292.4628874731
stared	4158	9344	12209.3632293364
going	63340	764352	12087.1492376406
came	44746	497421	11617.6538963053
shook	4750	13380	11501.4860027665
stood	12195	79091	11054.4158558602
come	66594	837909	10606.102210251
suppose	10085	60056	10439.6850151267
put	57085	698874	10250.8180259581
laughed	4453	16653	8205.85376148963
glanced	2691	5923	8036.47509317303
sat	10902	80218	7936.29321380891
gone	18333	175759	7431.80549208909
walked	8649	58650	7279.37730523819
round	7509	48623	6822.71905951852
shrugged	2106	4341	6630.724852836
murmured	1833	3214	6552.84966997072
tell	28845	337479	6211.33636190341
wanted	22020	239826	6118.52537765221

saw	24578	277836	6009.510152216
began	20662	222812	5931.05290179183
took	37164	468509	5855.58339550071
fucking	2995	10913	5669.41004471657
paused	2233	6186	5487.46584499645
leaned	2015	5023	5426.64645382841
whispered	2353	7170	5303.16929877595
saying	17688	188278	5293.0570930088
grinned	1616	3286	5143.35115144889
ai	3552	16549	5112.46629470961

Appendix 5

Appendix 5. Adjectives most typical of ukWaC.

LEMMA	NUMBER OF OCCURRENCES IN ukWaC	TOTAL NUMBER OF OCCURRENCES	LOG-LIKELIHOOD RATIO
online	516703	517213	45741.3726907257
available	1361747	1388705	31680.8501940919
key	478027	485796	14447.6914091056
digital	185104	186171	11493.0478715314
free	732455	750438	11445.0968196981
new	2847479	2952429	10871.0129350272
current	553955	567371	8905.29362055612
mobile	158667	160156	7712.91834692576
excellent	324480	331026	7288.43356694519
more	2575746	2675269	7266.05100253827
sustainable	114308	114951	7163.12011552942
full	889386	917728	6610.36293728605
global	201850	204956	6483.93275063241
local	1266053	1310375	6299.60031443562
top	468396	480847	6065.75507680867
relevant	330565	338436	5499.83612164654
fantastic	112921	114015	5376.7268294835
additional	309960	317288	5232.40558740225
unique	216623	220873	5078.52888063354
further	699516	721942	5075.80999325473
interactive	86685	87304	4899.92338690307
related	153506	156000	4633.61063285084
clinical	165935	168820	4607.29667344852
innovative	98353	99346	4555.43861734585
appropriate	399820	411225	4274.72694617517
ongoing	80552	81237	4153.50160651367
accessible	114067	115628	4131.06955304512
electronic	154565	157450	3898.24104963194
academic	194634	199020	3607.43342456125
creative	127053	129438	3178.03066237145
audio	63354	63946	3086.67906429088
professional	324017	333676	3026.40813809291
virtual	68543	69315	2932.69390615233
live	130446	133085	2913.94921602081
registered	82223	83394	2863.31324397647
affordable	51301	51690	2814.18592354093
inclusive	51230	51643	2721.03785862445
wide	473735	489885	2658.33161068887
disabled	132725	135596	2649.78513425849
original	343708	354592	2606.93511576702

specific	346987	358033	2580.48938856153
non	70426	71409	2501.75150708205
annual	255822	263361	2476.99322344436
comprehensive	141349	144668	2417.12247904114
subject	250075	257553	2311.46916146813
website	23400	23400	2293.23346447238
personal	467905	484420	2224.20350517966
diverse	77181	78484	2223.18070531309
dedicated	55102	55790	2168.03378044829
multiple	104496	106707	2167.43048584494

Appendix 6

Appendix 6. Adjectives most typical of the BNC.

