The ‘body’ and the ‘web’: The web as corpus ten years on

Maristella Gatto, Università di Bari

Abstract

In this article the web’s controversial nature as a corpus is explored on both theoretical and applicative grounds. More specifically, the article shows how the notion of the web as corpus has changed, during the past decade, the way we conceive of a corpus from the somewhat reassuring standards subsumed under the corpus-as-body metaphor, to a new more flexible and challenging corpus-as-web image. On the one hand the traditional notion of a linguistic corpus as a body of texts rests on some correlate issues such as finite size, balance, part-whole relationship, permanence; on the other hand the very idea of a web of texts brings about notions of non-finiteness, flexibility, de-centering and re-centering, and provisionality. In terms of methodology, this questions issues which could be taken for granted when working with traditional corpora such as the stability of the data, the reproducibility of the research, and the reliability of the results, but has also created the conditions for the development of specific tools that try to make the ‘webscape’ a more hospitable space for corpus research. By simply reworking the output format of ordinary search engines to make it suitable for linguistic analysis (e.g. WebCorp, KWiCFinder), or by allowing the creation of quick flexible small specialized and customized multilingual corpora form the web (e.g. BootCaT), or by crawling more controlled parts of the web for the creation of large web corpora (e.g. Wacky project, Google Books NGram Viewer), recently developed tools and resources are decidedly redirecting the way we conceive of corpus work in the new Millennium along those lines envisaged by Martin Wynne as characterizing linguistic resources in the 21st century, such as multilinguality, dynamic content, distributed architecture, virtual corpora, connection with web search (Wynne 2002: 1204). These issues are discussed in this article with examples concerning the most common ways to exploit the web’s potential as a corpus, investigating the impact that the web has had on corpus linguistics in the past ten years, and pointing to future developments in this field.

1 A body of texts? Corpus linguistics and the web

Ten years ago, in a brief seminal paper on "The Web as Corpus" attention was called for the first time on the value of the web as a linguistic resource:

> The corpus resource for the 1990s was the BNC. Conceived in the 80s, completed in the mid 90s, it was hugely innovative and opened up myriad new research avenues for comparing different text types, sociolinguistics, empirical NLP, language teaching and lexicography.
>
> But now the web is with us, giving access to colossal quantities of text, of any number of varieties, at the click of a button, for free. While the BNC and other fixed corpora remain of huge value, it is the web that presents the most provocative questions about the nature of language (Kilgarriff 2001: 344).

In his stimulating paper Kilgarriff was envisaging a deeper connection between corpus linguistics and the web than could at the time be safely recognized on the basis of the actual impact of the web on corpus linguistics. He was indeed voicing a hope that one day the largest collection of authentic machine-readable text could be used by linguists, freely and ubiquitously, as a corpus in its own right. And it was perhaps more than hope, rather certainty, that made him conclude his paper with the challenging and controversial statement "The corpus of the new millennium is the web" (Kilgarriff 2001: 345).

Certainly, as Kilgarriff himself implicitly acknowledged, the notion of the web as corpus was in the first place nourished by opportunistic reasons. But while the reasons for turning to the web as a corpus were no doubt mainly practical (size, open access, low cost) at the outset, there appear to have been also other less obvious reasons for taking the patently risky direction of thinking of the web as a linguistic corpus. It can be argued, indeed, that the linguists' interest in the web originated in qualitative considerations concerning the nature of the web itself. Language is indeed "at the heart of the Internet", Crystal (2006: 271) has argued; it is a "social fact", rather than simply a "technological fact", where "the chief stock-in-trade is language" (Crystal 2006: 271) and as such it could not but attract the linguistis' attention – in certain respects almost against their will. The very notion of the web as "accidental" (Renouf and Kehoe 2006) or "unwanted" (Mair 2007) corpus came out to some extent as the result of inescapable convergence between a social phenomenon existing independently from linguistic investigation (the web) and the corpus linguistics approach, where the web came to be seen as a huge amount of texts in electronic format which both "tantalize and challenge linguists and other language professionals" (Fletcher 2007: 27).

The way to treating the web as a linguistic corpus was nonetheless by no means straightforward and was often landmarked by false starts, disappointment and disillusion. The idea of considering the web as a corpus did not only presuppose a view of what a corpus *is*, but also entailed a redefinition of what a corpus *could be*. Starting from the well-known Latin etymology of the word, virtually any collection of more than one text can be called a corpus, and there is obvious consensus that the authenticity of language data and electronic format are the basic *sine qua non* of a corpus in the modern linguistics sense of the word. The term has however notoriously acquired more specific connotations than this simple definition implies (McEnery and Wilson 2006: 29) and issues such as representativeness, size, sampling, balance, design and purpose always entered the debate at different levels whenever the notion of 'corpus' was at stake. Accordingly, the idea of considering the World Wide Web as a ready-made corpus by virtue of its mere nature as a collection of authentic texts in machine readable format was for a long time called into question. Nonetheless, linguists from all over the world were increasingly attracted by the web not only as a source of language text for the creation of conventional (well designed and carefully constructed) corpora, but also as a corpus in its own right. While taking for granted the qualitative difference between the web and a corpus designed and compiled as an object of language study, it was soon deemed possible to bypass the "ontological" question relating to what a corpus is and focus on the practical question "Is corpus x good for task y?", as Kilgarriff and Grefenstette argued in their famous editorial for the 2003 special issue of *Computational Linguistics* on "The Web as Corpus":

> We wish to avoid the smuggling of values into the criterion of corpushood. McEnery and Wilson (following others before them) mix the question "What is a corpus?" with "What is a *good* corpus (for certain kinds of linguistic study)?", muddying the simple question "Is corpus *x* good for task *y*?" with the semantic question "Is *x* a corpus at all?". The semantic question then becomes a distraction, all too likely to absorb energies that would otherwise be addressed to the practical one. So that the semantic question may be set aside, the definition of corpus should be broad. We define a corpus simply as a "collection of texts". If that seems too broad, the one qualification we allow relates to the domain and contexts in which the word is used, rather than its denotation: A corpus is a collection of texts when considered as an object of language or literary study (Kilgarriff and Grefenstette 2003: 334).

By implicitly shifting the notion of "corpushood" to the intention of the researcher rather than seeing it as intrinsic to the text collection itself, Kilgarriff and Grefenstette contributed to the emergence of a scientific community determined to exploit the inestimable potential of the web "when considered as an object of language or literary study" (Kilgarriff and Grefenstette 2003: 334). Mainly committed to the practical task of seeing whether the web could be profitably used as a corpus, research carried out under the label "web-as-corpus", especially in the field of computational linguistics and NLP, was apparently limited only to answering practical questions, while in fact each new study in this controversial field was imperceptibly contributing to reshaping corpus linguistics as a whole in the light of the specific features of the web as a spontaneous, self-generating collection of texts.

