University Press on-line

Dipartimento di Studi Anglo-Germanici e dell'Europa Orientale
Università di Bari

Prima edizione 2008

Questo volume beneficia
di un contributo del CdA
dell'Università degli Studi di Bari
e di un contributo a carico dei fondi
di Ateneo (60%) dell'Università di Bari.

Università degli Studi di Bari
www.uniba.it
Editori Laterza
www.laterza.it
University Press on-line
www.universitypressonline.it

Maristella Gatto

From Body to Web

An Introduction to the Web as Corpus

Università di Bari • Editori Laterza

University Press on-line

Finito di stampare nel novembre 2008
Global Print srl - via degli Abeti, 17/1
20064 Gorgonzola (MI)
per conto della
Gius. Laterza & Figli Spa
ISBN 978-88-420-8854-7

Contents

To
those who sent me on my way,
those who let me go on my way,
and those who asked me
where I was going and why.

(from Michael Hedges, *Breakfast in the Field*)

Introduction

It is perhaps no exaggeration to maintain that the rise of corpus linguistics in recent years has brought about a revolution in the study of language whose impact is still to be fully acknowledged especially in everyday work with languages. By changing the «unit of currency» (Tognini Bonelli 2001: 1) of linguistic investigation, corpus linguistics has proved not only an invaluable way of exploring language structures and use, but it has also opened up new perspectives in language teaching, in the study of LSP, and in translation studies. More recently, increasing interest in the text/discourse dimension of corpus studies is showing the contribution that corpus research can make to the study of literary texts or to understand the relationship between discourse and society (Hoey et al. 2007).

While the contribution brought by corpus linguistics to «a qualitative change in our understanding of language» (Halliday 1993: 24) cannot be underestimated, it is still a matter of debate whether corpus linguistics should be regarded primarily as a method that can be applied in a variety of fields, or as a theory, «because it is in a position to contribute specifically to other applications» (Tognini Bonelli 2001: 1). To account for the special status of corpus linguistics as a methodology which is nonetheless «in a position to define its own set of rules and pieces of knowledge *before* they are applied», Tognini Bonelli devised the notion of «pre-application methodology» (2001: 3), thus paving the way for subsequent interest in the theoretical implications of corpus studies. Such focus on the theoretical consequences of corpus findings, which seem to hint at a possible theoretical status of the discipline as a whole, has led to the notion of «corpus linguistics

as a theoretical approach to the study of language» (Teubert 2005: 2), which means that «corpus linguistics is not merely seen as a methodology but as an approach (...) with its own theoretical framework» (Mahlberg 2005: 2).

For the purpose of the present study, such a view of corpus linguistics as an approach, indeed «a new philosophical approach» (Leech 1992: 106), seems to be the most fruitful to explore the complex relation between the theoretical implications of corpus linguistics, and the emergence of new methods to profit from the web's immense potential as a linguistic corpus. An approach, as the very word suggests, is «the act of drawing near» (*Oxford English Dictionary*), and corpus linguistics as an approach has undoubtedly brought us closer to what have been termed «pluralist» language models, which overcome the dualisms and monisms of the past and point to a new dimension where repeated language behaviour seems to bridge the gap between the brute facts of individual parole/performance/instance and the abstraction of langue/competence/system (Stubbs 2002: 238-242; Stubbs 2007: 127ff). Indeed, by allowing us to look at language from a new point of view, the corpus linguistics approach, does not only provide «new and surprising facts about language use», but also – and more crucially – «can help to solve paradoxes which have plagued linguistics for at least one hundred years (...) and therefore help to restate the conceptual foundations of our discipline» (Stubbs 2007: 127).

An approach needs, however, to be implemented through methods, where a method is seen as a specific way (from the Greek –οδοσ), a set of steps, a procedure for doing something. Thus, in a Janus-faced position between theory and method, the corpus linguistic approach seems to stand in an osmotic relation between a theory, i.e. a «contemplation» (*Oxford English Dictionary*), of the nature and reality of language, and several methods (i.e. ways) to explore that reality. With this in mind, one can see how then corpus linguistic approach can profit from the web as a pervasive medium. Corpus linguistics has increased our awareness of the fact that repetitions across the language behaviour of many speakers are a significant fact for the study of language, and one which can be fruitfully. As a social phenomenon whose currency is language, the web makes such repetitions immediately evident and readily available to any user, providing the linguist with countless instances of repeated social and shared linguistic behaviour. This vast amount of data only requires that appropriate methods are devised to exploit its significance from a corpus linguistic perspective.

It is against this background that the present study aims to deal with the emergence of a research field labelled «web as corpus». By no means itself a new theory or approach, web as corpus is best seen as an umbrella term for a number of methods that look at the web as their main resource to implement the corpus linguistic approach. The methods devised to exploit the web's potential from a corpus linguistics perspective must not be seen therefore as competing with other more traditional ways for corpus work, but as established a useful complement to more practices in corpus linguistics. Thus, while the notion of the web as corpus apparently questions fundamental issues within corpus linguistics, it in fact contributes to the growth of the research field as a whole, indirectly contributing to reshaping our view of language.

Nonetheless some of the questions raised by the very idea of considering the web as a corpus by virtue of its very nature as a 'body' of texts seem to deserve further investigation on both theoretical and applicative grounds, and this is what the present work aims to do. Given such a twofold focus the present work is mainly intended for people who have an interest using the web as a corpus, but also interested in the theoretical implications which the very idea of considering the web as a corpus inevitably raises. This suggests a wide audience including researchers or students of languages and linguistics, as well as language professionals of any kind, especially language teachers and translators.

The twofold focus is also reflected in the structure of the book. A first chapter exploring the theoretical implications of the emerging notion of the web as corpus is followed by four chapters discussing different methods to exploit the web's potential from a corpus linguistic perspective. The basic assumption is that while apparently only committed to the practical task of devising appropriate methods and tools, research carried out under the label web as corpus may have been contributing to reshaping the way we conceive of corpus linguistics as a whole in the new Millennium. **Chapter 1** revisits some key issues in corpus linguistics

such as «authenticity», «representativeness», «size» and «content» in the light of the web as a «spontaneous», «self-generating» collection of texts. The chapter also explores the new issues such as «dynamism», «reproducibility», «relevance and reliability» that the notion of the web as corpus possibly raises. **Chapter 2** focuses on search methods as an issue of major concern when the web is considered as a corpus in its own right with particular reference to the different roles played by the linguist in empirical inductive research based on corpora and on uses of the web for linguistic reference through ordinary web search engines are exemplified. **Chapter 3** and **Chapter 4** introduce tools devised to exploit the web's potential from a corpus linguistics perspective showing how different ways of using the web as a corpus do not only help to overcome the obvious limitations of the web as a linguistic resource but also provide linguistic information that is particularly appropriate in specific contexts and for specific tasks. Finally particular attention is given to the creation of several general reference mega-corpora from the web for such languages as Italian, German, English, Spanish, Chinese and Russian, in **Chapter 5**. These corpora fall into 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.

Without the aim of providing an exhaustive list, the methods and tools discussed in the applicative sections of this work were selected so as to exemplify different ways of using the web as a corpus, drawing on the possible meanings of the umbrella term web as/for corpus suggested by Baroni and Bernardini (2006: 13ff). The steps thus taken chart a process of decreasing dependence on the typical gateway to information on the web (i.e. ordinary search engines), from the web as corpus surrogate, to the web as corpus shop to the mega-corpus mini-web. This final step signifies a Copernican revolution in our way of conceiving of corpora, corpus tools and methods for corpus work under the impact the of web, with the notion of *web as corpus* apparently giving way to the new horizons of the *corpus as web*.

Acknowledgments

It is perhaps not uncommon, at the end of a demanding task, to feel overwhelmed not only by tiredness, but also by feelings of gratitude for those who have helped and supported you.

First of all, I wish to express my gratitude to Prof. Annamaria Sportelli for believing in this project from the start and encouraging me throughout. I would also like to thank prof. Vittoria Intonti, prof. Domenico Torretta and prof. Maristella Trulli, for generously contributing in different ways to the realization of this book.

Special thanks go to friends and colleagues for reading and commenting on earlier versions or parts of the book: Michaela Mahlberg, Sara Laviosa, and Federico Zanettin. To their friendship I owe as much as to their competence. Needless to say, whatever faults remain are entirely my own.

A word of thanks is also due to Gaetano Falco for endless discussions on controversial issues, to Sedley Proctor for language revision, and to Pierpaolo Basile for reading parts of this work from the point of view of information technology.

Finally, I wish to thank my family, to whose love I owe all the energy I spent for this work: my husband Antonello, with my children Francesco, Serena and little Stefano, who was born on the 7th June this year, thus providing me with the most convincing reason for bringing this work to an end.

From Body to Web

An Introduction to the Web as Corpus

Chapter I

Corpus Linguistics And The Web: Old and New Issues

Introduction

This chapter revises some key issues in corpus linguistics in the light of the properties of the web as a spontaneous, self-generating collection of texts, and explores some of the new issues which the emerging notion of the web as corpus seems to raise. The basic assumption is that the challenge of using the World Wide Web in corpus linguistics does not aim to push key questions onto the background, but rather «serves as a magnifying glass for the methodological issues that corpus linguists have discussed all along» (Hundt *et al.* 2007: 4). Section 1 and Section 2 discuss the relationship between corpus linguistics and the web as a 'body' of texts. Section 3 revisits authenticity, representativeness, size and content, in order to envisage the «changing face» of corpus linguistics under the impact of the World Wide Web, while Section 4 explores some new issues such as dynamism, reproducibility, relevance and reliability.

1. *Corpus linguistics and the web*

In the opening page of a recent volume on *Corpus Linguistics and the Web*, the editors wonder «why should anyone want to use other than carefully compiled corpora?» (Hundt *et al.* 2007: 1). This is indeed a legitimate question which is undoubtedly relevant to the object of the present study. Why should one even consider taking the risk of using a database as anarchic and chaotic as the World Wide Web? Why should linguists, translators, language professionals of any kind, turn their attention to a collection of texts whose content is largely unknown and whose size is hardly

measured? As the editors of the above mentioned volume argue, there are a number of obvious answers to these questions. For some areas of investigation within corpus linguistics, such as lexical innovation or morphological productivity, or for investigations concerning ephemeral points in grammar, one might need a corpus larger than the mega-size corpora of the BNC type. Then there are language varieties for which carefully compiled corpora are not always viable, but which are extensively represented on the web. Finally, the technological development itself has resulted in the emergence of new textualities, such as chat rooms, blogs and home pages, along with the re-mediation of traditional genres, which call for new attention, since they seem to blur old distinctions between written/spoken, formal/informal registers and require new categories of approach. Last, but not least, there is the problem of updating, which requires that profit and loss are carefully balanced in the compilation of a conventional corpus, which obviously runs the risk of being already out of date when it is finished (Hundt *et al.* 2007: 1-2).

Starting from these considerations, it is self evident that the notion of the web as corpus is, in the first place, nourished by practical and opportunistic reasons, which have resulted in different meanings of the expression *Web as Corpus*, corresponding to different ways to exploit the web's potential from a corpus linguistics perspective[1]. In rather general terms, most uses of the web in corpus linguistics research can be summed up under the label «web as/for corpus» (De Schryver 2002; Fletcher 2007), depending on whether the web is accessed directly as a source of online language data or as a source for the compilation of offline corpora. More precisely, in their introduction to the collection of papers resulting from the *First International Workshop on the Web as Corpus* (Forlì, 14th January 2005), Baroni and Bernardini (2006: 10-14) focus on four basic ways of conceiving of the web as/for corpus:

1. The web as a corpus surrogate: researchers using the web as a corpus *surrogate* use the web for linguistic purposes either via a standard commercial search engine, mainly for opportunistic reasons (e.g. as a reference tool for translation tasks), or through linguist-oriented metasearch engines (e.g. WebCorp or KwiCFinder);

2. The web as a corpus shop: researchers using the web as a corpus *shop*, select and download texts retrieved by search engines to create «disposable corpora» either manually or in a semi-automatic way (e.g. using a toolkit which allows crawling and extensively downloading from the web such as BootCaT);

3. The web as corpus proper: researchers using the web as a corpus *proper* purport to investigate the nature of the web, and more specifically they look at the web as a corpus that represents web English;

4. The mega-Corpus mini-Web: the most radical way of understanding the web as corpus refers to attempts to create a new object (mini-Web/mega-Corpus) adapted to language research and combining Web-derived (large, up-to-date, web-based interface) and Corpus-like features (annotation, sophisticated queries, stability).

This overview shows how rich the relationship between corpus linguistics and the web can be and undoubtedly testifies to the liveliness of this exciting research field. And yet this is probably not the whole story. 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 using the web as a resource for linguistic research. It can be argued, indeed, that if the web has been considered as the corpus of the new Millennium (Kilgarriff 2001), this must also be due to qualitative considerations concerning the nature of the web itself, so that there may have been deeper reasons for turning it into an object of linguistic investigation. We stand no longer «at the brink of a new age», as Nelson foresaw over 25 years ago (Nelson 1981), but deeply immersed in it. And in this new age, it is perhaps the web that presents «the most provocative questions about the nature of language» (Kilgarriff 2001). Language is indeed «at the heart of the Internet» (Crystal 2006: 271) and as a «social fact», rather than simply a «technological fact», where «the chief stock-in-trade is language» (Crystal 2006: 271), the web may paradoxically have been brought to the attention of many linguists as the largest text collection in the world almost against their will. Hence the convergence between a social phenomenon existing independently from linguistic investigation (the web) and the corpus linguistics approach, where the web is seen as a huge amount of texts

in electronic format which both «tantalize and challenge linguists and other language professionals» (Fletcher 2007: 27).

With advent of the web in the new Millennium, the relationship between (corpus) linguistics and information technology seems thus to have entered an exciting stage which can be envisaged in terms of a role reversal between 'giving' and 'taking'. While web technologies in the past drew extensively on language research through computational linguistics and natural language processing, it seems that today the relationship has been reversed in a sort of «'homecoming' of web technologies, with the web now feeding one of the hands that fostered it». (Kilgarriff and Grefenstette 2003: 336-7). In this context, it is to be expected that such characteristics as multilinguality and multimediality, dynamic content and distributed architecture, seen as very likely to become standards for linguistic resources in the 21st century (Wynne 2002: 1204), and which are all clearly linked to the emergence of the web as a key phenomenon of our times, should affect the way we conceive of a linguistic corpus. Even more crucially, perhaps, such changes can be seen from a wider perspective as signifying a deeper «measure of convergence of technologies and standards in several related fields having in common the goal of delivering linguistic content through electronic means» (Wynne 2002: 1207), and possibly mirror also changes taking place in society at large under the impact of the new technologies.

It is against this background that the web's nature as a body of texts, which provides the material basis for its controversial and intriguing status as a corpus, deserves some further investigation from the perspective of corpus linguistics as a whole.

2. *The web as corpus: a «body» of texts?*

The very idea of exploring the possibilities of treating the web as a linguistic corpus presupposes a view of what a corpus *is*, and possibly entails a redefinition of what a corpus *can be*. If we start from the Latin etymology of the word, virtually any collection of more than one text can be called a corpus, but the term has acquired more specific connotations in modern linguistics than this simple definition implies (McEnery and Wilson 2006: 29). Indeed, even

though some branches of linguistics have always been to some extent *corpus-based* (i.e. based on the study of a number of authentic texts), and concepts such as corpus and concordance have been for many years the daily bread of scholars studying the Bible or Shakespeare's works (Kennedy 1998: 14), corpus linguistics as a distinct research field is a fairly recent phenomenon which rests on certain basic assumptions about what a corpus is, and also, perhaps more crucially for the purpose of the present study, of what a corpus is not. «A corpus», according to Sinclair's seminal definition «is a collection of naturally-occurring language chosen to characterize a state or variety of a language» (Sinclair 1991: 171). In modern linguistics this also entails such basic standards as machine readable format, finite size, sampling and representativeness (McEnery and Wilson 2006: 29). Thus, Francis observes, an anthology such as *The Oxford Book of English Verse* cannot be properly considered a corpus, neither can, despite its name, the *Corpus Iuris Civilis* instigated by the Emperor Justinian in the 6th century (1992: 17). And while it is doubtful whether a collection of proverbs can be considered a corpus in its own right (Tognini Bonelli 2001: 53), it may eventually be considered as such if it is the object of linguistic research carried out using corpus linguistics tools. The notion of «corpus-hood»[2] seems therefore to defy simple definitions based on the corpus-as-object alone and is best approached from a wider perspective including considerations on both form and purpose (Hunston 2002: 2), the latter being a fundamental criterion in discriminating between a linguistic corpus and other collections of language texts (Francis 1992: 17).

Nonetheless, owing to the growing popularity, or rather «remarkable renaissance» (McEnery and Wilson 2001: 1), of corpus linguistics in the last decades of the 20th century, there has been in recent years greater and greater pressure to identify explicit criteria for corpus creation and corpus investigation and to define «good practice» in corpus work. Thus, each new study in corpus linguistics rests – implicitly or explicitly – on a definition of corpus that is based on fundamental criteria and standards. There is obvious consensus that «authenticity» of language data and «electronic format» are the basic *sine qua non* of a corpus in the modern linguistics sense of the word, while differences may emerge concerning other aspects. Regardless, however, of the definition

chosen as a starting point, issues such as representativeness, size, sampling, balance, design and purpose always enter the debate at different levels whenever the notion of corpus is at stake. Accordingly, the idea of considering the World Wide Web as a ready-made corpus by virtue of its very nature as a collection of authentic texts in machine readable format is called into question by the rigorous standards of corpus design. It is not surprising therefore that in a publication aimed at clarifying and spreading the «good practice» of corpus work, Sinclair explicitly declares that «[a] corpus is a remarkable thing, not so much because it is a collection of language text, but because of the properties that it acquires if it is well-designed and carefully-constructed», consistently denying corpus dignity to the web:

> The World Wide Web is not a corpus, because its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective (Sinclair 2005).

Notwithstanding doubts concerning the hypothesis of using the web as a corpus, made explicit by one of the founding fathers of contemporary corpus linguistics, linguists from all over the world have been increasingly turning their attention to 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. The relationship between corpus linguistics and the web seems thus to have become all but marginal, to the extent that this could well be envisaged as a new stage in the penetration of the new technologies in corpus linguistics. According to Tognini Bonelli, the computer was, at first, only a tool for speeding up processes and systematising data; then it offered a methodological frame by providing evidence of patterns of regularity which would never have been noticed or could not have been elicited by mere introspection; finally, information technology immensely contributed to the creation of new corpora, simplifying the work of corpus builders and potentially turning corpus linguistics from an area of investigation for specialists only to a research field virtually open to all (Tognini Bonelli 2001: 43ff.). Today, the web itself seems to claim the right of being considered as a corpus by virtue of its very nature as a collection of machine

readable and searchable authentic texts, thus opening up new perspectives and offering new challenges.

Acknowledging the undisputable role that the web can play, and has actually been playing so far in corpus linguistics, and acknowledging its potential as a ready-made corpus can however by no means obliterate the difference between the textual *mare magnum* which constitutes the web and a corpus where texts are gathered «according to explicit design criteria, with a specific purpose in mind, and with a claim to represent larger chunks of language selected according to a specific typology and with a specific typology» (Tognini Bonelli 2001: 2). While taking for granted the qualitative difference between the web and a corpus designed and compiled as an object of language study, it seems nonetheless still possible to break this *impasse* by pointing out the fundamental difference between attempts at answering the «ontological» question relating to what a corpus is – which implicitly points to a «deontological» notion of what a corpus should be – and more empirical, yet legitimate, attempts to test the web's potential as a corpus by answering the practical question «Is corpus *x* good for task *y*?». As Kilgarriff and Grefenstette have argued in their influential 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 corpus-hood. 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 (2003: 334).

Going for a «broad definition of corpus» as «a collection of texts when considered as an object of language or literary study», implicitly shifting the notion of «corpus-hood» to the intention of the researcher rather than seeing it as intrinsic to the text collection

itself, Kilgarriff and Grefenstette have 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). 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» has apparently been limited only to the practical question of seeing whether, as Kilgarriff put it, a web corpus is «good» for a certain task, while in fact each new study in this controversial field has imperceptibly contributed to reshaping the way we conceive of a corpus in the new Millennium. As a result it is perhaps no longer simply a matter of highlighting the advantages and disadvantages of using the web as a corpus, but it may also be necessary to reinterpret some key issues in corpus linguistics in light of the specific properties of the web as a spontaneous, self-generating collection of texts, and to explore the new issues that the emerging notion of the web as corpus possibly raises.

Drawing on the most common definitions of a corpus in the literature produced over the last fifteen years it is clear that some issues have become of paramount importance in determining the nature of a corpus as an object of language study; and if we do not simply equate a corpus to «an helluva lot of texts», as Leech suggests (Leech 1992: 106), these issues have to be carefully considered when exploring the position of the web as a corpus and have to be re-addressed from a new perspective. With this in mind, it can be argued that the challenge of using the World Wide Web in corpus linguistics does not aim to push key questions onto the background, but rather «serves as a magnifying glass for the methodological issues that corpus linguists have discussed all along» (Hundt *et al.* 2007: 4). Indeed, no development in web-as-corpus studies can question the fundamental tenets of corpus linguistics, neither is it the aim of research focused on the web as corpus to replace traditional corpus linguistics. As any other field of human knowledge, linguistic research can only profit from the force field created in the tension between traditional theoretical positions and the new tools and methods developed to meet practical needs, between the gravitational pull of existing standards and the promises of the web. It is therefore more than desirable, and probable, that corpus linguistics and studies on the web as

corpus will happily coexist for a long time, providing the linguistic community with a wider spectrum of resources to choose from.

3. *The corpus and the web: key issues*

While the emerging notion of the web as corpus apparently questions some basic standards in corpus linguistics, it provides in fact an opportunity to further explore some of the theoretical and methodological issues on which the good practice of corpus work rests. Such issues can be profitably revisited from the perspective of the web in order to envisage, if possible, the «changing face» of corpus linguistics under the impact of the World Wide Web. The most relevant for the purpose of the present study are authenticity, representativeness, size and content and will be discussed in the following pages.

3.1. *Authenticity*

It is a basic assumption of corpus linguistics that all the language included in a corpus is authentic, and certainly the most prominent feature of the web to have attracted the linguists' attention is its undisputable nature as a reservoir of authentic «purposeful language behaviour»: a collection of authentic texts produced in authentic human interactions by people whose aim «is not to display their language competence, but rather to achieve some objective through language» (Baroni and Bernardini 2006: 9). 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 comprised of authentic texts, which are the result of genuine communicative events, produced by people going about their normal business. 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), namely the revival of attention paid by linguists to authentic language in use, which can be well considered as a resurgence in popularity of early 20th century, *ante-litteram*, corpus-based methodologies (such as those by American structuralism and field linguistics) regaining prominence also owing to the new possibilities offered by computer data storage (Mc Enery

and Wilson 2001: 2 ff). Thus the web-as-corpus issue, as a more or less legitimate offspring of corpus linguistics, is one of the many outcomes of the prominence gained by language study grounded in empiric data rather than in introspection; of language study interested in what Saussure called *parole* rather than in *langue*; of focus on the *instance* as the only way to get evidence of an otherwise invisible *system*; of language study more focussed on visible actual *performance* than on invisible potential *competence*. If then, to paraphrase Sinclair, it has now become fashionable to look outwards to society rather than inwards to the mind (Sinclair 1991: 1) 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. Furthermore, as it is often the case with research domains related to some form of technology, there may have been mutual influence between the exponential growth of easily accessible authentic data in electronic format caused by the digital revolution and preference for empiricism, so that it is probably not far from the truth to state that also the web itself as a repository of huge amounts of authentic language in electronic format freely available with little effort has contributed to making the corpus linguistics approach so popular and virtually accessible to all.

It could be argued, then, that the authenticity of the web really makes it «a fabulous linguist's playground» (Kilgarriff and Grefenstette 2003: 345), a hitherto unavailable open space to explore the interplay between what is formally possible in language, and actual linguistic behaviour. Its very unplanned inclusiveness makes it a place where data gained through introspection can be tested against the background of what has been actually performed. A simple Google search for phrases could easily demonstrate, for instance, how the web can make repetitions across the language behaviour of many speakers immediately visible, thus highlighting not only the repetitive and routine nature of language, but also providing evidence for what Teubert has recently defined as the «autopoietic» and inherently «diachronic» nature of the discourse (Teubert 2007)[3].

In this context, however, it is important to remember that while authenticity is the most obvious strength in the similarity between a corpus and the web, it is also one of the latter's major flaws, and the

reason for being cautious. Owing to its nature as an unplanned unsupervised unedited collection of texts, authenticity in the web is often related to problems of «authoritativeness». Everyday experience suggests that authentic in the web often means inaccurate (misspelt words, grammar mistakes, improper usage by non-native speakers), i.e. not reliable from the linguistic point of view. As a consequence it is of crucial importance that linguists purporting to look at the web form a corpus linguistics perspective become familiar with some of its basic features so that they can profit from its potential without running the risk of being tangled in the web itself.

3.2. *Representativeness*

Closely related to authenticity is the «vexed question» (Tognini Bonelli 2001: 57) of representativeness, which can be considered – on the basis of the most widely accepted definitions of a corpus – as being perhaps even more important than authenticity in corpus design. Reference to some of the most frequently quoted definitions such as those by Francis, Biber or McEnery and Wilson, shows that representativeness is almost invariably mentioned as *the* key issue. According to Francis, for instance, a corpus is «a collection of texts assumed to be *representative* of a given language, dialect, or other subset of a language, to be used for linguistic analysis» (Francis 1992: 17, my emphasis), whereas Biber *et al.* state that:

> A corpus is not simply a collection of texts. Rather *a corpus seeks to represent* a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to *represent*. The *representativeness* of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalizability of the results of the research (Biber *et al.* 1998: 246, my emphasis).

Similarly, McEnery and Wilson see sampling and representativeness as the goal to which the four main characteristics of the modern corpus are subordinate:

> A corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a finite-sized body of machine-readable text, sampled *in order to be maximally representative* of the language variety under consideration (McEnery and Wilson 2001: 32, my emphasis).

A crucial issue in determining the value of a corpus as an object of linguistic study, the notion of representativeness concerns «the kind of texts included, the number of texts, the selection of particular texts, the selection of text sample from within texts, and the length of text samples» (Biber 1992: 174). It also entails careful considerations concerning the users of the language which a corpus aims to represent, a task which Sinclair aptly considered «hardly a job for linguists at all, but more appropriate for the sociology of culture» (Sinclair 1991: 13). More recently, Sinclair himself has stressed again the same point by stating that «the contents of a corpus should be selected [...] according to their communicative function in the community within which they arise». Putting it more simply, he has voiced these concerns through the following questions:

> What sort of documents do they write and read, and what sort of spoken encounters do they have? How can we allow for the relative popularity of some publications over others, and the difference in attention given to different publications? How do we allow for the unavoidable influence of practicalities such as the relative ease of acquiring public printed language, e-mails and web pages as compared with the labour and expense of recording and transcribing private conversations or acquiring and keying personal handwritten correspondence? How do we identify the instances of language that are influential as models for the population, and therefore might be weighted more heavily than the rest?» (Sinclair 2005).

Such detailing of criteria for representativeness seems however to have paradoxically also fostered greater awareness of the utopian and almost mystical nature of representativeness in corpus design, which is reflected in metaphors such as Kilgarriff's «Pandora's box» (2003: 333) or Leech's most recent «Holy Grail» (2007: 134-136). Undoubtedly a thorny issue, if it is not to be regarded, «as an act of faith» (Leech 1991: 27), representativeness must at least be seen as a work-in-progress. As Biber's concept of «cyclical fashion» (1993: 243) suggests, linguists should become aware of its «scalar» nature, i.e. of the possibility of reaching, and the necessity of aiming at, a certain degree of representativeness of data, even when it is clear that absolute representativeness is definitely out of reach (Leech 2007: 140).

As in the case of authenticity, the issue of representativeness in corpus linguistics is not devoid of implications in terms of linguistic theory as a whole. As Tognini Bonelli argues, «representativeness is the natural correlate of the language model upheld by Firth, Halliday and Sinclair» (Tognini Bonelli 2001: 57). According to these scholars, repeated events noticed in language samples at the level of individual performance are an essential element in the formulation of generalization about language, relating to what could be defined in Saussure's words as *langue* or in Chomsky's words as I-language. This is what makes representativeness so crucial. It is on the representativeness of the corpus that the value of generalizations about language informed by the corpus linguistic approach ultimately rest. This is also why early criticism of corpus linguistics by Chomsky was focused precisely on the issue of representativeness. The typical charge against corpus linguistics was precisely that, however large a corpus, the data would only be a small sample of a potentially infinite population, and that any corpus would be intrinsically unrepresentative and «skewed»:

> Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list (Chomsky 1962: 159; quoted in Aijmer and Altenberg 2004: 8).

While Chomsky's much quoted criticism of early corpus linguistics – which might be equally applied to any other type of scientific investigation based on sampling – was to some extent valid criticism, it can also be considered as a 'child of its times', and therefore 'more true' at the time of early corpora than today. Where Chomsky sees irreducible qualitative difference between limited performance and potentially unlimited competence, so that performance could never yield significant insight into competence, corpus linguistics has made us see how the two things are rather found on a continuum, with E-language representing «a crucial, indispensable manifestation of I-language» (Leech 2007: 135), and I-language made somehow discernible as a summation of E-language events tending to infinity. As Halliday has persuasively argued through his powerful weather/climate metaphor,

the relationship between «visible» instances of language use and the «invisible» system can be compared to the relationship between the weather, relating to the next few hours or days, and the climate as the summation of each day's weather (Halliday 1991: 41-42). With the possibility of handling many millions of words (or weather reports, to keep Halliday's metaphor), modern computerised corpora have greater possibilities than in the past of allowing insight into the language system (the climate), or at least into the social, public, shared features of that system, as pluralist positions such as those recently put forward by Stubbs suggest (Stubbs 2002; Stubbs 2007).

As to how these concerns are related to the web as corpus, this remains nonetheless a difficult question. Certainly the standard of representativeness is the one that most puzzles those who claim corpus dignity for the web. As Leech has recently argued, it seems that «the web as corpus makes the notion of a representative corpus redundant» (Leech 2007: 144), and indeed early research in the field, mainly within computational linguistics, seemed to dispense altogether with the notion of representativeness on the grounds that the researcher's effort could be more profitably devoted to the solution of practical problems (Kilgarriff and Grefenstette 2003: 343).

Exploring the web's potential for representativeness remains therefore the most crucial concern for scholar interested in investigating the value of the web as a corpus-like collection of texts. The real problem is that the notion of representativeness is the issue more closely bound up with the organic metaphor of the corpus-as-body, based as it is on the assumption that each part of a *body* can be representative of the whole. 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, its imbalances and potential unrepresentativeness impair its value as a language resource. While it is true that the web gives access to a wide range of genres, some of which are undeniably well-established in the written medium, such as academic writing, and others are newly evolving and closer to speech, such as blogs, it is also true that it gives little access to private discourse, such as everyday conversation, telephone dialogues, and the like (Leech 2007: 144-5). Furthermore, there are

other major areas of language use which are underrepresented, while certain varieties are definitely overrepresented. As a consequence, even if the web's textual universe, at least as far as the English language is concerned, largely overlaps with the non-electronic textual universe of the English language, its status as a «representative sample» is in Leech's opinion non-existent:

> It is a textual universe of unfathomed extent and variety, but it can in no way be considered a representative sample of language use in general (Leech 2007: 145).

As Leech's most recent contribution on the issue shows, it is clear that the notion of representativeness as it is generally conceived of in corpus linguistics can only pertain to corpora which have been designed and created out of selection from carefully chosen material. This is not the case with the web which already exists, independently of 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 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 actually represents – the international community at large in real time. In some way it could be argued, as recent research has, that the web 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. The increasing prominence of the web in contemporary culture, the very fact that it is «directly jacked into the culture's nervous system» (Battelle 2005: 2), along with its evident ability to mirror changes in real time thanks to its intrinsic dynamism, seems on the contrary to mitigate the problems arising from lack of representativeness in

the corpus linguistics sense of the word. Certainly the web cannot be considered a representative sample of language use in general, but its scope, variety, and above all its immense size seem to legitimize the opinion 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.

3.3. *Size*

Intrinsically related to representativeness, the issue of size is equally fundamental in determining the value of a corpus as an object of language study and affects the kind of generalizations that can be made out of corpus data.

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, if it is to be considered as a source of data for quantitative studies. As McEnery and Wilson suggest, the notion of corpus should by default imply «a body of text of a finite size» (2006: 30), whereas the web is by its very nature bound to perpetual expansion. This may have gained the web a reputation as «the ultimate monitor corpus» (Bergh 2005: 26), i.e. a corpus which – according to Sinclair's definition – «has no final extent because, like language itself, it keeps on developing» (Sinclair 1991: 25). As such, a monitor corpus can be a more opportunistic and less balanced corpus, one where «quantity of data replaces planning of sampling as the main compilation criterion» (Kennedy 1998: 61).

The main problems with the web as a corpus are thus related precisely to its being non-finite. The impossibility of estimating its size exactly results in uncertainties and doubts concerning its value as an object of scientific study, which in turn reflect more general anxiety about the impossibility of knowing the web as such in a satisfactory way. Unsurprisingly, indeed, one of the earliest attempts at establishing the exact size of the web, an effort which entails, as the author acknowledged, «difficult qualitative questions concerning the Web, and attempts to provide some partial quantitative answers to them» (Bray 1996) opened with a famous quote

from mathematician and physicist Lord Kelvin (1824-1907) on the poverty of knowledge which is not based on exact measures:

> When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of science (Lord Kelvin, quoted in Bray 1996: 993).

Puzzled with the qualitative implications of the quantitative questions at hand, scholars and Internet research groups have nonetheless long been engaged with the intrinsically frustrating task of «measuring the web». Even though it is self evident that all results are of a particularly ephemeral nature, an overview of research in the field can be useful to give us at least a sense of scope of the «colossal» corpus – as a now famous article in the *Economist* (20th January 2005) dubbed the web.

When it comes to the web's size, the basic question both from the perspective of information technology and from the perspective of linguistics, is simple and extremely complex at the same time. What is the current size of the web? And, more specifically, how many running words does the web contain when considered as a text corpus? Any answer to the above questions is by necessity approximate and ephemeral, owing to the intrinsically dynamic nature of the web. It is not surprising therefore that the difficulties faced by people engaged in the task of answering such questions have been compared from the outset to those of the cartographers of centuries past, struggling with the impossibility of mapping territories that were still largely unknown (Bray 1996: 993).

In a much quoted pioneering study published in *Nature*, Lawrence and Giles (1999) estimated that in the World Wide Web there were at least 800 million publicly accessible pages, amounting to 6 Tb of text (mark-up and white space excluded), which, according to calculation by Meyer *et al.* (2003), meant over 800 billion words. In 2000 a study conducted by Inktomi & Nec Research Institute verified that the web had grown to at least 1 billion unique documents (Inktomi 2000), while in 2002 Google claimed to index 3 billion pages. In 2003, on the basis of Google's

claim, Kilgarriff and Grefenstette estimated 20 Tb of non-markup text amounting to 2000 billion words of running text (2003: 337).

Regardless of the temporary nature of these figures, they are nonetheless indicative of the steady dramatic growth of the web as a repository not only of information but, more crucially for the linguist, of language text. Taking as a new starting point more recent estimates of the web's size amounting to at least 11.5 billion pages as of the end of January 2005 (Gulli and Signorini 2005), it could be calculated that the World Wide Web contains nearly 80 Tb of text, which amounts to a multilingual collection of texts of over eight trillion (8 000 billion) words[4]. Assuming that out of this huge collection of texts nearly one third were written in English, (see estimates for language distribution in 3.4.1.), there could be something in the range of over 2.500-3000 billion words of English on the web, «a virtual English supercorpus ready for use by enterprising linguists in all manner of language research» (Bergh 2005: 26). A more conservative but up-to-date and fairly reliable estimate, as far as English language only is concerned, is about 1 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).