LEMMA	NUMBER OF OCCURRENCES IN BNC	TOTAL NUMBER OF OCCURRENCES	LOG-LIKELIHOOD RATIO
er	21462	23756	115707.983586224
ll	16786	17347	97210.9027986477
erm	9055	9440	51892.5519850649
okay	10522	31839	25678.3558031432
little	47698	563582	13845.4193688351
soviet	5833	24995	10190.0163581817
bloody	6596	32453	9877.27926690663
political	29324	341743	8877.50129130138
right	28849	372321	6125.29202406955
sorry	7461	56652	6048.93573987524
black	18989	218653	5961.76940738643
old	56564	851795	5832.91062838751
male	8570	72854	5642.94791864939
gastric	2041	6678	4644.32385797405
economic	21067	270865	4534.63325862644
mum	5432	40808	4484.73381575219
dark	10771	119654	3760.23137826026
white	17498	226958	3639.66852767304
last	71552	1194000	3602.83941970122
french	13757	170066	3417.80773865656
pale	3115	20221	3242.0652116934
conservative	5516	50102	3170.04656869786
afraid	5537	50413	3165.88167468566
cold	9655	111227	3024.20976433897
dead	9527	110485	2923.24331227817
foreign	11351	139257	2895.52633846993
colonic	788	1572	2689.32567360227
royal	3866	32203	2647.29749578233
sudden	4091	35108	2642.89943167489
angry	3958	33676	2600.82419927853
considerable	9459	114048	2559.57714692531
social	36137	578324	2513.27455110589
industrial	10124	125677	2477.51900104747
sexual	6629	72250	2442.1487681326
long	40646	665158	2392.53035207045
much	28413	443245	2351.1069350679
nineteenth	2925	22446	2327.42756093179
terrible	4368	41213	2309.06808927529
difficult	21580	322828	2302.33201934791
ulcerative	748	1721	2287.5373862302

tall	4961	49822	2264.59920097392
oesophageal	838	2273	2244.01595809266
thin	5297	55011	2221.81922862715
anxious	2943	23457	2188.44241856978
same	61126	1057961	2184.25228117195
wrong	14864	209521	2159.14607664358
former	16647	241568	2093.33593241023
own	67032	1178270	2030.84674874134
certain	21741	333601	1996.24503502373
thick	4945	52255	1981.72515741021

Appendix 7

Appendix 7. *–ly* adverbs most typical of ukWaC.

LEMMA	NUMBER OF OCCURRENCES IN ukWaC	TOTAL NUMBER OF OCCURRENCES	LOG-LIKELIHOOD RATIO
currently	399441	406452	17184.927842343
fully	261454	270238	3882.11644215666
approximately	120148	122974	3585.30422471683
recently	305750	317904	2648.39726912626
directly	222667	231254	2153.6453517307
automatically	96600	99342	2093.33137945644
originally	130050	134485	1844.66678089971
truly	99248	102323	1787.30002769785
regularly	114689	118513	1727.07867734891
specifically	112399	116118	1726.57704562374
internationally	39957	40649	1720.46185045567
alternatively	66355	68083	1681.59147787613
locally	65597	67381	1539.59590098055
highly	210721	219663	1374.75079095394
typically	69796	71899	1335.87134626681
hopefully	62079	63916	1234.56135276177
actively	53428	54907	1212.8957910054
additionally	29703	30273	1147.43493184642
nationally	36988	37846	1122.82936388408
electronically	19263	19502	1090.88849546729
ideally	43764	44937	1053.71259044742
effectively	125215	130227	1047.29990789342
successfully	90421	93753	1016.21452714287
globally	13327	13436	951.832732404298
hugely	20423	20772	890.39530940088
unfortunately	110746	115295	832.137212307417
previously	153839	160630	822.189951030166
visually	25050	25622	777.365901279425
annually	37331	38419	765.480962809499
normally	175787	183900	728.301192155383
manually	15172	15417	697.597110190928
genetically	17642	17975	693.846075805983
extremely	146605	153272	663.290037664882
individually	35322	36410	642.959527994284
importantly	40093	41400	637.770312711891
correctly	52216	54099	627.428516873096
incredibly	27705	28482	612.925622862648
potentially	62388	64812	583.197372580689
primarily	74422	77517	529.119944398414
definately	5338	5356	501.581352507795

seamlessly	6229	6267	499.397458539608
critically	21189	21764	498.464661161637
jointly	33388	34508	493.17585605116
newly	63686	66344	445.450474946645
easily	192193	201814	436.985041867393
formerly	49502	51457	434.930994672589
daily	24193	24952	422.021358371122
personally	61333	63903	421.34931999303
especially	327858	345215	419.212141535709
externally	13160	13469	392.148304725707

Appendix 8

Appendix 8. *–ly* adverbs most typical of the BNC.

