# Building general- and special-purpose corpora by Web crawling

*Marco Baroni & Motoko Ueyama*

SSLMIT
University of Bologna
(baroni|motoko)@sslmit.unibo.it

**Abstract**

The Web is a potentially unlimited source of linguistic data; however, commercial search engines are not the best way for linguists to gather data from it. In this paper, we present a procedure to build language corpora by crawling and post-processing Web data. We describe the construction of a very large Italian general-purpose Web corpus (almost 2 billion words) and a specialized Japanese "blogs" corpus (about 62 million words). In both cases, we compare the corpora to existing newspapers corpora, and discuss examples of potential linguistic applications.

## 1. Introduction

Corpora (collections of language samples produced in natural contexts and without experimental interference) play an increasingly central role in various branches of linguistics and related disciplines. For example, corpora have been extensively used to bring actual usage evidence to theoretical and applied linguistic questions ([21]), in simulations of language acquisition ([10]), in lexicography ([24]) and in a large number of tasks in natural language processing ([20]).

While some corpora, such as the 100M words British National Corpus (BNC, [1]), have become widely used resources, corpus-based linguistic/language studies are in constant need of more data, for a variety of reasons. For example, most languages lack a large, balanced corpus comparable to the BNC. Moreover, because of the Zipfian properties of language ([2]), even a large corpus such as the BNC contains a sizable number of examples only for a relatively limited number of frequent words, with most words of English occurring only once or not occurring at all. The problem of "data sparseness" is of course even bigger for word combinations and constructions: bigrams such as *beautiful memories* and *nice memories*, which seem fairly plausible collocations on an intuitive basis, only occur once in the whole BNC. Even the pair *cherished memories*, which strikes us as a lexicalized collocation, only occurs 3 times in the corpus. Thus, for a study of the different connotations of adjectives such as *beautiful*, *nice* and *cherished* as modifiers of *memories*, one would need a larger corpus. The situation is much worse if one wants to study technical languages, or genres/registers that are not represented in the available corpora. Moreover, since language is in constant evolution, no corpus can represent truly contemporary usages for long. For example, the BNC does not contain any instance of the word *blog* and its derivatives, and, obviously, it does not contain samples of the blog genre.

For these reasons (lack of resources in language of interest; data sparseness problems; need to study sub-languages or recent usages), researchers have been increasingly interested in the Web as a potential source of linguistic data ([18]). The Web contains a huge quantity of textual data for an ever increasing number of languages, it contains many different genres and specialized texts, and, of course, it is a "renewable" source of language, as long as people post new data.

In this paper, we first discuss some general pros and cons of using the Web as a corpus, arguing that none of the cons is specific to Web corpora *per se* (section 2). We shortly review the main approaches that have been adopted to obtain linguistic data from the Web, arguing for linguistically targeted crawls as the only viable long-term solution (section 3). We then describe the methodology we are currently following to build large Web corpora of Western languages, focusing on Italian (section 4). In section 5 we describe an experiment in which we adapted our techniques to crawl a corpus of Japanese blogs. For both languages, we also present examples of certain linguistic questions, e.g., those pertaining to sentence construction variation in colloquial Japanese, for which the constructed Web corpora might be better suited than traditional newspaper corpora. We conclude (section 6) by sketching directions for further work.

## 2. Pros and cons of using the Web as a corpus

As we already mentioned, one of the main advantages of constructing Web-based corpora is size. Size matters in NLP: In an influential paper ([3]), Banko and Brill have shown that even a simple disambiguation algorithm outperforms more sophisticated methods when it is trained on a larger amount of

Thanks to the members of the WaCky mailing list and in particular Silvia Bernardini, Adam Kilgarriff and Serge Sharoff for stimulating discussions. Special thanks to Eros Zanchetta, who prepared and ran the Italian crawl, and contributed to the software used in the reported experiments.

data, and that, even after seeing a training corpus of one billion words, the size-dependent increase in performance does not show signs of having approached an asymptote. However, size matters also in other more theoretical fields of linguistics. For example, Mair ([19]) has shown that the Web, unlike the BNC, is large enough to allow a full study of the grammaticalization of *get* as a passive in English. Turney ([27]) showed how a simple algorithm relying on Web frequencies outperforms a much more sophisticated method trained on a smaller corpus in a synonym detection task. For very specialized sub-languages, the vastness of the Web might be crucial in another sense: even if we intend to build a small corpus, only a database as large as the Web will provide enough documents to be included in such corpus ([5]).

The second advantage of the Web is that it allows fast and cheap construction of corpora in many languages for which no standard reference corpus such as the BNC is available to researchers. This set does not include only so-called "minority languages", but also well studied languages such as Italian and Japanese. The results of [23] and [28] suggest that Web corpora built by a single researcher literally in minutes are, in terms of variety of genres, topics and lexicon represented, closer to traditional "balanced" corpora such as the BNC than to mono-source corpora, such as newswire-based corpora. Moreover, these corpora will tend to reflect more recent phases of a language than traditional corpora, that are often subject to a certain lag between the time of production of the materials that end up in the corpus and the publication of the corpus itself.