Whatever its exact size, it is clear that the web presents linguists with a collection of texts definitely larger than any other existing corpus and certainly larger than they need, which alters altogether the meaning of size as a basic corpus issue. In the early days of corpus linguistics, anyone could testify to what extent size mattered. It was a pains-taking task to reach the minimum size required for a corpus to yield significant evidence. Even today, when corpora are enriched by extensively downloading from the Web, the «Shandian paradox» of creating something which seems to be getting old at a faster pace than it grows is still everyday experience for corpus linguists, leaving corpus compilers with the strange feeling of working at something which could be deemed old or inadequate by the time it is released. When it comes to the web as corpus, however, the role played by size seems to be reversed. If early corpus linguists had to strive for size, with the dimension of the corpus being of prime concern to researchers constantly asking themselves «is the corpus we are building large enough?», the web as corpus revolution seems to push the problem to the other extreme by providing a collection of

texts which can be literally overwhelming in terms of running words, and hence potentially useless[5]. 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), looking at the web as a corpus in its own right makes linguists revise such slogans as «the larger the better». That bigger is better is a truth that cannot hold when big actually means gargantuan and uncontrollable as is the case with the World Wide Web. This is the case not only from a corpus linguistics perspective. Also from the point of view of information retrieval it has been explicitly argued that with the exponential growth of the Internet it is becoming disputable whether «bigger is better», even though it is undeniable that a large quantity of data accessible through the web can be of great help when seeking unusual or hard-to-find information (Sullivan 2005). The same applies to linguistics where «sheer quantity of linguistic information can be overwhelming for the observer» (Hunston 2002: 25). Nonetheless, there is also ample evidence, especially from Natural Language Processing research carried out using the web as a corpus, that probabilistic models of language based on very large quantities of data, even if very «noisy», are better than ones based on estimates from smaller and cleaner datasets (Keller and Lapata 2003; Nakov and Hearst 2005). Moreover research on the web as corpus has proved particularly useful in cases where other resources only provided sparse data (Keller and Lapata 2003). It is on the basis of such evidence that further research on the web as corpus has been encouraged, leading to the creation of specific tools and methods to fully exploit the web's real potential.

3.4. *Content*

The exponential growth of the web since its inception in the early 90s has also had a great impact on its content, another key issue to be explored when determining the web's value as a corpus.

A natural correlate of representativeness, which depends on decisions concerning what should go in and what should be left out of a corpus, the issue of content in corpus linguistics is often related, unfortunately, also to practical considerations such as text availability, copyright issues, technical problems (Hunston 2002: 27) which more often than not affect choice in designing corpora. Nonetheless the content of a corpus is what ultimately

determines the scope of generalizations that can be made out of corpus data. As Lee (2001: 37) argues,

> [it] is impossible to make any useful generalizations about «the English language» or «general English» since these are abstract constructions. Instead it is far easier and theoretically more sound to talk about the language of different genres of text, or the language(s) used in different *domains*.

Texts in traditional corpora are generally classified at least in terms of topic/domain and genre, an approach which is difficult to reproduce with the web, even though there seems to be no more universally accepted typology for classifying texts in traditional corpora than there is for the Internet (Sharoff 2007: 84). As far as the web as corpus is concerned, however, this issue becomes more *indigestibile* given the intrinsic difficulties of characterizing the web in any of its aspects. As pointed out by Chakrabati already in the late 90s, «the Web has evolved into a global mess of previously unimagined proportions. Web pages can be written in any language, dialect or style, by individuals with any background, education, culture, interest and motivation» (Chakrabati 1999: 54). And in the past few decades the World Wide Web has grown so big and in such an anarchic fashion, that it is virtually impossible to describe it in terms of its content (Grefenstette and Nioche 2000: 1).

Notwithstanding the patent impossibility of reaching any conclusive result, researchers have been trying to characterize the web through a number of parameters such as size, content and structure (O'Neill 2002). It is hardly surprising, however, that this remains a frustrating task bound to end up in failure. One might even conclude that the content of the web can only be envisaged *via negativa*, through what one possibly searches there and fails to find, rather than positively scanning all of its content. Moreover, when seen from a corpus linguistics perspective, a major flaw of the web seems to be its intrinsic irreducible *anarchism*, which does not only make the 100 million words *British National Corpus* comparatively resemble «an English country garden», but, more importantly, seems to put an end to any hope of relying on the web as an object of scientific enquiry. Nothing possibly voices better the puzzlement and bewilderment of the (corpus) linguist

when confronted with the web, than the words of one of the most influential supporters of the web as corpus:

> First, not all documents contain text, and many of those that do are not only text. Second, it changes all the time. Third, like Borges's Library of Babel, it contains duplicates, near duplicates, documents pointing to duplicates that may not be there, and documents that claim to be duplicates but are not. Next, the language has to be identified (and documents may contain mixes of language). Then comes the question of text type: to gain any perspective on the language we have at our disposal in the web, we must classify some of the millions of web pages, and we shall never do so manually, so corpus linguists, and also web search engines, need ways of telling what sort of text a document contains: chat or hate-mail; learned article or bus timetable (Kilgarriff 2001).

While these may sound like arguments against the web as corpus, for Kilgarriff this is precisely where the challenge really lies: «For the web to be useful for language study, we must address its anarchy» (Kilgarriff 2001). Anarchy is thus the original sin of a virtual space which, as its very name reveals, is *global* more than anything else on earth. In the World Wide Web anyone, regardless of country or language, is free to make information and services available, and this is achieved – significantly from a linguist's perspective – mainly through written texts produced and made available, often in real time, as authentic machine readable format texts. Despite therefore the limitations of any attempt to confront the anarchy of the web, and with no pretence at exhaustiveness, in the following pages the issue of content has been conveniently split into three basic components: language, topic, registers and genres. With reference to such issues current attempts at characterizing the web have been reported, with the only aim of giving an idea of scope of the web as a corpus from the point of view of content.

3.4.1. *Language* When confronted with the idea of the web as a linguistic corpus, most people would think of it mainly as a monolingual English language corpus, since English has established itself as the lingua franca of the Internet. On the contrary, one of the most interesting characteristics of the web is its multilinguality, which, from a corpus linguistics perspective, means that the

web contains virtually endless corpora in almost any language on earth. In Crystal's words,

> The Web is an eclectic medium, and this is seen also in its multilinguistic inclusiveness. Not only does it offer a home to all linguistic styles within a language; it offers a home to all languages – once their communities have a functioning computer technology (Crystal 2006: 229).

In the past few years several techniques have been implemented for estimating the number of words available through web browsers for given languages by applying to the web common techniques used to estimate the size of a language-specific corpus based on the frequency of commonly occurring words in the corpus itself. In their much quoted article published in 2000, Grefenstette and Nioche estimated English and non-English language use on the World Wide Web, thus providing the basis for further exploitation of the web as a multilingual corpus. Though clearly faced with a predominantly English language corpus, with over two-thirds of the pages written in English (Grefenstette and Nioche 2000: 2), the authors could already notice that non-English languages were growing at a faster pace than English. More recent estimates by Internet World Stat (2005), reported in a special issue of UNESCO *The New Courier*, clearly show that the World Wide Web is no longer going to be the predominantly English speaking world it used to be at the outset, since other languages are increasingly and significantly represented.

The exponential growth of non-English languages may be surprising but is easily explained with the growth of websites providing news in different languages (such as newspaper websites), of governmental official websites, and even of collaborative enter-

Fig. 1.1. Languages used on the Internet (Internet World Stats).

prises such as Wikipedia, together with the growing number of personal or corporative homepages or blogs. Nonetheless, the relative growth of languages other than English does not necessarily imply that access to the benefits of the Internet are more evenly distributed around the world, and the persistence of differences in the «weight» of individual languages on the web points to more general problems concerning the so called digital divide between rich and poor countries. The Internet seems to have rather disappointed all hopes for a «global village» where even «the poor countries would be able to benefit, with unprecedented ease, from a myriad of databases, from training, from online courses, all of which would provide access to the knowledge society and allow these countries to catch up progressively with the pack of prosperous nations» (Mouhobi 2005). The digital divide between first and third worlds is an issue, which is made even worse by the technical problems relating to the encoding of non-Latin alphabet using a system (ASCII codes) devised for Latin alphabets only (Crystal 2006). But while the problem of non-Latin alphabets on the Internet still calls for a solution to redress imbalances in access to the Internet, the web has paradoxically proved a language resource precisely for some «minor» or «endangered» languages (Ghani R. *et al.* 2001; De Schryver 2002; Scannel 2007; Zuraw 2006). As Fraser Gupta argues:

> There are some imbalances in access to the web (for example, Africa is especially underrepresented), but every day participation is extended, and, in any case, the web has given opportunities to writers all over the world who would previously never have had the opportunity to see their writing in print. Because of its wide reach, the web has also put writers and readers in touch with each other who would not otherwise have been able to share their writing. If we regard the investigation of written language as worthy of attention, we must accommodate the huge resources of written language that we have access to on the web (Gupta 2005).

As far as the present distribution of languages used on the web is concerned, the most recent estimates of the top ten languages report that (as of June 2008) English and Chinese were shown at 430 and 276 million Internet users respectively, as the most widely used languages, followed by Spanish, Japanese, French, German, Arabic, Portuguese, Korean, and Italian:

While interesting, these data are obviously not significant in

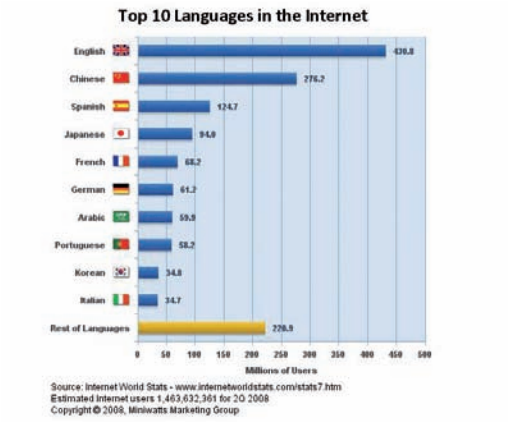


Fig. 1.2. Top 10 Internet languages (Internet World Stats).

themselves since they only provide a snapshot of a constantly changing reality. As the table below shows, it is mainly in terms of percentages that numbers can help us figure out how the World Wide Web has been changing as an increasingly global and multilingual environment, with changes in the relative growth of languages possibly mirroring changes taking place at different levels in society:

Top Ten Languages Used in the Web
(Number of Internet Users by Language)

TOP TEN LANGUAGES IN THE INTERNET	% of all Internet Users	Internet Users by Language	Internet Penetration by Language	Language Growth in Internet (2000 - 2008)	2008 Estimated World Population for the Language
English	30.5 %	430,802,172	21.1 %	203.5 %	2,039,114,892
Chinese	20.4 %	276,216,713	20.2 %	755.1 %	1,365,053,177
Spanish	6.8 %	124,714,378	27.6 %	405.3 %	451,910,690
Japanese	1.9 %	94,000,000	73.8 %	99.7 %	127,288,419
French	6.1 %	68,152,447	16.6 %	458.7 %	410,498,144
German	1.4 %	61,213,160	63.5 %	121.0 %	96,402,649
Arabic	5.4 %	59,853,630	16.8 %	2,063.7 %	357,271,398
Portuguese	3.6 %	58,180,960	24.3 %	668.0 %	239,646,701
Korean	1.1 %	34,820,000	47.9 %	82.9 %	72,711,933
Italian	0.9 %	34,708,144	59.7 %	162.9 %	58,175,843
TOP 10 LANGUAGES	78.2 %	1,242,661,604	23.8 %	278.3 %	5,218,073,846
Rest of the Languages	21.8 %	220,970,757	15.2 %	580.4 %	1,458,046,442
WORLD TOTAL	100.0 %	1,463,632,361	21.9 %	305.5 %	6,676,120,288

(*) NOTES: (1) Internet Top Ten Languages Usage Stats were updated for June 30, 2008. (2) Internet Penetration is the ratio between the sum of Internet users speaking a language and the total population estimate that speaks that specific language. (3) The most recent Internet usage information comes from data published by Nielsen//NetRatings, International Telecommunications Union, Computer Industry Almanac, and other reliable sources. (4) World population information comes from the U.S. Census Bureau. (5) For definitions and navigation help in several languages, see the Site Surfing Guide. (6) Stats may be cited, stating the source and establishing an active link back to Internet World Stats. Copyright © 2008, Miniwatts Marketing Group. All rights reserved worldwide.

Fig. 1.3. Internet users by language (Internet World Stats).

Taking the presence of Arabic on the Internet as an example, these estimates suggest that as of June 2008 there were 59,853,630 Arabic speaking people using the Internet, who represented 5,4 % of all Internet users. This means that out of an estimated 357,271,398 world population that speaks Arabic, 16.8 % use the Internet. Thus the number of Arabic speaking Internet users has grown by 2,063.7 % in the last eight years (2000-2008). Even allowing for errors in these figures, they clearly portray an ever-changing scenario. A significant comparison can be drawn with estimates dating back to 2000:

Tab. 1.1. *Web pages by language* (Pastore 2000)

Language	Web Pages By Language Web Pages	Percent of Total
English	214,250,996	68.39
Japanese	18,335,739	5.85
German	18,069,744	5.77
Chinese	12,113,803	3.87
French	9,262,663	2.96
Spanish	7,573,064	2.42
Russian	5,900,956	1.88
Italian	4,883,497	1.56
Portuguese	4,291,237	1.37
Korean	4,046,530	1.29
Dutch	3,161,844	1.01
Sweden	2,929,241	0.93
Danish	1,374,886	0.44
Norwegian	1,259,189	0.40
Finnish	1,198,956	0.38
Czech	991,075	0.32
Polish	848,672	0.27
Hungarian	498,625	0.16
Catalan	443,301	0.14
Turkish	430,996	0.14
Greek	287,980	0.09
Hebrew	198,030	0.06
Estonian	173,265	0.06
Romanian	141,587	0.05
Icelandic	136,788	0.04
Slovenian	134,454	0.04
Arabic	127,565	0.04
Lithuanian	82,829	0.03
Latvian	60,959	0.02
Bulgarian	51,336	0.02
Basque	36,321	0.01

These estimates published at the Millennium by Global Reach give us an idea of how the presence of each single language is subject to change. One cannot help noticing, for instance, the macroscopic growth of Chinese as a language for Internet communication, moving up from 4th position in these 2000 estimates to 2nd position in the more recent 2008 estimates, perhaps mirroring the country's economic growth. Other languages that have significantly changed position are Portuguese and Korean (9th and 10th position in 2000), which have now overtaken Italian, and Spanish (6th position in 2000 and now immediately following Chinese as the third language for Internet communication). Finally, and most significantly perhaps, one should notice the growth of Arabic, leaving its 27th position in 2000 to become the 7th language in the recent 2008 estimates – possibly a further consequence of the events surrounding September 2001.

The astonishing variety of languages on the web, and the impressive growth of non-English and also non-Western languages, show the importance of the web as a multilingual environment which is ready to reflect changes taking place in society at large. While variety should not make us forget the problems relating to the «digital divide», it is hardly questionable that from a linguistic point of view the web can be considered as a vast, dynamic, easy to access, multilingual language resource, whose significance is further enhanced by the astonishing diversity of the topics covered.

3.4.2. *Topics* The web has become such a pervasive communication medium that there seems to be no field of human activity that is not some way or other covered by it. This has probably made us forget that the web has its origins in the world of US defence and that only subsequently has it developed as a way to share knowledge and research within the academic world. Today it continues to play a major role in governmental and scientific communication, but its relative ease of use has meant that it did not take long before people outside research and political institutions began to develop other uses for the web, thus turning it into a major facilitator of personal communication, electronic commerce, publishing, marketing and much else (Day 2003: 12). The web has thus become a not only more inclusive, but also more chaotic, environment, in which

content that can be considered of great value (academic papers, literary texts, governmental documents) coexists with content which is of low quality, or worse (Day 2003: 6).

While diversity of web content implies that each page may range from a few characters to thousands of words «containing truth, falsehood, wisdom, propaganda, or sheer nonsense» (Chakrabarti *et al.* 1999: 54), several attempts have been made to implement some principles of classification based on topic[6]. This is generally performed through directories, which group web pages on the basis of content into a number of categories. An apparently trivial task, the classification of web pages in directories relating to their content is something which ordinary search engines have trouble coping with, given the intrinsic nature of the web as a democratic, or rather *anarchic*, space, apparently free of any form of organization and planning.

The earliest attempt at organizing the content of the web from the point of view of topics was made in the mid-90s by two Ph.D. students at Stanford University, Jerry Yang and David Filo, who created the Yahoo directory to help their friends locate useful web sites. In a matter of months their initially informal project was incorporated as a fully-fledged company. As Chakrabarti argues, the success of this enterprise was due to the attempt at «reviving the ancient art of organizing knowledge into ontologies» – an art which «descends from epistemology and philosophy and relates to the possibility of creating a tree like hierarchy as an organizing principle for topics» (Chakrabarti 2003: 7). The paradigm of browsing topics arranged in a tree is a pervasive one and the average computer user is generally familiar with hierarchies of this kind through directories and files. This familiarity, Chakrabarti suggests, carries over rather naturally to topic taxonomies.

If not entirely reliable in terms of coverage, directories are from the linguist's point of view a simple way to envisage the web's content in terms of topics. Even though the content of the World Wide Web can by no means be reduced to the web pages indexed by even the largest search engine, the wide range of topics can be easily seen at a glance through a survey of the «directories» listed by a search engine. Here are, for instance, the topics covered by Yahoo!:

Fig. 1.4. Yahoo! Directory (Reproduced with permission of Yahoo! Inc. ®)

And here are the directories listed by Google:

Fig. 1.5. Google Directory

As is well known, each topic label in the directory is in the form of a hyperlink and is to be considered as a gateway to subdirectories, in a sort of Chinese box structure. Thus, choosing «Engineering» from the directory «Science» in Yahoo the user is

projected into a new set of topics, which are in turn open doors to other topics:

Fig. 1.6. Yahoo! Categories (Reproduced with permission of Yahoo! Inc. ®)

From the categories listed above it is self-evident that the role directories can play in identifying subsets of web pages focussing on one topic cannot be underestimated. Each directory could – theoretically at least – be considered as a «virtual» corpus including texts all dealing with the same topic, so that to find evidence of usage for a certain phrase in a specific domain (e.g. Health) one could restrict search to the relevant directory instead of searching the whole web. On the other hand, however, it should be stressed that «topic» or «domain» are themselves controversial categories in corpus linguistics. As the EAGLES (1996) report argues, texts cannot be safely categorized on the basis of a limited list of topics, i.e. inevitably in terms of text-external criteria. Topic/domain classification should rather be based on text-internal evidence (such as vocabulary clustering), and this is certainly not the case with web topic categorization through directories. Recent studies are in fact questioning existing Internet classifications and working at new methods for establishing suitable categories for classifying web texts into domain and genre (Sharoff 2007).

Despite their attractiveness, web directories cannot therefore answer the linguist's need but partially. As recent research by Biber and Kurjian has shown, the categorization provided by search engine directories, while considered as indicative in terms of reference to general topical domain, still remains well behind the minimum standards required for linguistic research (Biber and Kurjian 2007). It is of crucial importance in fact for (corpus) linguists to identify texts on situational/linguistic grounds, i.e. to know what kind of texts are included in a corpus (even in a «virtual» one such as a search engine directory) in terms of register, genre, text type. A label generically referring only to topic or domain is not enough for a linguist to discriminate web content, and as the web grows in size and anarchy, classification of web pages by topic only seems to be insufficient to maintain acceptable standards of effectiveness even from the point of view of general purpose information retrieval. For this reason greater and greater interest is being paid to web categorization by genre as a complement to topic classification, and it is precisely on this issue that research under way in the field of information retrieval and the interests of the corpus linguistics community seem to have finally converged.

With the growth of the web as «a massive loosely organized library of information» (Boase 2005: 1), it has in fact become more and more evident that topic alone is insufficient for text categorization and that new forms of «representation» of information, including genre, are needed (Crowston-Kwasnik 2004). Information on the web now so vastly exceeds users' needs, that even a common search runs the risk of being frustrating just on the ground of genre/text-type. What are we looking for, for instance, when we ask a search engine to return results for San Francisco? Information on the city? Accommodation? Flights? People searching for information (whether general or specifically linguistic) do not simply look for a topic but also have implicit generic (i.e. genre-related) requirements (Kessler *et al.* 1997: 32). The problem of matching a user's query with the relevant answers in terms of text types and genres is anyway bound to remain an issue unless more information is integrated by web designers in their web pages, or unless users are taught to formulate more explicit and complex queries, or unless – and this is a crucial point – web information retrieval applications (i.e search

engines) themselves are modified so that they can focus not only on topic and content but also on features of form and style as a gateway to genres and text-types (Boase 2005: 3-4). As a category «ortoghonal to topic» (Boase 2005: 6) genre would thus make it possible to find documents that match the user's search terms from the point of view of topic, including (or excluding) documents of a given genre or genre cluster.

3.4.3 *Registers and genres* While the wide variety of languages and the extreme inclusiveness of topics which characterize the web have been of paramount importance in making it an object worthy of the corpus linguistics community's attention, it is self-evident that languages and topics can only partially represent the web's content in a way that is meaningful from the linguist's point of view. In order to take stock of the web as a textual corpus greater attention must therefore be paid to the issue of register and genre.

From a corpus linguistics perspective, discriminating texts in terms of genre, register and text-type is a fundamental concern (Lee 2001), and identifying web genres/registers would certainly pave the way towards a more methodologically sound use of the web as a corpus. Indeed, the very impossibility of getting to know anything about the web in this respect is one of the reasons why its representativeness from a corpus perspective has been radically questioned. While the linear organization of most paper documents and the apparent fixity of most traditional genres can still be reflected in traditional electronic corpora such as the BNC, a similar approach would be hardly applicable to a corpus of web documents, which are more complex and unpredictable than paper documents, and where the very notion of genre seem to be undergoing a process of democratization (Yates and Sumner 1997: 3; Santini 2005: 1). Furthermore, it could be argued, genre and register, along with related concepts of text-type, domain and style are themselves «somewhat confusing terms» (Lee 2001: 37). This makes it even more difficult to thoroughly map the web from this point of view.

Apart from the intrinsic difficulty of sorting out the truly amazing plethora of highly individualized documents which make up the web, when dealing with the web in terms of genres and registers some basic prejudices also need to be addressed. It

has for instance become something of a commonplace to think of the web as a writing space that is characterized mainly by ephemeral writing, while in fact diversity is the only keyword. Certainly many texts are created in real time and do not undergo any kind of editing, but others are faithful reproductions of historical or literary texts. Some of the writing on the web can give the impression that this is a place where traditional genres and text-types have been superseded by the new genres of electronically mediated communication, but there are still also many traditional genres which have been adapted to the new electronic environment, without losing their basic properties as genres (e.g. newspapers, academic articles...). Variation of register is thought to be set permanently at the informal end on the web, but web texts actually range from the most formal legal documents to quite informal blogs and chatrooms. Even though, as has been argued, all attention is drawn to the new – or not-yet-standard – text types and styles that have clearly emerged in recent years in computer mediated communication, these are in fact only a part of all the text available on the web (Crystal 2006: 84). From the point of view of registers and genres, the problem with the web is therefore not so much what it actually contains, but rather how to discriminate and take advantage of its sprawling content. In a world where «the dream of creating an information-rich society has become a nightmare of information overload» because information has exceeded the limits of our ability to process it and the advantages of a huge store of information as the World Wide Web seem to be outweighed by the difficulties of accessing them (Kwasnik-Crowston 2000), the problem of categorizing web content in terms of genre, register and text type as a complement to topical principles of classification has become a common priority for information retrieval, computational linguistics, and also corpus linguistics. However, methods for achieving this have not yet been fully established. Researchers involved in web genre analysis have long recognized the need for new categories of approach, based on the awareness that the web, as a dynamic environment, is more prone to centripetal forces which result in constant genre evolution (Santini 2007). On the one hand it is clear that the web hosts a number of traditional genres which have simply changed medium (from paper

to electronic) without any further modification of their intrinsic features (reproduced genres), on the other hand there are text-types and genres that are undergoing processes of *remediation* which make them either hybrid genres (adapted novel genres), or completely new genres (emerging genres). On the basis of this evolutionary pattern, and drawing on earlier classifications (Shepherd and Watters 1998; Crowston and Williams 2000), Santini (2007) proposes a list of five recognizable genres typologies:

1. reproduced/replicated genres,
2. adapted/variant genres,
3. emergent/novel genres,
4. spontaneous genres,
5. unclassified web pages.

In this new and varied context the notion of genre is clearly undergoing a process of transformation so that traditional criteria for text categorization and genre identification cannot hold. Previous clear-cut distinctions between spoken/written, formal/informal seem for example inappropriate ways to address the registers used in most instances of computer mediated communication, whose complexity rather requires multidimensional approaches. A recent trend within this research field is a definition of genre based on the notion of «bundle of facets» (Kessler *et al.* 1997: 32f). This means that rather than identifying genres and text-types on the web in terms of their adherence to *a priori* determined genres, automatic genre detection systems identify recurring patterns and generic cues in terms of facets (i.e basic attributes). Each genre label can thus be determined *a posteriori* as a post-coordination of facets, corresponding to a bundle of co-occurring features, such as first /third person, specialized vocabulary, domain. (Crowston-Kwasnik 2004: 6-7). It goes without saying that these methods for detecting web genres are proving more flexible and hospitable to the many hybrid genres which characterize the web as a new medium, and make research hope for a real possibility to implement genre categorization within web search systems, from which both information retrieval and corpus linguistics would greatly benefit.

4. *From «body» to «web»: new issues*

The attempt at analyzing the web as a linguistic corpus has already highlighted some characteristics such as constant change, non-finite size, anarchism, which in turn indicate the necessity of addressing 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, 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) seems likely 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. This is also changing notions of permanence/stability for corpora. As Wynne, again, argues:

> In the traditional model a corpus is carefully prepared, by taking a sample of the population of texts of which it aims to be representative, and possibly encoded and annotated in ways which make it amenable for linguistic research. The value and the reusability of the resource are therefore dependent on a bundle of factors, such as the validity of the design criteria, the quality and availability of the documentation, the quality of the metadata and the validity and generalisability of the research goals of the corpus creator (Wynne 2002: 1205).

In the new scenario the relative importance of design criteria may well change, as it might become the norm to create a collection of texts (the corpus) on an *ad hoc* basis – such as «all 17th century English fiction» or all «Bulgarian newspaper texts» – by simply choosing from within larger existing text archives (Wynne 2002: 1205).

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, de-centering and re-centering, 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. Some of these new issues will be briefly explored in the following pages.

4.1. *Dynamism*

An important characteristic of the web that has implications for its supposed nature as a corpus is its inherently dynamic nature. Firstly, the web is characterized by exponential growth, with new pages and sites appearing at a significantly high rate. Secondly, the content of existing documents is continually updated, so that sites and pages do not only frequently appear but also as frequently disappear. Thirdly, the very structure of the web is in constant flux, with new links between documents being continually established and removed (Risvik and Michelsen 2002). These factors have largely contributed to making the web the largest and most accessible information resource in contemporary society, but have also gained it a reputation for volatility. No doubt everybody has experienced such volatility through the so called «broken-link» problem – the most evident sign of the ever-changing nature of the web – 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 cre-

ated over time, the web as a whole is constantly changing, and precise estimates are not only difficult to produce but also, perhaps, useless. As Fletcher argues «studies of the nature of the web echo the story of the blind man and the elephant: each one extrapolates from its own samples of this ever evolving entity taken at different times and by divergent means» (Fletcher 2007: 25). It goes without saying, then, that existing studies of the web's fluidity can only give us a faint idea of what we really mean when we say that the web is «dynamic». A study carried out in 2004 on a selection of commercial, academic, governmental and media US sites, estimated for instance how many new pages were created every week calculating a «weekly birth rate» for web pages of 8%. Then the authors addressed the issue of «new content», finding that on average each week around 5% of the page content was really new, coming to the conclusion that nearly 62% of the content of new URLs introduced each week was actually new. Finally they estimated the life of individual web pages on the web, by combining their findings with results from a study by the Online Computer Library Center (OCLC 2002) to get a picture of the rate of change for the entire web. On the basis of the two sets of data the authors could speculate that only 20% web pages is still accessible after one year (Ntoulas *et al.* 2004).

While this low rate of «survival» of web pages has made historical archiving and long term access to web content a crucial concern, prompting the work of institutions such as the Internet Archive or initiatives such as the Wayback machine, it is also worth considering how such dynamism affects the web's potential as a reservoir of attested usage from the linguist's point of view. Web texts certainly change over time but there is no reason to assume that this perpetual change in content altogether alters the nature and composition of the whole. If the web is a huge collection of authentic texts which are the result of genuine human interactions, the source of each single «utterance» or «lexical item» may vary, but if these are to be considered as evidence of usage, it may not be of crucial importance whether the source is one or other document. On the contrary, evidence of attested usage found in different web pages at different times would testify to the social dimension of usage, providing evidence of the autopoietic and diachronic nature of discourse (Teubert 2007: 67)

and contributing to a richer notion of intertextuality (Kristeva 1986: 39), by making visible how

> (each) new text repeats to a considerable extent things that have already been said. Text segments, collocations, complex and simple lexical items which were already in use, are being recombined, permuted, embedded in new contexts and paraphrased in new ways (Teubert 2007: 78).

Thus, while the fluid nature of the web is often invoked as one of the main arguments against using the web as a corpus because a result computed today may not be exactly reproducible tomorrow, one is tempted to revive on the other hand a powerful analogy with water, as Kilgarriff (2001) does in his seminal 2001 *Web as Corpus*, arguing that nobody would demand that the chemical composition of water in a river is exactly the same at each experiment. Nonetheless river water is undoubtedly a legitimate object of scientific enquiry, and so is the web. As Volk suggests, we only have to learn how «to fish in the waters of the web» (Volk 2002: 9).

4.2. *Reproducibility*

One of the most obvious practical consequences for linguistic research of the web's dynamic nature is the impossibility to reproduce any experiment – a really serious problem since it is one of the basic requirements of scientific research that an experiment can be replicated/reproduced so that it can also be validated or, perhaps more crucially for the scientific method, invalidated. This also applies to corpus linguistics research which aims to be scientific. Accordingly, it is an implicit requirement of a corpus that it should be stable, so that «the results of a study can be *validated* by direct replication of the experiment» (Lüdeling *et al.* 2007: 10). While for traditional corpora this is, at least in principle, irrelevant, the problem of reproducibility and validation of experiments becomes a crucial issue when using the web as a corpus, especially when accessing it through ordinary commercial search engines. As Lüdeling again argues:

> [...] the web is constantly in flux, and so are the databases of all commercial search engines. Therefore, it is impossible to replicate an ex-

periment in an exact way at a later time. Some pages will have been added, some updated, and some deleted since the original experiment. In addition, the indexing and search strategies of a commercial engine may be modified at any time without notice (Lüdeling *et al.* 2007: 11).

Furthermore, the wild inconsistencies and fluctuations discovered in the result counts even for common English words via common search engines (e.g. Veronìs 2005) make us understand that any linguistic study based on the web as corpus necessarily calls for some form of validation. This has made the issue of reproducibility become one of the new key concerns, prompting research on the web as corpus particularly in terms of a reconsideration of tools and methods. Thus, while some researchers using the web as corpus via ordinary search engines simply validate their results by repeating the same search at distant intervals in time (Lüdeling *et al.* 2007: 11), others have opted for different methods of using the web as a corpus, i.e. by downloading the results of the queries submitted to a search engine so as to create a more stable, and hence verifiable, object.

4.3. *Relevance and reliability*

The dynamic and fluid nature of the web makes it an apparently unreliable environment for corpus-based research also from the point of view of relevance and reliability (Fletcher 2004), which have also become key concerns for corpus research based on the web, especially when web data are to be used as a basis of both quantitative and qualitative evidence (Rosenbach 2007: 168). In this contest relevance and reliability can be seen in terms of «precision» and «recall», two issues pertaining to information retrieval which are often mentioned as worth some consideration in studies concerning the web as corpus (Baroni and Bernardini 2004; Fletcher 2007; Lüdeling *et al.* 2007)[7].

The importance of precision and recall even in the most basic use of the web as corpus is easily explained. While any linguistic search carried out by means of specific software tools on any traditional stable corpus of finite size (such as the BNC) will certainly report *only* (precision) results exactly matching the query, and *all* (recall) the results matching the query, this is obviously not the case with the web, where recall is impaired by its unstable nature as a dynamic non-linguistically oriented collection of text, whereas precision is impaired by the intrinsic limitations, from the linguist's perspective, of search tools such as ordinary search engines. If locating an item in a corpus makes sense only assuming a minimum qualitative and quantitative accuracy in the search, a web search for linguistics purposes would make sense only assuming that the search should not return too many «wrong» hits, called *false positives* (precision), and should not miss too many correct items, called *false negatives* (recall). This is precisely what makes using search engine hits as a source of linguistic information, i.e. 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 – as we will see later – something 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, dramatically inflates frequency counts and makes numeric data obtained from hit counts on the web virtually useless from the point of view of statistics. Thus, while using page hit counts as indicative of the frequency of a given lexical item seems intuitively to be a cheap, quick and convenient way for researchers to obtain frequency data, the reliability of such data is seriously impaired by the instability of the web itself.

As to relevance and precision, these are impaired by the very strategies that enhance the power of search engines as tools for retrieving information (not specifically linguistic) from the web. In order to retrieve as many documents as possible matching a user's query, search engines usually perform some sort of normalization: searches are usually insensitive to capitalization, automatically recognize variants («white-space» finds «white space», «white-space» and «whitespace»), implement stemming for certain languages (as in «lawyer fees» vs. «laywer's fees» vs. «lawyers' fees»), ignore punctuation characters and prevent querying directly for terms containing hyphens or possessive markers (Lüdeling *et al.* 2007; Rosenbach 2007). While such features are undoubtedly

helpful when searching for information on the web, they certainly affect the search possibilities in terms of precision and relevance, and accordingly distort frequency data (Nakov and Hearst 2005). Indeed it is out of the necessity to counteract the limits in terms of relevance and reliability of any use of the web as corpus based on access via ordinary search engines that the most important tools for the creation of web corpora were born.

Conclusion

In this chapter some theoretical aspects relating to the emerging notion of the web as corpus have been explored by revising key issues in corpus linguistics in the light of the characteristics of the web as a spontaneous, self-generating collection of texts, and by hinting at some of the new issues which the very possibility of using the web as a corpus seems to rise. While the notion of a 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 the web as corpus introduces notions of non-finiteness, flexibility, de-centering and re-centering, provisionality, which do not only seem to be calling into question the «good practice» of corpus work, but may also be affecting the very notion of a linguistic corpus in more radical ways. All these issues will be further explored on applicative grounds in the following chapters.

Note

[1] In the present work both «web as corpus» and «web-as-corpus» have been used with reference to the research field as a whole. The hyphenated form «web-as-corpus» has been generally opted for as a noun modifier.

[2] The term «corpus-hood» is a neologism first used by Kilgarriff (2003: 334).

[3] It should be noted that Teubert (2007) makes explicit use of Google search to support his views.

[4] This calculation is based on an average page length of 7.3 Kb and on an average of 10 bytes per word (see Meyer *et al.* 2003; Kilgarriff and Grefenstette 2003).

[5] The problem of size is related to the so called Zipfian properties of language, according to which nearly 50% of the words in a corpus occur only once. This requires that a corpus is large enough to contain enough occurrences of 50% of its running words (Sinclair 2005, referring to Zipf 1935).

[6] The word topic is here used as a synonym for domain in the sense of subject field. Preference for the word topic was suggested to avoid confusion with the notion of «domain» in Internet technology, where the term refers to part of the URL used to identify a website.

[7] Precision and recall are defined as «measures of effectiveness of search systems». More specifically, «*precision* is the ratio of the number of relevant documents retrieved to the total number of documents retrieved, and *recall* is the ratio of the number of relevant documents retrieved to the total number of relevant documents» (Van Rijsbergen 1979).