Thus the notion of the web as corpus rather than pushing key questions onto the background has worked as a sort of "magnifying glass for the methodological issues that corpus linguists have discussed all along" (Hundt *et al*. 2007: 4) and has provided an opportunity to explore some of the main tenets on which the good practice of corpus work rests. The most relevant issues for the purpose of the present article will be discussed in the following paragraphs.

2 The web as corpus: Key issues

2.1 Authenticity and representativeness

One of the most obvious reasons for turning to the web as a corpus was from the very beginning its undisputable nature as a reservoir of authentic language behaviour. Whatever its size, and however easily available as a collection of texts in machine readable format, there would have been no reason for turning to the web as an object of linguistic study had it not been made only of authentic texts, which are the result of genuine communicative events. Attention to the web as a source of linguistic information must therefore be seen as deeply rooted in the context of that "growing respect for real examples" (Sinclair 1991: 5), which the advent (and diffusion) of the new technologies has but reinforced. If then, to paraphrase Sinclair (1991: 1), it has now become fashionable to look "outwards to society" rather than "inwards to the mind" in the search for linguistic evidence, the web seems to be there ready at hand just to provide such evidence of language use as an integral part of the society it mirrors.

When it comes to the web, however, authentic does not necessarily mean reliable. Indeed, authenticity is often related to problems of 'authoritativeness' and it is everyday experience that authentic in the web often means inaccurate (misspelt words, grammar mistakes, improper usage by non-native speakers),

owing to its nature as an unsupervised unedited collection of texts. This has notoriously impaired any pretence to representativeness for the web as a corpus. Closely related to authenticity, representativeness is indeed the standard that most puzzled those who wanted to claim corpus dignity for the web. Representativeness entails in fact considerations concerning what should go in and what should be left out of a corpus based on clear ideas concerning the users of the language which a corpus aims to represent. As a consequence, while the enormous size of the web and its inclusiveness apparently make it a gateway to a potentially representative heterogeneous amount of language events, the notion of representativeness as it has been generally conceived of in corpus linguistics can only pertain to corpora which have been designed and created out of a selection from carefully chosen material. And this is not the case with the web, which is already there, independently from the linguist's intentions, as the result of a wide range of (but not all) everyday activities which imply knowledge exchange, communication, interaction, and for which the web is proving more and more a privileged mode. Paradoxically, however, this is where its real potential for representativeness also lies. The web is not constructed by a human mind, but it is the direct result of a number of human interactions taking place – significantly from a linguist's perspective – mainly through written texts which in the very act of their production are made available worldwide as authentic machine readable texts. Accordingly, the web's textual content inevitably reflects – if not represents – the international community at large in real time, and it could be argued, indeed, that it can be considered as "an increasingly representative and unprecedented in scale machine-readable sample of interests and activity in the world" (Henzinger and Lawrence 2004: 5186). Even though such a view of representativeness is not necessarily significant from the point of view of language, it cannot be dismissed as altogether irrelevant to it. Certainly the web "can in no way be considered a representative sample of language use in general" (Leech 2007: 145), but its scope, variety, and above all immense size, seem to legitimize confidence that these characteristics can counterbalance the limits of representativeness, so that the web's impossibility of being representative of nothing else but itself does not altogether destroy its value as a source of linguistic information from a corpus linguistics perspective.

2.2 Size and content

Intrinsically related to representativeness, the issue of size was equally fundamental in determining the value of the web as a resource for language study. While enormous size and virtually endless growth are the most notable characteristics of the web when compared to traditional corpora, this is precisely where

its limitations as an object of scientific enquiry lie. The notion of corpus should by default imply "a body of text of a finite size" (McEnery and Wilson 2006: 30), so that exact computations can be carried out, results can be compared, and statistics developed. Accordingly, the web's non-finiteness results in the impossibility to carry out such tasks, and hence in uncertainties and doubts concerning its value from the point of view of scientific research.

As to the question of how many running words there are in the web when considered as a text corpus, recent estimates of the web's size amount to one trillion unique URLs (Alpert and Hajaj 2008) which makes any computation of pages, let alone words, virtually impossible and useless. As far as the English language only is concerned, a fairly reliable estimate can be found in the respectable lower bound of one trillion words, i.e. the size of the training corpus used by Google when releasing their Web1IT data set in September 2006 (Official Google Research Blog 2006). Regardless of the temporary nature of these figures, what all estimates of the web's size imply is that linguists are faced with a corpus definitely larger than any other existing corpus, which alters altogether the meaning of size as a basic corpus issue. In the early days of corpus linguistics, it was a pains-taking task to reach the minimum size required for a corpus to yield significant evidence, but when it comes to the web as corpus, the role played by size seems to be reversed, and the linguist is faced with a collection of texts which can be literally overwhelming in terms of running words, and hence useless. Thus, while Sinclair could safely suggest, in the early 90s, that "a corpus should be as large as possible, and should keep on growing" (Sinclair 1991: 18), this is a truth that cannot hold when large actually means gargantuan and uncontrolled as is the World Wide Web. And this is the case not only from a corpus linguistics perspective. Also from the point of view of information retrieval, it has long become doubtful and disputable whether "bigger" is "better", even though it is undeniable that a large quantity of data accessible through the web is of great help when seeking unusual or hard-to-find information (Sullivan 2005).

The exponential growth of the web in size also had a great impact on its content, so that this was undoubtedly another key issue in determining its value as a corpus. A natural correlate of representativeness, the issue of content becomes even more indigestible given the intrinsic difficulties of characterizing the web in any of its aspects. In the past few decades the World Wide Web has indeed grown in such an anarchic fashion that it is virtually impossible to describe it in terms of its content (Grefenstette and Nioche 2000: 237). Moreover, when seen from a corpus linguistics perspective, a major flaw of the web was found in its intrinsic irreducible *anarchism*, which made the 100 million word *British National Corpus* comparatively resemble "an English country gar-

den" (Kilgarriff 2001: 344), and, more importantly, seemed from the very beginning to toll the bell for any hope to use the web as a corpus on sound methodological bases. Anarchy has always been the original sin of a virtual space which, as its very name reveals, is global more than anything else on earth. Not only all sorts of ephemera coexist with literary masterpieces, as pornography does with governmental documents and promotional text, but online texts are often only fragments, stock phrases, hot lists, and come in a myriad of duplicates and near-duplicates which are not of use from a linguist's point of view (Fletcher 2004). Furthermore the web has increasingly become a repository for multimodal content, with video and audio files representing a non-negligible part of its content, and ranking very high in most search engines' result pages. Despite such indescribable scenario, the issue of content will be conveniently split into three basic components – languages, topics, and genres – in the following pages to give, at least, an idea of scope of the web as corpus from the point of view of its content.