The third advantage of Web corpora is that they can potentially contain a number of genres that are not present in traditional written sources. Phenomena such as blogging should be of interest to linguists since they generate vast amounts of written samples on a huge variety of topics that are spontaneously produced by non-professional writers. Furthermore, the Web provides plenty of samples of interactive communication that, while in written form, possesses some characteristics of oral communication ([25]). At the same time, as usage of the Web for various archival and practical purposes spreads, it is hard to think of traditional written genres that are not represented online.[1]

Web corpora, of course, pose also some problems. First, such corpora tend to contain a lot of noise, such as automatically generated non-linguistic material and duplicated documents. Second, perhaps more worryingly, since Web corpora are typically constructed with automated text mining methods, the researcher often does not have full control over what ends up in the corpus, and can only estimate the composition of the corpus with post-hoc methods. As a consequence of this, the corpus is likely to be defective in terms of meta-data information. Third, if a researcher plans to distribute a large Web corpus made of million of documents, (s)he will have a very hard time obtaining permission to use the documents from all the copyright holders.

Notice that none of these issues are unique to Web corpora. Rather, they come to the forefront with Web corpora because, using Web mining methods, one can collect a very large corpus in a very short time. If one collected a Web corpus of about 100M words spending the same amount of time and resources that were invested in the creation of the BNC, there is no reason to think that the resulting corpus would be less clean, its contents less controlled[2] or its copyright status less clear than in the case of the BNC. Vice versa, collecting a 1 billion word multi-source corpus from non-Web sources in a few days is probably not possible, but, if it were possible, the resulting corpus would almost certainly have exactly the same problems of noise,[3] control over the corpus contents and copyright that we listed above. Thus, we would like to stress that it is not correct to refer to the problems above as "problems of Web corpora"; rather, they are problems of large corpora built in short time and with little resources, and they emerge clearly with Web corpora since the Web makes it possible to build "quick and dirty" large corpora. It is a matter of research policy, time constraints and funding to determine if, for a certain project, it is better to invest considerable time and funds in building a thoroughly controlled, probably relatively small corpus, or if it is better (or: the only viable solution given external constraints) to build a possibly very large corpus that will probably suffer of all the problems above.

The very fact that it is relatively easy, very fast and cheap to construct Web corpora might actually provide innovative solutions to some of the issues described above. In particular, for certain lines of research it might be sufficient to share rapid corpus construction methods, rather than the data themselves, thus overcoming the copyright problems: a scientist could replicate the experiment of another scientist not on the same corpus, but on a corpus constructed according to the same criteria. This would be equivalent to what is commonly done in experimental fields of linguistics, such as phonetics, where scientists in different labs replicate experiments on comparable samples of subjects – not on the very same subjects!

[1] It has been pointed out to us that language on the Web might over-represent some groups of a language community, such as younger speakers, males, techno-savvy people. However, over-representation of certain groups seems a more general property of written language in general. Therefore, it seems part of the definition of what written language is about: 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.

[2] Serge Sharoff (personal communication) observes that a well constructed Web corpus might provide a straightforward operational answer to the eternal question of what is a "representative" corpus representative of: a Web corpus could be a corpus that samples, in the right proportions, the types of linguistic contents that an average user typically accesses online in a certain period of time.

[3] For example the LDC "gigaword" newswire corpora (`http://www.ldc.upenn.edu/Catalog/byType.jsp#Text.newswire`) are reported to suffer of serious duplication of text problems ([17]).

While, in general, we do not see any cons that are unique to Web corpora, there are, nonetheless, problems specific to particular Web-as-corpus methodologies, namely those that rely heavily on Google or other commercial search engines to obtain Web data. We turn now to a discussion of this and other approaches to using the Web as a corpus.

## 3. Three approaches to the Web as a linguistic corpus

### 3.1. Hit counts as frequency estimates

Probably the oldest and most widespread approach to using the Web for linguistic purposes is to issue a query to a search engine such as Google,[4] and to use the "hit count" reported by the search engine as an estimate of the frequency of occurrence of the searched string in the target language. This strategy has proven very successful in various tasks. To quote just one famous example, Turney ([27]) showed that a simple approach based on collecting hit counts of queries of the type `A NEAR B` to the AltaVista search engine performed better, in a synonym detection task, than a much more sophisticated method (latent semantic analysis) that used a traditional corpus as input.

However, despite its empirical successes, this approach is very problematic. First, the types of queries that one can issue to a search engine are very limited: for example, one cannot restrict the search on the basis of parts-of-speech, and one cannot use regular expressions. Indeed, since search engine users are typically interested in what their search terms refer to, rather than on their linguistic properties, search engines tend to perform a number of normalizations of the search terms that can be extremely annoying for linguists, such as ignoring case, stripping off diacritics, ignoring apostrophes and dashes. For similar reasons, search engines often ignore or produce very strange results when queried for function words. Thus, the types of research questions that one can approach with this methodology are *a priori* very limited: one can use search engine counts to look for the frequency of pre-selected content words and word combinations, and only if exact matching is not crucial to the research question.

Second, 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. A very dramatic illustration of this problem took place in April 2004, when AltaVista suddenly stopped supporting the `NEAR` operator, making Turney's original algorithm unusable. More recently, Jean Véronis has been reporting on his blog[5] a series of experiments pointing out various ways in which the counts of pretty much all the major search engines are unreliable and often obviously inconsistent.

In short, using search engine hits seems only appropriate for pilot studies, or in very restricted contexts, but it is not a viable long-term approach to using the Web as a corpus.