Chapter II

Challenging Anarchy. The Corpus and the Search

Introduction

In this chapter the potential and limitations of accessing the web as a ready-made corpus via ordinary search engines are introduced and discussed. In Section 1 some basic issues concerning web search are surveyed, before providing an overview of how search engines work in Section 2. In Section 3 some techniques for exploiting search engines for linguistic purposes are shown. In this Section the role and meaning which search engine advanced options can perform from the linguist's point of view are briefly discussed before introducing the use of complex queries for linguistic purposes in Section 4.

1. *The corpus and the search*

Despite an undisputedly controversial status as a linguistic corpus, the web has often been used as a corpus in recent years, especially in the field of computational linguistics. As a text collection of unprecedented size and scope, freely and ubiquitously accessible, the web was in most cases the obvious source to go to for the solution of typical Natural Language Processing tasks, soon gaining a reputation as «the largest data set that is available for Natural Language Processing» (Keller and Lapata 2003: 459)[1]. While using the web as a training and testing corpus has undoubtedly become common practice in computational linguistics, with an increasing body of studies showing that simple algorithms using web-based evidence can even outperform sophisticated methods based on smaller but more controlled data sources (Lüdeling *et al.* 2007: 7), awareness of problems and limitations of using such a

huge, anarchic and unstable database as a source of linguistic information remains. The first and most obvious problem is the total lack of control over the source of the texts, both from the point of view of their production and from the point of view of their content, which makes it very hard to assess the «authoritativeness» of a web page in terms of accuracy of content and representativeness of form (Fletcher 2004a: 275). Moreover, the content of the web, as a heterogeneous and non-sampled body of text is so varied that not only «unpolished ephemera abound along with rare treasures» (Fletcher 2004a: 275), but online documents often consist of mere fragments, stock phrases, hot lists, and come in a myriad of duplicates and near-duplicates which are not of use from a linguist's perspective. Then there is the already discussed problem of size. While working with a huge amount of data greatly enhances the chances of finding enough information for any item, this advantage is counterbalanced by the fact that dealing with too much data requires an enormous amount of processing power. Last, but not least, there are the thorny issues of representativeness, the impossibility of relying on any form of linguistically oriented utilities or linguistic metadata, and the notorious «volatility» of the «webscape» (Fletcher 2007). These are no doubt limitations which make the web a very inhospitable place for serious linguistic research.

While all these drawbacks are clearly acknowledged, it can nonetheless be argued that a certain feeling of excitement has been nourished in recent years by the extremely interesting results obtained at a practical level, so that skepticism and willingness to profit from the «promises and possibilities» offered by the web as a corpus seem to coexist in the linguistic community. As Christian Mair has recently argued, «attitudes towards the web as a corpus span the whole range from enthusiasm to distinct reserve», a stance clearly reflected in technical papers which generally focus «as much on the potential as on the hazards of using the web as a corpus» (Mair 2007: 235). In a time when «working styles in corpus-linguistic research are changing fast», and the traditional view of «close(d) communities of researchers forming around a specific corpus or set of corpora [...] is becoming increasingly problematical» (Mair 2007: 233), the web seems to have entered the scene as an unexpected novelty, providing (if nothing else) solutions to specific problems. It remains, however, an «accidental» corpus (Renouf 2004), which has imposed itself to the linguists' attention almost against their will. Perhaps nothing better than the recently coined phrase «unwanted corpus» (Mair 2007: 235) can capture the ambiguous mood of reluctant acceptance which in most cases surrounds research on the web as a corpus. As Mair again suggests, «the web will have to be used because it is there, but clearly it is not the corpus that linguists would have compiled» (Mair 2007: 236).

Reservations are even more crucial when it comes to the apparently widespread practice of using the web as a source of corpus-like linguistic information through an ordinary search engine. Such concern has recently found expression in a powerful *caveat googlator*:

> [W]e seem to be witnessing [...] a shift in the way some linguists find and utilize data – many papers now use corpora as their primary data, and many use internet data. These are clearly changes that are technologically induced, and in an era in which google is now a common verb, why not? I feel compelled to add, though, *caveat googlator*! In the culling of data from Medieval Greek texts for my dissertation, [...] I ran across some examples that I felt were best treated as having been altered [...]. Thus I considered them to be attested but ungrammatical – some examples obtained on the Internet in papers I read now strike me as quite the same, that is possibly produced by non-native speakers, or typed quickly and thus reflecting performance errors, and so on. I have no doubt that we will learn how to deal with this new data source effectively [...] (Joseph 2004: 382, quoted in Rosenbach 2007: 167).

The pun «googlator» is an overt allusion to the most popular web search engine, Google. Indeed, there are good reasons to be cautious when accessing the web as a linguistic resource via commercial search engines, beyond those listed by Joseph. While the quantity, diversity and topicality of web documents literally «tantalize» language professionals (Fletcher 2004a: 271), the problem of locating information that is linguistically both reliable and relevant by accessing the web through a search engine represents a real challenge which seems to outweigh all advantages. Firstly, search engines are not designed for use by linguists, since they impose a limit to the number of results that can

be displayed, and the results themselves are displayed in a format which is not suitable for linguistic analysis and are ranked according to algorithms which escape the user's control. Secondly, the very strategies that enhance the effectiveness of search engines for general purposes, such as normalization of spelling and lemmatization, clearly limit the effectiveness of web search for linguistic purposes, which of necessity requires greater precision.

So many obvious drawbacks, however, have not been enough to stop interest in the use of web data both as a source *for* the creation of conventional corpora and *as* a source of data readily amenable to linguistic analysis, at least as a complement to more controlled data sources. Interest in the web for linguistic reference is on the increase, however, also among language professionals outside the corpus linguistics community. Besides typical uses of the web as a source of multilingual encyclopaedic information, and as the platform for distribution of resources such as online glossaries and dictionaries, new uses seem to be emerging which under the disguise of a casual «let me see how many hits I find for this in Google» may be bearing the fruits of a linguistically aware use of the web. It is precisely against the background of this apparent intersection between web search as a common non-linguistically oriented practice and linguistic research informed by the corpus linguistics approach that it is perhaps useful at this stage to survey some basic concepts concerning web search itself, before exploring its meaning from the linguist's perspective.

Even at first glance, the act of searching for information through a digital database via an ordinary search engine and reading *vertically* through the results, which is what typically happens in the simplest web search, seems to be strikingly similar to what also happens when searching a corpus through a concordancer. Indeed the so called SERP (Search Engine Results Page) vaguely resembles a concordance list where the search item is highlighted within a small amount of co-text, even though this is by no means the same format as the KeyWordinContext typical of linguistically oriented tools. It is as if reading «vertically», «fragmented», and looking for «repeated events», which Tognini Bonelli sees among the features that set apart the act of

reading through a corpus from the common act of reading a text (2001: 3), is becoming everyday experience for more people beyond the corpus linguistics community. This superficial similarity cannot conceal, however, the fundamental difference between searching a deliberately collected static corpus using specific tools, and searching texts found on the Internet only by means of a commercial search engine. In this case the linguist has not only very scarce control over the *corpus* itself but also over the *search*, since web search tools are designed to address a variety of needs for which the linguistic form is only a means to an end, and not – as in the case of the linguist – the end itself. As pointed out by Bergh, «whereas standard corpora typically come with software specialized for searches for different linguistic forms, search engines on the web are designed to find contents, using the linguistic form *only as a means* to achieve that goal» (2005: 27, my emphasis). It would be an oversimplification, however, to think that the problem posed by search engines is simply that they are geared towards the retrieval of general information from the web rather than towards the extraction of specific linguistic information. The real problem is that the very needs which search engines address are evolving. It can no longer be assumed for instance that search engines are built so as to answer straightforwardly informational needs. On the contrary, as has been argued, in the web context the need behind the query is often not so much «informational» in nature, but rather «navigational» or «transactional» (Broder 2002)[2]. Accordingly, web search tools are evolving in the same direction, blending data from different sources and clustering results in the attempt to guess what the user is really looking for. This is what makes the web undoubtedly more useful for the average user, but rather less reliable for linguistic research. The latter could in fact be labelled – according to Broder's taxonomy – as exclusively and quintessentially «informational» and does not profit from strategies aimed at enhancing the ability of web search tool to meet other needs. It is therefore advisable that linguists determined to exploit its potential as a source of linguistic information via ordinary search engines become aware of some technical and theoretical aspects of web search, starting with the tools themselves.

2. *Crawling, indexing, searching, ranking: on search and search engines*

It is obviously beyond the scope of the present work to examine in detail the nature of web search engines, but a brief overview of how a web search engine actually works may be nonetheless useful in order to clarify the role it can play in supporting linguistically-oriented research. Most search engines are basically driven by what could be termed as a «text-based approach» (Battelle 2005: 20). More precisely, search engines are Information Retrieval (IR) systems «which prepare a keyword index for a given corpus, and respond to keyword queries with a ranked list of documents» (Chakrabarti 2003: 45). The engine thus only connects the keywords a user enters (queries) to another list of keywords (index) which represents a database of web pages, to produce a list of website addresses with a short amount of context[3].

Three major steps are involved in performing this task (crawling, indexing, searching) and these steps correspond to the parts of the search engine: the crawler; the indexing program; the search engine. The main task that a web search engine performs is to store information about a large number of web pages, which are retrieved from the World Wide Web itself by a web crawler, or «spider», in keeping with the general metaphor representing the Internet as a «web» of documents. This is the program used by the search engine services to scan the Internet identify new sites or sites that have recently changed. Once a new page is identified by the search engine's crawler, its content is analysed (i.e. words are extracted from titles, headings, or special fields called meta-tags) and indexed under virtually any word in the page – even though most search engines tend to exclude from the index particularly frequent words such as articles or prepositions, the so-called stopwords. All data about web pages are then stored in an index database for use in later queries, so that when a user enters one or more search terms into the search engine, the engine examines its index and provides a list of web pages matching the query, including a short summary containing the document's title and extracts from the text. In this phase a fundamental and challenging process is involved, i.e. determining the order in which the retrieved records should be displayed. This process is related to a relevance–ranking

algorithm which takes a number of factors into account, such as the popularity of the page (measured by the number of other pages linking to it), the number of times the search term occurs in the page, the relative proximity of search terms in the page, the location of search terms (for example, pages where the search terms occur in the title page get higher ranking), and even the geographical provenance of the query (which may prompt a bias to ranking higher those web sites which are closer to the user).

It goes without saying that the usefulness of a search engine ultimately depends on the relevance of the result set it gives back, even though relevant can mean something different to the linguist and to the average user. In such a huge text collection as the web there may be millions of web pages that include a particular word or phrase, and obviously some pages may be more popular or authoritative than others. It is precisely the method employed to rank the results, which varies widely from one engine to another, that determines the quality of the response. At the moment the most popular, and to some extent effective, ranking method seems to be Google's PageRank which is based on a computation of website popularity based on a criterion similar to the impact factor in scientific publications. As the Google technology page explains,

> Pagerank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote by page A for page B. But Google looks at considerably more than the sheer volume of votes or links a page receives; for example, it also analyzes the page that casts the vote. Votes cast by pages that are themselves «important» weigh more heavily and help to make other pages «important». Using these and other factors, Google provides its views on pages' relative importance. (*Google technology*: online).

An ingenious system for arranging search results, PageRank may be «slightly problematic from the point of view of corpus linguistics», since «the ranking of Web pages is likely to favour linguistic constructions which happen to occur on more popular pages, thereby risking a certain bias in studies based on language data mined by Google.» (Bergh 2005: 33) On the other hand, Google's ranking system, aimed as it is at weighing the «impor-

tance» of a web page, can be also seen as a contributing to reliability from a linguistic perspective, in so far as it seems to allow some confidence concerning the authoritativeness of the web pages that are ranked higher in the results page.

What is certain is that while most of the working of search engines and web content fall out of the linguist's control, the only part of the search that can be controlled is the query. A search always starts with the user's query, which is yet another fundamental difference between web search and ordinary corpus research. There is no way, via search engines, to start from the corpus itself to produce frequency lists, or to compare corpora in order to obtain keywords. Neither can the results be ordered according to criteria other than the ranking system used by the search engine itself. What is more, no further processing of the results can be done of those generally allowed by linguistic tools (e.g. sorting, statistics, clustering...), unless the web pages are downloaded into a corpus to be analysed with traditional corpus linguistics tools. The only thing that a search engine can do is produce a list of concordance-like strings for a given item, which display a number of occurrences of a certain word or group of words in context, with vague indications of frequency. This suggests quite different roles for the linguist in inductive empirical research on conventional corpora and when using the web as a corpus. As Sinclair (2005) suggests, it is the very «cheerful anarchy of the Web» that places «a burden of care» on the user, and while this is true with reference to the use of the web as a source of texts for conventional corpora (as Sinclair meant it), it is even more so in cases when the web is accessed as a corpus «surrogate». The problem for the linguist is then to turn awareness of all these limits into a resource. More specifically, the linguists' task is to learn how to *challenge* the anarchy of the web and the limited service provided by search engines, either by creating their own tailor-made search engines or by helping ordinary search engines understand what linguists are looking for. In the latter case it should be borne in mind that the relevance of the results and their usefulness for the linguist does not only depend on the nature of the database or on the ranking algorithm used by the search engine but can also crucially depend on the appropriateness of the query. Most searchers – and linguists may be no exception – are instead incredibly lazy,

generally typing in a few words and expecting the engine to bring back perfect results, ignoring that it is only the act of offering more data in the query that often dramatically improves the results (Battelle 2005: 23-25). The way most users generally approach the act of searching the web through ordinary search engines is thus often naïve, and this may account not only for much of the frustration experienced by web searchers, but also for the risk of misusing the web as a linguistic resource. In the following pages therefore some useful techniques for exploiting search engines for linguistic reference will be discussed.

3. *Challenging anarchy: the web as a source of «evidence»*

In spite of its intrinsic limitations, the web is often searched *as* a corpus via ordinary search engines, particularly as a source of evidence of attested usage. More specifically, as pointed out by Rosenbach,

> [a]s any other corpus the web can be used for ascertaining two types of evidence, i.e. qualitative evidence and quantitative evidence. Qualitative evidence is used to show that a certain form or construction is attested; quantitative evidence addresses the question of «how many» of these forms/constructions can be found in a corpus. (...) Drawing such data from the web, in this respect, is similar to «normal» corpus data, though there are some problems that are specific to web data (Rosenbach 2007: 168).

While there is no doubt that accessing the web via ordinary search engines as a source of attested usage is not the best way of using the web in corpus linguistics, and that it is one that forces the linguist to develop «workarounds» (Kilgarriff 2007: 147), this remains the most widespread method of using the web as a corpus. Yet it cannot be denied that «the argument that the commercial search engines provide low-cost access to the web fades, as we realise how much of our time is devoted to working with and against the constraints that the search engine imposes» (Kilgarriff 2007: 147-8). One really runs the risk of wasting time in becoming expert in the volatile syntax of search engines, or – as Kilgarriff suggests – of becoming a «gooleologist». Nonetheless a better knowledge of search engines and of the options they pro-

vide can not only help the linguist profit as much as possible from the most immediate way of access to linguistic information on the web, but can also contribute to a deeper understanding of the role each search option can play in helping the linguist elicit qualitative and quantitative evidence from the web, thus yielding further insight into the relationship between corpus linguistics and the web in more general terms. This is the reason why web search deserves some attention in the present work.

3.1. *An overview of web search options*

Even though apparently trivial, the task of searching the web for evidence of usage poses specific problems for the researcher, and requires that cautionary procedures are adopted both in submitting the query to the search engine and in interpreting the results. The user needs to take into account basic issues which are related to the peculiar nature of the web as a huge, dynamic, multilingual corpus, as well as other problems specific to our gateway to information on the web, the search engines. As to the latter point, the first step towards a competent use of search engines to access linguistic information is a deeper understanding of the search options they provide. A use of the web as a source for attestated usage is to some extent implicit in the query language of most search engines, which basically allows the user to look for web pages that contain (or do not contain) specified words and phrases, together with a rough indication of how many web pages satisfy these conditions. As Chakrabarti (2003: 45) explains, the simplest kind of query essentially involves a relationship beween terms and documents such as:

– documents containing the word X
– documents containing the words X and Y
– documents containing the word X but not the word Y
– documents containing the word X or the word Y

These conditions represent what is generally known as a Boolean search, which, in the context of web search, refers to the process of identifying those web pages that contain a particular combination of words. More specifically, Boolean search is used to indicate that a particular group of words must all be present (the Boolean AND operator) in the web pages retrieved by the

search engine; or that any word of a group is accepted (the Boolean OR operator); or that if a particular word is present in a web page that page is to be rejected (the Boolean NOT operator). While the explicit use of the Boolean operators AND, OR and NOT has been progressively downplayed in most search engines, and searchers may not even be familiar with them, Boolean search has been partially replaced in some engines by the use of menus in the advanced search mode or by a specific query syntax:

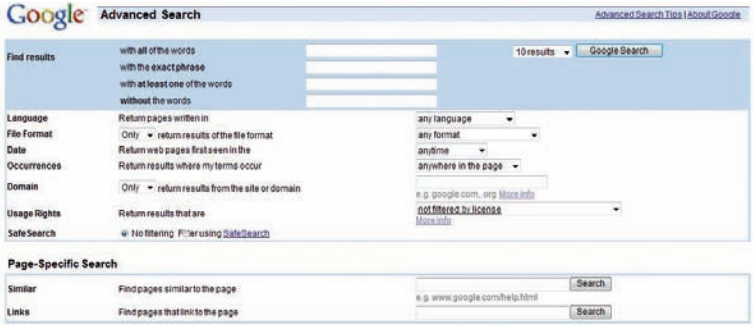


Fig. 2.1. Google's advanced search user interface.

Thus, by selecting «with all the words» from the pull down menu of a search engine you are implicitly using a Boolean AND; by selecting «at least one of the words» you are using a Boolean OR; by selecting «without the words» you are expressing the Boolean NOT. It is perhaps also useful to remember that the Boolean AND is by now a default in most search engines, so that whenever more terms are entered in the query box these are implicitly linked by an AND operator. Turning our attention more specifically to Google, one can also input the following symbols directly in the main query box:

+ (for the Boolean AND)
OR (for the Boolean OR)
- (for the Boolean NOT)

As the advanced search menu shows, Google, like most search engines today, also provides further search options: search for an

exact phrase, search for pages in a given language, or within a single domain (e.g. *.uk* or *.it* or *.org*), or within a specific time-span, and even restricted to a specific file type. As pointed out by Bergh, it is the very existence of these «slicing possibilities» that «accentuates Google's potential as a versatile tool for various forms of empirical language research» (2005: 34). From the linguist's point of view all these «conditions» can in fact be seen as a form of selection from among the wealth of texts that constitute the web and can to some extent be compared to the creation of a temporary sub-corpus relating to a specific language environment. This is the reason why understanding what is the effect of each single operator upon the web-corpus itself does not simply result in familiarity with the working of web search engines, or in a more effective exploitation of their potential, but can also give an indirect insight into the web-as-corpus question at a theoretical level.

3.2. *Simple search and the limits of «webidence»*

The evidence of attested usage that the web can provide has been labelled by Fletcher as «webidence» (2007: 36). In its most basic form this could entail a simple web search for a single word. A patently «low-tech» use of the web for linguistic purposes, a simple search for a single word is not however devoid of interesting implications, and provides a good starting point for exploring potential and limitations of the web as a source of quantitative and qualitative evidence. A case in point is spell-checking (Kilgarriff 2003: 332): by alternatively searching the web for two competing spelling forms, one can for instance come to the conclusion that the word hitting more matches is very likely to be the one spelled correctly, thus blending qualitative and quantitative evidence. An already widespread use of the web for linguistic reference, even if not directly connected with the corpus linguistics approach, using the web as a spellchecker can nevertheless highlight some typical problems arising when resorting to web search engines for linguistic purposes. The first problem obviously relates to the unknown size of the web, which makes it very difficult to interpret the relative «weight» of frequency of occurrence on the web. See for instance what happens if we input the word «himmunotherapy», half-guessing its spelling[4]:

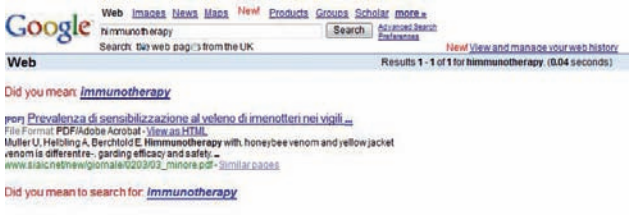


Fig. 2.2.

The misspelt word «himmunotherapy» finds only one match, while the search engine automatically suggests the correct spelling «immunotherapy», which in fact finds over 5 million matches:

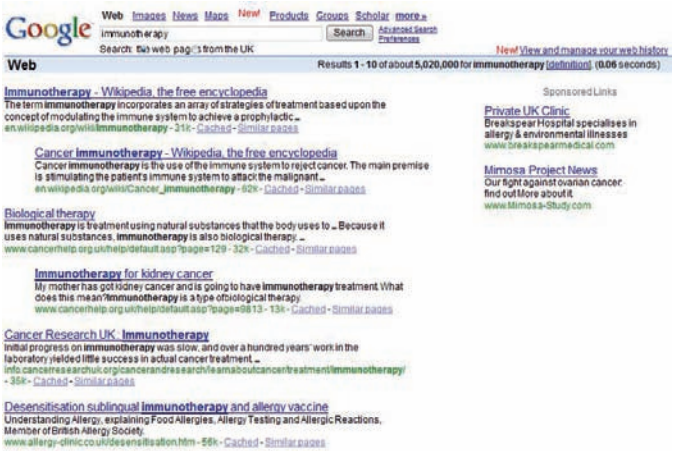


Fig. 2.3.

In this case, the significantly low number of matches for «himmunotherapy» could trigger doubts in the user's mind, and prompt checking strategies. In other cases, however, results yield no such evidence of unreliability of results, since the number of matches itself can be misleading. See for instance what happens with a search for the word «accomodation».

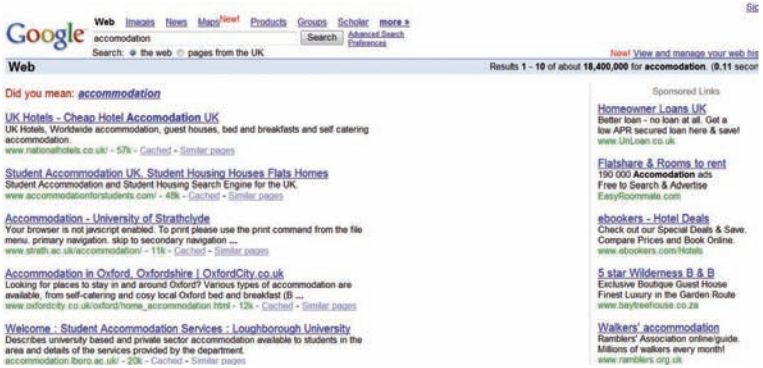


Fig. 2.4.

Over 18.400.000 matches cannot, at first glance, be considered as an invitation to further investigation, and the apparently high number could be simplistically taken as a sort of confirmation (webidence) that the word acco*m*odation not only does exist, but is also widely used, and can therefore be safely spelt with only one «m». As to the useful tip *Did you mean: accommodation?*, this is very likely to be disregarded if the user is not alerted to it. Checking results for the correct spelling «accommodation» shows instead that numbers are relevant only in relative terms, since the correct spelling «accommodation» finds 115.000.000 matches, a number about six times higher than the 184.000.000 found for the misspelt form! A more accurate investigation in the results provided by the first search, shows that the word «acco*m*odation» does not even appear highlighted (as is typical of search results) in the first four matches, which in turn contain non-highlighted correctly spelled forms. Despite clues undermining too easygoing acceptance of the misspelt form, further confusion for the user can come from the case of an official site (www.oxfordcity.co.uk) and an academic site (accommodation.lboro.ac.uk.) unexpectedly featuring an *Accomodation Homepage*.

Problems such as these are a warning that a lexical item's mere existence on the web can never be taken as sufficient condition to draw conclusions on usage, and that other parameters should be taken into account if willing to turn to the web as a source of qualitative/quantitative evidence. The web cannot be seen as a sort of

new *ipse dixit* from the linguistic point of view. Showing that a certain form is attested in a corpus – the web in this case – is one thing, but showing that it is linguistically acceptable is another. As far as the web as a corpus is concerned it is not the mere fact of «being there» that means something. This fact needs to be interpreted through an analysis of a number of other factors, such as the number of matches, the provenance of the results, the co-occurrence with other words, and the «authoritativeness» of each web page, as we will see in the following paragraphs.

3.3. *Advanced search as a solution to specific language questions*

Unlike searching the web-corpus for one single word, which does not seem to promise much to the linguist, searching for «phrases» is one of the most interesting possibilities offered by ordinary search engines. This is a very useful option because it can help the linguist exploit the web as a huge reservoir of attested usage especially for collocations and testing translation candidates.

3.3.1. *Collocation* The most common way to use phrase search for linguistic reference is to check whether a certain sequence of words actually occurs in a given language, as an indirect way to explore collocation. Defined by Firth as the tendency of some words to keep each other's company (Firth 1957: 11), collocation is one of the language phenomena more typically related to the use of language, and therefore one of the areas which even the most fluent speakers of a foreign language have difficulty in mastering completely.

As an area of language study which can profit from the possibility of testing stretches of language for attestation of usage, collocation is one of the most rewarding fields of application of the web as a corpus via ordinary search engines. A case in point is the example of «suggestive» as a potential collocate of «landscapes» in the phrase «suggestive landscapes». This phrase comes from an English tourism text written by an Italian speaker of English as a foreign language, and can be seen as representative of the «open choice» principle (Sinclair 1991: 109-110), according to which a language user may fall into the trap of considering any word virtually entitled to fill a slot in a text, provided that morpho-syn-

tactic constraints are observed. More specifically, this use of «suggestive» as a collocate for landscapes seems to be grounded on a common case of interference, or «shining through» of the source language into the target language (Teich 2003). While «suggestive» is not even mentioned in Italian-English bilingual dictionaries as a translation candidate for *suggestivi*, the latter is such a common collocate of the word *paesaggi* (meaning landscapes) in Italian, particularly in the context of tourism, that the English cognate word «suggestive» is very often mistakenly used as its direct equivalent. A web search for the phrase «suggestive landscapes» using Google's exact phrase match option[5] would seem to provide evidence of this:

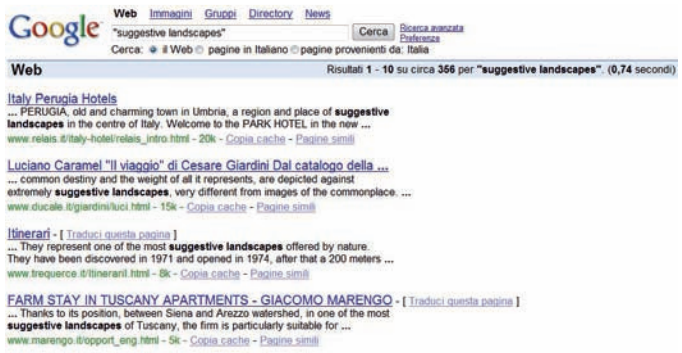


Fig. 2.5.

The problem then becomes how to interpret these results. In this case, 356 matches for «suggestive landscapes» could either be mistakenly considered as evidence of attested usage, especially if the user does not consider that 356 is quite a low number when compared with the size of the web in English, or instead provide support to claims about the unreliability of the web as a source of linguistic information. At closer inspection, the reliability of these results is called into question by their very provenance, almost invariably Italy, which suggests that the phrase is here very likely the result of a mistranslation from the Italian «paesaggi suggestivi». This could easily be checked by restricting the search to .uk only pages by means

of the domain restriction option provided by the search engine (see later in this chapter) which tellingly leads to the reported below results (Fig. 2.6) .

By slightly modifying the query the number of matches has dramatically fallen to three, two of which are again linked to Italy, thus providing clearer evidence of the inappropriateness of «suggestive» as a collocate for «landscapes».

Fig. 2.6.

3.3.2. *Testing translation candidates* In recent years, translation has increasingly become an «ideal field» (Bernardini 2006) for corpus application, and indeed this is one of the most interesting areas for exploring uses of the web as corpus via ordinary search engines. If, as it has been argued, «browsing» monolingual target language corpora throughout the translation process can help reduce the amount of shining through of the source language into the target language (Bernardini 2006), browsing the web for instances of attested usage could be a simple – yet not simplistic – way to test translation candidates on the basis of evidence provided from web.

A very interesting use of advanced search for linguistic purposes can be the evaluation of competing alternatives in terms of syntax within the noun phrase[6]. A case from the point of view of translation is the choice between premodification and postmodification, a potentially difficult one for a non-native speaker. The example discussed below, for instance, relates to the choice between «onset site» and «site of onset», as a translation candidate

for *sede di insorgenza*, in the context of a medical text about cancer. The preference for premodification or postmodification can of course be dependent on the register or genre of the whole text, a factor which could not be easily taken into consideration using the web as a corpus; nonetheless a preliminary exploration simply concerning frequency of usage can be carried out on the basis of data from the web. In this case mere figures (13,100 matches for «site of onset» vs. 836 matches for «onset site») would seem to favour postmodification:

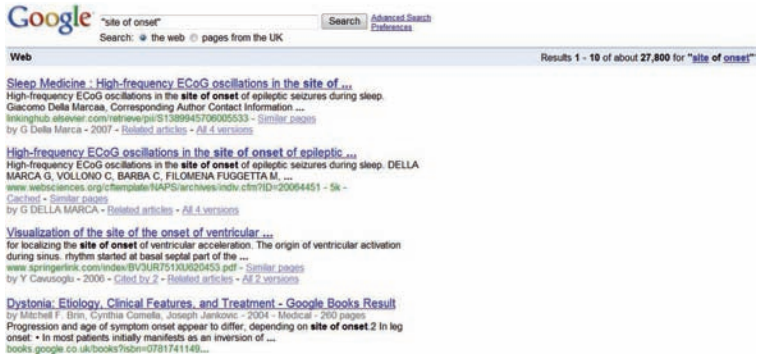


Fig. 2.7.

These results for «site of onset» do not give however clear evidence of the relevance of this phrase to the general topic of the text, i.e. cancer. Therefore Boolean search can be used heuristically to refine the query and enhance the possibility of hitting more relevant matches before jumping to conclusions concerning postmodification as the preferred mode to express this concept. By adding the word «cancer» to the search string it should be more likely that the results the engines gives back are from texts some way addressing this topic, and the new results seem indeed to confirm the hypothesis and prove to be definitely more relevant, providing new, clearer evidence of attested usage for «site of onset» in the context of «cancer»:

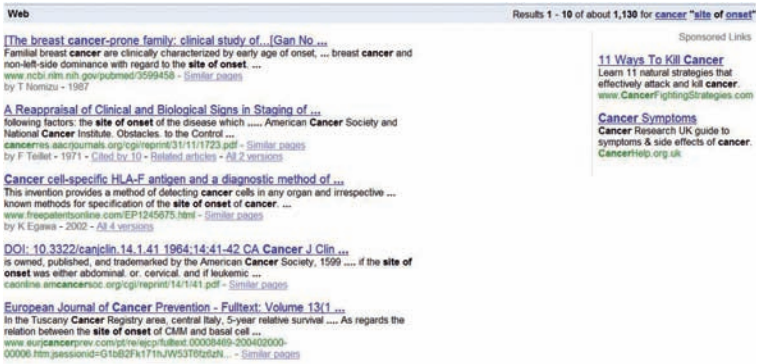


Fig. 2.8.

As to the alternative «onset site», the results for the search string **cancer "onset site"** are reported below:

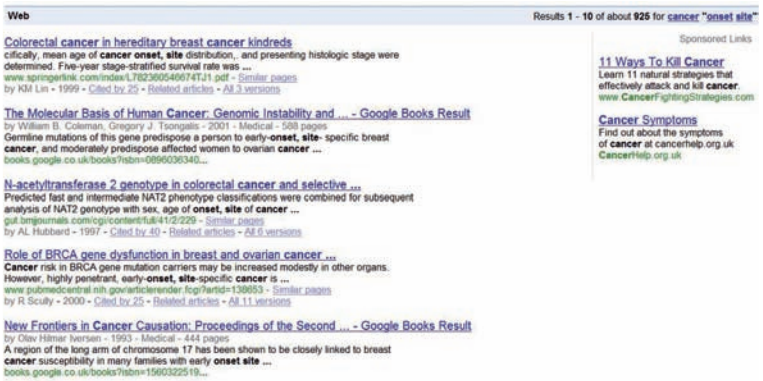


Fig. 2.9.

The low number of matches (213) seems to suggest that premodification may not be the preferred way to combine the two words. Moreover, it is of crucial importance to realize that in most of the occurrences reported the two nouns «onset» and «site» are separated by some punctuation marks, generally a comma. By quickly browsing through more result pages it becomes evident

that the majority of hits are for «onset, site» rather than for the noun phrase «onset site», thus further invalidating any assumption concerning attestation of usage for «onset site».

This last example provides a new opportunity to stress the importance of interpreting the results. Not only are the number of matches or the provenance of results of key importance, but also their relevance to the required context of use, and their reliability, i.e. the degree of precision with which they correspond to the user's query. It is equally important that the user remembers that not all tasks can be equally performed by search engines, which generally ignore punctuation and consider several variants as one (either by stemming or by ignoring special characters). This is what makes, for instance, an apparently similar use of web search engines aimed at discriminating between noun+noun constructions and noun+'s+noun constructions virtually impossibile (Lüdeling 2007; Rosenbach 2007) and calls for the necessity of tailor-made web search engines. Also worth mentioning is that the web has from time to time proved totally unreliable even with Boolean search. As shown by Vèronis (2005: online), Boolean logic is often defeated by totally absurd results for search strings including more than one item[7].

A further example concerns the evaluation of translation candidates for the Italian phrase «paesaggi aspri»[8]. In this case our starting point is the dictionary entry for the Italian «aspri» from one of the most prestigious and reliable bilingual dictionaries, which reveals at a glance the problem faced by the translator, even in an apparently trivial case such as this:

àspro, a. 1 (di sapore) sour; tart, bitter [...]; 2 (di suono) harsh; rasping; grating. 3 (fig. duro) harsh; hard; bitter; 4 (ruvido) rough; rugged [...]; 5 (scosceso) steep; 6 (di clima) severe; raw; harsh. • (ling.) [...] (Il Ragazzini 2006)

A first selection of translation candidates can be done on intuitive grounds, which makes it quite easy to exclude that the Italian word *aspri* could in this case be translated with «sour, tart» (relating to taste), or with «rasping» or «grating» (relating to sound). Among the translation equivalents offered by the bilingual dictionary the only sensible translation candidate would seem to be those

relating to points 3, 4 and 5 in the entry. More specifically, «harsh», «hard», «rough» and «rugged» seem to be the most plausible candidates. A fifth candidate can instead be obtained by introspection: searching for equivalence of effect the Italian «aspri» can in this case be replaced with another adjective (e.g. «forti») thus obtaining «strong» as a new translation candidate for «aspri». Choosing «landscape» and «scenery» as equivalents of «paesaggi», an evaluation of translation candidates for «paesaggi aspri» should therefore consider at least ten competing phrases, resulting from all possible combinations of «hard», «harsh», «rough», «rugged» and «strong» with both «landscape» and «scenery». While all formally possible, the ten combinations may in fact not be all actually used, and evidence for this can be drawn the web.

The following table reports results for the 10 competing alternatives. In order to improve chances of obtaining results both relevant and reliable, the query was progressively refined using both Boolean search and domain restriction. Thus the ten competing options were searched first through the whole web, and then by choosing only results from UK; finally, the words «travel OR tourism» were added to the query in order to enhance the possibility of retrieving pages related to the tourism sector. The table also reports data from the BNC as a contribution to an evaluation of web results:

Tab. 2.1.