LEMMA	NUMBER OF OCCURRENCES IN BNC	TOTAL NUMBER OF OCCURRENCES	LOG-LIKELIHOOD RATIO
suddenly	11008	78421	7110.74799552331
softly	2255	7706	4166.23195639272
hardly	8410	67740	4146.09442914197
quietly	3847	24847	2986.68963032729
slowly	7378	67523	2599.63438463595
certainly	18112	220008	2108.79698885782
sharply	2343	14355	1994.03523690972
merely	7417	74679	1931.16362902293
angrily	1040	3539	1929.37612977023
obviously	10663	119593	1826.68977100042
gently	3758	31748	1643.40807722752
drily	389	562	1539.53851159233
probably	26522	359347	1506.34675090431
actually	25440	343090	1498.65616045495
abruptly	1158	5575	1420.17791449865
coldly	530	1400	1261.67029807348
grimly	539	1462	1251.9198229716
stiffly	419	906	1190.86509877642
wearily	467	1221	1122.22478395917
impatiently	544	1706	1097.74823119155
bitterly	1045	5678	1077.69726865109
faintly	708	2935	1051.20902851019
crossly	243	339	989.666614855332
partly	5581	62138	984.12750146851
scarcely	1571	11390	974.846131859015
irritably	256	405	941.171942232129
huskily	213	275	925.439758112281
silently	1094	6806	906.448819527632
firmly	3815	39631	886.432752413182
nervously	644	2890	866.37422417865
badly	4176	45060	838.448821254073
anxiously	603	2658	829.90493516427
mentally	1905	16487	781.591099723128
ruefully	320	830	774.602603917032
briskly	455	1784	721.51711038594
tightly	1620	13584	721.254334625051
furiously	576	2736	719.207906599948
helplessly	406	1482	698.080745189508
wryly	330	999	689.734592337678
hastily	807	4936	688.336563607326

thoughtfully	661	3607	677.160692808025
lightly	1872	16995	673.157765056699
presumably	3200	34389	652.620838498868
casually	707	4159	643.504015052936
reportedly	1452	12334	625.391767712347
uncertainly	255	683	599.23745651262
cautiously	666	3971	592.878703402915
reluctantly	910	6503	581.477255016193
uneasily	352	1357	569.157605033152
warily	263	769	568.691418679485

Appendix 9

Appendix 9. Function words most typical of ukWaC.

WORD FORM	NUMBER OF OCCURRENCES IN ukWaC	TOTAL NUMBER OF OCCURRENCES	LOG-LIKELIHOOD RATIO
and	58468926	61090240	129455.320649785
for	21754049	22634768	94333.1720547443
your	5051063	5184384	91752.3661484529
will	8049591	8331389	67032.9625163988
our	3518059	3610673	64149.3738731802
is	22380449	23372228	52172.5846002407
are	11556268	12020974	51414.5915911336
this	11090811	11541936	46189.9696862638
or	8955907	9323186	35440.9883626599
can	5305265	5514729	25875.7647349056
the	115573265	121616477	23032.3048540639
of	59869219	62918539	22818.4918297408
on	15543561	16270750	21322.8831395429
with	13949929	14608453	17220.282505281
from	9309027	9733821	16412.2848934755
via	313054	317406	13434.340317018
has	5696578	5953437	11181.6974786208
us	1633582	1694546	10412.8744315196
any	2870038	2990874	9545.90190375249
by	10582509	11094364	9503.01405651815

Appendix 10

Appendix 10. Function words most typical of ukWaC.