As far as language distribution is concerned, people have long thought quite naturally of the web as an almost monolingual English language corpus. On the contrary, in their much quoted article published in 2000, Grefenstette and Nioche estimated that while the web actually was a predominantly English language corpus (66%), non-English languages were growing at a faster pace than English. In fact one of the most interesting intrinsic characteristics of the web is its multilinguality, which, from a corpus linguistics perspective, means that it contains virtually endless parallel and comparable corpora, in almost any written language on earth, covering every domain and many topics, registers and genres. As to the present distribution of languages used on the Web, recent estimates of the top ten languages (30th June 2010) report English and Chinese as the most widely used languages, followed by Spanish, Japanese, Portuguese, German, Arabic, French, Russian, and Korean; see Table 1:

Table 1: Top ten languages used in the web (source: www.internetworldstats.com)

Top Ten Languages Used in the Web (Number of Internet Users by Language)					
TOP TEN LANGUAGES IN THE INTERNET	**Internet Users by Language**	**Internet Penetration by Language**	**Growth in Internet (2000 - 2010)**	**Internet Users % of Total**	**World Population for this Language (2010 Estimate)**
English	536,564,837	42.0 %	281.2 %	27.3 %	1,277,528,133
Chinese	444,948,013	32.6 %	1,277.4 %	22.6 %	1,365,524,982
Spanish	153,309,074	36.5 %	743.2 %	7.8 %	420,469,703
Japanese	99,143,700	78.2 %	110.6 %	5.0 %	126,804,433
Portuguese	82,548,200	33.0 %	989.6 %	4.2 %	250,372,925
German	75,158,584	78.6 %	173.1 %	3.8 %	95,637,049
Arabic	65,365,400	18.8 %	2,501.2 %	3.3 %	347,002,991
French	59,779,525	17.2 %	398.2 %	3.0 %	347,932,305
Russian	59,700,000	42.8 %	1,825.8 %	3.0 %	139,390,205
Korean	39,440,000	55.2 %	107.1 %	2.0 %	71,393,343
TOP 10 LANGUAGES	1,615,957,333	36.4 %	421.2 %	82.2 %	4,442,056,069
Rest of the Languages	350,557,483	14.6 %	588.5 %	17.8 %	2,403,553,891
WORLD TOTAL	1,966,514,816	28.7 %	444.8 %	100.0 %	6,845,609,960

Notoriously, differences in the relative weight of individual languages on the web point to more general problems concerning the so called “digital divide” between rich and poor countries, and the growth of languages other than English does not necessarily imply that access to the benefits of the Internet are more evenly distributed on earth. Anyway, while the problem of unequal access to the Internet remains an issue, the web has paradoxically proved an excellent lan-

guage resource precisely for some "minor", or even "endangered" languages (Ghani *et al*. 2001; De Schryver 2002; Zuraw 2006; Scannell 2007).

The importance of language variety on the web from a corpus linguistics perspective is further enhanced by the astonishing diversity of the topics covered. Indeed, there seems to be no field of human activity that is not some way or other covered by the web, so that several attempts have been made to implement some principles of classification of the web's content based on topic. This was generally performed through directories, which group web pages on the basis of content into a number of categories. A quick glance to the directories listed in the relevant page by the most common web search engines was enough to suggest that each directory (and sub-directory) could be considered as a "virtual" corpus including texts about the same topic. In practice, however, and despite their attractiveness, web directories seemed to answer the linguist's need only partially (Biber and Kurjian 2007). Moreover, as the web grows in size and anarchy, classification of web pages only by topic seems to be insufficient to maintain acceptable standards of effectiveness also from the point of view of information retrieval. This is the reason why greater and greater interest has been recently paid to the issue of web categorization by register and genre as a necessary complement to topic classification (Kwasnik and Crowstone 2001; Santini 2005; Mehler *et al*. 2010) and this is certainly an area where research by linguists, and the search engine industry's agenda partly converge. It goes without saying that the possibility of automatically identifying web genres is certainly going to pave the way towards a more methodologically sound use of the web as a corpus.

3 From body to web: New issues

An important characteristic of the web that has had strong implications for its supposed nature as a corpus is its inherently dynamic nature, with new pages and sites appearing at a significantly high rate and the content of existing documents being continually updated, so that sites and pages do not only frequently appear but also as frequently disappear. Furthermore, the very link structure of the web is in constant flux, with new links between documents being constantly established and removed (Risvik and Michelsen 2002). While such dynamism ensures that the web is constantly updated, also as a source of linguistic information, these factors have made the web gain a reputation for volatility – which no doubt everybody has experienced through the so called "broken-link" problem symbolised by the well known *HTTP Error 404 Page not found* message.

With a large fraction of existing pages changing over time, and a significant fraction of changes due to new pages that are created over time, all the web is constantly changing. Nonetheless there is no reason to assume that this perpetual change definitely alters the nature and composition of the whole. And while its fluid nature is often invoked as one of the main arguments against using the web as a corpus, one is also tempted to revive a powerful analogy with water (Kilgarriff 2001: 343); nobody would demand that the chemical composition of water in a river is exactly the same at each experiment, and nonetheless river water is undoubtedly a legitimate object of scientific enquiry. As Volk suggested at the very beginning of the web as corpus 'adventure', we only have had to learn how "to fish in the waters of the web" (Volk 2002).

An obvious practical consequence for linguistic research of the web's dynamism is the impossibility to reproduce any experiment. This poses a really serious problem since it is one of the basic requirements of scientific research that an experiment can be reproduced so that it can also be validated or, perhaps more crucially for the scientific method, invalidated. While validation of experiments is in most cases trivial when working with conventional corpora, the issue becomes crucial when using the web as a corpus (Lüdeling *et al*. 2007: 10–12), especially via ordinary search engines. The problem has been empirically addressed by some researchers who simply validate their results by repeating the same web search at distant intervals in time; others have instead opted for the possibility of using the web as corpus in different ways, e.g. by automatically downloading the results of the queries submitted to a search engine so as to create a more stable, and hence verifiable, object.