	Web	Uk only	travel OR tourism	BNC
hard landscapes	2180	1440	116	0
hard scenery	485	43	2	0
harsh landscapes	826	441	193	1
harsh scenery	468	30	14	0
strong landscapes	507	173	18	1
strong scenery	138	6	1	0
rugged landscapes	66.300	10900	640	3
rugged scenery	51.700	12900	751	1
rough landscapes	1400	121	19	0
rough scenery	660	32	9	0

As the figures clearly show, web results outnumber results for the seven collocations which are not found in the BNC. By refining the query however there is a significant reduction of occurrences which seems to bring web data closer to BNC values. The only exception is for «rugged landscapes» and «rugged scenery» which are actually used in contemporary tourism discourse with increasing frequency despite low frequency in a general reference corpus such as the BNC. This seems to provide evidence of the capability of the web to capture changes in language use in – as the phrase goes – real time. In terms of the authoritativeness and reliability of the results, a quick glance at some of the web pages ranked higher in the results page shows that there are many edited sources, such as pages from the travel section of popular magazines (e.g. travel.guardian.co.uk) or published travel guides. These results also suggest that evidence of language use obtained by progressively refining the query towards greater complexity can be – despite all *caveats* – considered qualitatively reliable and quantitatively significant.

3.4. *Towards query complexity: other options*

3.4.1. *Language and domain* Among the opportunities offered by search engines to refine the user's query, the one most obviously related to a more specific use of the web for linguistic purposes is restriction by language. For most search engines results can be limited to one of over 40 languages, ranging from Arabic to Vietnamese, including Esperanto and Belarusian (see *Google advanced search*: online). This may not be however enough to boost the quality of the results, especially in the case of English, which is used by many non native speakers on the web. A different, and perhaps more useful, option is the provenance utility recently implemented by Google under the main query box:

Fig. 2.10.

A further, more flexible, option is restriction by domain. While the national top level domains (such as .it for Italy, .fr for France, .es for Spain, .ie for Ireland and so on) are no more than a «rough guide to provenance» (Fletcher 2007: 36), they can nonetheless contribute to the exploitation of the web as a multilingual corpus. Not only can domain restriction help lay bare phenomena like interference, as in the case of «suggestive landscapes» discussed above (see 3.3.1), but it can also provide quick access to parallel documents on the web. For instance, an Italian equivalent of an English word when a multilingual glossary is not available or does not contain an entry for it, can be found by asking a search engine to retrieve .it pages containing that word. See the following examples, reporting Italian pages featuring the English term «backscattered» and its translation as an adjective in the phrase «elettroni retrodiffusi».

Fig. 2.11.

While «the rough and ready regionally differentiated Google advanced mode search» has already started to provide the basis for research on change and variation in present-day English, as can be seen in studies by Christian Mair (2007: 233-247), the usefulness of domain restriction is not limited to national/regional domains. Also very useful is for instance searching only within academic domains in English speaking countries (such as .ac.uk and .edu), or within well known portals for the distribution of scientific journals, as an indirect way to control register. Thus in the search for «site of onset» and «onset site» carried out in the previous paragraph, further domain restriction to Elsevier, a well known portal for scientific publications, would have immediately

confirmed the unappropriateness of «onset site» (only one hit featuring, again, «onset, site» rather than «onset site»):

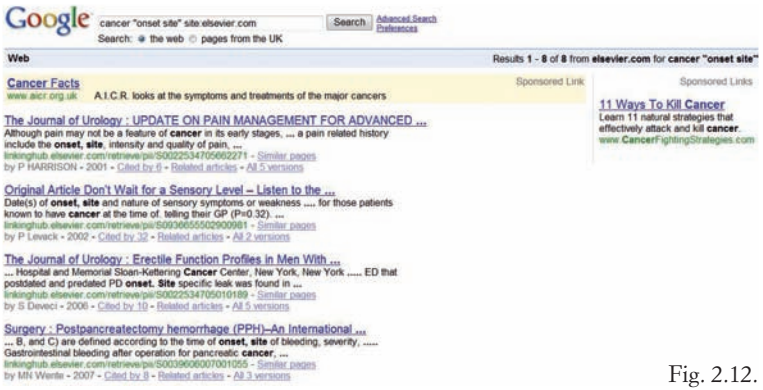


Fig. 2.12.

The alternative search for «site of onset» based on similar criteria produces 8 matches, all to be considered relevant and reliable because of the co-occurrence with cancer and of the «controlled» provenance of the results.

3.4.2. *Wildcards* Another meaningful option that can be exploited from a linguistic perspective is the search for an unspecified word in a certain position within a phrase. The unspecified word is represented by an asterisk «*», otherwise called a wildcard, in the search string. Unfortunately this option, referred to as «fill in the blank» in the advanced search tips provided by Google, is not fully supported by most search engines, including Google itself. Typically, one wildcard in the string should match one word only, so that multiple asterisks could be used as a sort of «proximity» search option. Google however has recently changed the processing so that one single asterisk does not necessarily match a single word. Despite lack of precision, however, the use of wildcards can be helpful not only for such tasks as checking phraseology and idioms (e.g. «a little neglect may * mischief» could be the string to submit to elicit «breed» if one is not certain about the last but one word), but also for highlighting areas of co-text which users may wish to

explore. This is particularly useful, for instance, when testing longer stretches of a text, or patterns, for evidence of attested usage. See what happens with the following translation of a medical text from Italian into English:

La frequenza del carcinoma a cellule squamose della mucosa orale è in rapido aumento; inoltre, il suo comportamento clinico è difficilmente prevedibile basandosi solo sui classici parametri istologici[9].

The apparently straightforward opening sentence of the source text poses some problems which cannot be solved only by reference to a specialized dictionary, and would rather benefit from access to a specialized corpus. The first clause, for instance, can be literally translated as «The frequency of squamous cell oral cancer is rapidly increasing», but a translator may have doubts relating to the phraseology. Do people really say that «the frequency of something is increasing»? Would people say this when talking about cancer?

To test this using the web as a corpus, the first step is to see whether the search string **"The frequency of * is increasing**" finds any matches. Almost 50,000 matches for such a long string seem indeed to be encouraging results, though it would be useful to refine the query by adding the word «cancer», to boost relevance, and by selecting known sites, e.g. *ac.uk* sites, for reliability. The new results seem to be only partially confirmatory: 9,000 hits for the search string **cancer "The frequency of * is increasing"** seem to suggest that the pattern is used in pages also containing the word cancer, but domain restriction to British academic sites results in a dramatic fall in the number of matches, whereas American academic sites still seem to provide evidence of attested usage:

cancer "The frequency of * is increasing" site: .ac.uk (4 matches)
cancer "The frequency of * is increasing" site: .edu (561 matches)

A further checking procedure is to change the position of the word «cancer» in the string, allowing for one word preceding it, given the high probability of cancer being referred to as a specific form

(breast cancer, lung cancer, etc.). The 3680 matches for the string **"the frequency of * cancer is increasing"** would seem again a confirmatory result, but restriction to pages from .ac.uk and .edu sites, as well as to reliable sources such as portals for the distribution of scientific publications (e.g. Elsevier or Pubmed) provides only 2 hits in both cases, which seems to suggest the opportunity of new checking procedures.

At this stage one could for instance test the pattern for an alternative to either frequency or increasing. A search for the string **"The * of * cancer is increasing"**, for instance, seems to suggest «incidence» as an alternative for frequency.

In fact **"The incidence of * cancer is increasing"** finds 22,900 hits in the whole web, 111 in sites ac.uk and 61 hits from one of the specific reliable websites (site: .pubmedcentral). This finally suggests that «The incidence of oral squamous cell cancer is increasing» as a suitable opening sentence for the target text.

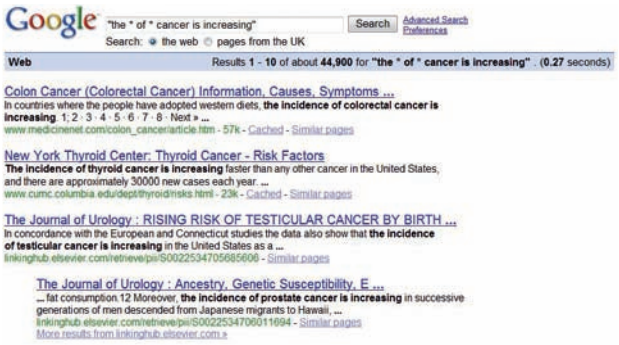


Fig. 2.13.

Similarly one could try to solve problems in the rest of the passage by checking other phrases for attestation of usage, such as «clinical behaviour», also testing its co-occurrence with the verb «predict» and «only on the basis of * parameters», along with a test for the phrase «histological parameters», and so on. By shifting a wildcard back and forth within the search string, by including and excluding dubious items, and by refining the query in

terms of domain, one can thus turn the web into a useful reservoir of attested usage for longer stretches of text and patterns.

4. *Query complexity: web search from a corpus perspective*

As the examples provided have hopefully shown, mastering the advanced search options offered by most search engines can really contribute to making the «webscape» a less anarchic and less inhospitable space for linguistic research. There seems in fact to be evidence that, in spite of its limitations, the web can be considered as a reliable source of quantitative and qualitative evidence of usage, provided that a few cautionary procedures are adopted in submitting the query and interpreting the results.

It remains doubtful, however, whether as web searchers linguists have actually learnt how to profit from the opportunities offered by more complex queries to enhance the relevance and reliability of their results. While at a theoretical level the World Wide Web has turned the search into a pervasive paradigm in our society, a «universally understood method of navigating our information universe» (Battelle 2005: 4), it seems that, at a more practical level, search by most Internet users – and linguists may be no exception – is still extremely naïve. Underestimating the role played by the query is to reduce the possibility of success. As Battelle suggests, «the query is the loadstone of search, the runes we toss in our ongoing pursuit of the perfect results» (2005: 27). And to the linguist the query can really be the place where the practice of web search and the linguist's theoretical approach to the web as a corpus can fruitfully interact.

The practical use that can be made of search engines' advanced options for linguistic purposes has been illustrated in the previous pages. A further step can now be taken by revising the role and meaning which web search can play from a linguistic perspective. Originally designed to enhance the power of search engines from the point of view of information retrieval, most options can in fact be seen as performing specific tasks that can be interpreted from the point of view of corpus linguistics. Even the very basic act of searching the web for a single word can be regarded as the instantaneous creation of a temporary finite subcorpus out of the virtually endless and incommensurable web-corpus. Push-

ing to the extreme Stubbs' idea that each single word or phrase «creates a mini-world or universe of discourse» (Stubbs 2001: 7), it can be argued that searching a word can be compared to the first step in corpus creation. It is as if, albeit only for a few moments, our virtually endless corpus, the web, complies with finite size, one of the fundamental criteria of corpus design. It will seem now obvious, for instance, how the search for the word «ecoturismo» would create a temporary virtual subcorpus from the web, all made up of texts written in Italian, someway or other relating to ecotourism. By contrast a search for the word «cancro» would not necessarily be as precise, and would in fact create a corpus of pages in Italian dealing with both the terrible disease and the zodiac sign. One should resort to the NOT operator (a minus sign «–» in Google query syntax) to refine the query by excluding references either to the horoscope (e.g. cancer –horoscope) or to the disease (e.g. cancer – patients – disease – treatment). Thus, while a search for a single word can be compared to the creation of a sort of sub-corpus, the search for two words (or more) can be read in terms of co-occurrence and contributes to the creation of a co-text for each search item. Similarly, the search for phrases, combined with the use of wildcards, can represent the search for collocates or patterns. Finally, language and domain restriction can indirectly be read in terms of constraints at the level of register or geographical variation. As highlighted in some of the examples provided so far, it is only by progressively refining the query towards greater complexity that linguists can contribute to improve the quality of the results. It is therefore on the process of refining the query that our attention can now be focussed.

The search for linguistic information from the web can be seen as a specific case of information retrieval. In this specific case the user's information need is related to language only and not relevant to the other activities which can be performed on the web, such as navigation and transaction. Thus the process involved in the creation of a complex query by the linguist can well be represented by the basic model for information retrieval as adapted for the web by Broder (2002: 4).

According to the basic model used in many standard information retrieval reference textbooks (e.g. Van Rijsbergen 1979) Information Retrieval can be represented as shown in figs 2.14 and 2.15:

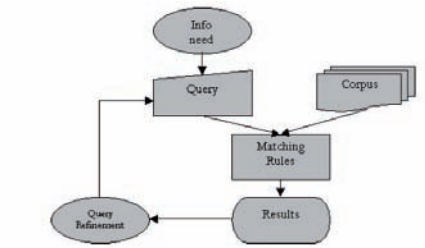


Fig. 2.14. The classic model for Information Retieval (Broder 2002).

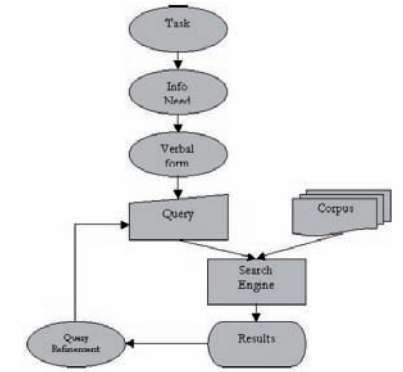


Fig. 2.15. The classic model for IR, augmented for the web (Broder 2002).

A user, driven by an information need, constructs a query in some query language. The query is submitted to a system that selects from a collection of documents (corpus) those documents that match the query as indicated by certain matching rules. A query refinement process might then be used to create new queries. When this classic model is adapted for the web the matching rules are provided by a search engine.

As Broder's model makes clear, the starting point is always an information need, which is generally associated with some task to be performed, and this need is verbalized and translated into a query posed to a search engine. By way of example, this basic algorithm will be now used to represent the procedure previously adopted to perform a specific task, i.e. the evaluation of translation candidates for the phrase «paesaggi aspri». In this case the user started from a

working hypothesis, in the form of a number of translation candidates, which in the query language of the search engine were expressed by the use of double quotes ("harsh landscapes", "hard landscapes") and so on. The query was submitted to the web/corpus and produced a different number of matches for each phrase. The user then proceeded to a refinement of the query by suggesting a co-occurrence with travel OR tourism and by restricting the search to .uk only sites, which got sharply different results for each phrase, thus giving indication of potentially more reliable translation candidates. A similar procedure was used to test «onset site» as a translation candidate for «sede di insorgenza» in the context of «cancer».

A less obvious use of query refinement can be aimed at devising complex queries capable of eliciting answers to a specific linguistic problem from the web as a corpus. Starting again from the already discussed example of «paesaggi aspri», we can see how a very interesting collocate for landscapes, as a translation candidate for «aspri», could be elicited by means of a single complex query that translates the linguist's need into the language of ordinary search engines.

In the case of «paesaggi aspri», the phrase was taken a tourism website dedicated to Sardinia[10]. The first step to be taken is therefore to recreate the context for this phrase in the search string, by asking the web search engine to give back pages containing the words travel or tourism (travel OR tourism). Then we can go on assuming «landscapes» as a *prima facie* translation for «paesaggi»: since our translation problem relates to the search of a collocate for landscapes in the English language, the search string will include the phrase **"* landscape"** (with a wildcard in the place of the adjective). Finally, given that the source text is about Sardinia, the query will search for pages also containing the words **Sardinia OR Sardegna**. A further step is to filter out linguistically unreliable pages by selecting only pages registered as .uk, which are more likely than others written by native speakers. The resulting search string is the following:

travel OR tourism "* landscapes" Sardinia OR Sardegna site: .uk

Here are the first five results out of the 556 retrieved for this search string.

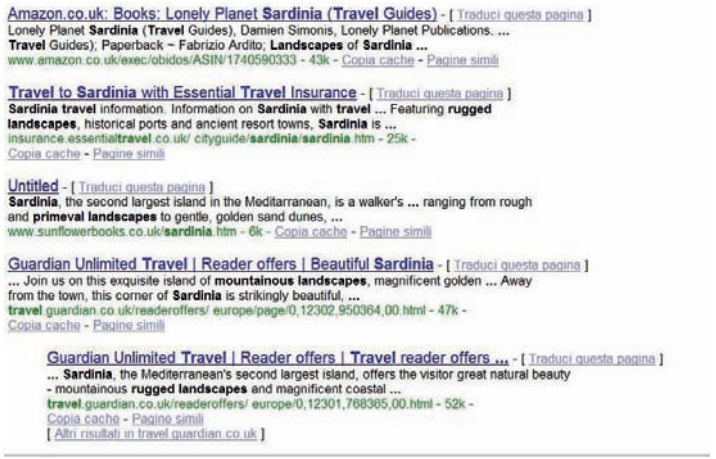


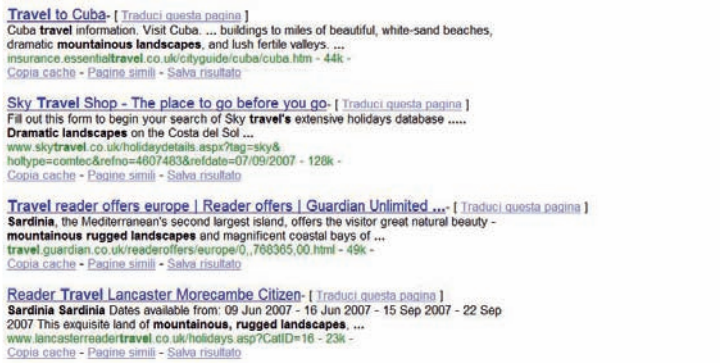
Fig. 2.16.

A quick glance at the results reveals a very interesting collocate for landscape in the phrase «rugged landscapes», which is a good translation equivalent for the source text's «paesaggi aspri», and a stunning result for more than one reason. Firstly «rugged» is a very appropriate adjective in this context and it is one that would not come naturally to the mind of a non-native speaker; evidence for this can be sought for again from the web by checking the frequency of «rugged landscapes» or «rugged scenery» in .it sites:

"rugged scenery" + tourism site: .it (7 matches)
"rugged landscapes" + tourism site: .it (4 matches)

Secondly, this specific case has also provided interesting results in terms of reproducibility. By submitting the same search string to a search engine twice more at a year interval similar results have been found, which seem to reinforce the idea that «rugged» is an adjective typically used by English native speakers

in a description of the Sardinian landscape somehow connected with tourism discourse. It is worth pointing out that the results obtained in subsequent searches for the same string all feature «rugged landscapes» among the results ranked higher by the search engine, but these are by no means in the same web pages. In fig. 2.17 are the results from a more recent search (September 2007), where, apart from a couple of irrelevant results (Cuba and Costa del Sol), Sardinian landscapes are almost invariably referred to as «rugged» and «mountainous», especially in the phrase «mounatainous rugged landscapes»:


Travel to Cuba- [Traduci questa pagina]
Cuba **travel** information. Visit Cuba. ... buildings to miles of beautiful, white-sand beaches, dramatic **mountainous landscapes**, and lush fertile valleys. ...
insurance.essentialtravel.co.uk/cityguide/cuba/cuba.htm - 44k -
Copia cache - Pagine simili - Salva risultato

Sky Travel Shop - The place to go before you go- [Traduci questa pagina]
Fill out this form to begin your search of Sky **travel's** extensive holidays database .....
Dramatic landscapes on the Costa del Sol ...
www.skytravel.co.uk/holidaydetails.aspx?tag=sky&holtype=comtec&refno=4607483&refdate=07/09/2007 - 126k -
Copia cache - Pagine simili - Salva risultato

Travel reader offers europe | Reader offers | Guardian Unlimited ...- [Traduci questa pagina]
Sardinia, the Mediterranean's second largest island, offers the visitor great natural beauty - **mountainous rugged landscapes** and magnificent coastal bays of ...
travel.guardian.co.uk/readeroffers/europe/0,,768365,00.html - 49k -
Copia cache - Pagine simili - Salva risultato

Reader Travel Lancaster Morecambe Citizen- [Traduci questa pagina]
Sardinia Sardinia Dates available from: 09 Jun 2007 - 16 Jun 2007 - 15 Sep 2007 - 22 Sep 2007 This exquisite land of **mountainous, rugged landscapes**, ...
www.lancasterreadertravel.co.uk/holidays.asp?CatID=16 - 23k -
Copia cache - Pagine simili - Salva risultato

Fig. 2.17.

Comparable results were obtained by submitting the same query to another search engine (www.alltheweb.com):

This seems to provide evidence of the fact that, in English, when thinking of the Sardinian landscapes one of the adjectives most likely to come to the speaker's mind is «rugged». While it is obvious that replicating the same search at fixed interval (or using different engines) is a dubious way to claim standards of «reproducibility» and «verifiability» for research carried on the web as corpus, the value of such results in term of a confirmation of the methodology adopted cannot be altogether dismissed.

The examples reported still represent however a very limited use of the web for corpus research, and should rather be seen as special cases representing the many opportunities offered by the web as a

source of linguistic information, which can be seen as only partially connected with the corpus linguistics approach as a whole. «Google linguistics», or rather «Googleology», as Kilgarriff has recently dubbed the practice of using web search engines for linguistic purposes, still remains «bad science» (2007). For specific corpus linguistics oriented tasks the web is an unsuitable and inhospitable corpus, if accessed through ordinary search engines only. To exploit its potential, the best thing to do is to turn to linguistically oriented tools, as we will do ourselves in the following chapters.

Conclusion

The issues discussed and the examples reported in this chapter suggest that the web can be used as a «quick-and-dirty» source of linguistic information, at least for specific tasks, provided that one is well-equipped to face the challenge posed by the web's «anarchy», and by the limits of ordinary search engines. It is indeed of crucial importance, when accessing the web as a corpus via ordinary search engines, that cautionary procedures are adopted not only in interpreting the results, but also in submitting the query. The complex query thus becomes the place where the corpus linguistics approach – as a way to conceive of language and not only as a method for investigating it – and the common practice of web search can significantly interact. It could be argued that it is by virtue of such interaction, and not so much in its own right, that the web can claim a status as a corpus despite so many obvious shortcomings.

Note

[1] Particularly for Machine Translation, Word Sense Disambiguation, Prepositional Phrase Attachment, the idea of using large text collections as an alternative (or complement) to sophisticated algorithms has become increasingly popular. A pioneer study by Grefenstette tested the possibility of using the web as a corpus to improve the performance of example-based machine translation and set up a model for further research (Grefenstette 1999). More recently, studies by Keller and Lapata have found a high correlation between the number of page hits found by a search engine (web frequency) for a given group of words and the frequency of the same group in a standard reference corpus like the BNC, as well as between web frequency and human plausibility judgments. This supports the hypothesis that web frequency can be used as a baseline for many Natural Language Processing tasks including machine translation

candidate selection, spelling correction, adjective ordering, article generation, noun-compound bracketing, noun compound interpretation, countability detection and prepositional phrase attachment (Keller-Lapata 2003; Lapata-Keller 2004). Research by Nakov and Hearst has confirmed encouraging results, with particular reference to the use of web counts for noun-compound bracketing and interpretation (Nakov and Hearst 2005b).

[2] According to Broder (2002) an «informational» need can be defined as the search for information assumed to be available on the web in a static form, so that no further interaction is predicted except reading. A «navigational» need is represented by the attempt to a reach a particular site that the user has in mind – either because it has been already visited or because the user assumes that such a site exists. A «transactional» need is the search for a site where further interaction (e.g booking a hotel or buying a book or downloading a file) can take place.

[3] For a simple overview of how search engines work see Hock 2007. A simple and detailed account of the way Google works can be also found in Bergh (2005: 29-34).

[4] When not otherwise stated, all Google searches where carried out in September 2007.

[5] These data are based on research carried out in May 2005.

[6] Tasks of this kind are common in Natural Language Processing (NLP) where web texts have started playing a major role to improve automatic parsing by offering data which may help the decisional process. Two typical examples are noun compound bracketing and prepositional phrase attachment (Volk 2002; Calvo and Gelbukh 2003; Nakov-Hearst 2005b).

[7] For further details on this topic see also the thread *Problems with Google Counts* in «Corpora List» (2005)

[8] Again, this is a typical NLP task using the web as corpus. One of the pioneer studies on the web as a corpus was in fact a study by Grefenstette who first used the web to improve the performance of example-based Machine Translation (Grefenstette 1999). His case study was based on the evaluation of translation candidates for the French compound *groupe de travail* into English. Starting from five translations of the word *groupe* and three translations for the word *travail* into English fifteen potential candidates for the translation of the compound were hypothesized. Only one, however, proved to have high corpus frequency (i.e. work group) both in a standard reference corpus and in the web, and this was therefore taken as the best translation candidate.

[9] The author wishes to thank Dr. Lucio Milillo for allowing her to quote from his Ph.D. thesis in Clinical Dentistry: L. Milillo, *Il ruolo della laminina-5 nel carcinoma orale: diagnosi, patogenesi, terapia*, Tesi di Dottorato di Ricerca Internazionale Multicentrico, Università di Bari, A.A. 2003-2004

[10] http://www.marenostrum.it/turismo-vacanze-sardegna/concerti-sardegna.html.

Chapter III

Webcorp: the Web as Corpus

Introduction

This chapter introduces WebCorp, one of the tools devised to make the web more useful for linguistic research. Thanks to a linguist-friendly user interface, WebCorp makes it easier to formulate linguistically useful queries to search engines (Lüdeling *et al.* 2007: 16) and returns results which are already tailored for linguistic analysis. Despite some limitations, depending primarily on the system's exclusive reliance on ordinary search engines and a rather limited storage/processing performance, WebCorp has already proved an excellent tool to obtain data for linguistic purposes, especially in a teaching context (Kübler 2003), and in the context of research on neologisms, rare or obsolete terms, and phrasal creativity (Renouf *et al.* 2007).

Section 1 and 2 provide background information on the tool and briefly comment on its technical features. Section 3 reports classroom activities based on the use of WebCorp that show how the tool can be used not only to obtain information otherwise requiring longer and more complex research activities, but also to offer students thought-provoking data to prompt classroom discussion and shift their attention from language to society and culture.

1. *Beyond ordinary search engines*

Ordinary search engines provide immediate but admittedly limited access to the enormous potential of the web as a ready-made corpus, and it is precisely such limitations that have prompted increasing interest in the development of specific tools and methods aimed at making the web a more hospitable place for linguistic research. A

number of projects facing the challenge posed by the *anarchism* of the web and by the limits of search engines as a gateway to linguistic information on the other are thus in progress. Such projects interpret in different ways the umbrella phrase «web as/for corpus», depending on the kind of access they provide to web data, the degree of dependence on existing commercial search engines, the stability and verifiability of results, and the flexibility and variety of the linguistically-oriented processing options offered. In this context, a useful distinction has more specifically been drawn between those tools which work as «intermediaries» between the linguists' needs and the information retrieval services already available on the web (pre-/post-processing systems), and tools which try to dispense with ordinary search engines completely, by autonomously crawling the web in order to build and index their own corpora (Lüdeling *et al.* 2007: 16). One of the most remarkable achievements in the former category is WebCorp (Kehoe and Renouf 2002), to which the present chapter is specifically devoted[1].

2. *WebCorp: using the web as a corpus*

Designed by the Research and Development Unit of English Studies (formerly at the University of Liverpool, now at the University of Birmingham) and available as a free service on the Internet, the WebCorp project (www.webcorp.org.uk) was established in the late '90s «to test the hypothesis that the web could be used as a large 'corpus' of text for linguistic study» (Morley 2006: 283). Today it is perhaps the most famous web concordancer, i.e. a suite of tools which provides contextualized examples of language usage from the web in a form tailored for linguistic analysis.

Searching a word or phrase using WebCorp is not in principle different from searching the web through an ordinary search engine, and the system's user interface is in fact very similar to the interfaces provided by standard web search tools. (Fig. 3.1.)

Strikingly different, however, is the format of the result page, which is presented to the user in the so-called Key-Word-In-Context (KWiC) format familiar to linguists, with the chosen word aligned and highlighted as «node word» within a context of between 1 and 50 words to the left and to the right. Neither the content of the web nor the search engine functions are modified, but

Fig. 3.1. WebCorp user interface.

only the output format, so that the result page of an ordinary search engine like Altavista or Google is transformed into a concordance table that can be immediately used to explore web data from a corpus linguistics perspective as in the table reported below:

Tab. 3.1. *A sample from Webcorp concordances for «landscape»*

ac.uk/ Sapling: architecture, planning	landscape	information gateway Web sites are
For larger scale site planning nformation	landscape	architects also use geographic
capable of altering the political	landscape	. The voting system broke down
significant changes in the political	landscape	appears to have little direct
Includes prehistoric and pre-Hispanic	landscape	design. ENVI SB477 M6 K57

Such a result is achieved through the simple architecture represented in fig. 3.2.

As the graph clearly shows, the starting point is the WebCorp user interface, which receives the request for linguistic information. The linguist's query is then converted into a format acceptable to the selected search engine, which finds the term through

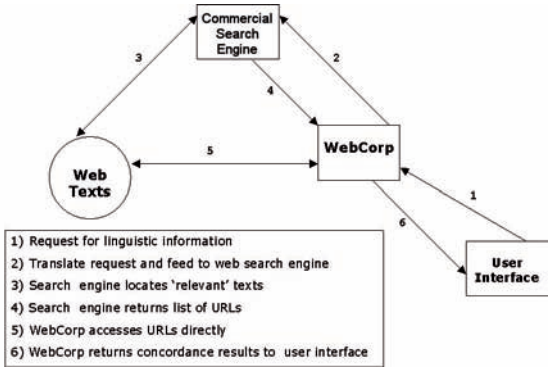


Fig. 3.2. Diagram of current WebCorp architecture (Renouf *et al.* 2007).

its index and provides a URL for the relevant text. The system temporarily downloads the text, extracts the search term and the appropriate linguistic context, collates it, and presents it to the user in the desired format (Renouf 2003; Renouf *et al.* 2005).

WebCorp can thus be seen as doing no more than adding «a layer of refinement to standard web search» (Kehoe and Renouf 2002), by framing some of the advanced options of commercial search engines into a new linguist-friendly environment, pre-processing the user's query before it is submitted to the search engine, and finally post-processing the results. The advantage of WebCorp lies therefore in the possibility it offers for a deeper linguistically-oriented exploitation of ordinary web search, which becomes particularly evident when turning to the system's Advanced Search Option interface. (Fig. 3.3.)

While some of the system's options clearly match the corresponding options offered by ordinary search engines (e.g. domain restriction, directory search), it is evident how WebCorp «makes it easier for linguists to formulate linguistically useful queries to search engines» (Lüdeling 2007: 16). Thanks to specific pre-processing functionalities the linguist's requests are in fact translated into complex queries (e.g. the search for a pattern is translated into a query containing wildcards), while post-processing functionalities of specific interest to linguists (e.g. KWiC format, computation of collocates, exclusion of stopwords, case sensitiveness)

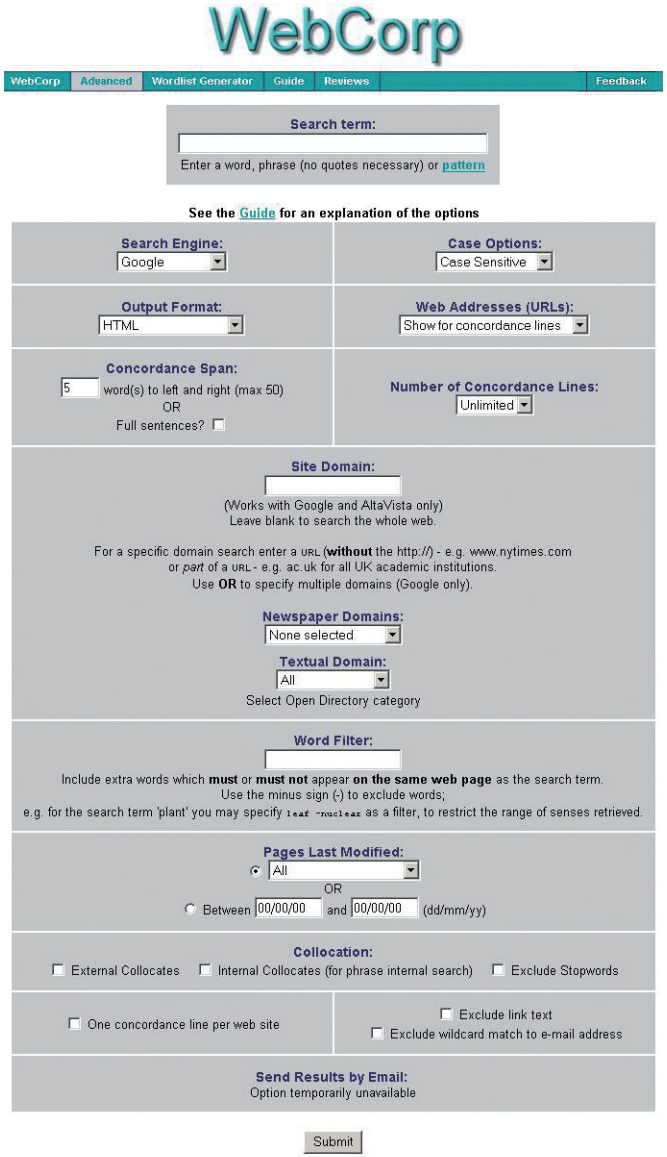


Fig. 3.3. WebCorp Advanced Search Option Interface.

transform the results into data similar to data obtained through a concordancer from conventional off-line corpora.

It is certainly beyond the scope of the present work to go into further technical details and to survey all the options offered by WebCorp – information that can be easily obtained from the tool's guide and in the growing body of research articles published in the past few years[2]. By way of example, here is a short list of its most important features, based on recent publications describing the system (Renouf et. al 2005; Morley 2006; Renouf *et al.* 2007). Key features include:

– the possibility of preselecting a site domain as an indirect way to specify language variety;
– a choice of 4 newspaper site groups (UK broadsheet, UK tabloid, French news, US news), to allow specification of register;
– a choice of textual domain based on the Open Directory categorisation, to control language register and probable topic range;
– a selection of data-subset according to last date of modification;
– restriction of the number of instances of an item to one per site, to avoid domination and skewing of results by one author or source;
– concordance filtering, so that the user can control which concordance lines will be processed by removing irrelevant concordances or duplicates;
– sorting left and right co-text;
– keyword extraction;
– removal of non-linguistic content, such as URLs, isolated hyperlinks, e-mail addresses and other distracters;
– use of a word filter, to improve recall or precision in search results, by allowing or suppressing particular words occurring in the same text as the main search term.

Despite such an extensive range of functions, WebCorp is nonetheless also characterized by some limitations. While ordinary search engines are able to process millions of search string matches, WebCorp is limited to treating results from a limited number of pages for reasons of processing speed (Bergh 2005: 28). This means that the proportion of potentially relevant web texts that is actually searched can be too low, and that recall can accordingly be rather poor (Renouf *et al.* 2007: 57). Moreover the system lacks the degree of processing and storage performance

which is required to meet the needs of its prospective users (Renouf *et al.* 2005), especially in the case of simultaneous use by more people, a condition which is to be considered a default for online tools. Finally, as a system subject to the technology of commercial search engines, WebCorp also suffers from typical limitations of web search such as ranking according to algorithms which the user cannot control; presence of duplicates in results, which need to be discarded manually; unreliable word count statistics; limited and/or inconsistent support for wildcard search.

Notwithstanding the above limitations, the WebCorp system has already proved an excellent tool to process web data for linguistic analysis, especially in the teaching context, where its user-friendliness and ease of access have made it a valuable resource from the start (Kübler 2004). Moreover, specific case studies concerning neologisms and coinages, rare or possibly obsolete terms and constructions, as well as phrasal variability and creativity, have also shown that in many cases the web can be a unique source of linguistic information which a tool like WebCorp can exploit to the full (Renouf *et al.* 2005; 2007).

By way of example an analysis of linguistic information obtained through classroom activities is reported below to demonstrate how using the web as a ready-made corpus through Webcorp can immediately improve students' language awareness, and also provide the basis for further explorations.