WORD FORM	NUMBER OF OCCURRENCES IN BNC	TOTAL NUMBER OF OCCURRENCES	LOG-LIKELIHOOD RATIO
she	352460	1830458	445067.026564789
he	640248	5352465	365539.230347591
her	303610	1894504	294647.503988712
had	421083	3728898	210468.787769171
was	883059	10234838	201643.449034021
i	847118	9999956	180268.764079308
his	409618	4211977	138099.599475272
it	1056305	14574824	107418.734610347
him	153313	1219944	96427.0597591273
were	313634	3926172	51750.8074473343
but	444604	6067210	47619.2541582331
they	420207	5733298	45028.5366147577
that	1115176	17709346	37778.033604401
could	160063	1859637	35759.5293508567
would	245685	3325380	27446.3517177092
herself	15869	74320	22759.6788231259
what	240696	3435065	19503.185874568
me	130150	1668948	19151.6509569969
like	109668	1359356	18832.6379033724
himself	28885	258508	13998.7707838935

REFERENCES

Aston, G. (2001) Text categories and corpus users: a response to David Lee. *Language learning & technology*. 5(3): 73-76.

Aston, G. and Burnard, L. (1998) *The BNC Handbook: exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Banko, M. and Brill, E. (2001) Scaling to very very large corpora for natural language disambiguation. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Online: http://research.microsoft.com/users/mbanko/ACL2001VeryVeryLargeCorpora.pdf [consulted: 04/12/2007].

Baroni, M. and Bernardini, S. (eds.) (2006) *WaCky! Working Papers on the Web as Corpus*. Bologna: GEDIT Edizioni.

Baroni, M. and Kilgarriff, A. (2006) Large linguistically-processed Web corpora for multiple languages. *Proceedings of EACL 2006, demo session*. 87-90.

Baroni, M. and Ueyama, M. (2006) Building general- and special purpose corpora by Web crawling. *Proceedings of the 13th NIJL International Symposium*. 31-40.

Baroni, M. and Bernardini, S. (2004) BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004*. 1313-1316.

Bayen, H. (2001) *Word frequency distributions*. Dordrecht: Kluwer.

Bernardini, S., Baroni, M. and Evert, S. (2006) A WaCky introduction. In Baroni, M. and Bernardini, S. (eds.) *WaCky! Working Papers on the Web as Corpus*. Bologna: GEDIT Edizioni. 9-40.

Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press.

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

Biber, D. (1988) *Variation across speech and writing*. Cambridge: Cambridge University Press.

Brekke, M. (2000) From the BNC towards the Cybercorpus: a quantum leap into chaos? In Kirk, J.M. (ed.) *Corpora galore: analyses and techniques in describing English. Papers from the 19th International Conference on English Language Research on Computerised Corpora*. Amsterdam, Atlanta: Rodopi. 227-247.

Broder, A., Glassman, S., Manasse, M. and Zweig, G. (1997) Syntactic clustering of the Web. *Proceedings of the Sixth International World-Wide Web Conference*. Online: www.hpl.hp.com/techreports/Compaq-DEC/SRC-TN-1997-015.pdf [consulted: 12/04/2007].

Burnard, L. (2007) *Users' reference guide to the British National Corpus (XML edition)*. Oxford: Oxford University Computing Services. Online: http://www.natcorp.ox.ac.uk/XMLedition/URG/ [consulted: 12/01/2007].

Cabré, M.T. (1999) *Terminology: theory, methods and applications*. Amsterdam, Philadelphia: John Benjamins.

Christ, O. (1994) A modular and flexible architecture for an integrated corpus query system. Budapest: COMPLEX'94. Online: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/ [consulted: 12/01/2007].

Clarke, C.L.A., Cormack, G.V., Laszlo, M., Lynam, T. R. and Terra, E.L. (2002) The impact of corpus size on question answering performance. *Proceedings of SIGIR '02*.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*. 19(1): 61-74.

Fairon, C., Naets, H., Kilgarriff, A. and de Schryver, G.-M. (eds.) (2007) *Building and exploring Web corpora. Proceedings of the WAC3 Conference*. Louvain: Presses Universitaires de Louvain.

Fantinuoli, C. (2006) Specialized corpora from the Web and term extraction for simultaneous interpreters. In Baroni, M. and Bernardini, S. (eds.) *WaCky! Working papers on the Web as Corpus*. Bologna: Gedit. 173-190.

Fetterly, D., Manasse, M., Najork, M. and Wiener, J. (2004) A large-scale study of the evolution of Web pages. *Software: Practice and Experience*. 34: 213-237.