### 3.2. Building corpora through search engine queries

Rather than relying on search engine counts, one can issue automated queries to the search engine (search engines such as Google and Yahoo! provide Web service APIs that allow users to perform a certain number of automated queries per day), retrieve the pages returned by the search engine, and process them to build a corpus.

This approach has been explored by various scholars, including [15, 5, 14], and very extensively by Sharoff ([23]) and Ueyama ([28]). Sharoff shows how corpora in various languages built by issuing queries for random combinations of frequent words to the Google search engine and retrieving and processing the pages found in this way have characteristics more similar to those of a balanced corpus like the BNC than to those of mono-source corpora. Ueyama applies the same methodology to Japanese, and shows, in particular, how repeating the procedure at different times produces corpora that are dramatically different in terms of the pages retrieved, but relatively comparable in terms of their qualitative make-up (distribution of genres and topics). On the other hand, different choices of query terms lead to corpora with very different make-ups. In particular, using basic vocabulary words as query terms leads to corpora that are characterized by a high proportion of personal, everyday life prose produced by non-professional writers, whereas query terms sampled from more formal sources lead to corpora that are characterized by public/scientific/technical topics and professional authors.

This approach is less heavily reliant on Google (or other search engines) than the one discussed in the previous section. Google is used to obtain a list of documents, but then these documents are retrieved and post-processed by the researcher (tokenized, POS-tagged, etc.) locally, so that the stability of the data will no longer depend on Google, the researcher has full access to the corpus and, with the appropriate tools, the corpus can be interrogated with sophisticated linguistic queries.

However, the approach is not devoid of problems. The set of pages that are retrieved is still dependent on the search engine matching and ranking criteria. Moreover, for obvious reasons search engines restrict the amounts of data that can be obtained by automated querying (e.g., in the case of Google one can maximally retrieve 10K result URLs per day – and, of course, not all the pages retrieved are appropriate for a corpus). Thus, while the vast amount of data available there is one of the main factors attracting linguists to the use of the Web, building truly large corpora (in the order of millions of documents) with this method is extremely impractical. More in general, providing a legal automated query interface is un-

[4] Here and below, we will often use Google as generic term for any commercial search engine.

[5] `http://aixtal.blogspot.com/`

likely to be a high priority of search engines, and it would not be too surprising if the search engine companies stopped supporting such services, or started charging for them. Thus, it is not clear that even this is a viable long term approach.[6]

### 3.3. Linguistic crawls of the Web

We believe that the only viable long term approach to constructing Web corpora is for linguists to perform their own crawls of the Internet. This makes linguists fully independent from commercial search engines, and provides full control over the whole corpus construction procedure. However this is also the most difficult approach to implement, especially if the target is a large corpus. Considerable computational resources are necessary to host a large scale crawl; the data produced by the crawl have to be "cleaned" (removing pages not in the target language or problematic for other reasons; stripping off html code and "boilerplate", discarding duplicates). The data must then be annotated, minimally, with POS tags and lemmas (adapting the annotation tools to the language of the Web). When the input data are in the order of hundreds of gigabytes, all these steps become far from trivial, not only from the point of view of linguistic quality of the results, but also in terms of time and efficiency.

Indeed, as far as we can tell, none of the previous projects of large linguistic crawls of the Web went through all the stages of the outlined procedure [7] The terabyte corpus described in [11] has not undergone language filtering nor other forms of post-processing, and it is not annotated with linguistic information. The English Academic Web site corpus of [26] has also not undergone any form of post-processing, besides simple tokenization. The Chinese corpus of [13] has been filtered in terms of language detection, but it is not processed in any other way.

We believe that the kind of resources and know-how necessary to build, annotate and maintain large Web corpora will be attained more easily by a community of linguists interested in this task. For this reason, in the last year we have launched the informal *WaCky* initiative[8], which currently involves about 15 linguists in our school and other institutions. A concrete result of this initiative has been the construction of large ($>$ 1 billion tokens) Web-mined corpora of German and Italian that have been thoroughly post-processed and annotated with basic morphosyntactic information. Experiments on indexing these corpora and constructing an appropriate Web-based query interface are currently in progress, while the construction of corpora of English and Russian is on the pipeline. The tools developed by the project are made freely available[9] in the hope of stimulating public sharing of resources and know-how.

The procedure we followed to build the German corpus and one case study in which such corpus was used for a lexicographic task are reported in [8]. In the next section, we describe the creation of the Italian corpus, and a preliminary analysis of its contents (more information on various aspects of the general procedure can be found in [8]).

## 4. Constructing a large general purpose Web-based corpus for Italian

The procedure described in this section was carried out on a server running RH Fedora Core 3 with 4 GB RAM, Dual Xeon 4.3 GHz CPUs and about 2.5 TB hard disk space.

### 4.1. Crawl seeding and crawling

The crawl must start from a set of pre-selected URLs (the crawl "seeds"). We obtained a list of URLs in Italian by random queries to Google for combinations of frequent Italian words, taken both from a basic vocabulary list[10] and from a corpus of articles from the Italian *la Repubblica* newspaper (used later as a term of comparison for the Web corpus), and retrieving maximally 10 pages per query.[11] We discarded repeated URLs and, to insure maximal sparseness, kept only one (randomly selected) URL for each domain. The resulting list of 5231 URLs was used to seed the crawl.