Finally, two more concepts that seem to have become worthy of the linguist's attention when it comes to the web as corpus are relevance and reliability (Baroni and Bernardini 2004; Fletcher 2004; Lüdeling *et al*. 2007), which relate to two aspects generally referred to in information retrieval as 'precision' and 'recall'.[1] While any linguistic search carried out by means of specific software tools on any traditional stable corpus of finite size (such as the BNC) would certainly report only (precision) results exactly matching the query, and all (recall) the results matching the query, it is patently not so with the web, whose unstable nature as a dynamic non-linguistically oriented collection of text fights against recall, whereas the intrinsic limitations, from the linguist's perspective, of search tools such as ordinary search engines challenge precision.

This is precisely what makes using frequency data from a search engine (the so-called "Google frequencies") as indicative of frequency of a given item in the web as corpus more problematic than it might seem at first glance. To assume a fairly high number of hits for a query as evidence of usage is not some-

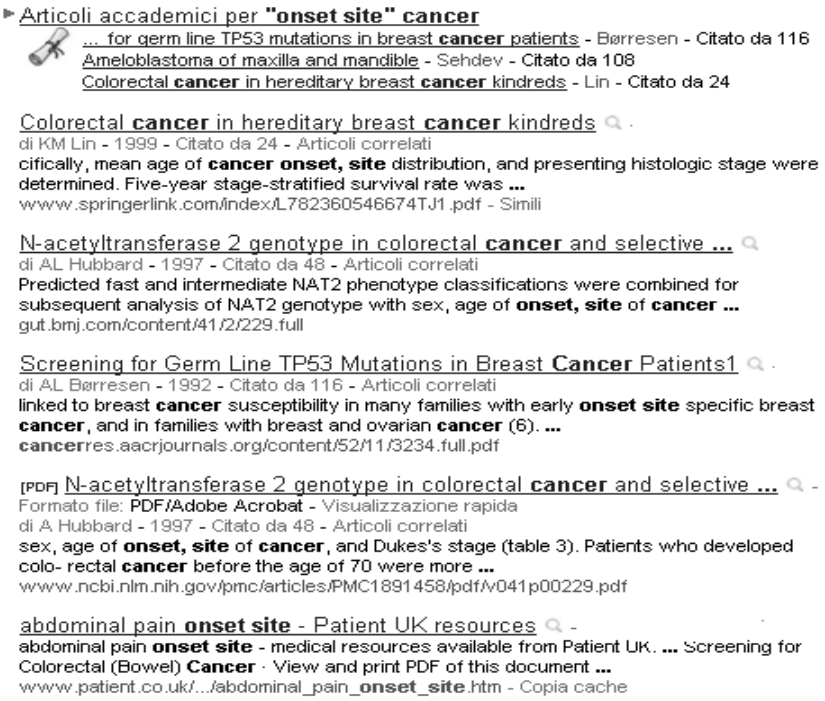
thing which can be taken for granted. For one thing, reliability and recall are made problematic by the huge number of duplicates and near-duplicates found by search engines, which ultimately depends on the very dynamic nature of the web. The presence of duplicates on the web, an issue generally alien to carefully compiled corpora, obviously inflates frequency counts dramatically, making numeric data obtained from hit counts on the web virtually useless from the point of view of statistics. Furthermore, search engines are themselves unreliable tools, subject to strange phenomena like "dancing" (Nakov and Hearst 2005) and often producing incredibly inconsistent results (Véronis 2005). On the other hand relevance and precision are impaired by the very strategies that enhance the power of search engines as tools for retrieving information (not specifically linguistic) from the web such as lemmatization, normalization of spelling, and so on. While such features are undoubtedly helpful when searching for general information on the web, they certainly affect the search possibilities in terms of precision and relevance (Lüdeling *et al*. 2007: 12–14).

All the issues so far surveyed indicate that even though the advantages of using the web as a corpus were evident from the very beginning, it was of crucial importance to devise new approaches to make the web more useful for corpus research. A brief survey on the following pages suggests that such development can be charted as a slow but steady process of decreasing dependence on ordinary websearch engines, pointing to a more fecund osmosis between corpus linguistics and the web.

4 Approaches, tools and methods: From web as corpus to corpus as web

The most obvious and immediate approach to the web as a corpus implies that the researcher queries the web as a corpus "surrogate" (Baroni and Bernardini 2006: 10–11) through an ordinary search engine and that hit counts are used as a source of qualitative/quantitative evidence of attested usage. While this approach may have proved successful at various tasks, it is the one that has most frequently foregrounded some of the basic uncertainties concerning the value of the web as an object of scientific enquiry in the past ten years. In this case, indeed, the linguist has not only very scarce control over the *corpus* itself, the web, but also over the search tools, i.e the search engines, which are clearly not designed for use by linguists. The result is more often than not an improved awareness of "googleology", as Kilgarriff (2007) would argue, rather than reliable linguistic information.

A first obvious problem is to interpret the meaning of frequency of occurrence on the web and assess the authoritativeness of the results. A second limit of web search is related to the already mentioned issues of relevance and reliability. This can be exemplified by a search aimed at evaluating *onset site* and *site of onset* as alternative wordings with reference to *cancer* in a medical text.[2] A Google search for *onset site* would find over 8000 matches, a result which is not significant in itself but requires further interpretation. What is particularly doubtful, besides the significance of mere Google frequency, is the relevance of the results to the specific context of *cancer.* Indeed, by adding the word *cancer* to the search string, the number of matches drops to 1790, all of which apparently irrelevant and/or unreliable. At closer inspection, in fact, one finds that only in one instance the word *onset* does really premodify *site*, while in most other cases the two nouns are separated by some punctuation mark:



At this stage one could further restrict the query to a known domain such as a portal for the online distribution of scientific journals (e.g. www.elsevier.com) or even to Google Books to boost the reliability of the results. These would indeed confirm the inappropriateness of *onset site* (featuring, again, *onset*, *site* rather than *onset site*), whereas an alternative search for *site of onset* produces a significant number of matches, all to be considered relevant and reliable because of the co-occurrence with *cancer* and of the controlled provenance of the results.