Fletcher, W.H. (2007) Implementing a BNC-Compare-able Web Corpus. In Fairon, C., Naets, H., Kilgarriff, A. and de Schryver, G.-M. (eds.) *Building and exploring Web corpora. Proceedings of the WAC3 Conference*. Louvain: Presses Universitaires de Louvain. 43-56.

Fletcher, W. (2004a) Facilitating the compilation and dissemination of ad-hoc web corpora. In Aston, G., Bernardini, S. and Stewart, D. (eds.) *Corpora and language learners*. Amsterdam: Benjamins. 273–300.

Fletcher, W.H. (2004b). Making the web more useful as a source for linguistic corpora. In Connor, U. and Upton, T. (eds.) *Corpus Linguistics in North America 2002*.

Grefenstette, G. (1999) The WWW as a resource for example-based MT tasks. Paper presented at the *ASLIB "Translating and the Computer" conference*.

Johansson, S. (1980) The LOB corpus of British English texts: presentation and comments. *ALLC journal*. 1: 25-36.

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

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

Kucera, H. and Francis, W.N. (1967) *Computational analysis of present-day American English*. Providence RI: Brown University Press.

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.

Lüdeling, A., Evert, S. and Baroni, M. (2007) Using Web data for linguistic purposes. In Hundt, M., Nesselhauf, N. and Caroline, B. (eds.) *Corpus linguistics and the Web*. Amsterdam: Rodopi. 7-24.

Mair, C. (2003) Tracking ongoing grammatical change and recent diversification in present-day standard English: the complementary role of small and large corpora. Paper presented at the *Annual ICAME Conference*.

Marek, M., Pecina, P. and Spousta, M. (2007) Web Page Cleaning with Conditional Random Fields. In Fairon, C., Naets, H., Kilgarriff, A. and de Schryver, G.-M. (eds.) *Building and exploring Web corpora. Proceedings of the WAC3 Conference*. Louvain: Presses Universitaires de Louvain. 155-162.

Manning, C. and Schütze, H. (1999) *Foundations of statistical natural language processing*. Boston: MIT Press.

McEnery, T. and Wilson, A. (2001) *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Olohan, M. (2004) *Introducing corpora in translation studies*. London, New York: Routledge.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) *A comprehensive grammar of the English language*. Harlow: Longman.

Rayson, P., Walkerdine, J., Fletcher, W.H. and Kilgarriff, A. (2006) Annotated Web as Corpus. In *Proceedings of EACL 2006*. 27-33.

Rayson, P., Berridge, D. and Francis, B. (2004) Extending the Cochrane rule for the comparison of word frequencies between corpora. In Purnelle, G., Fairon, C. and Dister, A. (eds.) *Le poids des mots. Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*. Louvain-la-Neuve: Presses universitaires de Louvain. 926-936.

Rayson, P., and Garside, R. (2000) Comparing corpora using frequency profiling. *Proceedings of Workshop on Comparing Corpora of ACL 2000*. 1-6.

Renouf, A., Kehoe, A. and Banerjee, J. (2007) WebCorp: an Integrated System for WebText Search. In Hundt, M., Nesselhauf, N. and Biewer, C. (eds.) *Corpus Linguistics and the Web*. Rodopi: Amsterdam. 47-67.

Resnik, P. and Smith, N. (2003) The Web as a parallel corpus. In *Computational linguistics*. 29 (3): 349-380.

Ringsletter, C., Schulz, K.U. and Mihov, S. (2006) Orthographic errors in Web pages: toward cleaner Web corpora. *Computational linguistics*. 32(3): 295-340.

Santini, M. (2007) Characterizing Genres of Web Pages: Genre Hybridism and Individualization. *Proceedings of the 40th Hawaii International Conference on System Sciences, poster session*. 1-10.

Santini M., Power, R. and Evans, R. (2006) Implementing a characterization of genre for automatic genre identification of Web pages. In *Proceeding of Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL 2006)*.

Santini, M. (2005) Genres In Formation? An Exploratory Study of Web Pages using Cluster Analysis. *Proceedings of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK 05)*.

Santorini, B. (1990) *Part-of-speech tagging guidelines for the Penn Treebank Project.* Technical report MS-CIS-90-47. Department of Computer and Information Science, University of Pennsylvania.