The crawl of pages in the `.it` domain was performed by the Heritrix[12] spider with a multi-threaded breadth-first crawling strategy, avoiding pages whose URLs end in a suffix cuing non-html data (`.pdf`, `.jpeg`, etc.)[13] We let the crawl run for about ten days, retrieving about 81GB of gzipped archives (the Heritrix output format).

[6] Linguist-friendly interfaces to search engines, such as *WebCorp* (`http://www.webcorp.org.uk/`), while useful in that they reformat/reorganize the data returned by the search engine in ways that are more conductive to linguistic research – *kwic* display, frequency lists – are not providing any more information than what is provided by the search engine, and, thus, they present the same problems we described in this and the previous section.

[7] It is likely that similar projects have also been carried out in private by some big companies (see, e.g., IBM's *webfountain*, `http://www.almaden.ibm.com/webfountain/`). However, since these private projects are not publicly documented and the resources they produce are likely to remain private, they have virtually no interest for the linguistic researcher community.

[8] `http://wacky.sslmit.unibo.it`

[9] `http://sslmitdev-online.sslmit.unibo.it/wac/wac.php`

[10] `http://www.bardito.com/language/italian_english_wordlist.html`

[11] Our procedure still relies on a commercial search engine for seeding the crawl. An alternative would be to sample random pages from a publicly available URL list, such as that of `http://dmoz.org`, along the lines of [13].

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

[13] We used a regular expression created by Tom Emerson. and available at `http://www.dreamersrealm.net/~tree/blog/?p=4`

### 4.2. Post-crawl cleaning

#### *4.2.1. First pass filtering*

Before further processing, we discarded all retrieved documents that were not of mime type `text/html`, and documents smaller than 5KB or larger than 200KB. We also removed all sets of documents with perfect duplicates in the collection (we noticed in a random sample that sets of documents that are identical before html-stripping tend to be linguistically uninteresting; they are typically warning messages or copyright statements from the same server, or similar kinds).

#### *4.2.2. Code removal and boilerplate stripping*

Besides html and javascript code, Web-pages often contain link lists, navigational information, fixed notices, and other sections poor in human-produced connected text. From the point of view of corpus construction, boilerplate identification is critical: Too much boilerplate will invalidate statistics collected from the corpus and will impair attempts to analyze the text by looking at keywords in context. Boilerplate is much harder to spot than code, that is relatively easy to identify with regular expressions. We used (a re-implementation of) the heuristic used in the Hyppia project BTE tool,[14] based on the idea that the content-rich section of a page will have a low html tag density, whereas boilerplate text tends to be accompanied by a wealth of html (because of special formatting, many newlines, many links, etc.) Thus, among all possible spans of text in a document, we pick the one for which the quantity $N(tokens) - N(tags)$ takes the highest value. We remove html tags after they are used for the count.

The boilerplate stripping method we adopted is based on general properties of Web documents, and thus relatively independent of language and crawling strategy. Moreover, our method is not demanding on memory and it can be run in parallel on different machines, since it analyzes one document at a time. This is not the case with other approaches we considered, which are based on searching for frequent strings in the whole collection. On the negative side, our method tends to return a content-rich *fragment* of each page. As such, the resulting corpus is only appropriate for those who are interested in large collections of unstructured samples of natural language, whereas it does not contain reliable data to study the structure of Web documents.

#### *4.2.3. Function word and pornography filtering*

Connected text in sentences is known to reliably contain a high proportion of function words ([2]), so, we reject pages that do not meet this criterion. This process works as a language filter, and also it eliminates pages that mostly contain word lists, numbers, and other non-linguistic material. In our experiment, we use a list of 411 Italian function words for this purpose.

We also tried to eliminate pages containing pornography (since they often contain random text, probably used to fool search engines), using a stop list of 146 words typical of pornographic sites, and eliminating documents that contained more than a certain number of types and tokens from this list.

The boilerplate stripping and filtering phase took about one week, and it resulted in a version of the corpus containing 4,433,146 documents for a total of about 19GB of uncompressed data.

#### *4.2.4. Near-duplicate detection*

Next, we looked for near duplicates, that is, documents that (while not identical) contain substantial overlapping portions. We use a simplified version of the "shingling" algorithm ([9]), implemented in perl/mysql. 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. We look for pairs of documents sharing more than $t$ n-grams, and we discard one of the two. The pairs are ordered by document ID. To avoid inconsistencies, we always remove the second document of each pair. Thus, if the pairs A-B, B-C and C-D are in the list, only the document A is kept; however, if the list contains the pairs A-C and B-C, only C is removed. We leave it to further research to devise efficient ways to identify clusters of near-duplicates.

More precisely, we extract 25 5-grams from each document, and we treat as near-duplicates documents that share at least two of these 5-grams. 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. Near-duplicate spotting took about one day, and it resulted in a corpus containing 1,875,337 documents, for a total of about 10GB of uncompressed data (notice the dramatic shrinking from the 81GB of *compressed* data we initially obtained from the crawl!)

### 4.3. Part-of-speech tagging/lemmatization and indexing

We performed part-of-speech tagging with the widely used TreeTagger,[15] re-trained on our own training data, and lemmatization using the free Morph-it! lexicon.[16]. Morphosyntactic annotation took about two days, and resulted in a corpus of

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

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

[16] `http://sslmitdev-online.sslmit.unibo.it/linguistics/morph-it.php`

about 1.9 billion tokens (about 31GB of data including annotation).

