# Using web data for linguistic purposes

*Anke Lüdeling,* Humboldt University Berlin

*Stefan Evert* University of Osnabrück

*Marco Baroni* University of Bologna

## Abstract

*The world wide web is a mine of language data of unprecedented richness and ease of access (Kilgarriff and Grefenstette 2003). A growing body of studies has shown that simple algorithms using web-based evidence are successful at many linguistic tasks, often outperforming sophisticated methods based on smaller but more controlled data sources (cf. Turney 2001; Keller and Lapata 2003).*

*Most current internet-based linguistic studies access the web through a commercial search engine. For example, some researchers rely on frequency estimates (number of hits) reported by engines (e.g. Turney 2001). Others use a search engine to find relevant pages, and then retrieve the pages to build a corpus (e.g. Ghani and Mladenic 2001; Baroni and Bernardini 2004).*

*In this study, we first survey the state of the art, discussing the advantages and limits of various approaches, and in particular the inherent limitations of depending on a commercial search engine as a data source. We then focus on what we believe to be some of the core issues of using the web to do linguistics. Some of these issues concern the quality and nature of data we can obtain from the internet (What languages, genres and styles are represented on the web?), others pertain to data extraction, encoding and preservation (How can we ensure data stability? How can web data be marked up and categorized? How can we identify duplicate pages and near duplicates?), and others yet concern quantitative aspects (Which statistical quantities can be reliably estimated from web data, and how much web data do we need? What are the possible pitfalls due to the massive presence of duplicates, mixed-language pages?). All points are illustrated through concrete examples from English, German and Italian web corpora.*

## 1. Introduction

Different kinds of data are needed for different linguistic purposes. Depending on the linguistic question or problem at hand, a researcher has to identify the data he or she needs. For many research questions, data from a standard corpus like the *British National Corpus* (BNC) are sufficient. But there are cases in which the data needed to answer or explore a question cannot be found in a standard corpus because the phenomenon under consideration is rare (sparse data), belongs to a genre or register not represented in the corpus, or stems from a time that the corpus data do not cover (for example, it is too new). In these cases, the web seems a good and convenient source of data.

In this paper we want to focus on the possibilities and limitations of using the web to obtain empirical evidence for different linguistic questions.[1] In principle, there are several options for using data from the web:

a) Searching the whole web through a commercial engine:

I. One can use the commercial engine, for example Google or AltaVista, directly.
II. One can add pre- and/or post-processing to the search engine, to refine query results etc. Examples are WebCorp (Kehoe and Renouf 2002) and KWiCFinder (Fletcher 2001).

b) Collecting pages from the web (randomly or controlled) and searching them locally:

III. One can construct a corpus automatically by downloading pages from the web. This can be done by running Google queries or by using one's own web crawler (Ghani et al. 2001, Baroni and Bernardini 2004, Träger 2005). The data can then be processed in any way necessary (cleaning up boilerplate (roughly: the templatic parts of the web page in which certain formatting information is coded, doing linguistic annotation etc.).
IV. One can collect a corpus by manual or semi-automatic selection of pages downloaded from the web, according to precisely specified design criteria. This procedure is not different in principle from building a corpus such as the BNC or Brown Corpus, and has the same advantages and disadvantages as these (except that there is much more material without strict copyright on the web, see e.g. Hermes and Benden 2005). An example of such a procedure is described by Hoffmann (this volume).

In section 2, we focus on the direct use of search engines (Option I) since this approach is taken by most researchers (if only for pragmatic reasons) and compare them to traditional corpora. As examples of the latter we look at the publicly available portion of the *DWDS-Corpus*[2] (http://www.dwds-corpus.de/) for German data and the *British National Corpus* (http://www.natcorp.ox.ac.uk/, Aston and Burnard 1998) for English data, both of which contain roughly 100 million tokens. The BNC represents a "traditional" synchronic balanced corpus. It contains samples of British English from a wide range of registers which were published or recorded in the early 1980s. The corpus is distributed together with specialized software for linguistic searches, but the full data are included in the distribution and can also be searched with other suitable tools. The *DWDS-Corpus*, on the other hand, can only be accessed through a web interface that limits the number of search results and the amount of context which can be obtained. It was compiled for lexicographic purposes and consists of 10 subcorpora, balanced samples from each decade between 1900 and 2000.[3]

The advantages and problems of the other solutions (II - IV) will be discussed in section 3. A conclusion and outlook is given in section 4.

## 2. Searching corpora and searching the web

In order to search a corpus, one needs

(a) a qualitative description of the items to be found that can be operationalized in the form of search conditions;
(b) a stable corpus (at least for the duration of the data acquisition, but ideally also in the long term, so that experiments can be replicated by other researchers),
(c) the necessary (linguistic) annotation so that the items of interest can be located according to the search conditions formulated in (a); a tool to perform the search with high precision and recall (a query processor or search engine), and
(d) the possibility to categorize search results according to meta-information such as genre and age of speaker.