Scannel, K.P. (2007) The Crúbadán Project: corpus building for under-resourced languages. In Fairon, C., Naets, H., Kilgarriff, A. and de Schryver, G.-M. (eds.) *Building and exploring Web corpora. Proceedings of the WAC3 Conference*. Louvain: Presses Universitaires de Louvain. 5-15.

Scott, M. (1996/2004) *Wordsmith tools*. Oxford: Oxford University Press.

Sharoff, S. (2007) Classifying Web corpora into domain and genre using automatic feature identification. In Fairon, C., Naets, H., Kilgarriff, A. and de Schryver, G.-M. (eds.) *Building and exploring Web corpora. Proceedings of the WAC3 Conference*. Louvain: Presses Universitaires de Louvain. 83-94.

Sharoff, S. (2006) Creating General-Purpose Corpora Using Automated Search Engine Queries. In Baroni, M. and Bernardini, S. (eds.) *WaCky! Working Papers on the Web as Corpus*. Bologna: GEDIT Edizioni. 63-98.

Sinclair, J. (2005) Corpus and Text - Basic Principles. In Wynne, M. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. 1-16. Online: http://ahds.ac.uk/linguistic-corpora/ [consulted: 12/01/2007].

Sinclair, J. (2003) Corpora for Lexicography. In van Sterkenberg, P. (ed.) *A practical guide to lexicography*. Amsterdam: Benjamins. 167-178.

Stubbs, M. (1996) *Text and corpus analysis. Computer-assisted studies of language and culture*. Oxford: Blackwell.

Varantola, K. (2003) Translators and disposable corpora. In Zanettin, F., Bernardini, S. and Stewart, D. (eds.) *Corpora in translator education*. Manchester: St. Jerome Publishing. 55-70.

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

Thelwall, M., Tang, R. and Price, E. (2003) Linguistic patterns of academic Web use in Western Europe. *Scientometrics*. 56(3): 417-432.

Ueyama, M. (2006) Evaluation of Web-based Japanese reference corpora: effects of seed selection and time interval. In Baroni, M. and Bernardini, S. (eds.) *WaCky! Working Papers on the Web as Corpus*. Bologna: GEDIT Edizioni. 99-126.

Ueyama, M. and Baroni, M. (2005) Automated construction and evaluation of Japanese Web-based reference corpora. In *Proceedings of Corpus Linguistics 2005*. Online: http://www.corpus.bham.ac.uk/PCLC/ [consulted: 12/02/2007].

Zanettin, F., Bernardini, S. and Stewart, D. (eds.) (2003) *Corpora in translator education*. Manchester: St. Jerome Publishing.

Zipf, G.K. (1935) *Psycho-biology of Language*. Houghton Mifflin: Boston, MA.

“RINGRAZIAMENTI”

Non so davvero da dove iniziare… Perché le persone che devo ringraziare sono tante, e ognuna a modo suo mi ha permesso di arrivare fino a qui. Ma partiamo dall’inizio.

Ringrazio la mia famiglia, mio padre Claudio, mia madre Magda, mio fratello Filippo. Grazie per il sostegno che non mi avete mai fatto mancare, in nessuna circostanza e per nessuna ragione. Credo che non troverò mai il modo di dire quanto siete veramente importanti per me, e quanto vi voglio bene. Ringrazio anche i miei nonni, Anna (le tue torte sono ormai storiche a Forlì, e non sai quanto io le apprezzi, per quello che sono e per quello che rappresentano), Romano, Celina e Arrigo. E ovviamente Daniela, per la disponibilità e l’amicizia, che va oltre l’essere parenti… Teresa, Gianni, Paolo, Iva: anche a voi va il mio grazie.

Ringrazio Erminia, con cui ho passato i miei anni più belli. Perché condividere una casa non è solo condividere quattro mura. Grazie di tutto, davvero. Sei speciale per me e spero che tu lo sappia…

Grazie a Chiara, perché so che c’è e ci sarà.

Natacha, che non deve chiedere come sto per capirlo. Perché quando ti vedo mi ride il mondo dentro. E ovviamente Alessandra, per questi tre anni passati insieme.