In order to use the corpus in an efficient and linguistically sophisticated way, indexing needs to be carried out with a program that allows fast retrieval and sophisticated queries (e.g., queries for flexible part-of-speech patterns). For these purposes, we use the IMS Corpus WorkBench[17], which is probably the most popular free indexing and retrieval toolkit for very large corpora. However, at the moment this software will not index corpora larger than about 450M tokens as a single database. Thus, we have to split the corpus into multiple databases for indexing purposes, which makes retrieval (through a wrapper that queries the sub-corpora) slower and harder.

Among our priorities for future work, we plan to develop a simple Web-based interface to make the corpus available to interested researchers.

### 4.4. Preliminary evaluation of the Italian corpus

We compared our Italian Web corpus (henceforth, itWaC) with the *la Repubblica* corpus ([7]), of about 380M tokens, which collects 16 years of an Italian daily newspaper. The *la Repubblica* corpus is probably the largest annotated Italian corpus currently available for research purposes.

First, as a sanity check on our procedure, we compared the 30 most frequent words from both corpora. Comfortingly, the overlap is very large: 29/30 words are shared (conjunction *ed* is in the top itWaC list only, and third person singular clitic *lo* is in the top *la Repubblica* list only).

Adopting the methodology of [23], we then extracted the 20 function words most characteristics of itWaC vs. *la Repubblica* and vice versa, based on the log-likelihood ratio association measure ([12]). Results are presented in Table 1.[18]

Strikingly, 7 out of the 20 words most typical of itWaC are second person forms, cuing the high level of interactivity that characterizes Web speech. We also notice the common Italian informal greeting *ciao* – again, a mark of conversational text. Finally, the forms *perchè* and *nn* (vs. *perché* and *non* in *la Repubblica*) are informal spelling variations of common words (the first a common spelling mistake, the second a typical Web abbreviation). Among the words most typical of *la Repubblica*, we find third person forms, past tenses, temporal adverbs and the complementizer *che* "that", all typical marks of a more formal, narrative, non-interactive style.

In ongoing work ([4]), we are studying the semantic properties of the Italian verbal prefix *ri-*, which is similar, although not identical, to the English prefix *re-*. Thus, *aprire* means "to open" and *riaprire* means "to reopen", or "to open again".

As part of this project, we extracted the verbs that never occur with the prefix *ri-* in the *la Repubblica* corpus; we ordered them by their frequency in the corpus; and we studied the most frequent ones. The inspection of these "*ri-* repelling" verbs gave puzzling results. There are verbs such as *restare* "to remain", *sembrare* "to seem/to look like" or *bisognare* "to need", for which our native speaker intuition suggests that, indeed, it would be hard to construct a *ri-* form, since they do not denote activities or events with clear end points that could be repeated or undone: *rirestare*, *risembrare* and *ribisognare* sound as odd as their English translations, "to reremain", "to reseem" and "to reneed". However, for other verbs such as *rispondere* "to answer", *raggiungere* "to reach" and *arrivare* "to arrive", the absence of *ri-* forms is surprising, since *ririspondere*, *riraggiungere* and *riarrivare* sound perfectly plausible.[19]

Very interestingly, if we look at the frequency of occurrence of the *ri-* forms of the same verbs in itWaC, we find that *ririspondere*, *riraggiungere* and *riarrivare* are well-attested (occurring 18, 4 and 4 times, respectively), whereas *risembrare*, *rirestare* and *ribisognare* are still unattested, in agreement with our intuitions. While these results are, of course, very preliminary, they point to the hypothesis that, with a cor-

| itWaC | |
|---|---|
| ed "and" | hai "you have (sg.)" |
| perchè "because/why" | tali "such (pl.)" |
| delle "of the (f. pl.)" | tuo "your (m. sg.)" |
| tale "such (sg.)" | vi "you (acc./dat. pl.)" |
| ti "you (acc./dat. sg.)" | nn "not" |
| cui "which" | nonché "as well as" |
| presso "at" | di "of" |
| ciao "hi" | tua "your (f. sg.)" |
| tu "you" | possono "they can" |
| te "you (acc./dat. sg.)" | ovvero "or rather" |
| **Repubblica** | |
| ha "has" | una "a/one (f.)" |
| ieri "yesterday" | due "two" |
| ma "but" | il "the (m. sg.)" |
| un "a/one (m.)" | suo "his/her (m. sg.)" |
| aveva "had" | dopo "after" |
| hanno "they have" | non "not" |
| era "was" | fa "makes" |
| che "that" | lui "he" |
| più "more" | si "it/her/himself" |
| perché "why/because" | adesso "now" |

Table 1: Most typical words of itWaC and *la Repubblica*

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

[18] We also tried content words, but results did not look very interesting: we simply found highly topic-dependent words at the top of the lists.

[19] English is more choosy about the verbs that can combine with *re-*, so it is not clear that most English speakers will find the translations of these prefixed verbs equally acceptable: "to reanswer/to rerespond", "to rereach" and "to rearrive". However, we asked 5 Italian native speakers for a judgment, and they all agreed on the fact that *rirestare*, *risembrare* and *ribisognare* sound odd (although acceptable in some contexts), whereas *ririspondere*, *riraggiungere* and *riarrivare* are fully acceptable.

pus as large as itWaC, we are reaching a size at which, at least for certain tasks, negative evidence (non-occurrence of a form in a corpus) can be taken as a linguistically meaningful fact.