3. *WebCorp in the classroom: the case of English for tourism*

In recent years increasing emphasis has been placed on corpus linguistics approaches to language teaching, especially with reference to translation and LSP discourse (Tognini Bonelli 2001; Bowker-Pearson 2002; Laviosa 2002; Zanettin *et al.* 2003; Sinclair 2004). Drawing on such seminal notions as «data-driven learning» and «discovery learning» (Johns 1991; Bernardini 2002), it can be argued that also using the web as a ready-made corpus through a simple tool like WebCorp can result in an extremely rewarding learning experience, which can be easily reproduced outside the classroom context.

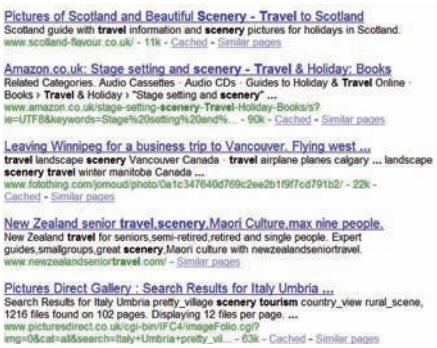
The following pages report classroom activities carried out with undergraduate students of English for Tourism in the A.Y. 2004-5 at the University of Bari. Although largely dependent on the teaching context, the choice of language for tourism proved a good starting point for more than one reason. This variety seems indeed to be one of the fields of enquiry where the web can be profitably used as a corpus – or where, at least, its uses as a corpus can be easily tested. The tourism industry has actually been a leader in the field of e-commerce for several years (Werthner – Klein 1999), with figures constantly on the increase, so that many acts of communication and economic transactions take place over the Internet. This suggests that the language of tourism available on the web can be considered reasonably «representative» for this specific domain.

3.1. *From «scenery» to «some of the most spectacular scenery»: exploring collocation and colligation*

This case study starts from an investigation of the collocational profile of the word «scenery» in the context of tourism discourse. To help the students appreciate the specific kind of linguistic evidence offered by WebCorp in this case, the warming-up phase for the activity consisted in the analysis of information on the word «scenery» derived from dictionaries with which the students were already familiar, such as the *Oxford English Dictionary* and the *Collins Cobuild English Dictionary*. Students observed that dictionary definitions perfectly explain the meaning of the word but provide limited information in terms of usage, even though, as a corpus-based dictionary, the *Collins Cobuild* suggests some typical phraseology and common collocates. Then the students were introduced to some basic corpus linguistics principles, before being shown the collocates for «scenery» provided by the *Oxford Collocations Dictionary*.

After commenting on the list of collocates provided by the dictionary, students were invited to use WebCorp as an alternative or complementary source of linguistic information. Before turning specifically to the tool, however, they were given the opportunity to consider what general information could be retrieved from the web through ordinary search engines such as Google and Altavista. Even using advanced search options and by specifying not only language (English), but also provenance (UK) pages, and imposing co-occurrence with the words «travel OR tourism», results obtained from

Google were not particularly encouraging, pointing instead to the shortcomings of the web as such for linguistic purposes:


Pictures of Scotland and Beautiful Scenery - Travel to Scotland
Scotland guide with travel information and scenery pictures for holidays in Scotland.
www.scotland-flavour.co.uk/ - 11k - Cached - Similar pages

Amazon.co.uk: Stage setting and scenery - Travel & Holiday: Books
Related Categories. Audio Cassettes · Audio CDs · Guides to Holiday & Travel Online · Books › Travel & Holiday › "Stage setting and scenery" ...
www.amazon.co.uk/stage-setting-scenery-Travel-Holiday-Books/s?ie=UTF8&keywords=Stage%20setting%20and%... - 90k - Cached - Similar pages

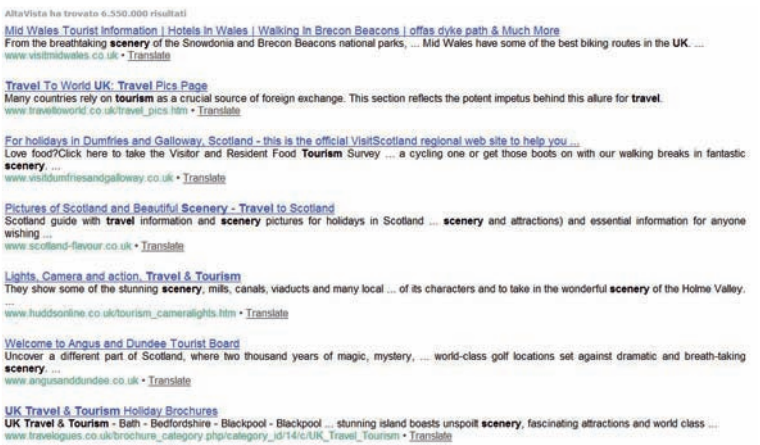
Leaving Winnipeg for a business trip to Vancouver. Flying west ...
travel landscape scenery Vancouver Canada · travel airplane planes calgary ... landscape scenery travel winter manitoba Canada ...
www.fotothing.com/jomoud/photo/0a1c347640d769c2ee2b1f9f7cd791b2/ - 22k - Cached - Similar pages

New Zealand senior travel,scenery,Maori Culture,max nine people.
New Zealand travel for seniors,semi-retired,retired and single people. Expert guides,smallgroups,great scenery,Maori culture with newzealandseniortravel.
www.newzealandseniortravel.com/ - Similar pages

Pictures Direct Gallery : Search Results for Italy Umbria ...
Search Results for Italy Umbria pretty_village scenery tourism country_view rural_scene, 1216 files found on 102 pages. Displaying 12 files per page. ...
www.picturesdirect.co.uk/cgi-bin/IFC4/imageFolio.cgi?img=0&cat=all&search=Italy+Umbria+pretty_vil... - 63k - Cached - Similar pages

Fig. 3.4.

None of the results produced by the search engine in this case seems to be relevant or particularly reliable, neither seemed the information they provide any useful from a linguistic point of view. Using a different engine did not result in better data. Here are, for instance, the results obtained from a similar search through Altavista:


AltaVista ha trovato 6.550.000 risultati

Mid Wales Tourist Information | Hotels in Wales | Walking in Brecon Beacons | offas dyke path & Much More
From the breathtaking scenery of the Snowdonia and Brecon Beacons national parks, ... Mid Wales have some of the best biking routes in the UK. ...
www.visitmidwales.co.uk • Translate

Travel To World UK: Travel Pics Page
Many countries rely on tourism as a crucial source of foreign exchange. This section reflects the potent impetus behind this allure for travel.
www.traveltoworld.co.uk/travel_pics.htm • Translate

For holidays in Dumfries and Galloway, Scotland - this is the official VisitScotland regional web site to help you ...
Love food?Click here to take the Visitor and Resident Food Tourism Survey ... a cycling one or get those boots on with our walking breaks in fantastic scenery. ...
www.visitdumfriesandgalloway.co.uk • Translate

Pictures of Scotland and Beautiful Scenery - Travel to Scotland
Scotland guide with travel information and scenery pictures for holidays in Scotland ... scenery and attractions) and essential information for anyone wishing ...
www.scotland-flavour.co.uk • Translate

Lights, Camera and action. Travel & Tourism
They show some of the stunning scenery, mills, canals, viaducts and many local ... of its characters and to take in the wonderful scenery of the Holme Valley. ...
www.huddsonline.co.uk/tourism_cameralights.htm • Translate

Welcome to Angus and Dundee Tourist Board
Uncover a different part of Scotland, where two thousand years of magic, mystery, ... world-class golf locations set against dramatic and breath-taking scenery. ...
www.angusanddundee.co.uk • Translate

UK Travel & Tourism Holiday Brochures
UK Travel & Tourism - Bath - Bedfordshire - Blackpool - Blackpool ... stunning island boasts unspoilt scenery, fascinating attractions and world class ...
www.travelogues.co.uk/brochure_category.php/category_id/14/c/UK_Travel_Tourism • Translate

Fig. 3.5.

Nonetheless it was evident that some of these pages could still be appropriate candidates for inclusion in an *ad hoc* corpus made up of web texts relating to tourism. To obtain relevant linguistic information concerning usage, however, it would have been necessary to go through all the basic stages of corpus compilation (even in the quick-and-dirty formula of Do-It-Yourself and disposable corpora put forward in Zanettin 2002; Varantola 2003) in order to explore the resulting corpus using specific tools. It is at this stage that students were given the option of using the WebCorp system.

After considering the meaning of each option in the system's advanced search interface, the students submitted a query for the word «scenery», also asking the system to return only pages including the words «travel OR tourism» (word filter) and from *.uk* sites (site domain). Here are the first few entries in the first page of the WebCorp output:

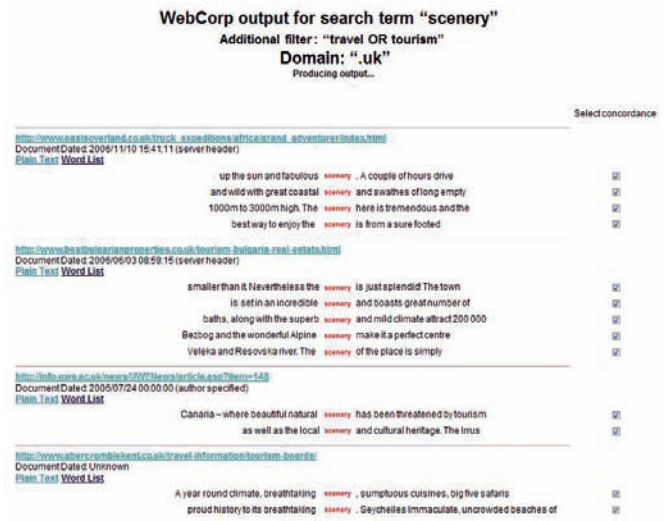


Fig. 3.6. A sample from WebCorp output for «scenery».

As the reported sample shows, students were this time presented with results displayed in clear KWiC format. A number of interesting collocates can be identified at a glance (e.g. fabulous, superb, breathtaking...) while the number of matches (258) seem

to provide a basis of manageable size for further considerations on the linguistic *behaviour* of the word «scenery». The concordance table is only a mouse-click away from the original web page, since each node word is a hypertextual link to the page itself, and reports such useful information as URL, date and even a link to a word list for each page. This makes it very easy for students to check the original webpage for relevance and reliability, and to discard irrelevant/unreliable results. Finally, the concordances produced can easily be re-sorted according to left/right co-text by means of a special button at the end of the page, to provide evidence of different patterns (See Appendix 1)

Beyond mere concordancing, however, the most important property of WebCorp is the possibility of producing a collocational profile for the node word, which is reported immediately after the concordance table. In this case it is crystal-clear to students that in texts including the words «travel OR tourism», taken – for the sake of greater, although by no means absolute, reliability – only from .uk sites, the word «scenery» is often accompanied by such words as beautiful (23), spectacular (22), stunning (19), breathtaking (13), dramatic (10), magnificent (7), as the table below clearly shows:

Top external collocates of "scenery" (excluding stopwords)

Word	Total	L4	L3	L2	L1	R1	R2	R3	R4	Left Total	Right Total
Beautiful	23	1	2	3	14	2	1			20	3
spectacular	22	2		4	15		1			21	1
stunning	19	2		5	12					19	0
wildlife	15	2	1	4		1	5	1	1	7	8
Coastal	15				14			1		14	1
Natural	14	2	4	1	5		1	1		12	2
breathtaking	13				12		1			12	1
History	12		1	3	1	4	1	2		5	7
mountain	11		1		7		1	1	1	8	3
Scotland	10	2	1	1		1	4		1	4	6
dramatic	10		2	2	6					10	0
Scottish	9		1	1	6			1		8	1
Variety	9	1	1	4			2	1		6	3
Best	9	4			4				1	8	1
Finest	8			1	7					8	0
beaches	7		2	2			1		2	4	3
Take	7	3	2				1	1		5	2
magnificent	7				6	1				6	1
Wild	6	1		1	2	1		1		4	2
Beauty	6	1	2	2					1	5	1

Key Phrases: spectacular scenery coastal scenery beautiful scenery stunning scenery breathtaking scenery finest scenery mountain scenery magnificent scenery Scottish scenery dramatic scenery wonderful scenery scenery history

Fig. 3.7. Webcorp table of collocates for the word «scenery».

These results were considered by students as immediately useful, in so far as they provided them with a wide a range of adjectives typically used in the description of scenery which they could compare with data gained from introspection, and with the data from the *Oxford Collocations Dictionary* previously considered. As to the reliability of these results, it is even more striking to consider the similarity that these data obtained from the web in a few minutes through WebCorp bear with the results from the BNC for the collocates of «scenery» in the subcorpus of *Miscellaneous* writing[3]:

Tab. 3.2. *Collocates for scenery from the BNC and from WebCorp output*

BNC	**WebCorp**
Adjectives in the immediate co-text (+5/-5 words) of «scenery» in the sub-corpus *Miscellaneous*	Search for: «scenery» filter: travel OR tourism domain: .uk excluding stopwords
Beautiful	Beautiful
Spectacular	Spectacular
Dramatic	Stunning
Breathtaking	Coastal
Magnificent	Breathtaking
Stunning	Dramatic
Coastal	Magnificent

This seems to suggest that however chaotic and anarchic, the web can provide, at least in specific situations, linguistic evidence which is comparable to evidence obtained from a conventional reference corpus. This finding could enhance confidence about the possibility of moving a further step towards the exploration of the different modifiers accompanying the word scenery. When for instance is «dramatic scenery» more suitable than «breathtaking scenery» or «stunning scenery»? How are these phrases used in the discourse of tourism? How can phraseology be further explored starting from such data?

Taking as an example «spectacular scenery», quite a common phrase in the language of tourist promotion, it is easy to see how

much students could learn by simply exploring the concordance lines produced for this phrase (for which one would find only 20 occurrences in the BNC). With nearly 180 concordances out of 200 pages, WebCorp in fact provides in fact enough data for a rewarding exploration. Here is an overview of the kind of linguistic information retrieved for this phrase in classroom activities by the students[4]:

1. verbs most frequently accompanying the phrase «spectacular scenery» are «boast» (8) and «enjoy» (10), pointing to phraseology of this kind:

as well as **boasting** a	**spectacular scenery**	of coastal walks, towns, beaches
Boasting award-winning beaches, unspoilt and	**spectacular scenery**	and a rich historical heritage
boasts some of the most	**spectacular scenery**	and few places can compare
out **enjoying** some of Devon's	**spectacular scenery**	is now on offer to
earth and **enjoy** the most	**spectacular scenery**	,breathtaking views, golden beaches, majestic
mention. In this area of	**spectacular scenery**	, you can **enjoy** walking, climbing

2. a «spectacular scenery» is something you appreciate best by walking rather than by driving (*walk** 12; *driv** 4);

3. it relates both to the coastal areas (*coast** 10, *island* 4, *cliff* 4, *beach* 15, *sand* 3, *sea**4), and to *mountain* or *countryside areas* (*rock** 3, *mountain** 10, *valley* 4, *lake* 3, *river* 1, *loch* 3, *park* 4);

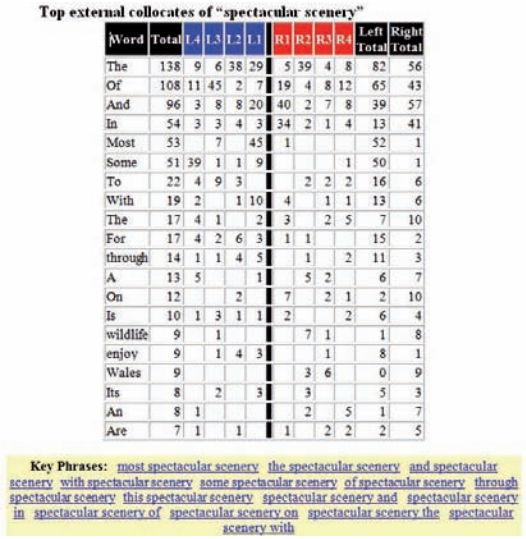
4. it has less to do with *cities* (0) than with *towns* (5) and *villages* (5)

5. it relates to a world of *unspoilt natural life* (*wildlife* 9, *unspoilt* 4, *natur** 5)

6. it also comprises *histor** (6) and *heritage* (5)

7. it evokes *variety*: *varied* (4), *divers** (3)

This is obviously only a fraction of the insight into usage that a detailed analysis of web concordances for «spectacular scenery» yields. Information of this kind could be complemented with exploration of the colligational profile of this phrase, obtained again with the help of the WebCorp system. By including function words (or «stopwords») in the count of collocates, for instance, the students could get a quite clear picture of what happens in the immediate co-text of our phrase in terms of colligation:



Top external collocates of "spectacular scenery"

Word	Total	L4	L3	L2	L1	R1	R2	R3	R4	Left Total	Right Total
The	138	9	6	38	29	5	39	4	8	82	56
Of	108	11	45	2	7	19	4	8	12	65	43
And	96	3	8	8	20	40	2	7	8	39	57
In	54	3	3	4	3	34	2	1	4	13	41
Most	53		7		45	1				52	1
Some	51	39	1	1	9				1	50	1
To	22	4	9	3			2	2	2	16	6
With	19	2		1	10	4		1	1	13	6
The	17	4	1		2	3		2	5	7	10
For	17	4	2	6	3	1	1			15	2
through	14	1	1	4	5		1		2	11	3
A	13	5			1		5	2		6	7
On	12			2		7		2	1	2	10
Is	10	1	3	1	1	2			2	6	4
wildlife	9		1				7	1		1	8
enjoy	9		1	4	3			1		8	1
Wales	9						3	6		0	9
Its	8		2		3		3			5	3
An	8	1					2		5	1	7
Are	7	1		1		1		2	2	2	5

Key Phrases: most spectacular scenery | the spectacular scenery | and spectacular scenery | with spectacular scenery | some spectacular scenery | of spectacular scenery | through spectacular scenery | this spectacular scenery | spectacular scenery and | spectacular scenery in | spectacular scenery of | spectacular scenery on | spectacular scenery the | spectacular scenery with

Fig. 3.8. Table of collocates for «spectacular scenery» including stopwords.

With the help of this table the students could find out by themselves that:

– the most frequent function word occurring to the left of spectacular scenery is «most» (45), followed by «the» (29), and then by «and» (20);

– «and» is twice as frequent in the immediate right co-text (R1 position) of «spectacular scenery» than in the left co-text (L1 position), which suggests a preference for the pattern «spectacular scenery and X» rather than «X and spectacular scenery»;

– «spectacular scenery» is very often followed (and seldom preceded) by «of» or «in», which are almost invariably followed by place names;

– other fairly frequent words preceding «spectacular scenery» are «through» and «with», pointing to such phrases as «through spectacular scenery» and «with spectacular scenery;

– «spectacular scenery» is used frequently in the pattern «some of the most spectacular scenery», as the table clearly shows when not only the frequency of colligates but also their position is considered.

In this specific case, exploring colligation proved to be of crucial importance. Unlike collocation, which students quite easily understand and whose immediate utility they readily acknowledge, colligation seemed in fact at first a less appealing concept to them. As however it turned out, the table quickly produced using WebCorp could really help the students *see* to what extent lexis and grammar constitute a unified whole, so that words and grammatical structures tend to co-select each other, and this indirectly contributed to their language competence in more general terms.

3.2. «*Dramatic landscapes*» *and* «*paesaggi suggestivi*»: *from collocation to semantic preference and beyond*

In this second example the starting point for the classroom activity was the comparison between the phrase «dramatic landscapes», a recurring phrase in English for tourism which is not always confidently used by learners, and the Italian phrase «paesaggi suggestivi». Both phrases are widely used, but none seems to have a direct equivalent in the other language, since neither «paesaggi drammatici» as an equivalent for «dramatic landscapes» nor «suggestive landscapes» as an equivalent for «paesaggi suggestivi» would sound as fluent to a native speaker. On intuitive grounds, however, the hypothesis was put forward that the two phrases could be used in similar contexts. It was therefore one of the aims of the comparison to see if, and to what extent, «dramatic landscapes» and «paesaggi suggestivi» could be considered as functionally equivalent.

As far as «dramatic landscapes» (for which one finds only 3 occurrences in the BNC) is concerned, the list of collocates produced by WebCorp seems to point to a marked preference for co-occurrence with words relating to historical and cultural heritage, such as «history», «architecture», «past», «ancient».

Here is an overview of further linguistic information retrieved by students for this phrase exploring all the concordance lines produced by the system:

– the phrase «dramatic landscapes» is related both to nature (*nature* 7, *coast** 6, *beach* 5, *mountain* 9), and to culture and history (*history* 10, *culture* 8, *architectur** 6, *heritage* 5)

Top external collocates of "dramatic landscapes" (excluding stopwords)

Word	Total	L4	L3	L2	L1	R1	R2	R3	R4	Left Total	Right Total
history	12	4		3	2	1	2			9	3
architecture	10		1	2			7			3	7
villages	9	1		6			1		1	7	2
views	9		2	1				1	5	3	6
past	9	6		1			1	1		7	2
stunning	8	2				1	5			2	6
beaches	8				1		3	4		1	7
	8	1	1	2		1	1	2		4	4
punctuating	6							6		0	6
people	6	1		5						6	0
000-year-old	6							6		0	6
quaint	6		6							6	0
Scotland	5	1		4						5	0
scenery	5	2		1			1	1		3	2
natural	5		3					1	1	3	2
rich	5	2	1				2			3	2
ancient	5							4	1	0	5
year-round	4		3					1		3	1
mountains	4						1	1	2	0	4
close-up	4							4		0	4

Fig. 3.9.

– it relates to a world of the past (*ancient* 6, *past* 5, *old* 2, *remote* 2)

– it evokes variety and contrast (*varied/variety* 4, *divers** 3, *combin** 4, *blend* 2, *mix* 2).

A similar analysis was then carried out on paesaggi suggestivi. It is in fact an obvious, and valuable, consequence of WebCorp's dependence on ordinary search engines that concordances can be produced for virtually any language available on the Internet.

Here are some of the data retrieved:

– the verbs most frequently accompanying the phrase «paesaggi suggestivi» are *offrire* 12, *regalare* 4, *ammirare* 3;

– the phrase has a semantic preference for both nature (*natura* 8, *coste* 4, *mare* 8, *monti* 4, *boschi* 3, *valli* 2, *vette* 2) and history (*stor** 12, *art** 5, *tradizion** 4, *memoria* ?);

– «richness» is another recurring semantic area, since *ricco/a di* and *arricchire* clearly emerge as recurring elements in the phraseology of «paesaggi suggestivi», which seems to indirectly point to the deep link between the undisputed wealth of historical and natural heritage in Italy, and the appeal of its landscapes and scenery.

At the level of colligation, some patterns seemed to emerge quite clearly. An interesting example is modalization through *consentire/potere* in co-occurrence with *ammirare/apprezzare*:

tratturi che consentono di ammirare	paesaggi suggestivi	senza causare inquinamento atmosferico ed
da dove si possono ammirare	paesaggi suggestivi	e concedersi tranquille passeggiate. Vicino
Europa, che permette di ammirare	paesaggi suggestivi	come il lago di Bolsena
appuntamenti che consentiranno di apprezzare	paesaggi suggestivi	e incontaminati, di gustare invitanti

or the frequent use of the preposition «dai» with the meaning of «with»:

costa del mar Adriatico dai	paesaggi suggestivi	, per non dimenticare dell'entroterra raccontata
flora e fauna e dai	paesaggi suggestivi	. Più a Sud si trovano
una delle zone d'Italia dai	paesaggi suggestivi	che la rendono tra le
ed attraversa zone selvagge dai	paesaggi suggestivi	dove vivono, nella natura intatta
sabbiose, coste e litorali dai	paesaggi suggestivi	e dal mare cristallino ed
flora e fauna e dai	paesaggi suggestivi	. Più a Sud si trovano

On the basis of their analysis the students concluded that certain similarities, especially concerning semantic preference (e.g. a tendency to co-occur with words relating to tradition, history and heritage), could support a relation of equivalence between the two phrases. Nonetheless the two items still displayed language-specific phraseology which suggest only partial coincidence. More specifically, the students noticed, «paesaggi suggestivi» seems to cover a wider spectrum, including features more typically associated with the phrase «spectacular scenery» or «breathtaking scenery». The task thus triggered further questions in the students, who decided to compare/contrast the Italian «paesaggi suggestivi» with such English phrases as «spectacular scenery/landscapes», «breathtaking scenery/landscapes», «dramatic scenery/landscapes», via further WebCorp searches.

In more general terms, it could be pointed out that apart from the learning outcomes outlined above, most students acknowledged the benefits of direct exposure to a large number of instances of authentic language use in a relatively short time, and in a learning context which could be easily replicated at home. This had indirectly resulted, in their opinion, in a feeling of greater familiarity with some aspects of this specific language variety. Moreover, as the students again acknowledged, the feeling of having taken part in the process of retrieving data from the web, rather

than being simply presented with off-line concordances, had significantly contributed to their involvement in the learning process, and accordingly, to its actual results.

3.3. *Not only scenery: the experience of tourists with disabilities*

As seen in the examples reported so far, WebCorp definitely represents a step forward in the attempt of exploiting the web's potential for linguistic ends, in so far as it is capable of transforming web data into an object amenable to analysis informed by the corpus linguistics approach. In many cases the results provided are however not only immediately useful but also, as has been argued, «thought-provoking» (Bergh 2005: 38). This was the case of the evidence provided by the concordance lines and collocational profile produced to prompt classroom discussion on the specific question of accessible tourism[5].

The starting point was WebCorp's collocational profile for the phrase «disabled tourists», as a specific group of people within the more general category of «disabled people»:

Top external collocates of "disabled tourists"

Word	Total	L4	L3	L2	L1	R1	R2	R3	R4	Left Total	Right Total
needs	12	2	2	8						12	0
information	12	2	2	6			1	1		10	2
access	9		4	5						9	0
facilities	9		4	5						9	0
visit	7	1					3	2	1	1	6
people	6						3	1	2	0	6
difficulties	6		3	3						6	0
travel	6	2	2	1					1	5	1
services	6			4		1		1		4	2
accessible	5		1	2					2	3	2
places	5				1			4		1	4
makes	5	1							4	1	4
improving	5	3	1						1	4	1
special	5	2	2						1	4	1
group	5			4				1		4	1
disabled	5		1					2	2	1	4
Information	4			4						4	0
service	4						2		2	0	4
offers	4	1		1			2			2	2
Tour	4		4							4	0

Fig. 3.10.

In this case, students could appreciate how the collocational profile provides evidence which is not only linguistically relevant

but also indicative in more general terms of all that is crucial to the holiday experience of the disabled. A key word in this respect is obviously «needs», which actually ranks first in the list of collocates. Other words which significantly occur very often in the immediate co-text of the phrase are «access/accessible», undeniably a key concept in the discourse of mobility for people with disabilities; and then «facilities»; «difficulties», «service/services», «special» and «group», all words which triggered students' reflections concerning the actual experience of the people involved. Interestingly, the second most frequent word immediately following «needs» is «information». It is indeed quite often the case that a specific need for people with disabilities in general, and for travellers/tourists with disabilities in particular, is obtaining precise information concerning facilities, options and services available in their holiday destination. Also significantly frequent in the immediate co-text of the phrase is the word «improving», which seems to point to a reality of work-in-progress in the field of accessible tourism.

As to the data revealed by a more detailed exploration of the concordance lines, it is also worth mentioning the recurring relation established between disabled tourists and other categories of citizens (elderly, older, senior unemployed, pensioners, working mothers, small children). As shown by the language data these are all people that share in some way or other the «special needs» of disabled tourists:

accessible for the **older and**	disabled tourists	. To give you a better
economic impact of **senior** and	disabled tourists	. Even that information was helpful
the area; the **elderly** and	disabled tourists	wanting to get the most
such as the **unemployed**, the	disabled tourists	, **pensioners**, etc. It appeared that

As to the colligational profile students noticed the occurrence of prepositional phrases introduced by «for» in patterns such as «access/information/service *for* disabled tourists», which represents «disabled tourists» mainly as beneficiaries/recipients rather than actors. Also frequent was the occurrence of prepositional phrases introduced by «of», almost invariably in phrases such as «the needs of disabled tourists».

Finally the data offered by the concordances also allowed some considerations about semantic prosody, closely related to an «unsatisfactory situation». This was made evident by a number of phrases containing negatives, such as:

unfriendly and **discriminating against**	disabled tourists	. The experience of disabled students
offered **less than nothing** for	disabled tourists	. But recently, our company worked
serious **lack of information** for	disabled; tourists	and that often disabled visitors
was still **largely inaccessible** to	disabled tourists	,» Hendi told Al-Ahram Weekly. «But
lack of information about where	disabled tourists	can visit, stay or eat
It is **unavailable to** the	disabled tourists	(inclined drifts, railways, stairs), The

In other cases, reference was simply to a situation of improvemente and/or work-in-progress:

Scotland **«can do more» for**	disabled tourists	MP points finger No support
constantly **improving** its facilities for	disabled tourists	and many places of interest
project to improve access for	disabled tourists	. The Heart of England Tourist
aim to improve facilities for	disabled tourists	By Soteris Charalambous PLANS are
strides in **improving** access for	disabled tourists	, so there's no reason to

Having analysed the data for «disabled tourists», students were invited to consider the alternative «tourists with disabilities», a phrase which, on the basis of corresponding «people with disabilities», has a wide currency in the discourse of accessible tourism. By examining the collocational profile of the corresponding phrase «tourists with disabilities», they noticed that apart from words such as accommodation, access, attract, inform provide, co-occurring with comparable frequency in the immediate co-text of both «disabled tourists» and «tourists with disabilities», the most interesting differences in the collocational profile seemed to be related to the absence of other categories such as senior citizens, families, elderly people, and to an increased frequency of terms relating to the provision of services/holidays for these travellers, no longer seen as members of a wider category sharing similar problems, but rather as specific customers, i.e. stakeholders in a specific economic activity:

Tourism that aims to attract	tourists with disabilities	from the **major world markets**
serve the **market segment** of	tourists with disabilities	. They operate basically in Cusco
for the **niche market** of	tourists with disabilities	, but his broadened market allowed

Thus, rather than providing the students only with answers, the activity had again triggered more questions. Why this difference in the collocational profile of apparently equivalent terms? What about issues of politically correct language in this field? Or, more specifically, when do people refer to «tourist with disabilities» and when is «disabled tourists» to be preferred? in which context? under which pragmatic constraints? These questions prompted further research by the students who, taking advantage of all the options provided by the system, went on refining their queries in an endless discovery journey.

Conclusion

The examples reported show that linguistic data obtained from WebCorp are tailored enough to meet specific needs on the linguist's part, thus confirming the hypothesis that web data can be a very useful resource for linguistic analysis. As a tool which requires no specific computer skills, which is extremely flexible and relatively quick in providing results, WebCorp proves particularly good in prompting classroom discussion on specific lexical items and in providing the starting point for data-driven and discovery learning activities. Although not exhaustive, the information obtained from web data in the context of the suggested classroom activities, does seem indeed to provide evidence that using WebCorp to produce quick *ad hoc* concordance lines can really contribute to students' awareness of specific language issues, and constitute the basis for a rewarding learning experience which they can easily repeat on their own.

In other respects, however, WebCorp remains a rather limited tool, which does no more than allow better exploitation of commercial search engines without removing their intrinsic shortcomings. It is precisely out of awareness of such limitations and «with an eye to the long-term sustainability of the WebCorp sys-

tem» (Renouf *et al.*: 58), that the Research and Development Unit of English Studies team at the University of Birmingham has been working in the past few years on the ambitious project of designing and assembling an independent linguistically-tailored search engine. Progress on this project can be followed at the following address: www.webcorp.org.uk/webcorp_linguistic_search_engine.html

Note

[1] Other renowned pre-/post-processing systems are KWiCFinder and WebKwic (Fletcher 2001) and the Linguist's Search Engine (Elkiss and Resnik 2004).

[2] The link «Publications» in WebCorp website is regularly updated with new publications.

[3] These results were obtained by accessing BNC through BYU-BNC interface (http://corpus.byu.edu/bnc). See Appendix 2 for a complete list of BNC results for this search.

[4] See Appendix 1 for the complete output produced by WebCorp.

[5] «Accessibile tourism» refers here to the specific market segment in the tourism industry addressing the tourism needs of people with disabilities.

Chapter IV

Bootcat: Building Corpora from the Web

Introduction

The present chapter introduces one of the most interesting tools devised in the attempt at making the web more useful as a corpus linguistics resource. Created as a suite of Perl programs freely available for download and further developed as a web service, BootCaT is a system capable of «bootstrapping», i.e. creating virtually ex-nihilo, specialized corpora and term lists from the web in a few minutes (Baroni and Bernardini 2004; Baroni *et al.* 2006). Section 1 introduces the tool as the natural development of the widespread practice of building Do-It-Yourself, «quick-and-dirty», disposable corpora (Zanettin 2002; Varantola 2003). Section 2 illustrates the compilation of a corpus made up of medical English texts on a specific topic (ORAL CANCER corpus), and discusses basic properties of the tool as well as the usefulness of the results for translation purposes. Finally, Section 3 reports data obtained with the creation of a second comparable corpus made up of Italian texts on the same topic (CANCRO ORALE corpus), showing to what extent the accessibility of the system and the relatively short time required for corpus compilation, make it an extremely useful tool for studies or tasks involving work with and across different languages. The examples reported in the present chapter, along with previous studies testing the tool in translation training and for terminology (Castagnoli 2006; Fantinuoli 2006) seem to provide evidence that the advantages of using BooCaT, especially in the context of specialized translation, largely outweigh potential limitations of the tool.

1. *BootCaT: the web as corpus «shop»*

While uses of the web as a corpus «surrogate» (Baroni and Bernardini 2006: 10) through a pre/post-processing tool like WebCorp undoubtedly represent a step forward in terms of exploitation of the web's potential from a corpus linguistics perspective, the system still displays some limitations typical of tools which mainly work on the output format of ordinary search engines and virtually download a temporary corpus for each searched item. It is therefore self-evident that using a tool like WebCorp to assist a language professional in a specific task requiring constant reference to one or more corpora becomes frustrating maybe when the task at hand requires repeated searches for several items.

A remarkable achievement in the attempt at making the web more useful as a corpus linguistics resource is BootCaT, which keeps the advantages of speed, size and topicality typical of the web while limiting some of its shortcomings. As its name promises, alluding to a well-known metaphor in the language of information technology[1], BootCaT is a suite of programs capable of creating virtually ex-nihilo specialized corpora and term lists from the web in a very short time. In its underlying «philosophy», the tool can be seen as the natural development of the widespread practice of building Do-It-Yourself, «quick-and-dirty» disposable corpora (Zanettin 2002; Varantola 2003), i.e. corpora created *ad hoc* from the web for a specific purpose, such as assisting a language professional in some translation task or in the compilation of a terminological database. As the creators of BootCaT have observed, such short-life corpora have indeed become basic resources for language professionals who routinely work with specialized languages (Baroni and Bernardini 2004). It is in fact often very difficult to find ready-made resources for highly specialized domains, and the very rate at which specific language domains grow, with new terms introduced virtually on a daily basis, seem to make standard reference corpora useless tools for tasks which must definitely rely on more focused and up-to-date text collections. On the other hand, the compilation of a web-based corpus through manual queries and downloads is notoriously an extremely time-consuming process and time investment of this kind, Baroni and Bernardini (2004) argue, is «particularly unjustified when the corpus which is

the final result of such effort is meant to be a single-use corpus», as is often the case with corpora created for a specific translation task, and not for a wider research project.

When creating a corpus from the web for a specific task, linguists generally query an ordinary search engine for a combination of search terms which are deemed relevant to the task at hand. In this case they take advantage of the options offered by the engine to focus the query, such as language or domain specification, selection of URLs, Boolean search, etc. (Pearson 2000; Zanettin 2002), and download the texts to create a small highly focused corpus to be explored with a concordancer. With BootCaT, rather then having the linguist manually querying the web, choosing relevant results to be included in the corpus, and finally performing the necessary format changes and archiving procedures, the whole process is automated by means of a suite of tools performing all these tasks together in a few minutes. It could be said, therefore, that the system has a bias towards *customization*, in the sense that it is primarily conceived as a tool helping language professionals build the corpus they need, whenever they need and as quickly as possible. It is this intrinsic feature that has perhaps suggested categorization of BootCaT under the label «web as *corpus shop*» (Baroni and Bernardini 2006: 11) by its creators. Certainly this is a very interesting feature from the point of view of its contribution to the changing face of corpus linguistics: by making the creation of *ad hoc* temporary corpora an easily achievable goal, BootCaT brings the reality of the web as a sort of virtual multilingual multipurpose corpus *on demand* a bit closer.