Every corpus search begins with a linguistic problem – the data are either used to explore a linguistic topic or to test a hypothesis that has been formulated by the researcher. As an example, consider the development of (German and English) non-medical *-itis*. A detailed discussion of the structural and quantitative properties of this suffix is given by Lüdeling and Evert (2005). Here, we chose it as an example because it is quite infrequent and there is some evidence that it has only developed recently. Therefore, standard corpora such as the BNC and the *DWDS-Corpus* will likely contain too few instances of non-medical *-itis* to support a thorough analysis.

In addition to medical *-itis,* which means 'inflammation' and combines with neoclassical stems denoting body parts (as in *arthritis* 'inflammation of the joints' or *appendicitis* 'inflammation of the appendix'), many languages have a non-medical version that is semantically derived from medical *-itis* but means something like 'hysteria' or 'excessively doing something', as illustrated in

(1) Possibly they are apt to become too ambitious – they rarely succumb to the disease of "fontitis" but are only too apt to have bad attacks of "linkitis" and "activitis". *(BNC, CG9:500)*
(2) Außerdem leide der Mann offensichtlich an Telefonitis, sagte am Donnerstag ein Polizeisprecher. *(DWDS-Corpus, o.A.[pid], Polizeibericht, in: Frankfurter Rundschau 06.08.1999, S. 31)*
'In addition, the man obviously suffers from telefonitis, a police spokesman said on Thursday.'

Types of questions that might be asked with respect to non-medical *-itis* are

- qualitative: With which bases does non-medical -*itis* combine?
- distributional: In which contexts are the resulting complex words used?
- quantitative: Is word formation with non-medical -*itis* productive?
- comparative: What are the differences (in structure or in use) between the English and the German affix? Is one of them more productive than the other?
- diachronic (recent change): When did non-medical -*itis* start to appear and what is its development?

First we need to formulate the search target. For all the research questions listed above we need to find instances of complex nouns ending in non-medical -*itis* in the given language. In most cases, we want to find all the noun *types* but it is not always necessary to obtain a complete list of their occurrences. For the quantitative studies, however, it is essential to identify all instances of each type so that type-token statistics can be computed. For the distributional studies, we also need some linguistic context and in most cases meta-information such as text type or age of speaker. The diachronic study requires a specific kind of meta-information, namely occurrence dates for all *itis*-tokens.

### 2.1 Reproducibility

In the next step, we need to find a suitable corpus. We do not address aspects of corpus design such as representativeness or balance (see Hunston, to appear), but rather focus on the issue of reproducibility. The corpus should be stable or grow in a controlled way (in the sense of a monitor corpus) so that the results of a study can be validated by direct replication of the experiment. Ideally, it should also be possible to test the *reproducibility* of the results by repeating the experiment on a different corpus that has been compiled according to the same criteria. For traditional corpora this is, at least in principle, possible by constructing a second comparable corpus. While often practically infeasible, it can be simulated by dividing up the corpus into two or more independent parts, to which the individual documents are assigned randomly. Results obtained on one of these parts can then be tested on the remaining parts. For corpora such as the *DWDS-Corpus*, which are only available via a web interface, the partitioning approach is usually difficult to implement (the only options provided by the *DWDS-Corpus* are partitioning by genre or by decade, so that the resulting sub-corpora are not truly comparable).

It should be immediately clear that being able to validate and reproduce findings is essential for any quantitative study, whose relevance depends crucially on the correctness and interpretability of the published numbers. It may be less obvious, though, why these issues also play a role for qualitative studies. Usually, a "qualitative" researcher is interested in finding examples of a specific construction or usage, which are then evaluated against a theory. Any example that exhibits the desired properties and is acceptable to native speakers can be used.

This superficial view is clearly inadequate, considering e.g. the qualitative description of the suffix *-itis.* Any claims made about the set of possible bases are invalidated when a replication (or repetition) of the experiment brings up contradictory examples.[4] Reproducibility is even more important when the interpretation of corpus examples depends on meta-information (which cannot be inferred from a simple example sentence, even by a native speaker) or a larger context (which cannot be included in a published report), as is typically the case for comparative and distributional studies.

When using the web as a corpus – especially when it is accessed through a commercial search engine – it is virtually impossible to test for reproducibility. Obviously, one cannot construct a second comparable corpus, a "shadow web", within the necessary time-frame for a synchronic analysis. While it would in principle be possible to divide the web pages collected by a search engine into random subsets in order to simulate repetition of an experiment, no commercial search engine currently offers such functionality.[5] One plausible solution is to perform experiments on a corpus that is compiled from the web in a controlled way. Then, additional comparable corpora can be constructed in the same way to test reproducibility of the results. This procedure is basically equivalent to regular corpus building and shares its limitations with respect to the amount of data that can be collected, cf. Option III in section 1. Another solution, which can – at least in principle – make use of the full amount of text available on the web, is to build a database of web documents (similar to that of a commercial search engine) that is fully under the control of linguistic researchers. It would then be easy to partition this database into random subsets of any size.