## 5. Constructing a corpus of Japanese blogs

Except for limited experiments ([28]), our previous corpus construction exercises focused on European languages (English, German, Italian), whose writing system is encoded in the latin-1 character set. Here, we report our first extended experience in creating a Japanese Web-derived corpus. Japanese is typologically very different from the Indo-European languages on which we previously worked. A difficulty presented by this language is that its orthographic system can be encoded in at least four distinct character sets (utf-8, shift-jis, euc-jp, iso-2022-jp). Moreover, there is no orthographic marking of the boundary between words.

We decided to focus on constructing a corpus of blogs. Blogging is a new genre not attested in traditional sources, and a genre that highlights how the Web contains samples of informal, colloquial expressions and constructions that are hard to come by in typical written language corpora, as we will show below.

### 5.1. Crawl seeding and crawling

We selected 4 popular Japanese blog services: Ameba, Goo, Livedoor and Yahoo. For each site, we manually picked 10 pages that contained links to blogs, and extracted the blog URLs in them with regular expressions, obtaining a list of 1399 URLs used to seed the crawl. Heritrix' crawl was constrained to pages whose URL matched patterns corresponding to the blog area of the various sites (e.g., for Yahoo, we only downloaded pages from the `blogs.yahoo.co.jp` sub-domain). We let the crawl run for about 1 day, retrieving about 380MB of gzipped archives.

### 5.2. Post-crawl filtering

As with Italian, we discarded documents of mime type different from `text/html`, documents outside the 5-200KB range, and sets of perfect duplicates. For the remaining documents, we extract the character set from the `charset` declaration in the html code,[20] and we convert all documents to utf-8.

We considered the possibility of writing ad-hoc boilerplate stripping scripts for the blog servers we crawled, but we then decided to use the same html density heuristic we described above for Italian, both because blog pages, as we found out in preliminary experiments, display a large variety of structures (due to different "skins" and such) and because we are interested in general solutions that will work with many different Web-text types, in view of our plans to build a general purpose Japanese Web corpus. The main issue with the html density heuristic is that it requires a rough tokenization of the html document, in order to compute the difference between number of tokens and number of types. With Western writing systems, we can perform a rough but fast tokenization by splitting on white space. This is not possible with Japanese data. At the same time, the use of a proper lexicon-based Japanese tokenizer already in the post-processing stage would make the phase much slower (not important for the current experiment, but crucial if we want to scale up). Moreover, we do not know of lexicon-based tokenizers that can handle html/javascript code and other peculiarities of Web documents. Instead, we simply chunked the text in the documents (excluding code) into fragments of 4 characters each, and treated each of these chunks as a token for our token-minus-tags score. Informal experimentation suggests that this method provides reasonable results, although output appears more noisy than in Italian.

Because of the same tokenization issues, we did not attempt to recognize near duplicates. We simply used md5 fingerprints to discard perfect duplicates.

### 5.3. Part-of-speech tagging/lemmatization and indexing

After post-crawl cleaning, we were left with a corpus of 28,530 documents (about 250MB of data). The corpus was converted into euc-jp, since both the annotation tool and the indexer showed problems handling utf-8. We tokenized and annotated the corpus with ChaSen ([22]), adding base form and part-of-speech information. The corpus processed in this way contains 61,885,180 tokens. The corpus was then indexed using the IMS Corpus WorkBench.

### 5.4. Exploring the corpus

We compare our blog corpus (henceforth, jBlogs) to the JENAAD corpus ([29]), which samples from the years 1989-2001 of the Yomiuri daily newspaper (although this was developed as a parallel Japanese/English corpus, we only use the Japanese texts). JENAAD is one of the few Japanese written text corpora freely and publicly available. For the corpora to be comparable, we re-tokenized the *JENAAD* texts using ChaSen with the same parameters used for jBlogs. The resulting version of the corpus contains 4,698,561 tokens.

First, we compared the distribution of parts-of-speech, as illustrated in figure 1.

Two striking differences emerge from the plot. First, jBlogs has a much higher ratio of symbols. By qualitative inspection, these are in part problematic tokens, such as html code, and in (apparently) a larger part the so called *kao-moji* ("face charac-

[20] In discussions on the WaCky mailing list, this method has been reported to be very unreliable for languages such as Chinese. However, in our previous experiments with Japanese (see in particular [28]) we found that the proportion of problematic documents is very small.