2. *WebBootCat and Medical English: the ORAL CANCER corpus*

Created by Baroni and Bernardini, BootCaT was born as a suite of Perl programs freely available for download at the Scuola Superiore di Lingue Moderne per Interpreti e Traduttori, University of Bologna website (http://sslmit.unibo.it/~baroni/bootcat.html). Despite extensive use for corpus creation, research on terminology and to assist translation tasks (Baroni and Bernardini 2004; Baroni and Ueyama. 2004; Sharoff 2006; Castagnoli 2006; Fantinuoli

2006), the tool was apparently not sufficiently user-friendly for non-technical people, since installing and running the program required a little more than basic computer system skills. In 2006, therefore, a new tool based on BootCaT was launched, WebBootCat, as «a web service for quickly producing corpora for specialist areas, in any of a range of languages, from the web» (Baroni *et al.* 2006). Through a clear web-based user interface, now available through the Sketch Engine website (www.sketchengine.co.uk), WebBootCaT has made the procedure of compiling and downloading disposable corpora from the web a really simple task. With the new web interface the user no longer needs to download or install any software, but rather uses the program which is installed on a remote server. The same server also keeps a copy of the corpus created by the user, which can be loaded into and analysed through a specific corpus query tool, the Sketch Engine (Kilgarriff *et al.* 2004), or downloaded in .txt format to one's own personal computer for analysis with other tools (e.g. Wordsmith Tools). In the following pages, the procedure used to compile a corpus of medical texts dealing with a specific disease («oral squamous cell cancer» or OSCC) is reported by way of example.

2.1. *From «seeds» to corpus: the bootstrap process*

The only thing WebBootCaT needs to start is a number of key words which the linguist considers particularly likely to occur in the specialized domain for which a corpus is going to be built. As already noticed in the present study, any choice of words can indeed be seen as evocative of «a mini-world or universe of discourse», as Stubbs (2002: 7) reminds us, and this is probably what triggered in the authors of the BootCaT system the idea that a handful of words could be enough to create from scratch, i.e. to bootstrap, a linguistic corpus focused on whatever domain required.

The words chosen to start the process are called «seeds» (Baroni and Bernardini 2004) and are transformed by the system into a set of automated queries submitted to an ordinary search engine. The search engine then retrieves and downloads relevant pages, post-processes them, and finally produces a corpus from which a new word list is extracted containing new terms to be used as seeds to build a larger corpus, and so forth.

In the present case study the compilation of a corpus on «oral squamous cell cancer» started from the four terms «oral», «squamous», «cell», and «cancer», which were used as seeds assuming that each term could to some extent be considered as a keyword for this specific domain:

Fig. 4.1.

As the user interface of WebBootCaT shows, all the linguist has to do is to key in the chosen seed terms, which are then randomly combined by the system and turned into Google query strings. The system automatically downloads as text (i.e. in .txt format) the top pages returned for each query (ten by default), to make up the first nucleus of the corpus. From this nucleus, a wordlist is created and a list of keyword terms is extracted, by comparing the frequency of occurrence of each word in the list with its frequency of occurrence in a reference corpus[2]. The keywords extracted are then turned into new seeds to be used in random combinations to build a larger corpus via more automated queries. This recursive procedure can be repeated several times, i.e. until the corpus reaches the desired size, though, as the system's creators suggest, two or three times is generally enough (Baroni and Bernardini 2004).

A key feature of BootCaT is that, although mainly automated, the process of corpus creation and term list extraction is clearly divided into different phases, allowing the user to interact with the system throughout the process. At each phase the user can in fact control several important parameters, such as the number of queries issued for each iteration, the number of seeds used in a single query, the number of pages to be retrieved. It is also possible to pre-view web pages that are going to be included in the corpus, and so exclude undesired pages before they are further processed. The latter is a particularly important option because it can really contribute to enhancing the relevance/reliability of the pages which finally make up the corpus. As the sample reported below shows, in the case of our ORAL CANCER corpus many of the pages selected in the first run came from.org or.gov sites, with some.com sites leading to web pages devoted to health information, and from portals dedicated to specialized journals such as PubMed (www.ncbi.nlm.nih.gov). These pages were considered as fairly reliable/relevant, while other pages required further inspection:

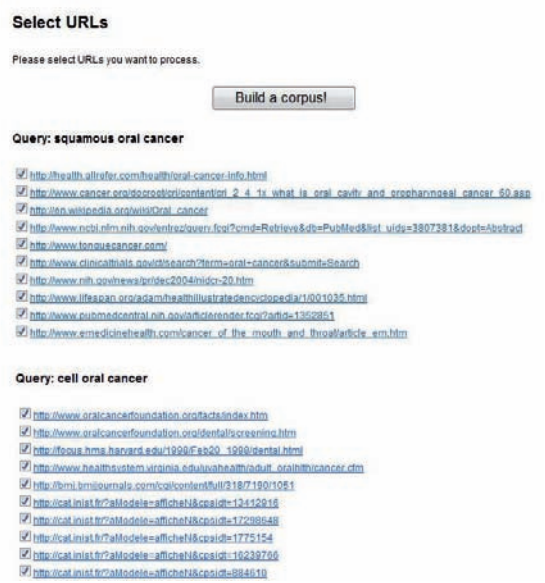


Fig. 4.2. A sample from the «Select URLs» page produced by WebBootCaT in the first run.

In the case of dubious or suspect pages, checking for relevance/reliability is quite easy because the original page is only one mouse-click away, so the user can have a quick look at it before deciding whether it should be included or excluded from the corpus. This was the case, for instance, with a number of results from the site http://cat.inist.fr which appeared at first non-convincing since their relevance to the topic and reliability in terms of language usage could not be easily guessed from the website address alone. By simply clicking on the link, however, it turned out that the address was that of a French portal for scientific information (Institut de l'Information Scientifique et Technique) leading to a specific journal article, which was both reliable and relevant, as the following example clearly reveals:

Fig. 4.3.

Once suspect links have been checked for relevance/reliability the process of corpus creation can start. It is not the purpose of the present study to go into further technical detail concerning

the pre/post-processing work «going on behind the scenes». It is perhaps useful, though, to consider at least the following key features (Baroni *et al.* 2006):

– the system uses the seeds to send a number of queries (ten by default) to Google, each containing a randomly selected triple of the seed terms;
– each query returns up to 100 hits, the top ten of which are taken by the system;
– the system also filters out very short (less than 5 kB) and very long (over 2MB) web pages on the assumption that these rarely contain useful samples of language;
– duplicate and near-duplicate web pages are deleted, while the remaining pages are further processed to filter out the so called boilerplate (HTML markup, javascript, navigation bars, etc.).

The decisive importance of the post-processing performed by the system can hardly be overemphasized, and will be readily acknowledged by anyone who has attempted to use the web as a corpus either through ordinary search engines or through a simpler tool like WebCorp. By filtering out duplicates and near duplicates and by excluding pages which, on the basis of size alone, can be assumed to contain little genuine text (Fletcher 2004b), the system does perhaps more, if not better, than the linguist manually can do, and all this in a shorter time. The result is a clean enough text collection which comes to the user in a few minutes as a basis for the iterative process:

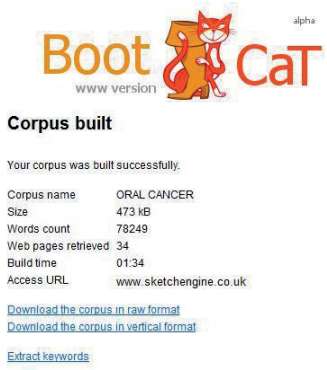


Fig. 4.4.

The table above reports information relating to the first run. The corpus compiled by the system can be downloaded or directly accessed through the Sketch Engine website. Clearly visible is the «Extract keywords» option: by clicking on the link the user is provided with a set of key terms that can be turned into new seeds, if considered appropriate. Here is a sample of single-word key terms extracted by the system from the provisional 78.000 tokens ORAL CANCER corpus after the first run:

Fig. 4.5. A sample from the list of single-word terms from the provisional 78.000 token ORAL CANCER corpus.

As the sample clearly shows, there are almost no terms which seem to be irrelevant or that could seriously impair the value of results in the following phases. On the contrary, most terms seem to be definitely relevant and suitable as new seed terms. Some were nonetheless deselected before running the process a second time, such as Roman numerals referring to stages of the disease, and the name of a specific portal for life sciences and biomedical bibliographic information.

With nearly one hundred new seeds, the system was ready to run again and produce a second corpus of 238,000 tokens, obtained in 4 minutes. This in turn provided interesting new key terms:

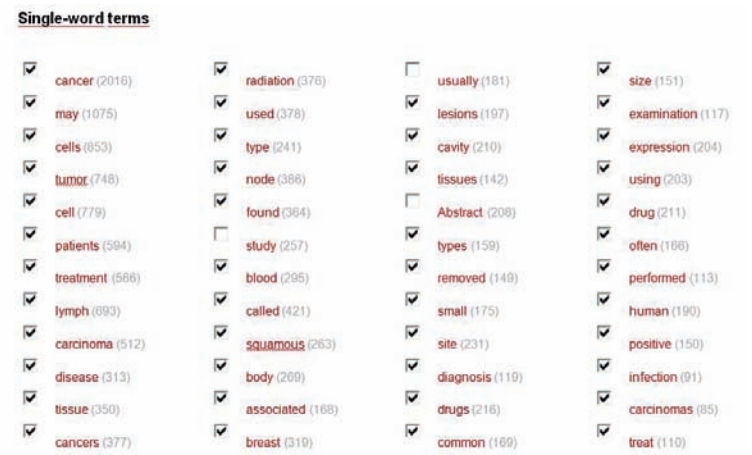


Fig. 4.6. A sample from the list of single-word terms extracted from the provisional 238.000 token ORAL CANCER corpus.

As the sample reported shows, key terms that were considered irrelevant or that could falsify the results were again deselected before running the process for the third time. This was the case, for instance, of such words as «study» or «abstract». While confirming that the corpus included a fair number of reliable texts such as scientific/academic articles, these terms were crossed out as no longer necessary to boost reliability/relevance in the following phases.

Besides extracting single-word terms, BootCaT is also designed for extracting multi-word terms. For the system's purposes multi-word terms are specifically defined by Baroni and Bernardini (2004) as sequences of words that:

– contain at least one of the terms extracted in the first phase;
– do not contain stop words; may contain connectors, (such as of, of the, and... but never at the edges nor adjacent to each other;have frequency above a certain threshold (dependent on length);
– cannot be part of longer multi-word terms orcontain shorter multi-word terms having similar frequency.

Here is a sample of the list of multi-word terms retrieved by the system in its second run:

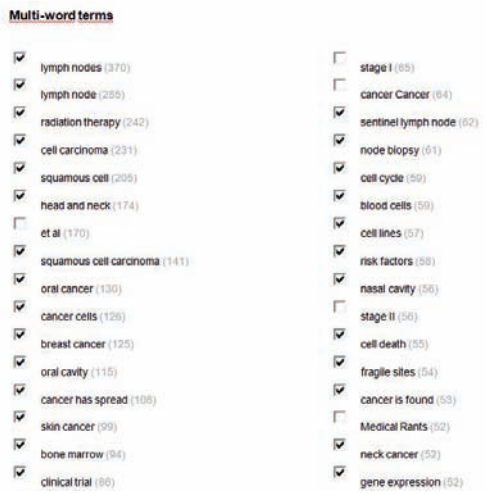


Fig. 4.7. A sample from the list of multi-word terms extracted from the provisional 238.000 token ORAL CANCER corpus.

The multi-word terms extracted appear not only relevant to the domain but also mostly well-formed, thus providing further evidence of the system's reliability. The list includes specific two- or three-word terms such as «bone marrow», «skin cancer», «breast cancer», «cell lines», «blood cells» and «sentinel lymph node» as well as easily recognizable collocations such as «risk factors» and clusters that point to distinct phraseology (as in the case of «cancer has spread» or «cancer is found»).

As in the previous phase, some terms were excluded before running the process again, for the third time, such as «et al.», another clue of the presence of texts belonging to the specific genre of the scientific article, «cancer Cancer» (an ill formed compound), «stage II» (which could result in unnecessary focus on one specific stage of the disease) and «Medical Rantal» (the name of a blog). All the single- and the multi-word terms selected by the user were then turned into new automated Google queries in order to complete the process.

Using the procedure described, a corpus of 444,231 tokens was built following a cyclical process, in three phases, taking less than 10 minutes in all. This was considered a large enough corpus

to allow a rewarding exploration of phraseology as samples from the list of single and multi-word terms reported below suggest:

Fig. 4.8.

Fig. 4.9.

Evidence for the usefulness of data so quickly and easily obtained from the web as a source of linguistic information is provided in the following paragraph.

2.2. *The ORAL CANCER corpus and translation practice*

Having explored the process of corpus compilation using WebBootCaT, we can now see to what extent the data obtained contribute to the solution of specific translation problems. To this end, the corpus was downloaded in .txt format for analysis offline with the Wordsmith Tools (Scott anno). It should be reminded however, that the corpus could have been also explored online using the Sketch Engine, the corpus query tool installed on the SketchEngine website, which currently hosts WebBootCaT.

The importance of using corpora for translation hardly needs to be explained, and virtually any aspect of translation theory and practice can benefit from the use corpora. More specifically, as far as specialized translation is concerned, corpora can be extremely useful for checking terms and collocations and identifying text-type specific forms (Ohlan 2004: 172). It is with reference to such tasks that our corpus was tested in the present work.

The basis for the case study was the translation of an International Ph.D. thesis in Clinical Dentistry[3]. Here is a sample passage from the Italian text:

> Il carcinoma squamoso (SSC) presenta una tendenza all'infiltrazione loco-regionale con l'interruzione della membrana basale e l'invasione del tessuto stromale sottostante fino al raggiungimento delle reti linfatiche ed ematiche.

In the excerpt reported above, some multi-word terms such as «infiltrazione loco-regionale» «membrana basale», «tessuto stromale» e «reti linfatiche ed ematiche», seemed at first to pose translation problems which could not be solved only on the basis of information provided by medical bilingual dictionaries, or by the many glossaries available on the Internet. A solution was therefore sought for in our ORAL CANCER corpus.

The first problem which reference to the corpus helped solve rather quickly was related to the term «membrana basale». The Italian adjective «basale» is generally translated as «basal» in English (*Dizionario Medico Italiano-Inglese Inglese-Italiano* Garzanti 1987; *Multilingual Glossary of technical and popular medical term*, online). The bilingual dictionary also reports the term «basilemma» for the compound «membrana basale». This suggests that the

only translation candidates for the term «membrana basale» are «basal membrane» and «basilemma». However, no instance of «basilemma» was found in our ORAL CANCER corpus and, out of the 146 occurrences for the word «basal», no instance was found for «basal membrane», «basal» being mainly used as a modifier for the word «cell», especially in «basal cell carcinoma» (see Appendix 3). By looking instead at concordances for «membrane», our corpus suggested «basement membrane» as a plausible equivalent for «membrane basale», including the phrases «infiltrate/invade/breaking through/spread through the basement membrane», which could all be considered as suitable equivalents for the Italian «con l'interruzione della membrane basale»:

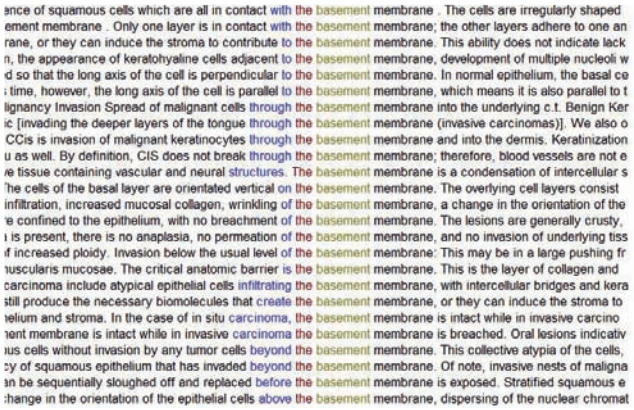


Fig. 4.10. A sample of concordances for «membrane» (L1-L2).

The very fact that such a large number of occurrences come from a corpus meant to be domain-specific suggests that we are faced in this case with specific terminology in the context of a specific topic. As to further evidence of attestation of usage, it was very useful to double-check the bilingual dictionary, which actually reported «basement membrane» in the English into Italian section, suggesting «membrane basale» as its Italian equivalent. It could be argued therefore that useful information was already there in traditional resources, but the translator could retrieve it only via further research through a corpus.

Reference to the corpus was also helpful in the search for translation equivalent for «tessuto stromale» and «reti linfatiche ed ematiche», which resulted in translation with «stromal cells» and«lymphatic and blood vessels», rather than «stromal *tissue*» «lymphatic and blood *networks*» on the basis of corpus evidence.

Less straightforward was finding a solution for the problems posed by «infiltrazione loco-regionale». The only instance of loco-regional found in our corpus was in fact not enough to support «loco-regional infiltration» as a translation candidate for «infiltrazione loco-regionale», nor was significant evidence for a different wording provided by the concordance lines for «infiltration». Analysis of the immediate right co-text of both «regional» and «local», however, highlighted the presence of words which could be considered as synonyms of infiltration (such as «invasion», «spread» and «metastasis»). The concordance of «local» in particular seemed to suggest «local spread» and «local invasion» as suitable translation candidates.

The English ORAL CANCER corpus was also used as a resource to improve the target text in more general terms. In the opening sentence of the quoted paragraph, for instance, the Italian «presenta una tendenza» is a typical example of phraseology which might lead the translator into the trap set up by false-friends. The word «presents» displays in fact patterns of usage, in our ORAL CANCER corpus, which only partially coincide with the Italian «presenta».

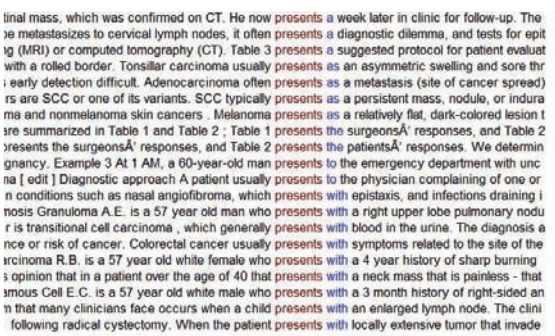


Fig. 4.11.

A solution was found therefore by exploring the left co-text of the word «tendency». In this case evidence from the corpus suggests «has» as a good translation equivalent for «presenta».

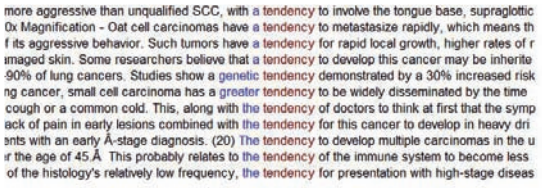

more aggressive than unqualified SCC, with a tendency to involve the tongue base, supraglottic
0x Magnification - Oat cell carcinomas have a tendency to metastasize rapidly, which means th
f its aggressive behavior. Such tumors have a tendency for rapid local growth, higher rates of r
imaged skin. Some researchers believe that a tendency to develop this cancer may be inherite
90% of lung cancers. Studies show a genetic tendency demonstrated by a 30% increased risk
ng cancer, small cell carcinoma has a greater tendency to be widely disseminated by the time
cough or a common cold. This, along with the tendency of doctors to think at first that the symp
ack of pain in early lesions combined with the tendency for this cancer to develop in heavy dri
ents with an early Â-stage diagnosis. (20) The tendency to develop multiple carcinomas in the u
r the age of 45.Â This probably relates to the tendency of the immune system to become less
of the histology's relatively low frequency, the tendency for presentation with high-stage diseas

Fig. 4.12.

Thus, on the basis of information obtained from our ORAL CANCER corpus, a first draft translation was revised into a more fluent and accurate target text:

Source Text:
Il carcinoma squamoso (SSC) presenta una tendenza all'infiltrazione loco-regionale con l'interruzione della membrana basale e l'invasione del tessuto stromale sottostante fino al raggiungimento delle reti linfatiche ed ematiche.

Target Text 1 (draft)
Squamous cell cancer (SSC) presents a tendency to loco-regional infiltration with the interruption of the basal membrane and invasion of the underlying stromal tissue, until reaching the lymphatic and blood networks.

Target Text 2 (revised)
Squamous cell cancer (SSC) has a tendency for local spread, breaking through the basement membrane and invading the underlying stromal cells, thus reaching the lymphatic and blood vessels.

Referring to the corpus was also extremely useful to find equivalents for specific terms pertaining to methods discussed in the Ph.D under translation. The phrase «colorazione con immunoistochimica», for instance, occurring 7 times in the source text, could be considered as highly specific terminology for which an equivalent could not be found in common references such as dictionaries and glossaries. As a term not necessary relating to the domain of «oral cancer», finding an English equivalent for it on the basis of

evidence provided by our ORAL CANCER corpus could not be taken for granted. Again, however, the corpus performed very well. A search for «immunoistochem*» yielded in fact immediate evidence of fair a number of occurrences (60), including «immunohistochemical staining» (11) which was then taken as a translation candidate for «colorazione con immunoistochimica»:

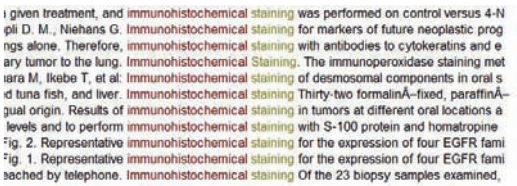

given treatment, and immunohistochemical staining was performed on control versus 4-N
pli D. M., Niehans G. Immunohistochemical staining for markers of future neoplastic prog
ngs alone. Therefore, immunohistochemical staining with antibodies to cytokeratins and e
ary tumor to the lung. Immunohistochemical Staining. The immunoperoxidase staining met
iara M, Ikebe T, et al: Immunohistochemical staining of desmosomal components in oral s
d tuna fish, and liver. Immunohistochemical staining Thirty-two formalinÂ–fixed, paraffinÂ–
gual origin. Results of immunohistochemical staining in tumors at different oral locations a
levels and to perform immunohistochemical staining with S-100 protein and homatropine
Fig. 2. Representative immunohistochemical staining for the expression of four EGFR fami
Fig. 1. Representative immunohistochemical staining for the expression of four EGFR fami
eached by telephone. Immunohistochemical staining Of the 23 biopsy samples examined,

Fig. 4.13.

Elsewhere in the source text, the translator was faced with the need to solve problems related to genre specific phraseology. This is the case for instance of the sentence beginning with «Da una attenta disamina della letteratura emerge...». Using the ORAL CANCER corpus as a resource, a first draft literal and clumsy translation with «From an attentive review of the literature...» was in fact replaced with the plainer «Review of the literature indicates...», thus resulting in a more fluent target text.

With no pretence at exhaustiveness, the examples reported seem to suggest that our ORAL CANCER corpus proved more than useful, both as a source of evidence of attested usage to test translation candidates and to elicit solutions to translation problems. It goes without saying, however, that what these examples aim to prove is not the usefulness of corpus data for the solution of translation problems in general, which – as already suggested – hardly needs to be demonstrated; they rather aim to test the «performance» of a corpus quickly and easily obtained through the automated process described in the first part of this chapter, whose real usefulness for a specific translation task could not be taken *a priori* for granted.

3. *Comparing corpora: «diagnos»* in English and Italian*

To further assess the value of the corpora compiled using WebBootCat, a second comparable corpus on «oral squamous cell cancer», this time composed of Italian texts, was created following the

same criteria as those followed for the English ORAL CANCER corpus. The two corpora were then used to explore phraseology in the two languages. The basic idea was that the two data sets obtained could provide a good basis for the creation of a glossary to be used throughout a translation task similar to the one referred to in the previous paragraph, without requiring a prohibitive investment in time.

The Italian CANCRO ORALE corpus is a 260.460 token corpus obtained using the words «cancro», «orale», «cellule» e «squamose» as seed terms. It was compiled in four phases, going through the same steps described for the creation of the English corpus, including single-word and multi-word terms extraction. In the process, many similarities between the two corpora emerged, especially concerning key terms. If we compare, for instance, the list of key terms automatically extracted by the system in the second run of the process for the compilation of the English ORAL CANCER corpus (see fig. 4.9, p. 112) and the list of key terms extracted in the second run of the process for the compilation of the Italian corpus, similarities seem to be self-evident:

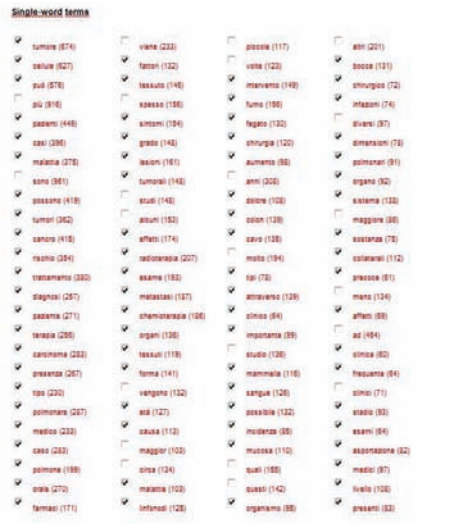


Fig. 4.14.

The similarities were confirmed at the end of the process when wordlists obtained from the two corpora using the Wordsmith Tools

were compared. By way of example here are the first 45 content words in each list (numbers to left and right refer to the position in the frequency list and to the number of occurrences respectively):

7	cancer 4605	26	tumore 1014
16	may 1833	27	cellule 986
17	oral 1795	33	può 799
18	radiation 1746	35	trattamento 720
19	cell 1642	37	pazienti 684
20	treatment 1528	39	rischio 588
23	patients 1379	43	tumori 537
25	cells 1346	44	malattia 531
26	therapy 1289	46	cancro 512
27	lung 1247	47	terapia 505
28	carcinoma 1211	49	anni 483
33	can 1155	51	orale 472
53	surgery 705	52	casi 464
54	cancers 684	71	radioterapia 327
55	head 682	75	dolore 314
56	used 681	78	chirurgia 309
59	disease 646	79	tessuto 309
60	tumors 637	81	farmaci 304
65	lymph 539	82	tipo 302
66	use 539	83	caso 295
67	mouth 531	85	medico 292
71	blood 505	86	due 290
73	survival 497	89	prima 275
74	chemotherapy 494	90	lesioni 274
79	information 472	91	sintomi 270
81	*body 458*	96	fattori 256
84	*clinical 452*	98	linfonodi 247
86	should 446	100	possibile 242
88	nodes 441	102	collo 239
90	*tobacco 422*	106	modo 231
91	patient 417	107	cavo 230
94	*called 415*	108	grado 225
97	small 391	111	effetti 220
99	stage 390	112	tempo 220
100	type 389	113	chemioterapia 219
102	lesions 377	119	donne 205
103	tissue 370	120	sopravvivenza 203
106	common 365	121	test 202

111	effects 342	122	tumorali 199
112	smoking 341	124	meno 192
113	include 327	125	mammella 190
114	health 326	127	forma 186
115	medical 326	128	tessuti 186
117	node 323	129	numero 184
118	cavity 322	130	secondo 184

As anybody with a knowledge of the two languages can immediately appreciate, most words in one list have their equivalent in the other list (e.g.: cancer/cancro; may/può-possono; oral/orale; treatment/trattamento; surgery/chirurgia and so on). Closer inspection of the complete word lists in both languages reveals that they are fairly consistent with each other, even though equivalent words occupy different positions in the two lists, depending on their relative frequency, and assuming that equivalence needs to be postulated also between words with different grammar functions: e.g. lung = polmone (noun)/polmonare (adjective); cell=cellula (noun)/cellulare (adjective), and so on (see Appendix 4 for a longer sample of the two word lists). Comparing the two wordlists seems to suggest, therefore, that the two corpora could well provide a basis for the creation of a specific glossary and/or phraseological dictionary. It is of course not the purpose of the present study to discuss methods for glossary creation and term extraction based on corpora, which are discussed elsewhere in detail (e.g. Pearson and Bowker 2002), and have been dealt with also with specific reference to the use of BootCaT (Castagnoli 2006; Fantinuoli 2006). By way of example, however, the following pages report information retrieved from the English and Italian corpora for DIAGNOS*, to illustrate the kind of linguistic information that can be derived from the two data sets.

3.1. *DIAGNOS* in the English ORAL CANCER corpus*

A search for DIAGNOS* in the English ORAL CANCER corpus yields in the first place evidence of a number of different realizations:

– diagnosis (290)
– diagnosed (126)
– diagnostic (47)

– diagnose (27)
– diagnoses (12)

By way of example, here is some insights into usage of the two most frequent forms «diagnosis» and «diagnosed», as suggested from an analysis of concordance lines. It goes without saying, again, that it is not the purpose of this work to provide an exhaustive analysis of the data reported but rather to show how consistent they are.

3.1.1. *DIAGNOSIS* In the ORAL CANCER corpus, the noun «diagnosis»:

– is often premodified by such words as definitive/delayed/differential/definitive (L1):

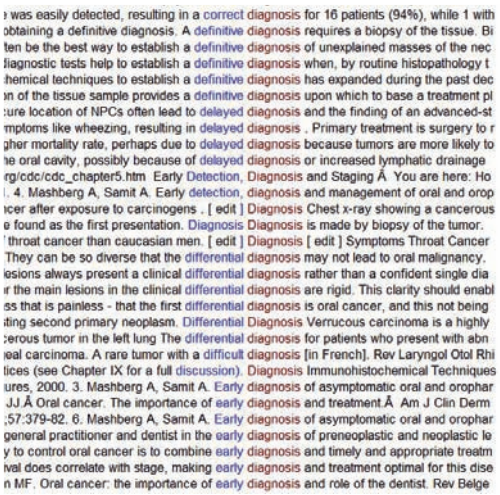


Fig. 4.15.

– has a tendency to co-occurr with time references (years, weeks, time), especially in L2 position;
– is often accompanied by such verbs as «confirm» and «establish»;

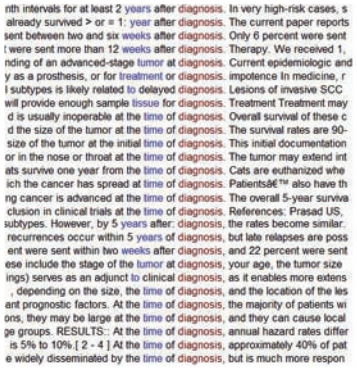


Fig. 4.16.

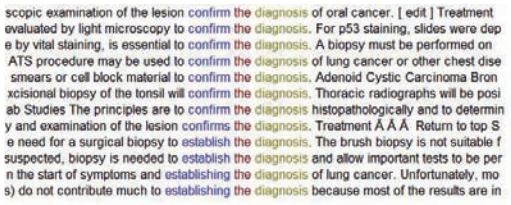


Fig. 4.17.

– is often found in association with the words «treatment» and «staging», especially in R2 position:

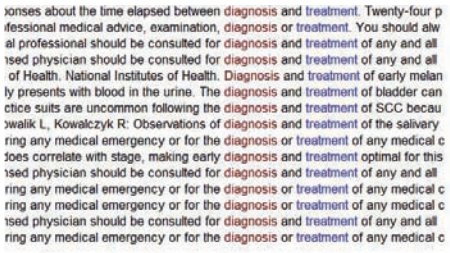


Fig. 4.18.

3.1.2. *DIAGNOSE* In our corpus, the verb, «diagnose» appears as almost invariably used in the past participle (see Appendix X for the complete concordance table), and often occurs in the patterns «N + (BE) diagnosed with» and «N + (BE) diagnosed in» where N is a noun referring to human beings (patients, people...)

in the first pattern and to the disease (tumour, cancer, cases...) in the second pattern:

– N + (BE) diagnosed with:

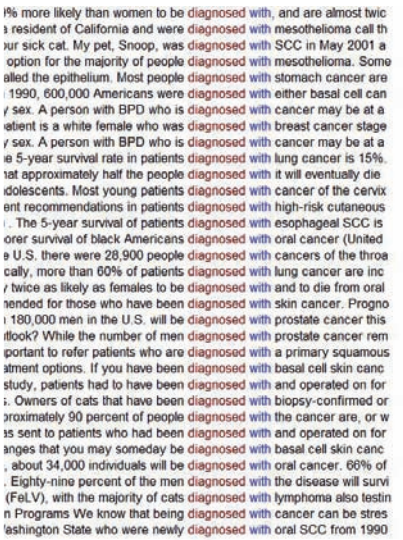


Fig. 4.19.

– N + (BE) diagnosed in:

Fig. 4.20.

The base form of the verb is mainly used, instead, in the infinitive form «to diagnose», in such phrases as «BE used to diagnose» or «failure to diagnose»:


er than food, that is used to prevent, diagnose, treat or relieve symptoms
Biologicals may be used to prevent, diagnose, treat or relieve of sympto
ing these tools will allow clinicians to diagnose the etiology with more exa
phy Medical/Legal Pitfalls: Failure to diagnose a malignant lesion could b
side the standard of care. Failure to diagnose SCC may lead to substanti
een set for cases in which failure to diagnose SCC has led to death. Fail
ces Medical/Legal Pitfalls Failure to diagnose correctly because of inade
n who examines tissues and fluids to diagnose disease in order to assist i
tion agency and are not intended to diagnose, or treat any disease or m
tion agency and are not intended to diagnose, or treat any disease or m
. All skin biopsy samples obtained to diagnose SCC must reach at least t
)5 The Buccal Mucosa Cells Test To Diagnose Lung Cancer Filed under:
f CT scanning of the chest, trying to diagnose lung cancers earlier. These
get. Other tests that can be used to diagnose this type of cancer include
nsity measurements may be used to diagnose osteoporosis, to see how
levels through the bone. It is used to diagnose osteoporosis (decrease in
on. In low doses, x-rays are used to diagnose diseases by making picture
nsity measurements may be used to diagnose osteoporosis, to see how
nsity measurements may be used to diagnose osteoporosis, to see how
nsity measurements may be used to diagnose osteoporosis, to see how
levels through the bone. It is used to diagnose osteoporosis (decrease in
ough studies are looking for ways to diagnose lung cancer earlier, no tes
even mortality, and physicians who diagnose and treat SCC are held le

Fig. 4.21.

3.2. *DIAGNOS* in the Italian CANCRO ORALE corpus*

Occurrences of «diagnos*» from the Italian CANCRO ORALE corpus provide evidence, again, of several different forms:

Diagonosi (382)
Diagnostico /Diagnostici/ Diagnostiche (89)
Diagnosticato / Diagnosticata/ Diagnosticati/Diagnosticate (50)
Diagnosticare (23)
Diagnostica (40)

On the basis of frequency data, it seems that the most frequent form is the noun «diagnosi», which is comparatively more frequent in the Italian corpus than in the English one. Bearing in mind that the Italian corpus is smaller, this is a datum which could be accounted for with a preference for nominal style in Italian, leaving room for further investigations.

3.2.1. *DIAGNOSI* Analysis of the immediate right co-text (R1) of the word «diagnosi» reveals frequent co-occurrence with such adjectives as «precoce», «definitiva», «accurata», «precisa». Here is a sample of concordances for «diagnosi precoce», «precoce» being the first collocate of «diagnosis» in our corpus.

In the left co-text (L2-L1), one notices a fair number of occurrences for the patterns «della diagnosi», «nella diagnosi», «alla diagnosi», «dalla diagnosi» e «per la diagnosi».