While validation of experiments is in most cases trivial for traditional corpora (provided that the corpus data and the search technology used are publicly available), the web is constantly in flux, and so are the databases of all commercial search engines. Therefore, it is impossible to replicate an experiment in an exact way at a later time. Some pages will have been added, some updated, and some deleted since the original experiment. In addition, the indexing and search strategies of a commercial engine may be modified at any time without notice. For instance, some unsettling inconsistencies have recently been discovered in Google's result counts for common English words. Shortly afterwards, the Google counts for many words (and especially those of more complex queries) began to fluctuate wildly and unpredictably as Google's engineers struggled to remove the inconsistencies.[6] Archiving efforts such as the Internet Archive's Wayback Machine (http://www.archive.org/) cannot solve this problem either. Despite the enormous size of its database,[7] the Wayback Machine covers a much smaller portion of the web than e.g. Google (Bill Fletcher, p.c.). It is difficult to estimate the true relevance of the replication problem: only experience will show how much the results produced by commercial search engines fluctuate over time (e.g. by tracking the web frequencies of different search engines for the same search terms over the course of several years).

A short digression seems to be called for at this point: Some researchers see the brittleness of web data more as an opportunity than as a problem. These researchers repeat their Google searches a few months after the original study. Provided that the results are overall the same, they claim that they have demonstrated the reproducibility of their experiment by repeating it on a different "snapshot" of the web. In doing so, they have succumbed to the statistical fallacy of using a non-independent data set for validation. While there can be no doubt that Google's database changes substantially over the course of a few months, the second snapshot will still contain almost all the web pages from the first one, except for those that were modified or deleted in the meantime.[8] It is therefore very unlikely that search results would change drastically during this time, except when the phenomenon being studied is more or less restricted to newly-indexed web pages (e.g. a new word that is coined and becomes popular in the time between the two experiments). Substantial changes usually indicate that the engine's indexing or search technology has been replaced by a different implementation, as noted above.

## 2.2 Corpus search

In this section, we look at the problem of locating the desired items in the corpus with high accuracy, the "corpus search". The two aspects of search accuracy are 'precision' (i.e. the search does not return too many "wrong" hits, called 'false positives'; see also Meurers 2005) and 'recall' (i.e. the search does not miss too many correct items, called 'false negatives'). While it is always necessary to achieve minimum levels of precision and recall, the precise requirements – and which of the two is more important – depend on the type of research question. Purely qualitative studies, where every example is evaluated by a native speaker, do not require a very high level of either precision or recall, although the manual work involved may become prohibitively time-consuming if too many false positives are returned. Low recall is problematic only when the search misses important instances that would support or contradict the hypothesis to be tested, and it is mitigated by large corpus size (especially when searching the web as a corpus). For quantitative studies, on the other hand, the correctness of the underlying frequency counts is paramount. Low precision can, in principle, be compensated by checking the result lists manually, provided that this is feasible both technically (i.e. full lists of results are available) and practically (i.e. it does not take too much time). For web data, these conditions are usually not met (see section 3.1). In any case, there is no way of correcting for low recall, which may lead to unpredictable errors in the frequency counts (since it is usually also impossible to estimate the level of recall that has been achieved).

The accuracy of a corpus search depends both on the range and the quality of linguistic annotations (including pre-processing steps such as identification of word and sentence boundaries) and on the search facilities offered by the software that is used. In the following, we will discuss these factors together, since the available annotations and search facilities are usually tightly coordinated: Corpora with rich annotations are often shipped with a specialized search tool that is geared to exactly the type and depth of annotation offered. It makes little sense for the Google database to include annotations that cannot be utilized by its search engine. The main purpose of this discussion is to compare the search possibilities and accuracy of traditional corpora (represented by BNC and *DWDS-Corpus*) with those of the web as a corpus (represented by Google). In doing so, we use the research questions on non-medical *-itis* outlined at the beginning of section 2 as a case study.

The basic requirement is to locate all complex nouns with the suffix *-itis* in the corpus. The corpus search has to be followed by manual inspection of the results in order to distinguish between medical and non-medical *-itis*. Since none of the corpora considered here are annotated with morphological structure,[9] we approximate the desired search condition by matching words that end in the string <itis>, regardless of whether it is a complete morpheme or not. Both the BNC and the *DWDS-Corpus* provide options for searching substrings of words. This method has perfect recall, but it will also return false positives such as *Kuwaitis*. Since both corpora include part-of-speech tagging, precision can be improved by searching only for instances tagged as nouns.[10] After manual validation, we find the following *-itis* nouns in the BNC that are clearly non-medical: *activitis, baggitis, combinitis, compensationitis, dietotectalitis, faxitis, fontitis, idlitis, lazyitis, leaguetableitis, linkitis, Pygmalionitis, ruggitis, taffyitis*, and *toesillitis*. Interestingly, some of them (*toesillitis, ruggitis*) are formed in direct analogy to medical terms and do not conform to the ‘doing too much’ semantics postulated above. We have now obtained a small set of qualitative evidence that can be used to describe the properties of non-medical *-itis*, such as the fact that non-medical *-itis* combines with native stems or names (medical *-itis* only combines with neo-classical stems). Similar results can be found for German (Lüdeling and Evert 2005).