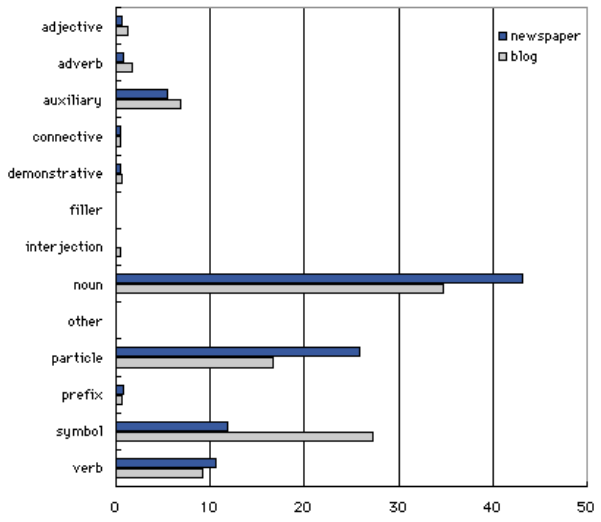


Figure 1: Percentage POS distribution in JENAAD (newspaper) and jBlogs (blog) corpora.

ters"), that is, the Japanese rich inventory of smiley-like symbols. These are split into individual characters by ChaSen (at least if default parameters are used), resulting in an inflation of the symbol category. Second, jBlogs has a much lower proportion of tokens tagged as particles than JENAAD. It is possible that one of the causes of this is the presence of particle dropping, a typical mark of informal speech, shortly discussed below.

Next, we looked for the most typical words of the two corpora, with the log-likelihood ratio score (as illustrated above in section 4.4 for Italian), focusing on the words tagged as verbs and nouns. For verbs, we examined the forms that the score ranked as the top 50 most typical of each corpus. In jBlogs, we found a majority of verbs characteristic of everyday communication, such as *taberu* "eat", *kaku* "write" and *hanasu* "talk/speak", whereas the verbs typical of JENAAD are all high register verbs, mostly pertaining to journalistic/narrative settings, such as *shimesu* "show", *motomeru* "seek", *susumeru* "proceed" and *mitomeru* "approve". Interestingly, we find register-dependent near synonym pairs such as the two verbs for "to say", *iu* (neutral) vs. *noberu* (formal), the first typical of jBlogs, the second of JENAAD. Given these patterns, we were surprised to find the very common verb *suru* "to do" among the most typical words of JENAAD. To further investigate this, we examined the contexts in which this verb appears in the two corpora, and in particular at the morpheme immediately preceding *suru* in 200 randomly selected occurrences from each corpus. It turns out that, in the large majority of cases, *92%*, *suru* in JENAAD occurs in Sino-Japanese compounds, such as *happyoo-suru* "announce". Such compounds are another mark of formal, journalistic style. These complex verbs made of a Sino-Japanese noun followed by *suru*, while not rare, are less common in the jBlogs sample (58%), where *suru* is also frequently attested in other contexts, e.g., preceded by native nouns or loanwords (24%), where it hardly occurs in JENAAD (1.5%).

For nouns, we inspected the top 150 most typical words in each corpus, classifying them into lexical classes. The major jBlogs types are blog/Internet jargon (e.g., *taitoru* "title"), first and second person pronouns, names with honorific suffixes, informal kinship terms (*mama* "mama", *danna* "husband"), temporal terms relating to the present and near future (*kyoo* "today", *ashita* "tomorrow", *mainichi* "every day"), as well as terms related to the events of personal live (love, death, happiness) and loanwords. The typical JENAAD nouns are lexically more uniform: Mostly, they are Sino-Japanese (hence higher register) words relating to social, economic or political issues (*keezai* "economics", *kaikaku* "reform", *gooi* "agreement").

As an example of the kind of research that would be difficult to carry out on a traditional newspaper corpus such as JENAAD, but for which jBlogs would provide a wealth of relevant examples, we can mention the topic of variation in particle dropping in informal Japanese (see [16] and references there). Particle dropping is an optional process. Both *sakana-o taberu?* ("fish-ACC eat") and *sakana-∅ taberu* ("fish eat") are acceptable informal ways to ask "Do you eat fish?". It is of interest to linguists to try to determine what makes one or the other choice more likely, and a corpus of informal texts can be an ideal source for such investigation ([16] use spoken conversational data, that are much harder to collect in large amounts).

In order to investigate whether our corpora do contain cases of optional particle dropping, we collected random samples of 100 sequences of a noun immediately followed by a verb from both corpora, with no intervening particle (the noun in this context is a candidate for particle dropping, since it could be the particle-less subject or object of the verb). The examination of the sentences sampled from JENAAD does not reveal a single case of optional particle dropping (the nouns turn out to be verbal predicates, parts of complex verbs, and other types where no particle is required). In contrast, 51 out of 100 cases in the jBlogs sample illustrate particle dropping. At the same time, it is easy to find, in the same corpus, plenty of instances in which particles are not dropped. For example, for the common expression *gohan-(o) taberu* "meal-(ACC) eat", we find 514 occurrences without the accusative particle, and 565 occurrences with the particle. Thus, the corpus would provide plenty of examples to study the contextual factors that make particle dropping more or less likely.

## 6. Conclusion

In this paper, we discussed the potential advantages of Web data for linguistic work, and we argued that such data should be accessed by crawling the Web directly, rather than using commercial search engines. We then illustrated the practical

work necessary to build Web corpora, describing a large Italian Web corpus collection project and a project that targeted the construction of a corpus of Japanese blogs. In both cases, comparisons with existing newspaper corpora suggested that Web data are characterized by a more colloquial, interactive style. In both cases, we also presented short examples of the sort of linguistic analysis that Web data are good for, either because of size (*ri-* in Italian), or because of the genres they represent (particle dropping in Japanese).