Valentino, per le serate insieme e non solo. Per la bella amicizia, per cui mi ritengo molto fortunato.

Luna, che anche se è lontana, so che riesce a sentirmi.

Simone. Non saprò mai come ringraziarti abbastanza, per l’infinita pazienza e il continuo supporto.…

Per ultimi, ovviamente non per importanza, ringrazio i miei relatori: Silvia Bernardini, per aver creduto in me e per tutto quello che ha sempre fatto per aiutarmi a migliorare questo lavoro; e Marco Baroni, per il supporto fondamentale a questa tesi, e per avermi introdotto e guidato nel mondo della riga di comando. Infine volevo ringraziare Eros Zanchetta per il suo fantastico lavoro sui server e per l’aiuto con ukWaC, e Federico Gaspari per l’attentissima lettura che ha fatto di questa tesi.

ABSTRACTS

4.3 Riassunto

Lo scopo del presente lavoro è quello di presentare e valutare un nuovo corpus di lingua inglese. Il corpus, chiamato ukWaC (in vista del fatto che è un Corpus derivato dal Web campionando siti dal dominio .UK), contiene circa due miliardi di parole. ukWaC è stato costruito con l'intenzione di fornire una risorsa aggiornata e di grandi dimensioni, che sia paragonabile, in termini di "bilanciamento" e varietà di materiali linguistici, a corpora di riferimento tradizionali, e in particolare al British National Corpus (BNC), uno standard affermato per l'inglese britannico.

Come nel caso di tutti i corpora costruiti attraverso procedure semi-automatiche, tuttavia, la possibilità di controllare il materiale che confluisce nel corpus finale è limitata, il che rende la valutazione a posteriori un compito cruciale al fine di vagliare la reale composizione del corpus. Viene pertanto proposto e applicato un metodo di valutazione, che consiste nel paragonare ukWaC al BNC.

Per quanto riguarda la struttura del lavoro, il Capitolo 1 presenta un'introduzione a due aspetti della linguistica dei corpora che si rivelano centrali per il presente studio. Da un lato viene fornita una breve introduzione generale alla disciplina, che offre una descrizione del ruolo dei corpora negli studi linguistici e delinea alcuni dei criteri tradizionalmente coinvolti nella progettazione di corpora di riferimento. Dall'altro lato, il Capitolo 1 esplora la nozione di "Web as corpus", prendendo in considerazione i vantaggi e i potenziali svantaggi connessi all'uso di dati tratti dal Web, nonché diversi metodi attraverso i quali la Rete può essere sfruttata per scopi linguistici. Vengono inoltre forniti due esempi di come tali approcci siano stati applicati alla costruzione di risorse (WebCorp e WaC).

Il capitolo 2 discute le ragioni per cui ukWaC può essere visto come una valida alternativa alle risorse esistenti, tra cui il fatto che è un corpus stabile, di grandi dimensioni e potenzialmente bilanciato. Viene poi descritta in dettaglio la procedura seguita per raccogliere, ripulire e annotare i dati.

Il Capitolo 3 si concentra sulla procedura di valutazione, che nel nostro caso consiste in un confronto tra ukWaC e il BNC, preso come modello di riferimento di corpus generale. In particolare, vengono confrontate diverse liste di frequenza, ognuna delle quali comprende tutte le parole che appartengono alle principali classi di parti del discorso (nomi, aggettivi, verbi, avverbi con suffisso *-ly* e parole grammaticali). I risultati dell'analisi sembrano indicare che sussistono certe differenze tra i due corpora. Si riscontra in ukWaC una proporzione relativamente alta di testi legati al Web, al tema dell'istruzione e dei servizi pubblici, nonché di testi pubblicitari, e una relativa mancanza di testi narrativi e di trascrizioni del parlato. Nonostante queste differenze, tuttavia, numerosi tipi testuali e domini semantici non emergono come caratteristici di nessuno dei due corpora, il che sembra confermare la validità delle strategie di campionamento adottate durante la costruzione di ukWaC.