For a more comprehensive and detailed account, it would be desirable to find more instances of these words (most of them occur just once or twice in the BNC and it is often difficult to derive their precise semantics from the examples) as well as additional *-itis* nouns (so that we can make valid generalizations about the set of possible bases). Using the web as a corpus, we should be able to obtain both substantially more *-itis* types and more tokens for each type.[11] Unfortunately, Google and other commercial search engines do not support any form of substring search, so it is impossible to obtain a list of all *-itis* nouns on the web. Thus, even this qualitative and exploratory study can only be performed on a traditional corpus, not on the web as corpus via a standard search engine. What can be done is to run web searches for the noun types found in the BNC in order to find more instances of them. Interestingly, for *Pygmalionitis* and *toesillitis* Google returns exactly the same example as in the BNC (from a poem and a best man’s speech, respectively), though in the latter case it is found on several different web pages, so a frequency of 10 is reported.[12]

In order to perform a quantitative study such as measuring the productivity of non-medical *-itis*, it is essential to have a complete list of types with reliable frequencies, to which a statistical model can then be applied. The frequency data obtained from the BNC and the *DWDS-Corpus* are highly accurate once the "wrong" types have been filtered out manually. Precision can be improved even further when all instances of the remaining types are checked as well, although this is often too time-consuming in practice.

Using frequency data from a search engine ("Google frequencies") is much more problematic. For one thing, all search engines perform some sort of normalization: searches are usually insensitive to capitalization ("poles" and "Poles" return the same number of matches), automatically recognize variants ("white-space" finds *white space*, *white-space* and *whitespace*) and implement stemming for certain languages (as in *lawyer fees* vs. *laywer's fees* vs. *lawyers' fees*, see Rosenbach, this volume). While such features can be helpful when searching information on the web, they may also distort the frequency counts. It is possible to deactivate some, but not all of these normalizations. However, this requires a detailed knowledge of the query syntax, which may change whenever Google decides to update its software (cf. the remarks on brittleness in section 3.1). Another serious problem has already been demonstrated by the example of *toesillitis* above, where 8 of the 10 pages found by Google are duplicates of the same best man's speech.[13] Such duplication, which is much more common on the web than in a carefully compiled corpus, may inflate frequency counts drastically. Manual checking could in principle be used to correct the frequency counts, both for normalization and for duplication, but it is prohibitively time-consuming (since the original web pages have to be downloaded) and is hampered by artificial limits that Google imposes on the number of search results returned.

## 2.3 Meta-data

Comparative studies rely on meta-data like mode (spoken vs. written), language, origin (dialect), genre, information about the demographic properties of the speaker, etc. to categorize search results. Statistical tests are then applied to the resulting frequency tables in order to detect systematic differences between the categories. Three requirements must be satisfied so that meaningful answers can be found with this procedure: (i) the corpus must contain a sufficient amount of data from all relevant categories; (ii) the corpus must be annotated with accurate meta-data (which have to be accessible through the search tool); and (iii) the total number of tokens in the corpus that belong to each category must be known. The BNC satisfies all three criteria, since its file headers provide rich meta-data that can be used for a broad range of comparative studies. The *DWDS-Corpus* also contains a certain (though smaller) amount of meta-information, but there is only limited access to this information via its web interface. In particular, requirement (iii) is not fulfilled.

For the web as corpus, it is reasonable to assume that all categories of written language are represented to some extent. However, there are no explicit meta-data, at least not of the kind required for linguistic research. The only possibilities for categorizing (or filtering) search results are by

- language: Google’s automatic classifier currently distinguishes between 35 languages;
- domain name: this has sometimes been used to approximate geographic location (national domains) or even dialect (e.g., ‘.com’ vs ‘.co.uk’), but is an extremely unreliable indicator (www.google.com, www.google. co.uk, www.google.de, www.google.it, etc. all refer to the same cluster of computers[14]), see also Fletcher (this volume) on problems of regional results;
- file format (HTML, PDF, Word, PowerPoint, etc.): this has presumably little linguistic relevance, except for highly specialized studies; and
- date: whether a web page has been updated within the last 3, 6 or 12 months.

In addition to these limitations on the available meta-data and their accuracy, requirement (iii) cannot be satisfied (except by extrapolation from the search results for a large set of very general words).

Diachronic studies can be seen as a particular type of comparative analysis based on a special kind of meta-data, namely date of occurrence (publication or recording). Of the three alternatives considered here, only the *DWDS-Corpus* provides the necessary information to answer a diachronic research question. Using the *DWDS-Corpus*, Lüdeling and Evert (2005) show that the non-medical use of *-itis* (in German) is not new, the first occurrences in the corpus are from 1915 (*Spionitis* ‘excessive fear of spies) but that it became much more productive and changed qualitatively in the 1990s. Neither the BNC nor the web could be used for such a diachronic study: Many traditional corpora, such as the BNC, are designed to be synchronic, so that diachronic analysis is only possible when a comparable corpus with material from a different time is available. While the web is an inherently diachronic resource, it has only existed for a short time span so far, and the available date information is highly unreliable. A recent date shown by Google may indicate that a page that has existed for years has only now been discovered by its crawler, or that minor (cosmetic) changes have been made to an old page. Conversely, many recent pages contain copies of novels, plays, poems, songs, etc. that were first published decades or centuries ago.