While we believe that there is no major theoretical or computational roadblock to the development of large Web-based corpora, our current research shows that there are many practical issues to be dealt with, if we want to build high quality Web corpora. This is particularly clear in the Japanese experiment, where the POS distribution revealed an anomalously high proportion of the symbol category, often corresponding to various forms of Web-specific data that were not filtered out during post-processing and/or that were not processed correctly by ChaSen. More generally, since various steps of the procedure that we used for Italian and other Western languages rely on efficient tokenization of the input, it is not clear that the procedure will scale up well for Japanese and other languages that do not mark word boundaries orthographically, if not at the price of quality of the results.

Another big issue is that standard corpus query indexing and search systems such as the Corpus WorkBench, at the moment, are not ready to handle corpora the size of our Italian Web corpus (less than 2 billion tokens), whereas in principle we could scale up to even larger sizes. This puts us in the frustrating situation of having potentially very useful data, but being able to use them and share them only in very limited ways.

To conclude on an optimistic node, however, we believe that our preliminary experiments show that even the current Web corpora, despite all their noise and indexing issues, can be very useful in various areas of linguistic research.

## 7. References


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

[2] Baayen, H. (2001) *Word Frequency Distributions*. Dordrecht: Kluwer.

[3] Banko, M., and Brill, E. (2001) 'Scaling to very very large corpora for natural language disambiguation', *Proceedings of ACL-01*.

[4] Baroni, M. (In press) 'Il prefisso *ri-* in italiano.' In *Studi linguistici offerti a Laura Vanelli*, R. Maschi, N. Penello and P. Rizzolatti (eds.) Udine: Forum.

[5] Baroni, M., and Bernardini, S. (2004) 'BootCaT: Bootstrapping corpora and terms from the web', *Proceedings of the Fourth Language Resources and Evaluation Conference*.

[6] Baroni, M., and Bernardini, S. (eds.) (2006) *WaCky! Working papers on the Web as Corpus*. Bologna: Gedit.

[7] Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G., and Mazzoleni, M. (2004) 'Introducing the La Repubblica corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian' *Proceedings of the Fourth Language Resources and Evaluation Conference*, 1771-1774.

[8] Baroni, M., Kilgarriff, A., Matiasek, J., Neubarth, F., and Trost, H. (Submitted) 'Fast construction of very large linguistic corpora by Web mining.'

[9] Broder, A., Glassman, S., Manasse, M., Zweig, G. (1997) 'Syntactic clustering of the Web'. *Proceedings of the Sixth International World-Wide Web Conference*.

[10] Clark, A., Cussens, J., Sakas, W., Xantho, A. (eds.) (2005) *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition of ACL 05*. New Brunswick: ACL.

[11] Clarke, C., Cormack, G., Laszlo, M., Lynam, T., and Terra, E. (2002) 'The impact of corpus size on question-answering performance.' *Proceedings of the 25th ACM SIGIR Conference on Research and Development in Information Retrieval*.

[12] Dunning, T. (1993) 'Accurate methods for the statistics of surprise and coincidence.' *Computational Linguistics* 19(1), 61-74.

[13] Emerson, T., O'Neil, J. (2006) 'Experience building a large corpus for Chinese lexicon construction.' In [6].

[14] Fletcher, B. (2004) 'Making the Web more useful as a source for linguistic corpora'. In *Corpus Linguistics in North America 2002*, U. Connor and T. Upton (eds.) Amsterdam: Rodopi.

[15] Ghani, R., Jones, R., Mladenic, D. (2001) 'Using the Web to create minority language corpora'. *Proceedings of the 10th International Conference on Information and Knowledge Management.*

[16] Endo Hudson, M., Kondo, J., and Sakakibara, Y. (In press) 'Zero-marked topics, subjects, and objects'. In *Japanese/Korean Linguistics 14*, T. Vance (eds.) Stanford: CSLI.

[17] Kilgarriff, A. (2005) 'The Web as corpus: An overview', paper presented at the *Giornata di Studi sulla Rete come Corpus*.

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

[19] Mair, Ch. (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*.

[20] Manning, Ch., and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge MA: MIT Press.

[21] McEnery, T., and Wilson, A. (2001) *Corpus Linguistics, 2nd edition*. Edinburgh: Edinburgh University Press.

[22] Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., and Asahara, M. (2000) *Morphological analysis system ChaSen version 2.3.3 manual.* NIST Technical Report.

[23] Sharoff, S. (2006) 'Creating general-purpose corpora using automated search engine queries', in [6].

[24] Sinclair, J. (2003) 'Corpora for lexicography', in *A practical guide to lexicography*, P. Van Sterkenberg (ed.) Amsterdam: John Benjamins.

[25] Storrer, A., and Beißwenger, M. (to appear) 'Corpora of computer-mediated communication', in *Corpus linguistics: An international handbook*, A. Lüdeling and M. Kytö (eds.). Berlin: Mouton de Gruyter.

[26] Thelwall, M. (2005) Creating and using Web corpora. *International Journal of Corpus Linguistics* 10(4), 517-541.

[27] Turney, P. (2001) 'Mining the Web for synonyms: PMI-IR versus LSA on TOEFL'. *Proceedings of ECML 2001*, 491-502.

[28] Ueyama, M. (2006) 'Creation of general-purpose Japanese Web corpora with different search engine query strategies', in [6].

[29] Utiyama, M., and Isahara, H. (2003) 'Reliable measures for aligning Japanese-English news articles and sentences. *Proceedings of ACL 2003*, 72-79.