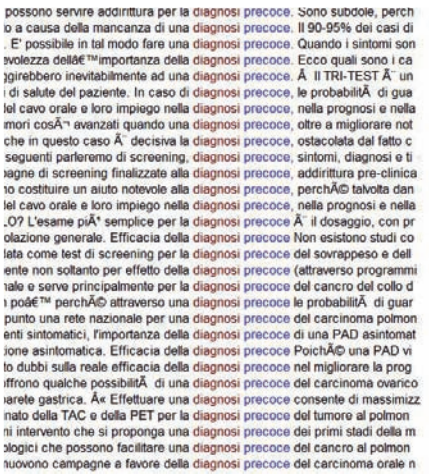

possono servire addirittura per la diagnosi precoce. Sono subdole, perch
o a causa della mancanza di una diagnosi precoce. Il 90-95% dei casi di
. E' possibile in tal modo fare una diagnosi precoce. Quando i sintomi son
volezza dellâ€™importanza della diagnosi precoce. Ecco quali sono i ca
girebbero inevitabilmente ad una diagnosi precoce. Â Il TRI-TEST Ã¨ un
di salute del paziente. In caso di diagnosi precoce, le probabilitÃ di gua
lel cavo orale e loro impiego nella diagnosi precoce, nella prognosi e nella
mori cosÃ¬ avanzati quando una diagnosi precoce, oltre a migliorare not
che in questo caso Ã¨ decisiva la diagnosi precoce, ostacolata dal fatto c
seguenti parleremo di screening, diagnosi precoce, sintomi, diagnosi e ti
agne di screening finalizzate alla diagnosi precoce, addirittura pre-clinica
o costituire un aiuto notevole alla diagnosi precoce, perchÃ© talvolta dan
lel cavo orale e loro impiego nella diagnosi precoce, nella prognosi e nella
.O? L'esame piÃ¹ semplice per la diagnosi precoce Ã¨ il dosaggio, con pr
olazione generale. Efficacia della diagnosi precoce Non esistono studi co
lata come test di screening per la diagnosi precoce del sovrappeso e dell
ente non soltanto per effetto della diagnosi precoce (attraverso programmi
nale e serve principalmente per la diagnosi precoce del cancro del collo d
poâ€™ perchÃ© attraverso una diagnosi precoce le probabilitÃ di guar
punto una rete nazionale per una diagnosi precoce del carcinoma polmon
enti sintomatici, l'importanza della diagnosi precoce di una PAD asintomat
ione asintomatica. Efficacia della diagnosi precoce PoichÃ© una PAD vi
to dubbi sulla reale efficacia della diagnosi precoce nel migliorare la prog
ffrono qualche possibilitÃ di una diagnosi precoce del carcinoma ovarico
arete gastrica. Â« Effettuare una diagnosi precoce consente di massimizz
nato della TAC e della PET per la diagnosi precoce del tumore al polmon
ni intervento che si proponga una diagnosi precoce dei primi stadi della m
logici che possono facilitare una diagnosi precoce del cancro al polmon
nuovono campagne a favore della diagnosi precoce del carcinoma orale n

Fig. 4.22.

Also evident is a tendency of the word «diagnosi» to co-occur with time references («anni», «momento», and also «età»), which can be compared to similar behaviour of the word «diagnosis» in the English corpus:

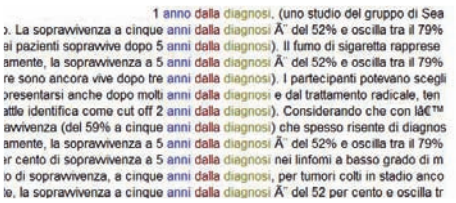

1 anno dalla diagnosi. (uno studio del gruppo di Sea
). La sopravvivenza a cinque anni dalla diagnosi Ã¨ del 52% e oscilla tra il 79%
ai pazienti sopravvive dopo 5 anni dalla diagnosi). Il fumo di sigaretta rapprese
amente, la sopravvivenza a 5 anni dalla diagnosi Ã¨ del 52% e oscilla tra il 79%
re sono ancora vive dopo tre anni dalla diagnosi). I partecipanti potevano scegli
presentarsi anche dopo molti anni dalla diagnosi e dal trattamento radicale, ten
attle identifica come cut off 2 anni dalla diagnosi). Considerando che con lâ€™
avvivenza (del 59% a cinque anni dalla diagnosi) che spesso risente di diagnos
amente, la sopravvivenza a 5 anni dalla diagnosi Ã¨ del 52% e oscilla tra il 79%
r cento di sopravvivenza a 5 anni dalla diagnosi nei linfomi a basso grado di m
o di sopravvivenza, a cinque anni dalla diagnosi, per tumori colti in stadio anco
le, la sopravvivenza a cinque anni dalla diagnosi Ã¨ del 52 per cento e oscilla tr

Fig. 4.23.

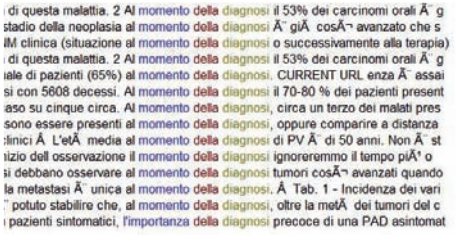

di questa malattia. 2 Al momento della diagnosi il 53% dei carcinomi orali Ã¨ g
stadio della neoplasia al momento della diagnosi Ã¨ giÃ cosÃ¬ avanzato che s
iM clinica (situazione al momento della diagnosi o successivamente alla terapia)
di questa malattia. 2 Al momento della diagnosi il 53% dei carcinomi orali Ã¨ g
ale di pazienti (65%) al momento della diagnosi. CURRENT URL enza Ã¨ assai
si con 5608 decessi. Al momento della diagnosi il 70-80 % dei pazienti present
aso su cinque circa. Al momento della diagnosi, circa un terzo dei malati pres
sono essere presenti al momento della diagnosi, oppure comparire a distanza
:linici Â L'etÃ media al momento della diagnosi di PV Ã¨ di 50 anni. Non Ã¨ st
izio dell osservazione il momento della diagnosi ignoreremmo il tempo piÃ¹ o
si debbano osservare al momento della diagnosi tumori cosÃ¬ avanzati quando
la metastasi Ã¨ unica al momento della diagnosi. Â Tab. 1 - Incidenza dei vari
potuto stabilire che, al momento della diagnosi, oltre la metÃ dei tumori del c
pazienti sintomatici, l'importanza della diagnosi precoce di una PAD asintomat

Fig. 4.24.

Another similarity with the English equivalent is the collocation with such verbs as «effettuare», «formulare», «emettere», «confermare», which can be considered as equivalents for «establish» and «confirm»:

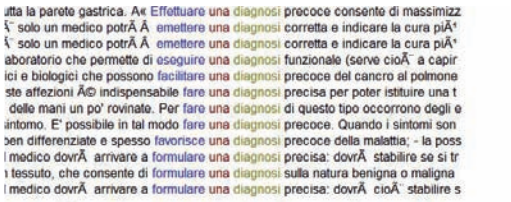

utta la parete gastrica. A« Effettuare una diagnosi precoce consente di massimizz
Ã¯ solo un medico potrÃ Â emettere una diagnosi corretta e indicare la cura piÃ¹
Ã¯ solo un medico potrÃ Â emettere una diagnosi corretta e indicare la cura piÃ¹
aboratorio che permette di eseguire una diagnosi funzionale (serve cioÃ¨ a capir
ici e biologici che possono facilitare una diagnosi precoce del cancro al polmone
ste affezioni Ã© indispensabile fare una diagnosi precisa per poter istituire una t
delle mani un po' rovinate. Per fare una diagnosi di questo tipo occorrono degli e
intomo. E' possibile in tal modo fare una diagnosi precoce. Quando i sintomi son
ben differenziate e spesso favorisce una diagnosi precoce della malattia; - la poss
l medico dovrÃ arrivare a formulare una diagnosi precisa: dovrÃ stabilire se si tr
n tessuto, che consente di formulare una diagnosi sulla natura benigna o maligna
l medico dovrÃ arrivare a formulare una diagnosi precisa: dovrÃ cioÃ¨ stabilire s

Fig. 4.25.

3.2.2. *DIAGNOSTICARE* Turning to the Italian verb «diagnosticare», it should be noted that the past participle form behaves rather differently from the equivalent English form. As the following concordance lines clearly show, it is the disease (tumori, casi) that is «diagnosticato». The basic pattern is «essere/venire diagnosticat*», preceded or followed by the subject (the disease), and often accompanied by time adverbials (e.g. «Ogni anno vengono diagnosticati circa 2 nuovi casi...»):

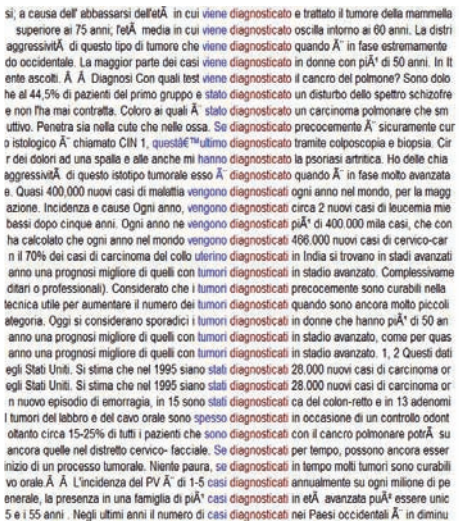

si, a causa dell' abbassarsi dell'etÃ in cui viene diagnosticato e trattato il tumore della mammella
superiore ai 75 anni, l'etÃ media in cui viene diagnosticato oscilla intorno ai 60 anni. La distri
aggressivitÃ di questo tipo di tumore che viene diagnosticato quando Ã¨ in fase estremamente
do occidentale. La maggior parte dei casi viene diagnosticato in donne con piÃ¹ di 50 anni. In It
ente ascolti. Â Â Diagnosi Con quali test viene diagnosticato il cancro del polmone? Sono dolo
he al 44,5% di pazienti del primo gruppo e stato diagnosticato un disturbo dello spettro schizofre
e non l'ha mai contratta. Coloro ai quali Ã¨ stato diagnosticato un carcinoma polmonare che sm
uttivo. Penetra sia nella cute che nelle ossa. Se diagnosticato precocemente Ã¨ sicuramente cur
o istologico Ã¨ chiamato CIN 1, questâ€™ultimo diagnosticato tramite colposcopia e biopsia. Cir
r dei dolori ad una spalla e alle anche mi hanno diagnosticato la psoriasi artritica. Ho delle chia
aggressivitÃ di questo istotipo tumorale esso Ã¨ diagnosticato quando Ã¨ in fase molto avanzata
e. Quasi 400,000 nuovi casi di malattia vengono diagnosticati ogni anno nel mondo, per la magg
azione. Incidenza e cause Ogni anno, vengono diagnosticati circa 2 nuovi casi di leucemia mie
bassi dopo cinque anni. Ogni anno ne vengono diagnosticati piÃ¹ di 400.000 mila casi, che con
ha calcolato che ogni anno nel mondo vengono diagnosticati 466,000 nuovi casi di cervico-car
n il 70% dei casi di carcinoma del collo uterino diagnosticati in India si trovano in stadi avanzati
anno una prognosi migliore di quelli con tumori diagnosticati in stadio avanzato. Complessivame
ditari o professionali). Considerato che i tumori diagnosticati precocemente sono curabili nella
tecnica utile per aumentare il numero dei tumori diagnosticati quando sono ancora molto piccoli
ategoria. Oggi si considerano sporadici i tumori diagnosticati in donne che hanno piÃ¹ di 50 an
anno una prognosi migliore di quelli con tumori diagnosticati in stadio avanzato, come per quas
anno una prognosi migliore di quelli con tumori diagnosticati in stadio avanzato. 1, 2 Questi dati
egli Stati Uniti. Si stima che nel 1995 siano stati diagnosticati 28.000 nuovi casi di carcinoma or
egli Stati Uniti. Si stima che nel 1995 siano stati diagnosticati 28.000 nuovi casi di carcinoma or
n nuovo episodio di emorragia, in 15 sono stati diagnosticati ca del colon-retto e in 13 adenomi
I tumori del labbro e del cavo orale sono spesso diagnosticati in occasione di un controllo odont
oltanto circa 15-25% di tutti i pazienti che sono diagnosticati con il cancro polmonare potrÃ su
ancora quelle nel distretto cervico- facciale. Se diagnosticati per tempo, possono ancora esser
inizio di un processo tumorale. Niente paura, se diagnosticati in tempo molti tumori sono curabili
vo orale.Â Â L'incidenza del PV Ã¨ di 1-5 casi diagnosticati annualmente su ogni milione di pe
enerale, la presenza in una famiglia di piÃ¹ casi diagnosticati in etÃ avanzata puÃ² essere unic
5 e i 55 anni . Negli ultimi anni il numero di casi diagnosticati nei Paesi occidentali Ã¨ in diminu

Fig. 4.26.

As to the infinitive form «diagnosticare», it tends to occur in final clauses such as «a/per diagnosticare» (e.g. «serve a/si usa per diagnosticare»).

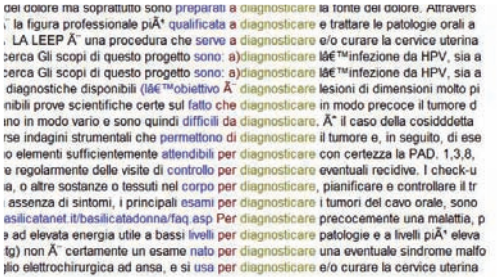

dei dolore ma soprattutto sono preparati a diagnosticare la fonte del dolore. Attravers
la figura professionale piÃ¹ qualificata a diagnosticare e trattare le patologie orali a
LA LEEP Ã¨ una procedura che serve a diagnosticare e/o curare la cervice uterina
erca Gli scopi di questo progetto sono: a)diagnosticare lâ€™infezione da HPV, sia a
erca Gli scopi di questo progetto sono: a)diagnosticare lâ€™infezione da HPV, sia a
diagnostiche disponibili (lâ€™obiettivo Ã¨ diagnosticare lesioni di dimensioni molto pi
nibili prove scientifiche certe sul fatto che diagnosticare in modo precoce il tumore d
no in modo vario e sono quindi difficili da diagnosticare. Ã¨ il caso della cosiddetta
rse indagini strumentali che permettono di diagnosticare il tumore e, in seguito, di ese
o elementi sufficientemente attendibili per diagnosticare con certezza la PAD. 1,3,8,
e regolarmente delle visite di controllo per diagnosticare eventuali recidive. I check-u
a, o altre sostanze o tessuti nel corpo per diagnosticare, pianificare e controllare il tr
assenza di sintomi, i principali esami per diagnosticare i tumori del cavo orale, sono
asilicatanet.it/basilicatadonna/faq.asp Per diagnosticare precocemente una malattia, p
ad elevata energia utile a bassi livelli per diagnosticare patologie e a livelli piÃ¹ eleva
tg) non Ã¨ certamente un esame nato per diagnosticare una eventuale sindrome malfo
plio elettrochirurgica ad ansa, e si usa per diagnosticare e/o curare la cervice uterina

Fig. 4.27.

The concordances provided also reveal some shortcomings of our Italian corpus. This is the case of the sentence reported among the concordances for «diagnostocat*» (Fig. 4.26): «circa 15-25% di tutti i pazienti che sono diagosticati con il cancro polmonare», which is obviously the outcome of a mistranslation from the English «nearly 15-25% of all the patients diagnosed with lung cancer», probably the results of a machine translation.

Conclusion

On the basis of the examples reported so far it can be argued that, despite some obvious limitations, WebBootCaT performs well for several tasks, especially when the time spent in creating the corpora and the usefulness of the information that can be retrieved are considered. Furthermore, the fact that the process of corpus compilation takes place on a remote server and that corpus data could be analysed both offline, by downloading it on one's own computer, and online, using the corpus query tool (Sketch Engine) installed on the remote server, really make WebBootCaT a telling example of how some aspects of corpus work have been changing under the impact of the web. With WebBootCat, corpus work is not only relying more and more on a «distributed architecture», thus embodying one of the changes envisaged by Martin Wynne with reference to the changing face of linguistic resources as a whole in the 21st century (Wynne 2002), but also appears to be moving towards «mass-customization», a keyword in contemporary society.

Note

[1] A bootstrap is a leather or fabric loop on the back or side of a boot to help pull it on. By extension bootstrap means «self-reliant and self-sustaining: relying solely on somebody's own efforts and resources» and «starting business from scratch: the building of a business from nothing, with minimum outside capital». In information technology the word synonymous with start up procedure (MSN Encarta 2007: online).

[2] The reference corpora used by BootCaT for key term extraction are large general corpora developed from the web using similar methods on a larger scale. The system currently includes five reference corpora (English, German, French, Italian and Spanish) of about 500 million words in average. (Baroni *et al.* 2006).

[3] The author wishes to thank Dr. Lucio Milillo for allowing her to quote from his Ph.D. thesis in Clinical Dentistry: L. Milillo, *Il ruolo della laminina-5 nel carcinoma orale: diagnosi, patogenesi, terapia*, Tesi di Dottorato di Ricerca Internazionale Multicentrico, Università di Bari, A.A. 2003-2004.

Chapter V

Exploring Large Web Corpora: from *Web as Corpus* to *Corpus as Web*

Introduction

This final chapter explores one of the most radical ways of understanding the relationship between corpus linguistics and the web. This corresponds to the «mega-corpus mini-Web» category among the possible meanings suggested by Baroni and Bernardini for the umbrella phrase web as/for corpus (2006: 13) and relates to the creation of large general-purpose corpora from the web via automated web crawling.

Drawing on descriptions by the creators of the 2 billion word itWaC corpus of Italian, Section 1 briefly introduces large general-purpose web corpora as a new object possessing both web-derived and corpus-like features. Secyion 2 describes the Sketch Engine as a web-based corpus query tool through which a number of recently compiled web corpora, including itWaC, can be accessed and explored. Finally, Section 3 paragraph reports «sketches» for the words «natura» and «nature», obtained from the itWaC and ukWaC corpora respectively, as an example of the variety of linguistic information that can be derived from the resources and tools described in the chapter.

1. *Large web corpora and corpus query tools*

In their collection of papers resulting from the First International Workshop on the Web as Corpus (Forlì, 14th January 2005), Baroni and Bernardini argue that «the most radical way of understanding the expression Web as a corpus refers to attempts to create a new object, «a sort of mini-Web (or mega-corpus) adapted to language research» (Baroni and Bernardini 2006: 13). This

object, they suggest, should be characterized by both web-derived and corpus-like features, to answer the widely-felt need for a resource that combines the potential for size, variety and topicality offered by the Web with the reliability of conventional corpora and corpus tools. This seems to represent a stage when linguists finally come to terms with the limitations of the web as a linguistic resource and come to view such limitations as a sort of 'necessary evil'which needs to be addressed if one is willing to exploit to the full the web's otherwise enormous potential. More specifically, the typical disadvantages of web corpora are accepted, assuming that none of these disadvantages are specific to web corpora per se (Baroni and Ueyama 2006: 31) but are rather simply *foregrounded* by such corpora, while they are in fact common to all «quick and dirty» large corpora:

> 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 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, 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 (Baroni and Ueyama 2006: 32).

As to the advantages, apart from size and timeliness, a fundamental advantage of creating large corpora from the web is that this allows «fast and cheap construction of corpora in many languages for which no standard reference corpus such as the BNC is available to researchers». Such languages, Baroni and Ueyama observe, do not simply include so-called «minority languages», but also well studied languages such as Italian and German (2006: 32). As to disadvantages, the only one which seems to be unique to web corpora is related to the necessity of accessing web data through ordinary search engines. This is the reason why large general corpora from the web are generally created via automated crawling, which makes the linguist as independent as possible from commercial search engines, allowing a certain degree of control over the corpus construction procedure. This is however a more difficult approach to using the web as a corpus than the ones described so far in the present study. Dispensing with commercial search engines and performing an autonomous crawl of the web obviously requires considerable computational skills and resources. Then there is the problem of cleaning the data produced by the crawl (removing undesired pages; discarding duplicates; removing mark-up language and other features typical associated with web documents). Finally, if the result is meant to be a very large corpus, the data should be annotated so as to allow analysis through specific corpus query tools.

It is nonetheless out of conviction of the feasibility of such a project that in the past few years a number of large general corpora from the web have been compiled, including the itWaC corpus of Italian and the ukWaC corpus of English. The basic steps involved in the compilation of large general purpose corpora from the web via automated crawling are described in Baroni and Bernardini (2006), Baroni and Kilgarriff (2006) and – as far as the itWaC and ukWaC corpora in particular are concerned – in Baroni and Ueyama (2006) and in Ferraresi (2007)[1]. These steps can be summed up as follows:

– Selecting «seed» URLs and crawling
– Data cleaning
– Annotation

Crawling the web for the compilation of a large general corpus requires a number of pre-selected URLs (or crawl «seeds») to start form. This means that the process starts with a program retrieving the pages corresponding to the seed URLs, extracts new URLs from the links in the retrieved pages, follows the new links to retrieve more pages, and so on. While for special purpose corpora as the ones created using BootCaT it seems to be relatively easy to find seed terms (and hence URLs), this is obviously not the case with a general-purpose corpus where, as Baroni and Bernardini point out, one would ideally have a number of «representative» URLs to

start from (2006: 16). It is in fact self-evident that, as the starting point of the whole process, this is the step most closely related to problems of representativeness, an issue, as already argued in the present work, extremely controversial with the web (Leech 2007). Here, again, «the fact that the notion of 'representativeness' of a corpus (and how to measure it) is far from well-understood» (Baroni and Bernardini 2006: 16, quoting Kilgarriff and Grefenstette 2003) complicates matters at a theoretical level. When it comes to web corpora it seems that the problem can only be addressed on applicative grounds, and in this context *post-hoc* methods to evaluate the composition of corpora have emerged as new crucial concerns (Baroni and Ueyama 2006: 32; Ferraresi 2007: 43; Sharoff 2007), replacing design based on *a priori* criteria, which was of paramount importance for traditional corpora. Accordingly, the apparently totalizing concept of representativeness is addressed through the related (and relative) concepts of «balance» and «unbiasedness»[2] (Kilgarriff and Grefenstette 2003; Ciaramita and Baroni 2006).

For the compilation of the itWaC corpus of Italian, the seed URLs were retrieved from Google with combinations of words extracted both from traditional newspaper corpora and from «basic vocabulary» lists for language learners, to ensure that both higher/public and lower/private registers were included[3]. The resulting list of over 5000 URLs was used to start a crawl which went on for nearly 10 days (Baroni and Ueyama 2006: 34).

Once the crawl is over, the linguist is presented with a vast set of HTML documents which have to be post-processed and cleaned before being converted into a linguistic corpus. The first step entails identifying and discarding potentially irrelevant documents on the basis of size, i.e. both small documents (below 5Kb) and large documents (over 200Kb) are removed on the assumption that they tend to contain little genuine text (Fletcher 2004). Then, removal of perfect duplicates is performed. In this phase not only the duplicates of a given document but also the document itself is removed, since it is an overt policy in the compilation of large web corpora to privilege precision over recall – a strategy which can be afforded owing to the vastness of the web (Baroni and Kilgarriff 2006). Besides removing duplicates, which are rather easy to identify, the cleaning

process also includes removal of near-duplicates, i.e. documents that differ only in trivial details, such as a date or a header.

After this phase the «noise» of non-linguistic material is removed from the documents. This generally means separating genuine language text from HTML code and from the so-called boilerplate, i.e. linguistically uninteresting material such as navigation information, copyright notices, advertisement, link lists, fixed notices and other sections lacking in human-produced connected text (Baroni and Kilgarriff 2006). The reasons for data cleaning need no explanation. As pointed out by Baroni and Bernardini (2006: 20) it is highly unlikely that one wants the bigram «click here» to come up as the most frequent bigram in a corpus of English, unless the aim of the corpus is precisely the study of the linguistic characteristics of web pages per se, which in turn corresponds to an altogether different way of conceiving of the web as a corpus, i.e. the «Web as corpus *proper*» (Baroni and Bernardini 2006: 13).

The final step is to filter for language and pornography. Even though web crawling for large corpora generally takes place within one single domain (.it in the case of the itWaC corpus, .uk for ukWaC, *.de* for the deWaC German corpus, and so on) and this should, ideally at least, by itself ensure that most pages are in the desired language, other strategies are generally adopted for filtering out pages in languages different from the target language. One such strategy is based on the assumption that connected text should contain a high proportion of function words (Baayen 2001, quoted in Ferraresi 2006: 38). Therefore in the compilation of the itWaC corpus a further step in the cleaning of the data was represented by the removal of pages which did not contain sufficient occurrences of function words. This process also worked as a language filter and further contributed to the removal of pages that mostly contained word lists, numbers, and other non-linguistic material (Baroni and Ueyama 2006: 35).

The importance of removing pages containing pornography is also generally acknowledged and stressed. This is done not «out of prudery», as Kilgarriff and Baroni argue, «but because they tend to contain randomly generated text, long keyword lists and other linguistically problematic elements» (2006). For the itWaC corpus, a stop list of 146 words typical of pornographic sites was used to identify and eliminate documents containing more than a

certain number of pornographic words. The whole filtering phase took about one week (Baroni and Ueyama 2006: 35).

The last step in the process is lemmatization and part-of-speech (POS) annotation of the corpus. Given the size of such web corpora this task has to be performed through automated machine-learning techniques. In the case of the itWaC corpus POS tagging was performed with the widely used TreeTagger and lemmatization using the free Morph-it! lexicon. Morphosyntactic annotation of the itWaC corpus took about two days, and resulted in a corpus of about 1.9 billion tokens (Baroni and Ueyama 2006: 35-36).

Using the procedure described above, a number of large general-purpose corpora from the web have been complied in a relatively short time. This is indeed a remarkable achievement whose success however also requires that adequate tools are devised to exploit the full potential of such corpora as sources of linguistic information. The minimum requirement for the tools is of course that «users must be able to browse the query results (displayed with varying amounts of context), sort the matches according to different criteria, and look at random subsets of the results to get a broad overview» (Baroni and Bernardini 2006: 35). Given that the user is very likely, in the case of large web corpora, to be presented with an overwhelming set of results, it is also desirable «to reduce and structure the massive amounts of data brought up by the corpus query, such as computing frequency lists, identify collocations, etc...» (Baroni and Bernardini 2006: 35). Many linguists could in fact lack the necessary technical skills to access and query such large web corpora, while copyright problems could refrain the compilers from publicly distributing the corpora for offline analysis (Baroni and Kilgarriff 2006). This seems to suggest the opportunity of adopting an advanced user-friendly web interface that allows linguists to do actual research on the corpus (including the possibility of saving settings and results across sessions) while allowing the compilers to make the corpus widely available through their servers. The requirements of corpus tools specifically designed for large web corpora seem thus to be making corpus search more and more similar to web search, to the extent of signifying a Copernican revolution from the seminal notion of

web as corpus to the new horizons of *corpus as web*. As Baroni and Bernardini argue, discussing the project for a corpus query tool specifically designed for large web corpora (Wacky query engine) and commenting on the 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. For these reasons, we conjecture that the success of the WaCky query engine and its acceptance among linguists will hinge on its ability to offer a similarly userfriendly, intuitive and familiar interface. As in the case of Google, a Web interface has the potential to satisfy all three criteria. 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 perhaps worth emphasizing that the authors are here not simply advocating the development of new corpus tools, but also indicating a shift in the expectations of users, as a consequence of a growing and widespread familiarity with ordinary web search. This seems to point to a metamorphosis in our 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. Some of these changes can be partly seen at work in the Sketch Engine, which will be briefly described in the following pages.

2. *The Sketch Engine: an overview*

As corpora become larger and require more sophisticated tools, 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 this new context, corpora and corpus tools are apparently undergoing a process of transformation that seems to be related to similar changes taking

place in society at large as far as the distribution of goods and resources, including linguistic resources, are concerned. While corpora and tools like BNC and the Wordsmith Tools can be considered finite *products* in a conventional way, in the sense that they are goods reproduced in several copies which can be sold and purchased, this is no longer the case with some of the recently compiled large web corpora and web-based corpus query tools, for which it would be in fact more correct to talk about *services*. Furthermore, as the notion of «mega-corpus mini-web» becomes a reality, 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 the best way to introduce the Sketch Engine in the present survey. The service provided by the Sketch Engine website can in fact be seen as a telling example of a different way of conceiving the basic activities of accessing, distributing and querying corpora. More specifically, the service provided through the Sketch Engine website makes a number of large web corpora available for online analysis and exploration, besides allowing the creation of smaller specialized corpora. Corpus analysis can be performed using a web-based corpus query tool, 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. 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 outline some of its key functions, namely the generation of «Concordances», the «Word Sketch» function, and the «Sketch Difference» function[4].

2.1. *Generating concordances*

The Sketch Engine mainly works on a number of pre-loaded corpora for several languages, including the BNC, besides allowing the exploration of customized corpora created using the tools

made available through the website itself. As the Home Page user interface shows, the first step for the user is therefore to select one

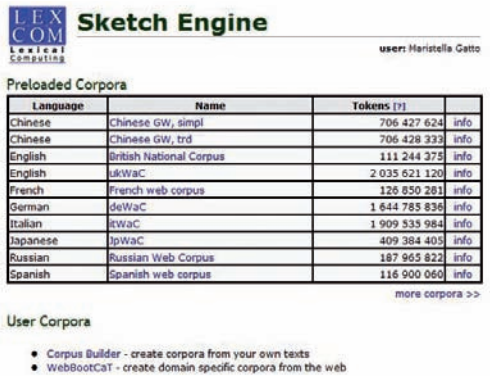


Fig. 5.1.

of the corpora made available by the service, or one of the tools for the creation of customized corpora:

The basic function provided by the Sketch Engine to explore each of these corpora is the generation of concordances. Here is,

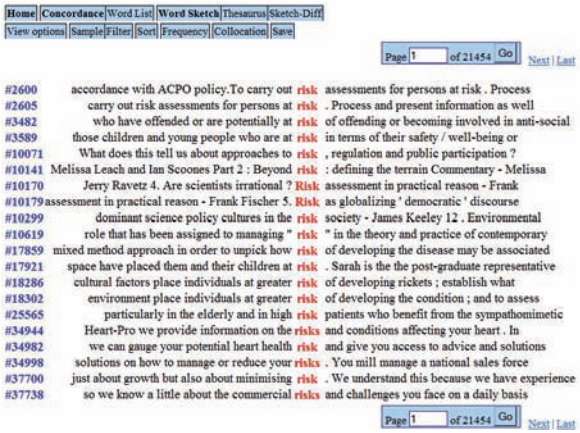


Fig. 5.2. A sample of concordances for the lemma RISK from the itWaC corpus.

for instance, a sample of concordances for the 429063 occurrences of the lemma RISK in the ukWaC corpus:

Besides reporting concordances in clear KWiC format, the concordance page features a number of buttons which allow further exploration. The «Sample» button, for instance, can be used to create a random sample of the concordances, an option that is particular useful when the number of hits is particularly high, as in this case. The «Sort» button can be used for a simple sort (sort by node word, or one position to the left, or one position to the right); more complex sort procedures can be specified through an advanced sort screen. The «Filter» button allows to a specify a word or lemma whose presence or absence is a condition to be satisfied before the concordances are displayed. Here is, for instance, a sample of concordances for RISK from the BNC using the word «cancer» as a positive filter, i.e. displaying only those lines where Risk occurs with «cancer»:


Home | Concordance | Word List | Word Sketch | Thesaurus | Sketch-Diff
View options | Sample | Filter | Sort | Frequency | Collocation | Save
First | Previous Page 4 of 12 Go Next | Last

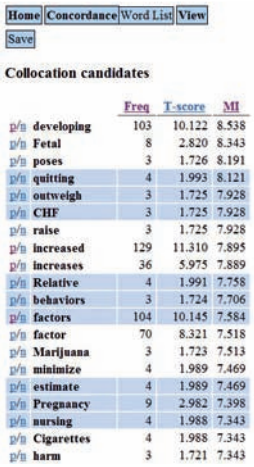
J2Y	America have since concluded that the *cancer*	risk	was negligible . A group of apple
HWS	important information about their *cancer*	risk	. All patients with colitis should be followed
EDK	investigations , the evidence for these *cancer*	risks	is still not conclusive , although it is
AN9	conservation measures to coal prices to *cancer*	risk	estimates . The hearings provide
K27	danger here.Nuclear workers told *cancer*	risk	is n't worth the worry . And , a
FSY	incidence at necropsy with the cumulative	risk	of detection of clinical *cancer* and the
J3G	there was between a fivefold and eightfold	risk	of *cancer* in children if their fathers
GU5	1977 ; Pentreath , 1980) . The estimated	risk	factors for radiation-induced fatal *cancers*
HU2	who used cimetidine in 1977-81 . An excess	risk	of gastric *cancer* was apparent in the first
HWW	of women with breast *cancer* had an excess	risk	of prostate cancer . The Icelandic report
FT2	for lung *cancer* excluded a twofold excess	risk	from atenolol . We can not exclude risks
ANX	Stuart Nachtwey of NASA estimates the extra	risk	of *cancer* over an 18-month voyage , including
HU3	long period should be considered in a final	risk	assessment for colorectal *cancer* in man
HJ4	weeks of intense sunbathing is at greater	risk	of developing skin *cancer* than a construction
HU2	activity . Populations at greater	risk	of colorectal *cancer* have been found to
HU2	pernicious anaemia patients with the greatest	risk	of developing gastric *cancer* could be identified
FSY	found women with anal *cancer* to have a high	risk	of having had cervical intraepithelial
HWS	is a common feature in subjects at high	risk	for colorectal *cancer* . The presence of
C94	incidence of this form of *cancer* . Those at high	risk	are eating a lot of fat and very little
AL4	Correspondent A NEW way to detect women at high	risk	of developing cervical *cancer* is announced

First | Previous Page 4 of 12 Go Next | Last

Fig. 5.3. A sample of concordances for RISK from the BNC using the word «cancer» as a positive filter.

Finally the «Collocation» button generates a list collocates for the node word, which can be sorted according to a number of parameters set by the user. Here is a sample of the collocates for RISK from the ORAL CANCER corpus created with WebBootCat (see Chapter 4) and accessed online through the Sketch Engine.

For each collocate the system also allows immediate visualization of recurring patterns. This can be obtained by simply clicking on the letter «p» in the «p/n» button left of each collocate.

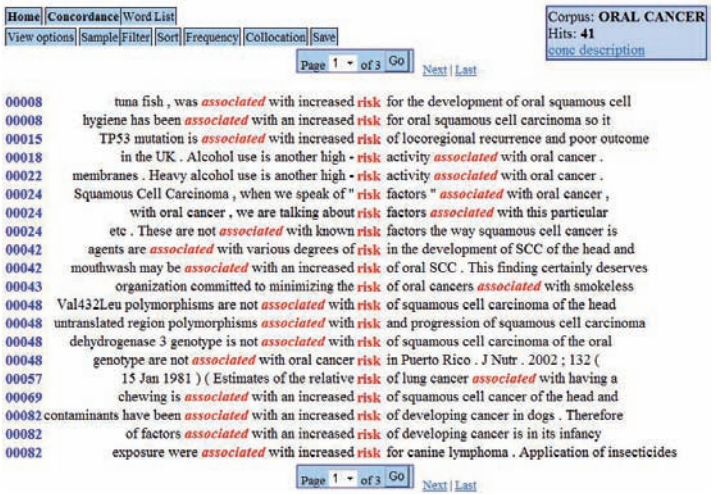

Home | Concordance | Word List | View
Save

Collocation candidates

		Freq	T-score	MI
p/n	developing	103	10.122	8.538
p/n	Fetal	8	2.820	8.343
p/n	poses	3	1.726	8.191
p/n	quitting	4	1.993	8.121
p/n	outweigh	3	1.725	7.928
p/n	CHF	3	1.725	7.928
p/n	raise	3	1.725	7.928
p/n	increased	129	11.310	7.895
p/n	increases	36	5.975	7.889
p/n	Relative	4	1.991	7.758
p/n	behaviors	3	1.724	7.706
p/n	factors	104	10.145	7.584
p/n	factor	70	8.321	7.518
p/n	Marijuana	3	1.723	7.513
p/n	minimize	4	1.989	7.469
p/n	estimate	4	1.989	7.469
p/n	Pregnancy	9	2.982	7.398
p/n	nursing	4	1.988	7.343
p/n	Cigarettes	4	1.988	7.343
p/n	harm	3	1.721	7.343

Fig. 5.4. A sample of the collocates for RISK from the ORAL CANCER corpus.