To summarize: For many linguistic research questions, such as the ones discussed with regard to non-medical *-itis*, there is no perfect corpus at the moment. The BNC is not diachronic and probably (if the productivity findings for German carry over to English) too old. The *DWDS-Corpus*, while it is diachronic and provides occurrence dates, is not yet stable enough and can only be searched through a web interface. While the necessary data is available on the web, there are not enough meta-data, the data are changing constantly, and the commercial search facilities are not useful to linguists. In the next section we therefore want to discuss other options for querying the web.

## 3. How to improve on Google

We discussed in some detail the problems of commercial search engines as tools for linguistic analysis. In this section, we shortly review current attempts to “improve on Google”, by making web data more suited for linguistic work. We can distinguish between systems that pre-process queries before they are sent to search engines and post-process the results to make them more linguist-friendly; and systems that try to dispense with search engines completely, by building and indexing their own web corpora.

### 3.1 Pre- and post-processors

Probably the most famous pre-/post-processing system is WebCorp (Kehoe and Renouf, 2002). Other tools in this category include KWiCFinder (Fletcher 2001) and the very sophisticated Linguist’s Search Engine (Elkiss and Resnik 2004). Here, we focus on WebCorp, but the main points of our discussion apply (albeit possibly in different ways) to any tool that relies on a commercial search engine as its data source.

WebCorp is a web-based interface to search engines such as Google and AltaVista, where the user can specify a query using a syntax that is more powerful and linguistically oriented than the one of the search engines. For example, it is possible to use wildcards such as * meaning “any substring” (as in: “*ing”). Moreover, WebCorp organizes the results returned by the search engine in a clean “keyword in context” format, similar to that of standard concordancing programs. Just like such programs, WebCorp also offers various result processing options such as tuning the kwic visualization parameters (e.g. larger / smaller windows), the possibility of retrieving the source document, word frequency list generation, computation of collocation statistics, etc.

A tool such as WebCorp makes it easier for linguists to formulate linguistically useful queries to search engines. For example, as we discussed above, search engines do not provide substring search options, e.g. the possibility of looking for all words that end in <itis> (“*itis”). WebCorp, by contrast, supports substring queries (see above). Moreover, WebCorp and the other tools provide post-processing functionalities that are obviously of great interest to linguists (e.g. the possibility of extracting a frequency list from a retrieved page). However, ultimately these tools are interfaces to Google and other search engines, and as such 1) they are subject to all the query limitations that the engines impose, 2) they cannot provide information that is not present in the data returned by the engines, and 3) they are subject to constant brittleness, as the nature of the services provided by the engines may change at any time. It is worthwhile looking at these three problems in more detail.

In terms of the first problem, the most obvious limitation is that search engines do not return more than a small, fixed number of results for a query. WebCorp cannot return more results than the search engine. As a matter of fact, WebCorp will typically return *fewer* results than the underlying engine, since it has to filter out results that do not match the user's query. For example, the search "*itis" (tried on WebCorp on April 18, 2005) did not return any results although, as we saw above, at least some of the *-itis* words from the BNC are also present in Google's database. The search "I like *ing" (tried on WebCorp on March 27, 2005) returned only 10 matches (3 of them from the same page). What probably happened here is that WebCorp had to query Google for "I like *" or "I like", and then go through the 1,000 pages returned by Google (the maximum for an automated query), looking for the small fraction of pages that contain the pattern "I like *ing". While precision is high (all contexts returned by WebCorp do indeed match the wildcard query), this comes at the cost of very low recall. In this example, recall is so low that it would have been better to use a traditional corpus such as the BNC (where the same "I like *ing" query returned 295 hits).

The situation is made worse by the fact that WebCorp (or any similar tool) does not have control over the Google ranking. If we can only see, say, 10 instances of a certain syntactic construction, we would probably prefer to see a random sample of the pages in which it occurs, or perhaps 10 pages that are "linguistically authoritative". Instead, the set of pages returned from a search engine will be the "best" according to criteria – such as popularity and topical relevance – that are not of particular linguistic interest (see also Fletcher, this volume).

The second problem with pre-/post-processors is that, if some information is not available through the search engine, it is very hard (and often impossible) for tools such as WebCorp to provide it to the user. Thus, most obviously, since the search engines do not allow queries for syntactic information (e.g. part of speech), such queries are not available through WebCorp either. More generally, any "abstract" query that is not tied to a specific lexical (sub-)string will either be impossible or, if the post-processor performs heavy filtering on the search engine output in order to simulate the query (as in the case of the "I like *ing" query above), it will result in very low recall.