This reveals that even the apparently trivial task of searching the web for evidence of usage poses specific problems for the researcher, which require that cautionary procedures are adopted both in the interpretation of the results and in submitting the query to the search engine. More precisely, while the working of search engines and the web's content fall out of the user's control, the only part of the search which the linguist can control is the query, indeed "the loadstone of search, the runes we toss in our ongoing pursuit of the perfect results" (Battelle 2005: 27). The query can be the place where the practice of web search and a linguist's theoretical approach to the web as a corpus can significantly interact.[3]

Another typical use of the web as a corpus 'surrogate' is by means of specific tools like WebCorp (www.webcorp.org.uk), a well known web concordancer capable of transforming the result page of an ordinary search engine into a concordance table that can be profitably used to explore web data from a corpus linguistics perspective. However limited, the system provides basic sorting facilities and statistics, such as the computation of collocates, and has already proved successful at various tasks (Renouf and Kehoe 2006; Gatto 2009).

A further achievement in the attempt at making the web more useful for linguistic research is BootCaT (Baroni and Bernardini 2004), a tool that can be seen as the natural development of the widespread practice of building up "Do-It-Yourself" or disposable corpora (Zanettin 2001; Varantola 2003). In this case the web is seen as a "corpus shop" (Baroni and Bernardini 2006: 11), the place to go to for quick-and-dirty corpus creation. The only thing BootCaT needs to start the process is a number of key words which the linguist considers likely to occur in the specialized domain for which a corpus is going to be built. These words (seeds) represent the first step towards corpus creation: the seeds are transformed into a set of automated queries submitted to an ordinary search engine, which in turn produces a list of results from which a new list of terms is extracted. These new terms are used as seeds to build a larger corpus via more automated queries, and so forth. The linguist can control various aspects of the cyclical process by means of a number of options including the possibility to manually filter – and possibly deselect – URLs before web pages are included in the corpus. An apparently trivial option, the possibility of manually filtering the

URLs signals a sort of continuity with conventional corpus creation where linguists are generally careful in choosing their text wisely. By checking URLs, one is probably doing more than simply limiting the risk of filling a corpus with undesired web pages. Indeed the act of controlling the source of each web page can be seen as a sort of post-hoc action in terms of representativeness and design, whereby the user determines, albeit in a very limited way, what should go into and what should be left out of the corpus on the basis of external criteria (such as the institution, organization or even the genre to which the URL's domain can indirectly refer). The average time spent in building the corpus (generally less than 15 minutes in all), its decidedly task-oriented nature, as well as the fact that it can be analysed either online (through a specific corpus query tool) or off-line (by downloading it to one's own personal computer), make this a good example of how some aspects of corpus work can change – and are actually changing – under the impact of the web. More specifically, a tool like BootCaT greatly enhances the possibility of exploiting to the full the web's nature as a multilingual dynamic environment and its inclusiveness in terms of topic coverage, without renouncing such standards of corpus work as the stability of the data set to be explored, the reproducibility of the search, and the possibility of performing more specific linguistic oriented analysis with corpus query tools.

Finally particular attention deserves the creation of several mega-corpora from the web for languages such as Italian, German, English, Spanish, Chinese and Russian, and many more – all available for free through the Wacky Project website. These corpora fall in the "mega-Corpus – mini-Web" category in the map drawn by Baroni and Bernardini for Web as/for Corpus research (Baroni and Bernardini 2006: 13), and seem to bridge the gap between the corpus linguistics community and those researchers who are fascinated by the promises and possibilities offered by the web as a corpus but are not going to give up high methodological standards. As corpora that have been created through a semi-automated process based on crawling, filtering, deletion of duplicates and near duplicates, and that are connected with the creation of specific query interfaces (Baroni and Bernardini 2006; Baroni and Ueyeama 2006) these corpora have actually moved a further step in the process of reconfiguration of the notion of a linguistic corpus in the web era. They do represent a really new object, characterized by both web-derived and corpus-like features, whose aim is to answer the widely-felt need for corpus resources that combine the potential for size, variety and topicality offered by the Web with the reliability of conventional corpora and corpus tools. What these new resources, and the tools that have accompanied them, largely signify is a shift from the idea of the *web as corpus*

to the new notion of *corpus as web*. As Baroni and Bernardini openly recognized, a few years ago, commenting on Google's popularity among linguists:

> The enormous popularity that Google enjoys among linguists can only in part be explained by the fact that it makes an unprecedented amount of language data available. We believe that an equally important role is played by the fact that Google search is easy to use and can be accessed through a familiar user interface, presents results in a clear and tidy way, and that no installation procedure is necessary. (…) In other words, we should not only use the Web as a corpus, but also present the *corpus as web*, i.e., provide access to Web corpora in the style of a Web search engine (Baroni and Bernardini 2006: 37).

It is indeed worth emphasizing that rather than simply advocating the development of new corpus resources tools, the authors were indicating a shift in the expectations of corpus linguists, as a consequence of a growing and widespread familiarity with ordinary web search. This seems to point to a metamorphosis in the way of conceiving of corpora and corpus tools under the impact of the web, which in turn brings about interesting changes also as far as the basic activities of accessing, distributing and querying corpora are concerned. As the notion of "mega-corpus mini-web" becomes a reality, the tendency for a working scenario where the linguist no longer downloads corpora and tools to his/her personal computer but rather works from any computer on data and query tools made available through a remote server has become more typical and desired than it was with traditional corpora. In such a context, even the basic act of reading, interpreting and drawing conclusions from concordance lines can become a problem. However refined and detailed, mere concordancing and statistics relating to collocates, clusters or patterns may be no longer enough with corpora where words can have thousands of occurrences and the plethora of data with which the linguist is likely to work definitely requires some form of summarising. This changing scenario is perhaps exemplified at its best by the Sketch Engine, a really telling example of a different way of conceiving the basic activities of accessing, distributing and querying corpora. The Sketch Engine service makes a number of large web corpora available for online analysis and exploration, which can be performed using a web-based corpus query tool, namely the Sketch Engine, which contributes to a thorough exploration of concordance lines by supporting complex queries and by providing statistics relating to the collocational profile and to the grammatical relations that each word in the corpus participates in. Interesting examples concern the kind of information that can be obtained, in a matter of seconds, for such complex and frequent words as

man, *woman* or *nature* and *culture*. By way of example one could mention data retrieved for such a frequent and indeed significant word as *culture*; see Table 2:

Table 2: Word sketch for *culture*

Home | **Concordance** | Word List | **Word Sketch** | Thesaurus | Sketch-Diff

Turn on clustering | More data | Less data | Save

culture **ukWaC freq = 161537**

object_of 29548 **1.4**		subject_of 18986 **1.5**		adj_subject_of 2830 **1.5**		a_modifier 63673 **2.8**		n_modifier 25393 **1.8**	
foster	333 50.47	influence	197 34.98	supernatant	14 41.93	popular	3327 64.07	youth	975 55.82
promote	740 42.89	collide	29 29.81	alive	42 31.19	Western	881 55.34	pop	542 55.04
experience	542 42.15	be	7677 27.83	evident	37 30.01	organisational	755 53.9	blame	238 53.51
create	1191 41.89	have	2156 26.99	different	141 28.02	contemporary	1265 53.02	tissue	657 52.96
change	986 41.8	flourish	30 25.4	prevalent	16 25.34	indigenous	371 47.92	yob	106 50.73
embrace	178 37.76	fascinate	24 24.0	such	135 24.91	different	3580 47.7	long-hours	36 50.45
understand	498 36.44	tend	62 23.24	first-hand	9 21.36	visual	905 47.59	compensation	425 48.44
develop	973 36.02	emerge	66 22.61	alien	13 21.1	entrepreneurial	222 46.8	cell	976 46.11
rave	58 35.96	thrive	25 22.52	conducive	9 20.98	American	1201 46.51	postmodern	98 45.3
embed	119 35.75	evolve	42 21.68	dependent	19 20.37	diverse	669 46.11	material	1451 43.26
reflect	361 35.29	permeate	17 21.61	rich	26 20.2	western	499 45.29	enterprise	368 41.27
celebrate	231 35.21	shape	46 20.9	rife	8 19.87	dominant	366 44.55	dependency	141 40.62
strengthen	187 34.87	exist	90 20.73	important	61 19.83	Japanese	462 42.35	consumer	449 40.26
pervade	53 34.21	inspire	46 20.21	positive	28 18.93	Chinese	499 41.98	anti-fraud	31 38.08
preserve	157 32.45	seem	114 19.74	superior	13 18.59	colorful	101 40.95	□compensation	15 37.61
prevail	73 32.32	worship	14 18.42	negative	18 18.11	vibrant	266 40.29	celebrity	161 37.13
respect	116 31.34	dominate	38 18.34	accessible	19 17.65	Jewish	450 40.16	post-modern	42 36.87
permeate	44 31.25	become	220 18.3	diverse	16 17.54	ancient	553 39.84	drinking	171 36.74
learn	492 31.09	survive	41 18.24	hostile	9 17.34	corporate	583 38.53	consumerist	28 36.07
shape	134 31.06	mean	96 18.08	central	23 16.73	modern	898 38.45	hip-hop	60 35.55
nurture	64 30.71	underpin	26 18.04	essential	23 16.5	organizational	135 38.41	vitro	60 35.53

As this sample from the word sketch for *culture* reveals at a glance, *culture* occurs 161,537 times in the 1.5 bn word web corpus of English ukWaC. In this corpus *culture* shows a clear tendency to occur as object of such verbs as *foster*, *promote*, *experience*, *create*, *change*, and each collocate is the gateway for potentially endless exploration of concordance lines and phraseological patterns which provide insight not only into questions of language use but also into discourse and society (see also Gatto 2010; Gatto forthcoming).

It is of course not the purpose of the present study to explain in detail how the Sketch Engine works but it is perhaps useful to focus on one fundamental opportunity offered by the tool, which is definitely related to its connection with web search: the user can at any moment go back to the original text, by clicking on the doc.id left of the concordance line, and find out more about the real com-

municative situation in which the lexical item, collocation or pattern under analysis was used. Of course this opportunity is seriously impaired by the volatility of the webscape, because some pages may have been removed or changed. Nonetheless whenever this shift back to the real life communicative event does take place, the linguist can experience one of most significant changes occurring in corpus linguistics under the impact of the web: by offering the linguist such a dynamic, flexible, corpus of living texts, these recently developed resources and tools make us experience to the full a shift from corpus-as-body to corpus-as-web.

In this context particular attention deserve also a number of experiments which have been using parts of the web as a corpus both to gain information concerning language use, and as a source of data for the interpretation of cultural dynamics. This is the case for instance of online services that exploit the huge amount of printed texts that make up the database of Google Books, which is treated as a sort of diachronic corpus. The new Google Books Ngram Viewer interface allows to search 500 billion words of text to see changes in frequency of words and phrases, by turning a simple query into a complex task in which time-span, case-sensitivenss and other detailed criteria contribute to eliciting quick answers to otherwise complex issues. Just consider, by way of example, the data obtained in a matter of seconds about the use of capitalized *Hopefully*, possibly as sentence initial, by simply querying the Books Ngram Viewer:

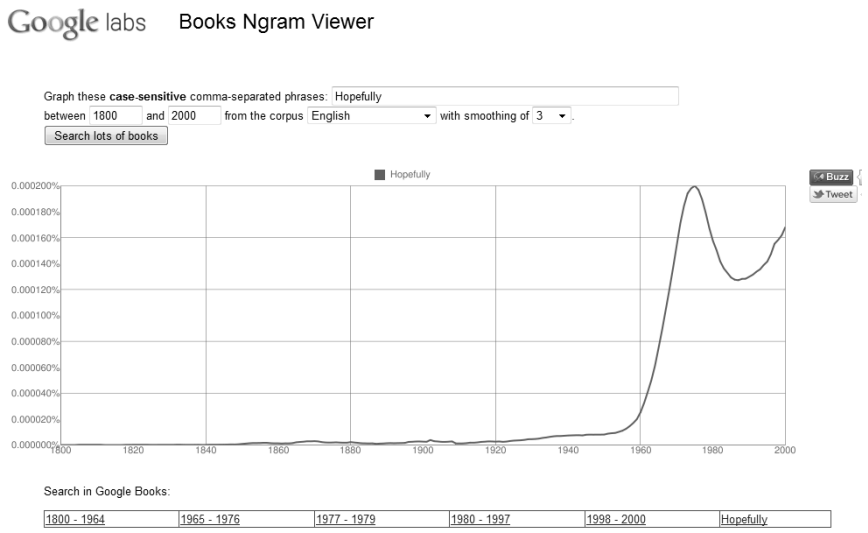


Figure 1: A Books Ngram Viewer query: Hopefully

The graph hardly needs an explanation, revealing as it is, at a glance, of the behavior of capitalized *Hopefully*. Obviously, the relevance and reliability of these data would require careful controls before being used as evidence for scientific purposes, since – as the authors themselves suggest – the greatest challenge lies in their interpretation. Which books do actually make up the database? Which constraints have determined its composition? Which varieties of English are more prominently represented? These are all issues which pertain to the web as corpus as a whole and suggest that all research in this field requires some form of post-hoc evaluation of the results. Nonetheless it cannot be denied that the contribution that a resource like this could provide in the context of linguistic research, could not have been even imagined without such a fecund and long awaited for interaction between corpus linguistics and the web.