Il Capitolo 4 conclude suggerendo alcune direzioni di ricerca future. Innanzitutto è previsto un miglioramento del corpus attraverso un processo di ulteriore ripulitura dei dati, che ci auspichiamo contribuisca a fare di ukWaC una risorsa di largo utilizzo per lo studio della lingua inglese. Inoltre, sulla base dell'esperienza maturata nel presente lavoro, si suggerisce la necessità di individuare un metodo più completo di valutazione dei corpora tratti dal Web, che integri l'approccio descrittivo, come quello adottato nel presente studio, con compiti più orientati all'uso pratico di tali risorse.

4.4 Résumé

Le but de ce mémoire est de présenter et évaluer un nouveau corpus de langue anglaise. Ce corpus, appelé ukWaC (puisqu'il s'agit d'un Corpus tiré du Web à travers un échantillonnage de sites dans le domaine .UK), contient environ deux milliards de mots. ukWaC a été construit avec l'intention de fournir une ressource actuelle et de grandes dimensions qui soit comparable, en termes de « balancement » et de variété des matériaux textuels, à des corpora traditionnels et en particulier au British National Corpus (BNC), qui représente un point de repère très connu pour l'anglais britannique.

Toutefois, comme c'est le cas pour tous les corpora construits grâce à des procédures semi-automatiques, la possibilité de contrôler les textes qui sont inclus dans la version finale du corpus est limitée. Cela implique que l'évaluation à posteriori joue un rôle central afin de déterminer la composition réelle du corpus. Par conséquent, ce mémoire propose et applique à ukWaC une méthode d'évaluation, qui consiste principalement **à** le comparer au BNC.

Pour ce qui est de la structure de cet étude, le Chapitre 1 présente une introduction à deux aspects de la linguistique de corpus qui ont une importance primordiale pour nos objectifs. D'un côté, l'on introduit les principes fondamentaux de la discipline, par le biais d'une brève analyse du rôle des corpora dans les études linguistiques et des critères qui sont traditionnellement pris en compte quand il s'agit de construire des corpora de type général. De l**'**autre côté, on explore la notion de « Web as corpus ». En particulier, on prend en considération les avantages et les désavantages potentiels liés à l'emploi de données tirées du Web, aussi bien que les différentes méthodes à travers lesquelles la toile peut être employée pour des buts linguistiques. En outre, on fournit deux exemples de comment ces approches ont été appliquées à la construction de ressources (WebCorp et WaC).

Le Chapitre 2 discute les raisons pour lesquelles ukWaC peut être considéré comme une alternative valable aux ressources existantes. A savoir, il s'agit entre autres d'un corpus stable, de grandes dimensions, et potentiellement balancé. Par la suite l'on explique en détail la procédure suivie pour construire, nettoyer et annoter le corpus.

Le Chapitre 3 se concentre sur la procédure d'évaluation, qui dans notre cas implique une comparaison entre ukWaC et le BNC, pris comme modèle de corpus général. En particulier, on compare différentes listes de fréquence, dont chacune comprend tous les mots qui appartiennent aux classes principales de parties du discours (noms, verbes, adjectives, adverbes avec le suffixe *–ly* et mots-outils). Il apparaît que les résultats de l'analyse montrent certaines différences entre les deux corpora. A savoir, on relève en ukWaC une proportion relativement élevée de textes publicitaires et de textes liés au Web, aux thèmes de l'université et des services publiques, ainsi qu'une relative

absence de textes narratifs et de transcriptions du parlé. Toutefois, malgré ces différences, nombre de types textuels et de domaines sémantiques n'émergent pas comme étant typiques des deux corpora, ce qui pourrait confirmer la validité des stratégies d'échantillonnage adoptées pour la construction de ukWaC.

Le Chapitre 4 conclut en suggérant de futures directions de recherche. D'abord, on envisage d'apporter des améliorations à ukWaC grâce à un nettoyage ultérieur des données, ce qui, nous l'espérons, contribuera à rendre ukWaC une ressource très utilisée dans l'études de la langue anglaise. De surcroît, sur la base de l'expérience maturée au cours de ce travail, on suggère la nécessité d'identifier une méthode plus complète d'évaluation des corpora tirés du Web, qui puisse intégrer une approche descriptive telle celle qui a été adoptée pour ce mémoire, avec des tâches plus orientées à l'emploi pratique de ces ressources.