Perhaps the most serious problem with systems that rely on search engines is their inherent brittleness. Search companies are constantly up-dating their databases and changing their interfaces. These changes imply that experiments done with a tool such as WebCorp are never truly replicable (because of changes in the databases). For example, the query "I like *ing" was repeated on April 18, 2005 (about 3 weeks after the first experiment) and returned only 8 results instead of 10. More dramatically, none of the functionality supported by the tools is guaranteed to work forever. For example, in March 2004, various features of KWiCFinder stopped working all of a sudden because the underlying search engine (AltaVista) had discontinued support for the relevant functionality (such as proximity queries). As another example, some features of WebCorp depend on the asterisk as a whole word wildcard in Google phrase queries. As of April 2005, it is not clear that Google will continue to support this syntax. Even if it does, the developers of WebCorp stated in recent postings to the Corpora mailing list that they intend to switch to their own search engine, in order to eliminate the brittleness problem (and more generally to avoid reliance on search companies whose priorities, of course, have little to do with helping linguistic research).

## 3.2 A search engine for linguists

This leads us to an alternative, more drastic way to try and "improve on Google", i.e. building one's own corpus directly from the web instead of relying on an existing search engine. Except for very small corpora, the process of downloading web pages to build the corpus (and any post-processing that is applied) must be automated. If the resulting corpus is in turn made available for querying through a web interface, one can speak of a proper "search engine for linguists" (Volk 2002, Kilgarriff 2003, Fletcher 2004, this volume). In principle, this is the optimal approach to using the web as corpus, given that it provides full control over the data (whose importance has been discussed in section 2). However, crawling, post-processing, annotating and indexing a sizeable portion of the web is by no means a trivial task.

It is telling that, even though the idea of building a linguist's search engine has been around for at least 3 years, to this date the only projects that have produced concrete results involved (relatively) small-scale crawls. For example, Ghani et al. (2001) sent automated queries to the AltaVista engine using words "typical" of specific languages and retrieved the pages found by the engine in order to build corpora of minority languages. Baroni and Bernardini (2004) used a similar approach (relying on Google instead of AltaVista) to create specialized language corpora for terminographical work. Sharoff (submitted) applied the tools developed by Baroni and Bernardini to build general corpora of English, Russian, Chinese and German text that are similar in size to the BNC. Studies of this sort have concrete results (e.g. Baroni and Bernardini's tools are publicly available and have been used in a number of terminological projects; Sharoff's corpora can be queried at http://corpus.leeds.ac.uk/internet.html), they demonstrate how various types of corpora can be created very rapidly using the web, and they provide useful material for the comparison of web data with traditional corpora. However, as one of the main reasons to use the web instead of a traditional corpus is to have access to an enormous database, small-scale corpus creation is not a satisfactory solution.

In what follows, we shortly review the main steps that would be necessary to build a linguist's search engine with a large database, highlighting the problems that must be solved at each step.

### 3.2.1 Crawling

A crawler is a program that traverses the web by following hyperlinks from one page to another. In our case, the crawler should download pages containing text, such as HTML pages, but also PDF and MS Word documents. The set of URLs used to initialize the crawl and various parameter settings of the crawler (e.g. the number of pages to be downloaded from each domain) will have strong effects on the nature of the corpus being built. Several tools that are freely available can perform efficient crawling (e.g. Heritrix: http://crawler.archive.org), but a broad crawl of the web will require considerable memory and disk storage resources. One argument that is often brought forward in favour of web corpora (as opposed to traditional static corpora) is that they offer language that is constantly "fresh" and open up the possibility of diachronic studies (cf. the discussion in section 2.3). To deliver on these promises, the linguist's search engine should do periodic crawls of the web. Thus, the issues of memory and storage are multiplied by the number of crawls to be performed (efficiency and computational power issues in all the following steps are of course also affected by the need to keep the corpus up-to-date).

### 3.2.2 Post-processing

Once a set of web pages has been crawled and retrieved, one has to strip off the HTML and other "boilerplate". The character encoding and language of each page must be identified. "Linguistically uninteresting" pages (e.g. catalogues and link lists) must be discarded. Identical and – much more difficult – "nearly identical" pages have to be identified and discarded (according to some criterion for when two pages are too similar to keep them both). None of these tasks is particularly difficult *per se*, and there is a large amount of literature in the Information Retrieval and www research community on topics such as near-duplicate detection (see, e.g., Broder et al. 1997). However, even "solved" problems such as language identification or near-duplicate detection require considerable computational resources and very careful implementations if they have to be applied to very large datasets, such as crawls that contain terabytes of data.

### 3.2.3 Linguistic encoding

Part-of-speech tagging, lemmatization, possibly automated categorization in terms of topic and other parameters are among the features that could really make the difference between a normal search engine and a specialized linguistic search engine. Again, it is not difficult to find tools to perform such tasks for many languages, but we will need very fast computers, very smart implementations and/or a lot of patience if we have to tag terabytes of data.

### 3.2.4 Indexing and retrieval

In our experiments, even a very efficient tool for indexing linguistic corpora such as the IMS Corpus WorkBench (CWB, ref. http://cwb.sourceforge.net/) has problems encoding corpora larger than about 500 million tokens. Thus, in order to index a corpus that contains many billions of tokens, one must either develop a new, extremely efficient indexer or design a distributed architecture in which the corpus is split into multiple sections that can be indexed separately. In turn, this complicates the retrieval process, which must pool the relevant information from several indexes. Based on our experience with CWB and large corpora, we also believe that retrieval would be much slower than on Google. However, this would probably not be seen as a major problem, as long as the information that can be retrieved is much more attractive to linguists than what is offered by Google.