5 Conclusion

The attempt at surveying the key issues that have accompanied the 10-year history of the web as a linguistic corpus highlights some characteristics such as constant change, non-finite size, anarchism, which clearly point to some radically new issues if the hypothesis of treating the web as a corpus is to be pursued on sound methodological bases. It is worth stressing, however, that some of these new issues are to some extent to be considered as not specifically related to the web as corpus but rather as a natural consequence of the impact of the new technologies on linguistic resources as a whole. Some of these issues can in fact be related to the changes envisaged by Wynne (2002: 1204) as likely to occur in the way we conceive of language resources in the 21st century: multilinguality and multimodality, dynamic content, distributed architecture, connection with web searching. While it is clear that a corpus is by no means the same as a text archive, for which Wynne envisaged the above mentioned changes, these new characteristics of language resources are clearly linked to the shift from real to virtual and with the emergence of the Web as a key phenomenon in contemporary society, and thus inevitably relating also to the web as corpus. More specifically, Wynne's idea of an inescapable shift towards virtual corpora is enlightening. The old scenario of the researcher "who downloads the corpus to his machine, installs a program to analyse it, then tweaks the program and/or the corpus mark-up to get the program and the corpus to work together, and finally performs the analysis" (Wynne 2002: 1205) now coexists with, and is possibly going to be replaced by, a new model where replicating digital data in a local copy and installing the software to analyse the data becomes redundant, as all the processing can be done over the network.

These emerging issues seem to affect the very notion of corpus in radical ways, prompting a shift away from the somewhat 'reassuring' conventional features subsumed by the corpus-as-body metaphor itself, to a new corpus-as-web metaphor. While the notion of linguistic corpus as a body of texts rests on some correlate issues such as finite size, balance, part-whole relationship, stability, the very idea of a web of texts brings about notions of non-finiteness, flexibility, decentering and re-centering, and provisionality. This calls into question, on methodological grounds, issues which could be instead taken for granted when working on conventional corpora, such as the stability of the data, the reproducibility of the research, and the reliability of the results. What seem to be also changing are notions of permanence/stability for corpora. In the traditional model the value and the reusability of a corpus are dependent on a bundle of factors, such as the validity of the design criteria, the quality and availability of the documen-

tation, the quality of the metadata and the validity and generalisability of the research goals of the corpus creator (Wynne 2002: 1205). However the relative importance of such criteria may change as the norm could become the creation of customized corpora on an *ad hoc* basis by simply choosing from within larger existing text archives, or the creation of smaller short-life specialized corpora from the web through specific tools. This seems to suggest that the "changing face of corpus linguistics" (cf. the title of Renouf and Kehoe 2006) can be really seen as the outcome of a wider process of redefinition in terms of flexibility, multiplicity, and mass-customization which corpus linguistics is undergoing along with other fields of human activity, in a sort of "convergence" (Wynne 2002: 1207) of technologies and standards in several related fields whenever the common goal of sharing, distributing and querying linguistic content through electronic means is at stake.

Notes

1. For a more specific definition of 'precision' and 'recall' from the perspective of Information Retrieval, see Van Rijsbergen (1979).
2. When not otherwise stated, results are from web searches carried out in November 2010.
3. Some of these aspects are dealt with also in Gatto (2009).

References

Alpert, Jesse and Nissan Hajaj. 2005. We knew the web was big… Available at http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.

Baroni, Marco and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, 1313–1316. Lisbon: Elda. Available at http://sslmit.unibo.it/~baroni/publications/lrec2004/bootcat_lrec_2004.pdf.

Baroni, Marco and Silvia Bernardini (eds.). 2006. *Wacky! Working papers on the web as corpus.* Bologna: Gedit.

Baroni, Marco and Motoko Ueyama. 2006. Building general- and special-purpose corpora by web crawling. In *Language corpora: Their compilation and application (Proceedings of the 13th NIJL International Symposium, Tokyo)*, 31–40. Available at http://clic.cimec.unitn.it/marco/publications/bu_wac_kokken_formatted.pdf.

Battelle, John. 2005. *The search: How Google and its rivals rewrote the rules of business and transformed our culture*. London: Nicholas Brealey.

Biber, Douglas and Jerry Kurjian. 2007. Towards a taxonomy of web registers and text types: A multi-dimensional analysis. In M. Hundt, N. Nesselhauf and C. Biewer (eds.). *Corpus linguistics and the web*, 109–131. Amsterdam and New York: Rodopi.

Crystal, David. 2006. *Language and the internet*. 2nd edition. Cambridge: Cambridge University Press.

De Schryver, Gilles-Maurice. 2002. Web for/as corpus: A perspective for the African languages. *Nordic Journal of African Studies* 11 (2): 266–282. Available at http://tshwanedje.com/publications/webtocorpus.pdf.

Fletcher, William H. 2004. Facilitating the compilation and the dissemination of ad-hoc web corpora. In G. Aston, S. Bernardini and D. Stewart (eds.). *Corpora and language learners*, 271–300. Amsterdam: Benjamins.

Fletcher, William H. 2007. Concordancing the web. Promise and problems, tools and techniques. In M. Hundt, N. Nesselhauf and C. Biewer (eds.). *Corpus linguistics and the web*, 25–45. Amsterdam and New York: Rodopi.

Fletcher, William H. Forthcoming. Corpus analysis of the World Wide Web. In C.A. Chapelle (ed.). *Encyclopedia of applied linguistics*. Oxford, UK: Wiley-Blackwell.

Gatto, Maristella. 2009. *From 'body' to 'web'. An introduction to the web as corpus*. Roma-Bari: Laterza Universitypressonline.

Gatto, Maristella. 2010. From language to culture and beyond. Building and exploring comparable web corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, Malta, 22nd May 2010. Available at:
http://www.fb06.uni-mainz.de/lk/bucc2010/documents/Proceedings-BUCC-2010.pdf

Gatto, Maristella. In press. Sketches of culture from the web. A preliminary study. In *Proceedings of the 24th AIA Conference*, Rome, 1–3 October 2009.

Ghani, Rayid, Rosie Jones and Dunja Mladenic. 2001. Mining the web to create minority language corpora. In *Proceedings of the 10th International Conference on Information and Knowledge Management*, Atlanta, 5–10 November 2001, 279–286. Available at
http://www.accenture.com/SiteCollectionDocuments/PDF/4.pdf.