### 3.2.5 Query interface

Powerful languages to retrieve information from a corpus are already available – e.g. the CWB *corpus query processing* language. A language of this sort would probably also be adequate for linguistic queries on the indexed web data, although, once again, particular attention must be paid to issues of efficiency (e.g. if a query is matched by 10 million kwic lines, the query interface has to provide highly efficient functionalities to work with the sheer amount of data that is returned).

## 4. Conclusion

A generalization emerges from the analysis of the various steps: While there is no major theoretical / algorithmic roadblock to the realization of a linguist's search engine, its implementation requires major computational resources and very serious, coordinated, high-efficiency programming – a far cry from the "do it yourself with a Perl script on whatever computer is available" approach typical of corpus linguistics.

There are also legal issues to be addressed. It is true that what we as linguists would be doing is not different from what Google and the other search engines have been doing for a decade, apparently without legal hassles. However, there are some worrying differences between linguists and Google: the linguist's search engine will "modify" the original pages (e.g. by adding POS information) in a much more radical way than Google does for cached pages; the linguist's engine would not provide "free advertising" as a high Google placement does; and the typical équipe of linguists is unlikely to have access to the same expensive legal expertise that Google can have. Even if the concrete legal threats are probably minor, they may have negative impact on fund-raising – and, as we just saw, such process is unlikely to be successful without the kind of computational and human infrastructure that requires a lot of funds.

It is very likely that the next few years will see the birth of one or more search engines for linguists. These engines will solve some of the problems we discussed in this paper: They will likely provide sophisticated query options, such as full substring search support ("*itis"), linguistic annotation (e.g. part of speech tagging), reliable meta-data, and they will not suffer from brittleness. In order to achieve these concrete goals, it is probably unavoidable that such engines, at least for the near future, will have to give up some of the most attractive characteristics of Google: Their databases will not nearly be as large nor have comparable cross-linguistic coverage, and (because of efficiency / storage constraints and to avoid brittleness) they will probably not be updated very frequently. Thus, for good or for bad, it is likely that this first generation of linguist's search engines and the underlying web corpora will look like oversized versions of the corpora we know (billions of words rather than hundreds of millions of word), solving some of the sparseness problems of current corpora, but still far away from exploiting all the dynamic linguistic potential of the web.

Despite the problems we highlighted, we are not pessimists. Indeed, two of the authors of this paper are involved in *WaCky* (*__W__eb as __C__orpus __k__ool __y__nitiative*, http://wacky.sslmit.unibo.it/), an informal initiative to rapidly build 1-billion-token proof-of-concept web corpora in three languages and a toolkit to collect, process and exploit such large corpora. However, we believe that – in order to go beyond roadmaps and manifestos, towards the concrete creation of a linguist's search engine – it is extremely important to be aware that this is a very difficult task and that this search engine will not be able to solve all the problems of corpus linguists. Too much optimism may lead to sour disappointment and unfair backlashes towards what is undoubtedly one of the most exciting perspectives in corpus linguistics today.

## Notes

[1] Whether the web can be viewed as a corpus is currently the object of much debate, since corpora are often defined as collections that have specific design criteria. This is not the topic of this paper (but see Kilgarriff and Grefenstette 2003 for a discussion). We are not interested in web data as an object of study (we will not study web English or Google English, for example); we are also not interested in data mining applications like Turney (2001), or other computational linguistic applications that use web data, as for example machine translation (Way and Gough 2003). We will also not argue for the general usefulness of corpora in linguistic research (see e.g. Meurers 2005).

[2] This corpus was compiled as a resource for the creation of a large German dictionary, the *Digitales Wörterbuch der deutschen Sprache*.

[3] At the moment, the publicly available portion of the *DWDS-Corpus* is slightly different from this core corpus because of legal problems.

[4] For our research question (qualitative description of words with non-medical -*itis*) we do not run into the problem of having to judge grammaticality (since we are looking for occurrences of a word formation process

in a changing process). For many other issues the difference between 'occurrence' and 'grammaticality' would have to be discussed.

5 It is also unlikely that such an option will be added in the future because it is irrelevant (perhaps even detrimental) for the search engines' target audience. The only possibility is to filter documents by their file type, language, or the internet domain they originate from (eg '.edu' vs. '.com' vs. '.org'), none of which can be expected to produce comparable subsets of the web (cf. Ide, Reppen and Suderman (2002), who express surprise at the fact that the language found in the domains '.edu' and '.gov' does not correspond to a balanced sample from general American English).

6 See http://aixtal.blogspot.com/2005/03/google-snapshot-of-update.html and pages referenced there for an entertaining and illuminating account of these events (accessed on 17 April 2005, but if these pages go off-line, you may still be able to retrieve them from Google's cache).

7 In October 2001, the archive had a size of over 100 terabytes and was growing at a rate of 12 terabytes per month (http://www.archive.org/about/wb_press_kit.php, accessed on 17 April 2005).