Grefenstette, Gregory and Julien Nioche. 2000. Estimation of English and non-English language use on the WWW. In *Proceedings of the RIAO (Recherche d'Informations Assistée par Ordinateur),* Paris, 12–14 April 2000, 237–246. Available at h ttp://arxiv.org/ftp/cs/papers/0006/0006032.pdf.

Henziger, Monika and Steve Lawrence. 2004. Extracting knowledge from the World Wide Web. In *Proceedings of the National Academy of Sciences* 101: 5186–5191. Available at http://www.pnas.org/content/101/suppl.1/5186.full.

Hundt, Marianne, Nadja Nesselhauf and Carolin Biewer (eds.). 2007. *Corpus linguistics and the web*. Amsterdam and New York: Rodopi.

Kilgarriff, Adam. 2001. Web as corpus. In P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds.). *Proceedings of the Corpus Linguistics 2001 Conference*, UCREL, Lancaster, 342–344. Available at http://ucrel.lancs.ac.uk/publications/CL2003/CL2001%20conference/papers/kilgarri.pdf.

Kilgarriff, Adam. 2007. Googleology is bad science. *Computational Linguistics* 33 (1): 147–151. Available at http://www.kilgarriff.co.uk/Publications/2007-K-CL-Googleology.pdf.

Kilgarriff, Adam and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics* 29 (3): 333–347. Available at http://acl.ldc.upenn.edu/J/J03/J03-3001.pdf.

Kwasnik, Barbara H., Kevin Crowston, Michael Nilan and Dmitri Roussinov. 2001. Identifying document genre to improve Web search effectiveness. *Bulletin of the American Society for Information Science & Technology* 27 (2): 23–26. Available at http://www.asis.org/Bulletin/Dec-01/kwasnikartic.html.

Leech, Geoffrey. 2007. New resources or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf and C. Biewer (eds.). *Corpus linguistics and the web*, 133–150. Amsterdam and New York: Rodopi.

Lüdeling, Anke, Stefan Evert and Marco Baroni. 2007. Using web data for linguistic purposes. In M. Hundt, N. Nesselhauf and C. Biewer (eds.). *Corpus linguistics and the web*, 7–24. Amsterdam and New York: Rodopi.

Mair, Christian. 2007. Change and variation in present-day English: Integrating the analysis of closed corpora and web-based monitoring. In M. Hundt, N. Nesselhauf and C. Biewer (eds.). *Corpus linguistics and the web*, 233–248. Amsterdam and New York: Rodopi.

McEnery, Tony and Andrew Wilson. 2006. *Corpus linguistics*. 2nd edition. Edinburgh: Edinburgh University Press.

Mehler, Alexander, Serge Sharoff and Marina Santini (eds.). 2010. *Genres on the web. Computational models and empirical studies*. London and New York: Springer.

Nakov, Preslav and Marti Hearst. 2005. A study of using search engine page hits as a proxy for n-gram frequencies. In *Proceedings of RANLP '05*. Available at http://biotext.berkeley.edu/papers/nakov_ranlp2005.pdf.

Official Google Research Blog. 2006. All our n-gram are belong to you. Available at http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

Renouf, Antoinette and Andrew Kehoe (eds.). 2006. *The changing face of corpus linguistics*. Amsterdam and New York: Rodopi.

Risvik, Knut Magne and Rolf Michelsen. 2002. Search engines and web dynamics. *Computer Networks* 39: 289–302. Available at http://www.idi.ntnu.no/~algkon/generelt/se-dynamicweb1.pdf.

Santini, Marina. 2005. Web pages, text types and linguistic features: Some issues. *ICAME Journal* 30: 67–86. Available at http://icame.uib.no/ij30/ij30-page67-86.pdf.

Scannell, Kevin P. 2007. The Crúbadán Project: Corpus building for under-resourced languages. In C. Fairon, H. Naets, A. Kilgarriff and G-M. de Schryver (eds.). *Building and exploring web corpora (WAC3–2007)*, 5–16. Louvain-la-Neuve: Presses Universitaires de Louvain. Available at http://borel.slu.edu/pub/wac3.pdf.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sullivan, Danny. 2005. Search engine sizes. Available at Search Engine Watch, http://searchenginewatch.com/showPage.html?page=2156481.

Top Ten Languages – Internet World Statistics (June 2010). Available at http://www.internetworldstats.com/stats7.htm.

Van Rijsbergen C.J. 1979. *Information retrieval*. Available at http://www.dcs.gla.ac.uk/Keith/Preface.html.

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

Véronis, Jean. 2005. Web: Google's missing pages: Mystery solved? *Technologies du Langage*. Available at http://blog.veronis.fr/2005/02/web-googles-missing-pages-mystery.html.

Volk, Martin. 2002. Using the web as corpus for linguistic research. In R. Pajusalu and T. Hennoste (eds.). *Tähendusepüüdja. Catcher of the meaning. A festschrift for Professor Haldur Õim*. Tartu: University of Tartu. Available at http://www.ifi.unizh.ch/cl/volk/papers/Oim_Festschrift_2002.pdf.

Wynne, Martin. 2002. The language resource archive of the 21st century. In M. González Rodríguez and C. Paz Suárez Araujo (eds.). *Proceedings of the Third International Conference on Language Resources and Evaluation,* 1204–1208. Available at https://nats-www.informatik.uni-hamburg.de/intern/proceedings/2002/LREC/pdf/271.pdf.

Zanettin, Federico. 2002. DIY corpora: The WWW and the translator. In B. Maia, J. Haller, and M. Ulrych (eds.). *Training the language services provider for the new millennium*, 239–248. Porto: Faculdade de Letras, Universidade do Porto. Available at http://sites.google.com/site/federicozanettinnet/dbpublications.

Zuraw, Kie. 2006. Using the web as a phonological corpus: A case study from Tagalog. In A. Kilgarriff and M. Baroni (eds.). *Proceedings of the 2nd International Workshop on Web as Corpus (EACL06)*, 59–66. Available at http://www.aclweb.org/anthology/W/W06/W06-1709.pdf.

Websites

BootCaT, bootcat.sslmit.unibo.it/

Google Books N-Gram Viewer, http://ngrams.googlelabs.com/

Internet World Stats, http://www.internetworldstats.com

Official Google Research Blog, http://googleresearch.blogspot.com

Sketch Engine, Sketchengine, http://sketchengine.co.uk

WaCky Project, wacky.sslmit.unibo.it/doku.php

WebCorp, http://www.webcorp.org.uk

KWiCFinder, www.kwicfinder.com