8 The second snapshot may even include many pages that were deleted and are no longer accessible, but are still available in Google's cache.

9 Much less for allomorphs – so it is not possible to search for non-medical *-itis* directly.

10 Making use of the fully automatic part-of-speech tagging of these corpora may result in a loss of recall, though, especially when there are systematic tagging errors in the data.

11 Keller and Lapata (2003: 467) estimate that the English part of the web indexed by Google is at least 1000 times larger than the BNC.

12 www.google.com, 17 April 2005.

13 The remaining two pages are a different version of the joke on which the speech is based, and a list of common misspellings of the word *tonsillitis* (www.google.com, 17 April 2005).

14 Tested on 17 April 2005 with the nslookup utility.

## References

Aston, G. and L. Burnard (1998), *The BNC Handbook.* Edinburgh: Edinburgh University Press.

Baroni, M. and S. Bernardini (2004), ‘BootCaT: bootstrapping corpora and terms from the web’, in: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, Lisbon.

Broder, A.Z., S.C. Glassman, M.S. Manasse and G. Zweig (1997), ‘Syntactic clustering of the web’, in: *Sixth International World-Wide Web Conference*, Santa Clara, California.

Elkiss, A. and P. Resnik (2004), *The Linguist’s Search Engine User's Guide*. Available at: http://lse.umiacs.umd.edu:8080/lseuser (March 29, 2005).

Fletcher, W.H. (2001), ‘Concordancing the web with KWiCFinder’, in: *Proceedings of the 3$^{rd}$ North American Symposium on Corpus Linguistics and Language Teaching*, Boston. Draft version at http://kwicfinder.com/FletcherCLLT2001.pdf (March 22, 2005).

Fletcher, W.H. (2004), ‘Facilitating the compilation and dissemination of ad-hoc web corpora’, in: G. Aston, S. Bernardini and D. Stewart (eds.) *Corpora and Language Learners*. Amsterdam: Benjamins. 275-302.

Fletcher, W.H. (this volume), ‘Concordancing the web: promise and problems, tools and techniques’.

Ghani, R., R. Jones and D. Mladenic (2001), ‘Mining the web to create minority language corpora’, in: *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM)*, 2001

Hermes, J. and C. Benden (2005), ‘Fusion von Annotation und Präprozessierung als Vorschlag zur Behebung des Rohtextproblems’, in: B. Fisseni, H.-C. Schmitz, B. Schröder and P. Wagner (eds.) *Sprachtechnologie, mobile Kommunikation und liguistische Ressourcen. Beiträge zur GLDV-Tagung 2005 in Bonn*. Volume 8 of *Computer Studies in Language and Speech*. Frankfurt: Peter Lang.

Hoffmann, S. (this volume), ‘From web-page to mega-corpus: the CNN transcripts’.

Hunston, S. (to appear), ‘Collection Strategies and Design Decision’, in: A. Lüdeling and M. Kytö (eds.) *Handbook of Corpus Linguistics* (HSK / Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science). Berlin: Mouton de Gruyter.

Ide, N., R. Reppen and K. Suderman (2002), ‘The American National Corpus: more than the web can provide’, in: *Proceedings of the Third Language Resources and Evaluation Conference* (LREC), Las Palmas, Spain. 839-844.

Kehoe, A. and A. Renouf (2002), ‘WebCorp: applying the web to linguistics and linguistics to the web’, in: *Proceedings of the WWW 2002 Conference*, Honolulu.

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

Kilgarriff, A. (2003), ‘Linguistic search engine. Abstract’, in: *Proceedings of the Workshop on Shallow Processing of Large Corpora 2003*, Lancaster. 412.

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

Lüdeling, A. and S. Evert (2005), 'The emergence of productive non-medical -*itis*. Corpus evidence and qualitative analysis', in: S. Kepser and M. Reis (eds) *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*. Berlin: Mouton de Gruyter. 350-370.

Meurers, D. (2005), 'On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German', in: *Lingua*, 115 (11): 1619-1639.

Rosenbach, A. (this volume), 'Exploring constructions on the web: a case study'.

Sharoff, S. (submitted), *Open-Source Corpora: Using the Net to Fish for Linguistic Data.*

Träger, S. (2005), *Korpora aus dem Netz – Die Erstellung eines Fachkorpus aus Webseiten.* MA Thesis, Humboldt-Universität zu Berlin.

Turney, P. (2001). 'Mining the web for synonyms: PMI-IR versus LSA on TOEFL', in: *Proceedings of the 12th European Conference on Machine Learning (ECML-2001).* Freiburg. 491-502.

Volk, M. (2002), 'Using the web as corpus for linguistic research', in: R. Pajusalu and T. Hennoste (eds.) *Tähendusepüüdja. Hatcher of the Meaning. A Festschrift for Professor Haldur Õim*. Tartu: University of Tartu.

Way, A. and N. Gough (2003), 'wEBMT: developing and validating an example-based machine translation system using the world wide web', *Computational Linguistics*, 29 (3): 421-457.