Here are, by way of example, patterns for the collocation RISK + ASSOCIATED, again from the ORAL CANCER corpus:


Home | Concordance | Word List
View options | Sample | Filter | Sort | Frequency | Collocation | Save
Corpus: **ORAL CANCER**
Hits: **41**
conc description
Page 1 of 3 Go Next | Last

00008	tuna fish , was *associated* with increased	risk	for the development of oral squamous cell
00008	hygiene has been *associated* with an increased	risk	for oral squamous cell carcinoma so it
00015	TP53 mutation is *associated* with increased	risk	of locoregional recurrence and poor outcome
00018	in the UK . Alcohol use is another high -	risk	activity *associated* with oral cancer .
00022	membranes . Heavy alcohol use is another high -	risk	activity *associated* with oral cancer .
00024	Squamous Cell Carcinoma , when we speak of "	risk	factors " *associated* with oral cancer ,
00024	with oral cancer , we are talking about	risk	factors *associated* with this particular
00024	etc . These are not *associated* with known	risk	factors the way squamous cell cancer is
00042	agents are *associated* with various degrees of	risk	in the development of SCC of the head and
00042	mouthwash may be *associated* with an increased	risk	of oral SCC . This finding certainly deserves
00043	organization committed to minimizing the	risk	of oral cancers *associated* with smokeless
00048	Val432Leu polymorphisms are not *associated* with	risk	of squamous cell carcinoma of the head
00048	untranslated region polymorphisms *associated* with	risk	and progression of squamous cell carcinoma
00048	dehydrogenase 3 genotype is not *associated* with	risk	of squamous cell carcinoma of the oral
00048	genotype are not *associated* with oral cancer	risk	in Puerto Rico . J Nutr . 2002 ; 132 (
00057	15 Jan 1981) (Estimates of the relative	risk	of lung cancer *associated* with having a
00069	chewing is *associated* with an increased	risk	of squamous cell cancer of the head and
00082	contaminants have been *associated* with an increased	risk	of developing cancer in dogs . Therefore
00082	of factors *associated* with an increased	risk	of developing cancer is in its infancy
00082	exposure were *associated* with increased	risk	for canine lymphoma . Application of insecticides

Page 1 of 3 Go Next | Last

Fig. 5.5. A sample of patterns for the collocation RISK + ASSOCIATED from the ORAL CANCER corpus.

Finally, information about the source-text of a particular concordance line can be obtained by clicking the document-id code at the left-hand end of the relevant line.

While in principle not different from information that can be obtained by querying a corpus with ordinary tools like Wordsmith, the way information from several corpora can be accessed using the Sketch Engine makes it a good example of the corpus-as-web metaphor. Whether the linguist is querying the BNC, or one of the new large web corpora such as ukWaC, or one of the customized corpora created by the user thorough WebBootCaT, the service proves quick, flexible and user-friendly in a way that reminds those Google-like features which should apparently characterize the shift from *web as corpus* to *corpus as web*.

2.2. *Word Sketches*

Besides producing concordances and providing information on the collocational profile of a word, in a way not dissimilar from other typical corpus tools, the Sketch Engine is specifically designed for offering the linguist «word sketches», i.e. «one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour» (Kilgarriff 2004 *et al.*). More specifically, a «word sketch» reports a list of collocates for each grammatical pattern so that, for each collocate, the user can see the corpus contexts in which the node word and its collocate co-occur (Kilgarriff *et al.* 2004). To provide a comprehensive sketch for whichever word a user inputs, the Sketch Engine needs in the first place to start from the corresponding lemma and the correct word class. This implies that to perform this specific function the Sketch Engine presupposes an annotated corpus as its basic input. By way of example, here is the Word Sketch entry form compiled so as to obtain a sketch for the lemma PAESAGGIO from the itWaC corpus:

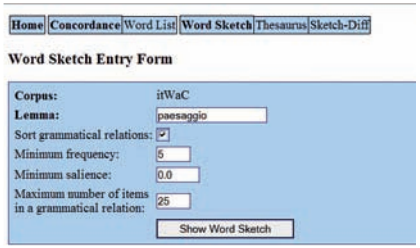
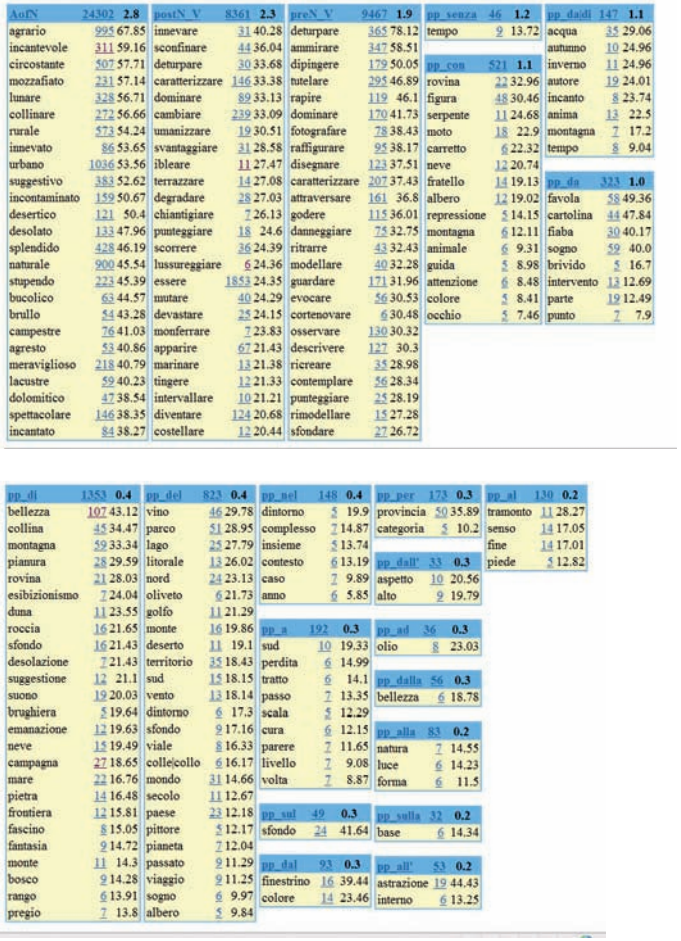


Fig. 5.6.

And here is a sample from the tool's output:



AofN	24302	2.8
agrario	995	67.85
incantevole	311	59.16
circostante	507	57.71
mozzafiato	231	57.14
lunare	328	56.71
collinare	272	56.66
rurale	573	54.24
innevato	86	53.65
urbano	1036	53.56
suggestivo	383	52.62
incontaminato	159	50.67
desertico	121	50.4
desolato	133	47.96
splendido	428	46.19
naturale	900	45.54
stupendo	223	45.39
bucolico	63	44.57
brullo	54	43.28
campestre	76	41.03
agresto	53	40.86
meraviglioso	218	40.79
lacustre	59	40.23
dolomitico	47	38.54
spettacolare	146	38.35
incantato	84	38.27

postN_V	8361	2.3
innevare	31	40.28
sconfinare	44	36.04
deturpare	30	33.68
caratterizzare	146	33.38
dominare	89	33.13
cambiare	239	33.09
umanizzare	19	30.51
svantaggiare	31	28.58
ibleare	11	27.47
terrazzare	14	27.08
degradare	28	27.03
chiantigiare	7	26.13
punteggiare	18	24.6
scorrere	36	24.39
lussureggiare	6	24.36
essere	1853	24.35
mutare	40	24.29
devastare	25	24.15
monferrare	7	23.83
apparire	67	21.43
marinare	13	21.38
tingere	12	21.33
intervallare	10	21.21
diventare	124	20.68
costellare	12	20.44

preN_V	9467	1.9
deturpare	365	78.12
ammirare	347	58.51
dipingere	179	50.05
tutelare	295	46.89
rapire	119	46.1
dominare	170	41.73
fotografare	78	38.43
raffigurare	95	38.17
disegnare	123	37.51
caratterizzare	207	37.43
attraversare	161	36.8
godere	115	36.01
danneggiare	75	32.75
ritrarre	43	32.43
modellare	40	32.28
guardare	171	31.96
evocare	56	30.53
cortenovare	6	30.48
osservare	130	30.32
descrivere	127	30.3
ricreare	35	28.98
contemplare	56	28.34
punteggiare	25	28.19
rimodellare	15	27.28
sfondare	27	26.72

pp_senza	46	1.2
tempo	9	13.72

pp_con	521	1.1
rovina	22	32.96
figura	48	30.46
serpente	11	24.68
moto	18	22.9
carretto	6	22.32
neve	12	20.74
fratello	14	19.13
albero	12	19.02
repressione	5	14.15
montagna	6	12.11
animale	6	9.31
guida	5	8.98
attenzione	6	8.48
colore	5	8.41
occhio	5	7.46

pp_dai/di	147	1.1
acqua	35	29.06
autunno	10	24.96
inverno	11	24.96
autore	19	24.01
incanto	8	23.74
anima	13	22.5
montagna	7	17.2
tempo	8	9.04

pp_da	323	1.0
favola	58	49.36
cartolina	44	47.84
fiaba	30	40.17
sogno	59	40.0
brivido	5	16.7
intervento	13	12.69
parte	19	12.49
punto	7	7.9

pp_di	1353	0.4
bellezza	107	43.12
collina	45	34.47
montagna	59	33.34
pianura	28	29.59
rovina	21	28.03
esibizionismo	7	24.04
duna	11	23.55
roccia	16	21.65
sfondo	16	21.43
desolazione	7	21.43
suggestione	12	21.1
suono	19	20.03
brughiera	5	19.64
emanazione	12	19.63
neve	15	19.49
campagna	27	18.65
mare	22	16.76
pietra	14	16.48
frontiera	12	15.81
fascino	8	15.05
fantasia	9	14.72
monte	11	14.3
bosco	9	14.28
rango	6	13.91
pregio	7	13.8

pp_del	823	0.4
vino	46	29.78
parco	51	28.95
lago	25	27.79
litorale	13	26.02
nord	24	23.13
oliveto	6	21.73
golfo	11	21.29
monte	16	19.86
deserto	11	19.1
territorio	35	18.43
sud	15	18.15
vento	13	18.14
dintorno	6	17.3
sfondo	9	17.16
viale	8	16.33
colle|collo	6	16.17
mondo	31	14.66
secolo	11	12.67
paese	23	12.18
pittore	5	12.17
pianeta	7	12.04
passato	9	11.29
viaggio	9	11.25
sogno	6	9.97
albero	5	9.84

pp_nel	148	0.4
dintorno	5	19.9
complesso	7	14.87
insieme	5	13.74
contesto	6	13.19
caso	7	9.89
anno	6	5.85

pp_a	192	0.3
sud	10	19.33
perdita	6	14.99
tratto	6	14.1
passo	7	13.35
scala	5	12.29
cura	6	12.15
parere	7	11.65
livello	7	9.08
volta	7	8.87

pp_sul	49	0.3
sfondo	24	41.64

pp_dal	93	0.3
finestrino	16	39.44
colore	14	23.46

pp_per	173	0.3
provincia	50	35.89
categoria	5	10.2

pp_dall'	33	0.3
aspetto	10	20.56
alto	9	19.79

pp_ad	36	0.3
olio	8	23.03

pp_dalla	56	0.3
bellezza	6	18.78

pp_alla	83	0.2
natura	7	14.55
luce	6	14.23
forma	6	11.5

pp_sulla	32	0.2
base	6	14.34

pp_all'	53	0.2
astrazione	19	44.43
interno	6	13.25

pp_al	130	0.2
tramonto	11	28.27
senso	14	17.05
fine	14	17.01
piede	5	12.82

Fig. 5.7a-b.

Clearly reported in each column are the words that typically combine with PAESAGGIO in a particular grammatical relation. The «AofN» lists reports adjectives that frequently accompany the PAESAGGIO/I, the «postN_V» column reports verbs that fre-

quently follow PAESAGGIO/I, the «preN_V» column reports verbs that frequently precede PAESAGGIO/I, and so on. When compared with information that could be obtained for the same word from the same corpus only through concordance lines the qualitative difference of information obtained through word sketches hardly needs to be demonstrated. The «sketch» for PAESAGGIO is useful and thought-provoking, indicative as it is – at a glance – of phraseological patterns, such as «paesaggio agrario», «paesaggio incantevole», «paesaggio circostante», «paesaggio urbano», but also «paesaggio da favola» or «paesaggio da cartolina».

An invaluable option provided by the tool is the possibility of switching at any time between Word Sketch mode and the Concordance mode, so that for each pattern a number of examples are available at a mouse-click. Thus, if interested in examples for the phrase «paesaggi di bellezza», the user only has to click on the number next to «bellezza» in the pattern «pp_di» to be shown all the 107 concordances for this collocation:

#3028094	clima eccezionale tutto l' anno , e vanta un	paesaggio	di straordinaria bellezza . Il clima è
#7835955	Pesaro . Un percorso tra arte e fede in un	paesaggio	di grande bellezza . Da non perdere : il
#27739792	dei dislivelli . Però si attraverseranno	paesaggi	di rara bellezza naturalistica , incastonati
#35963668	un viaggio di circa 9 ore ma vedrai dei	paesaggi	di una bellezza unica) altrimenti devi
#78915204	motore . Questa regione offre una varietà di	paesaggi	di una bellezza eccezionale , con valli
#81833267	ammirare le incantevoli cime del " Lagorai " e	paesaggi	di incomparabile bellezza . La struttura
#89914116	ruggenti e laghi pedemontani sconfinati ,	paesaggi	di bellezza superlativa , sono gli elementi
#92363041	vegetazioni di un tempo . Il paese gode di un	paesaggio	di rara bellezza e suggestione , posto
#108693421	natura , questo percorso si snoda in un	paesaggio	di rara bellezza . Si parte dalle caratteristiche
#143632358	millenni trascorsi , la sostituzione dei	paesaggi	di consolidata bellezza con panorami dominati
#177660386	montagne che vi si specchiano formano un	paesaggio	di una bellezza irreale : i disagi affrontati
#183927571	che dà vita a numerose cascate d' acqua e	paesaggi	di strabiliante bellezza , da visitare
#203368497	d' arte e di storia , cui fan da cornice	paesaggi	di prorompente bellezza . Gli influssi
#207050470	montata sul tetto , fermandosi ad ammirare i	paesaggi	di straordinaria bellezza offerti dal deserto
#207226090	la natura , che proprio in estate offre	paesaggi	di rara bellezza . Diverse mete proposte
#212987770	medioevali , siti archeologici , pievi e	paesaggi	di una bellezza ancora intatta . L' azienda
#217403448	, della poesia e dello struggimento dei	paesaggi	di rara bellezza , dell' incanto e della
#343026777	di chilometri . Offre di volta in volta	paesaggi	di una bellezza serena oppure selvaggia
#352422289	elementi che contribuiscono a rendere il	paesaggio	di eccezionale bellezza , caratterizzandolo
#406964930	strada ed arriviamo a salire sempre più in un	paesaggio	di eccezionale bellezza sino a quasi la

Fig. 5.8.

All patterns can of course be further explored using the options already illustrated for concordance lines, such as sorting or using filters.

In the case of Word Sketches the information provided by the system is definitely different from that obtained from an ordinary concordancer and this is to some extent related to the changes brought about by a closer relationship between corpus linguistics and the web. Although not specifically designed for interaction

with the web, a system that can provide such a summary of a word's behaviour is an appropriate answer not only to the need of processing large amounts of data, as in the case of large web corpora, but also to the desire of exploiting the inclusiveness, variety and accessibility of web data, without renouncing high standards of linguistic investigation. Furthermore, the very fact that both the corpus query tool and the corpora made available for analysis are offered as an integrated web-based service seems to make the Sketch Engine a good example of what it might mean to present the corpus as web, rather than simply using the web as a corpus.

2.3. *The Sketch Difference function*

With the Sketch Difference function the user has the opportunity to compare sketches for two similar words. The patterns and combinations shared by the two items are highlighted, while patterns and combinations typical of one or the other are contrasted. Here is, by way of example, a comparison between SCENERY and LANDSCAPE from ukWac:

Common patterns

scenery	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	landscape

subject_of	409	434	4.1	1.3
inspire	22	119	3.0	6.3
surround	263	97	6.3	4.8
impress	10	5	4.1	3.1
characterise	6	10	3.2	3.9
attract	7	9	1.1	0.9

a_modifier	1565	835198	3.7	2.3
breathtaking	992	197	10.2	7.1
spectacular	1902	330	10.0	7.1
stunning	1510	355	9.6	7.1
magnificent	615	142	8.6	6.0
beautiful	1861	949	8.4	7.3
coastal	607	209	8.3	6.4
dramatic	619	389	8.2	7.2
urban	7	866	1.1	7.8
historic	7	777	1.1	7.7
rugged	141	249	7.5	7.5
rural	85	829	4.3	7.4
unspoilt	119	144	7.4	6.8
varied	146	362	6.4	7.3
picturesque	152	148	7.2	6.5
majestic	91	49	7.2	5.0
volcanic	34	189	5.4	7.1
alpine	83	30	7.0	4.6
glorious	141	56	6.9	4.3
wonderful	450	162	6.8	5.1
awe-inspiring	59	16	6.8	3.8
rolling	34	123	3.8	6.7
lush	40	148	5.6	6.7
agricultural	5	278	1.1	6.6
mountainous	45	114	6.2	6.5
wooded	9	128	3.6	6.5

and/or	8767	215422	2.2	1.5
seascape	6	284	4.2	8.6
wildlife	440	836	7.5	8.2
portrait	35	453	3.4	8.0
costume	264	5	7.3	1.5
habitat	27	320	3.4	6.8
monument	14	172	3.3	6.5
heritage	49	256	4.4	6.5
geology	29	94	5.4	6.4
coastline	68	70	6.4	5.9
scenery	6	149	2.0	6.3
archaeology	11	126	3.2	6.3
architecture	63	260	4.3	6.2
beach	298	104	6.0	4.4
tranquillity	23	11	5.7	3.8
flora	31	64	5.2	5.7
amenity	12	91	3.1	5.7
climate	77	209	4.1	5.4
mountain	141	69	5.2	4.1
waterfall	22	9	5.2	3.2
landmark	17	49	4.1	5.2
landscape	149	99	5.1	4.4
abundance	26	8	5.0	2.8
painting	7	122	1.0	5.0
culture	156	315	4.0	5.0
lochs	13	6	5.0	2.9

Fig. 5.9.

As this sample from the Sketch Engine report shows, collocates for which the two lemmas share patterns and combinations are sorted according to salience scores and coloured according to dif-

ference between the scores, so as to emphasize 'preference' for one or other lemma. Thus, from the «a_modifier» (adjective_modifier) column one can deduce a preference for adjectives such as «breathtaking», «spectacular», «stunning» and «magnificent» to modify SCENERY rather than LANDSCAPE, which in turn is more typically modified by such adjectives as «urban», «historic» or «agricultural».

Patterns or combinations that can be considered unique to one or other lemma are reported in separate columns. Here are, for instance, the «scenery only» and «landscape only» patterns given by the Sketch Engine:

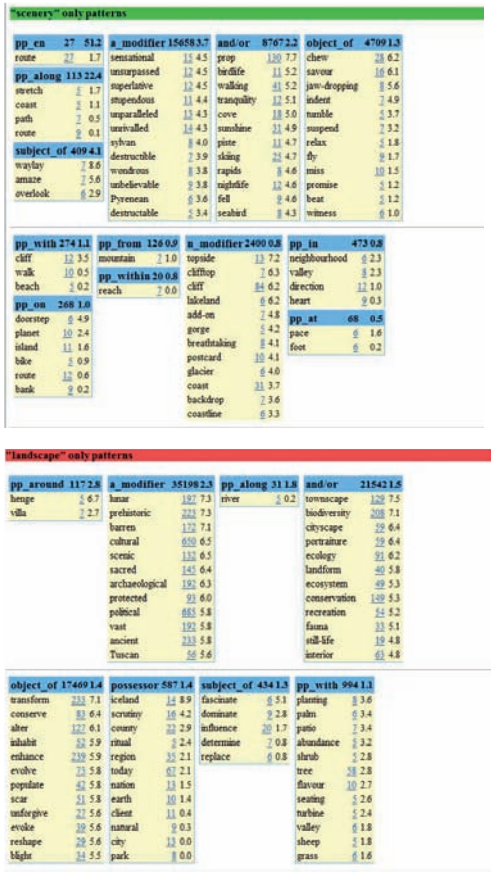


Fig. 5.10a-b.

Again, the qualitative difference between information that could be obtained by simply exploring concordance lines is self evident, especially if one considers that these data were obtained by summarizing (in a matter of seconds) information relating to the 25445 occurrences of «scenery» and the 110908 occurrences of «landscapes» in the ukWaC corpus. These examples definitely testify to the rich potential for linguistic information stored in the interaction between large web corpora and sophisticated web-based corpus query tools, but also indicates the great potential for linguistic information that lies in the possibility of processing more data when the tool performing the analysis can contribute information of this kind.

As to the immediate usefulness of the data, this can hardly be overemphasized. By comparing, for instance, the quickly produced sketch difference for LANDSCAPE and SCENERY with the word sketch produced for PAESAGGIO in the previous paragraph, one can see at a glance how the word «paesaggio» in Italian covers phraseological patterns which in English are covered by either «landscape» or «scenery», such as «paesaggio agricolo/urbano» and «agricultural/urban landscape» vs «paesaggio mozzafiato/ spettacolare» and «breathtaking/spectacular».

3. *Exploring large web corpora. Sketches of* NATURA *and* NATURE

The brief overview of the Sketch Engine functions can only suggest the scope and variety of information which can be gained by exploring mega-corpora from the web using a tool summarizing data in a way that is meaningful from the linguist's point of view. By way of example, this closing paragraph reports information derived from a brief analysis of sketches for NATURA and NATURE, obtained from the itWaC and ukWaC corpora. Again, the main purpose is not so much to provide an exhaustive analysis of the data, but rather to give an idea of the insight into language, and possibly into culture, provided by corpora which have been created with an automated procedure from the web.

3.1. *NATURA*

The word NATURA occurs 333722 times in the itWaC corpus, clearly a number of occurrences which could hardly be explored

without a tool contributing to the extraction of linguistic information. According to data reported by the Sketch Engine, in the itWaC corpus NATURA shows a clear tendency (126605 occurrences) to occur in the pattern Adjective + N (1st column), the first modifier in order of statistical significance being «incontaminata», followed by a number of other adjectives connecting NATURA to the legal and economic domain (e.g. «privatistico», «pubblicistico», «giuridico», «patrimoniale», «tributario», «economico, «finanziario», etc.) or to the philosophical domain (e.g. «umano», «divino», «naturata», «naturans»,...). Other words taking on again the meanings connected with «incontaminata», and therefore pointing to a more *concrete* reference to landscape, are «selvaggio» and «lussureggiante». A less dominant, yet significant, set of collocates preceding the noun NATURA in the Verb + N pattern (4th column) cluster around the concept of respect and suggest such phrases as «rispettare... preservare... salvaguardare... la NATURA».

Also worth exploring in the same pattern is a tendency of NATURA to co-occur with verbs pointing to mental processes in such patterns as «chiarire, rivelare, svelare, capire, conoscere, specificare, comprendere, precisare, scoprire la natura». Indeed most verbs preceding the NATURA can be seen as pointing to its abstract meaning as a synonymous of «reality» or «characteristic»:

natura itWaC freq = 333722

AofN	126605	2.7	pp_dell'	6162	2.3	postN_V	35585	1.8	preN_V	44313	1.7	pp_del	17600	1.7
incontaminato	1027	71.39	uomo	989	47.69	morire	1134	53.08	mutare	414	45.26	terreno	369	37.09
privatistico	888	67.38	appalto	218	42.27	risarcitoriare	167	48.02	chiarire	397	41.2	rapporto	893	36.93
pubblicistico	439	55.37	attività	677	34.54	regolamentare	536	47.74	amare	564	40.33	dato	615	30.32
giuridico	3527	53.82	incarico	149	32.51	impugnatoriare	35	40.74	avere	7645	40.27	cespite	26	27.62
rivisto	708	53.64	anima	139	32.38	subordinare	223	39.1	cambiare	971	38.89	suolo	113	27.35
rerum	241	53.51	attivita'attività	132	30.84	rotarianare	14	34.86	humare	95	37.98	rischio	306	27.09
umano	6059	52.13	handicap	75	28.86	enfiteuticare	14	32.95	imitare	133	37.85	bene	328	27.08
vario	3792	50.69	atto	269	28.25	vincolare	105	31.34	rivelare	428	35.76	prodotto	375	25.9
morto	810	48.41	infermità	30	27.13	ribellare	44	31.07	svelare	176	35.08	reato	150	25.49
ordinatorio	126	48.36	embrione	54	26.89	dotare	129	29.13	capire	449	34.12	denaro	94	24.32
divino	1136	47.44	i.i.	5	25.41	afflittire	12	28.11	conoscere	740	34.05	fenomeno	165	23.01
selvaggio	687	47.38	intervento	224	25.25	divinare	35	27.73	rispettare	415	34.0	contratto	249	22.35
provvedimentale	124	46.8	oggetto	128	23.54	gratiare	15	27.59	specificare	232	33.82	servizio	615	22.16
regolamentare	1028	46.5	affare	73	22.15	prescrittire	16	26.89	comprendere	721	33.34	problema	379	22.1
intrinseco	476	45.56	invalidità	28	21.78	essere	6786	26.64	riconoscere	587	33.2	legame	72	21.03
rigoglioso	192	44.97	opera	142	21.55	schiettamere	15	25.61	sicardare	12	33.14	provvedimento	218	20.79
patrimoniale	720	43.76	amianto	19	20.96	preminentemere	11	25.38	alterare	137	32.42	conflitto	100	19.62
naturans	35	43.51	assicurazione	41	20.8	bomandare	6	25.25	precisare	215	32.35	male	81	19.46
naturato	33	42.38	animo	33	20.31	naturare	8	25.09	scoprire	380	32.35	lavoro	550	18.91
deorum	56	42.0	operazione	77	19.55	stipendiare	35	25.0	rivestire	189	32.0	materiale	136	18.73

years

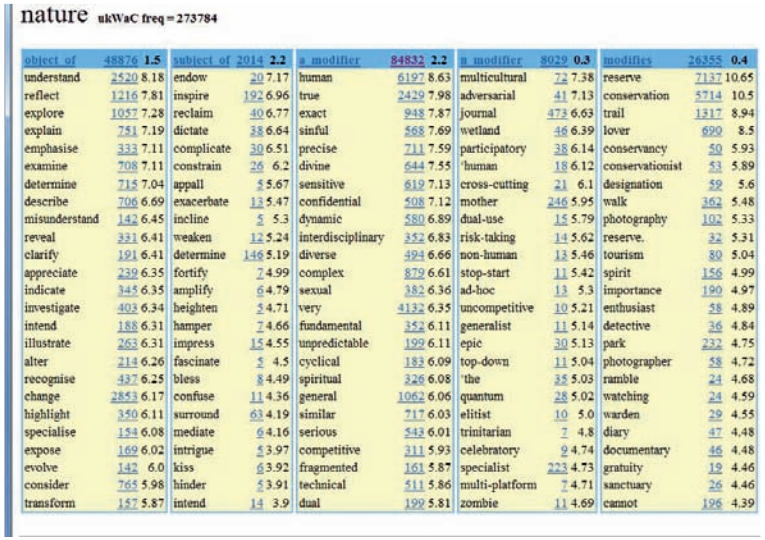
Fig. 5.11.

Another recurring pattern, «natura del/dell'», seems to be linked to man («uomo», «anima», «animo») or activities («attività», «atto», «intervento», «operazione», «lavoro», «servizio») as well as to areas apparently related to the legal or economic domain («appalto, «assicurazione, «rapporto, «rischio», «prodotto», «provvedimento», «reato», «denaro») and to health (handicap, infermità, embrione). These contextual features and different patterns seem to provide evidence that web data cover in this case a wide variety of lexical realizations perfectly corresponding to the meaning of the word as reported in Italian Dictionaries (e.g. Zanichelli 2005). Corpus data, however, apparently enrich dictionary meaning by providing phraseology hinting at more topical uses of the word: «natura incontaminata» and «preservare la natura», for instance, seem to be related to contemporary concerns about environmental protection, while «natura dell'embrione» might be related to recent ethic concerns about scientific research relating to infertility or to the use of stem cells obtained from embryos.

3.2. *NATURE*

The word *nature* occurs 273784 times in the ukWaC corpus. The sketch reported by the Sketch Engine shows that the word tends to occur as object of verbs such as «understand, reflect, explore, examine, reveal, investigate» (see the Object_of pattern in the 1st column), which seem to point to a level of high abstraction for the meaning of the NATURE, as already seen for the Italian corpus.

The pattern Adjective + N (4th column) is characterized by such words as «human», «true», «divine», pointing to the spiritual/philosophical meaning of the word NATURE, whereas no instance is reported of adjectives similar to the ones co-occuring with NATURA in Italian, such as «incontaminato» or «rigoglioso». This seems to suggest that the word *nature* does not necessarily cover the same semantic area of its Italian dictionary equivalent, at least as far as its concrete meaning related to the idea of landscape is concerned. The only collocates of NATURE which seem to point to a meaning of the word connected with the idea of landscape are those in which NATURE premodifies such words as «reserve, protection, trail, park, tourism» (5th column),



nature ukWaC freq = 273784

object of	48876	1.5	subject of	2014	2.2	a_modifier	84832	2.2	n_modifier	8029	0.3	modifies	26355	0.4
understand	2520	8.18	endow	20	7.17	human	6197	8.63	multicultural	72	7.38	reserve	7137	10.65
reflect	1216	7.81	inspire	192	6.96	true	2429	7.98	adversarial	41	7.13	conservation	5714	10.5
explore	1057	7.28	reclaim	40	6.77	exact	948	7.87	journal	473	6.63	trail	1317	8.94
explain	751	7.19	dictate	38	6.64	sinful	568	7.69	wetland	46	6.39	lover	690	8.5
emphasise	333	7.11	complicate	30	6.51	precise	711	7.59	participatory	38	6.14	conservancy	50	5.93
examine	708	7.11	constrain	26	6.2	divine	644	7.55	'human	18	6.12	conservationist	53	5.89
determine	715	7.04	appall	5	5.67	sensitive	619	7.13	cross-cutting	21	6.1	designation	59	5.6
describe	706	6.69	exacerbate	13	5.47	confidential	508	7.12	mother	246	5.95	walk	362	5.48
misunderstand	142	6.45	incline	5	5.3	dynamic	580	6.89	dual-use	15	5.79	photography	102	5.33
reveal	331	6.41	weaken	12	5.24	interdisciplinary	352	6.83	risk-taking	14	5.62	reserve.	32	5.31
clarify	191	6.41	determine	146	5.19	diverse	494	6.66	non-human	13	5.46	tourism	80	5.04
appreciate	239	6.35	fortify	7	4.99	complex	879	6.61	stop-start	11	5.42	spirit	156	4.99
indicate	345	6.35	amplify	6	4.79	sexual	382	6.36	ad-hoc	13	5.3	importance	190	4.97
investigate	403	6.34	heighten	5	4.71	very	4132	6.35	uncompetitive	10	5.21	enthusiast	58	4.89
intend	188	6.31	hamper	7	4.66	fundamental	352	6.11	generalist	11	5.14	detective	36	4.84
illustrate	263	6.31	impress	15	4.55	unpredictable	199	6.11	epic	30	5.13	park	232	4.75
alter	214	6.26	fascinate	5	4.5	cyclical	183	6.09	top-down	11	5.04	photographer	58	4.72
recognise	437	6.25	bless	8	4.49	spiritual	326	6.08	'the	35	5.03	ramble	24	4.68
change	2853	6.17	confuse	11	4.36	general	1062	6.06	quantum	28	5.02	watching	24	4.59
highlight	350	6.11	surround	63	4.19	similar	717	6.03	elitist	10	5.0	warden	29	4.55
specialise	154	6.08	mediate	6	4.16	serious	543	6.01	trinitarian	7	4.8	diary	47	4.48
expose	169	6.02	intrigue	5	3.97	competitive	311	5.93	celebratory	9	4.74	documentary	46	4.48
evolve	142	6.0	kiss	6	3.92	fragmented	161	5.87	specialist	223	4.73	gratuity	19	4.46
consider	765	5.98	hinder	5	3.91	technical	511	5.86	multi-platform	7	4.71	sanctuary	26	4.46
transform	157	5.87	intend	14	3.9	dual	199	5.81	zombie	11	4.69	cannot	196	4.39

Fig. 5.11.

thus resulting in such patterns as «nature reserve» (apparently the most frequent collocation) or «nature tourism».

This seems to point to a gap between the behaviour, and hence the meanings, of NATURA and NATURE which are apparently perfect equivalents in Italian and English. Such differences, which are to some extent genre- and domain-specific, have been partly explored by Manca (2002) with reference to tourism discourse. It is this supposed gap, for instance, that accounts for lack of correspondence between typical phraseology in the language of tourism in Italian and in English, as is the case with such phrases as «circondati dalla natura» or «la tranquillità della natura» in which «natura» cannot be translated with «nature» but rather with its hyponym «countryside». This gap, which apparently lays bare interesting differences at the level of context of culture, might deserve further exploration for which the huge amount of data made available by such corpora as ukWaC and itWaC, and the information provided by the Sketch Engine might prove extremely appropriate. It seems therefore that the latest development within this exciting and promising research field is definitely opening up new horizons for linguistic research.

Conclusion

As this last chapter has shown charting the latest achievements of the web as corpus, the development of new methods for doing corpus linguistics is not simply a matter of new corpora and new tools, but rather of changing ways of conceiving of corpora and corpus tools under the impact of the web and of web search. While the notion of the *web as corpus* might be giving way to a complementary view of *corpus as web*, other significant changes are apparently occurring in terms of availability and distribution of tools and resources, such as for instance the shift of both from products to services. This seems indeed to connect the changes taking place in contemporary corpus linguistics to similar changes taking place in society at large, and possibly represents the best way to sum up the real significance of the achievements presented in our itinerary.

Note

[1] Information about large web corpora for other languages can be found in the Sketch Engine (www.sketchengine.co.uk) and the Wacky project (wacky.sslmit.unibo.it) websites.

[2] The notion of «unbiasedness» is based on the comparison of the word frequency distribution of a corpus to those of deliberately biased corpora. (Ciaramita and Baroni 2006).

[3] Detailed information on this aspect of corpus creation is reported in the Wacky project website: http://wacky.sslmit.unibo.it/doku.php?id=seed_words_and_tuples.

[4] For a comprehensive overview of the Sketch Engine see *Getting started with the Sketch Engine*, http: //trac.sketchengine.co.uk/wiki/SkE/GettingStarted.

Conclusion

The steps taken throughout the book have shown how the notion of the web as a corpus is to some extent grounded on a *migration* of issues (Gatto forthcoming) between corpus linguistics and information technology, under whose impact the way we conceive of a corpus has been moving away from the somewhat reassuring standards subsumed under the *corpus-as-body* metaphor, to a new *web-as-corpus* image, and possibly moving a further step towards the new horizons of the corpus-as-web. On the one hand the notion of a linguistic corpus as a body of texts rests on some related issues such as finite size, balance, part-whole relationship and 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 help make the «webscape» a more hospitable space for corpus research. By either exploiting to the full the potential of ordinary search engines, or by reworking their output format to make it suitable for linguistic analysis (e.g. WebCorp), or by allowing the creation of quick, flexible, small, specialized and customized multilingual corpora from the web (e.g. WebBootCaT), these tools seems to be 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: multilinguality, dynamic content, distributed architecture, virtual corpora, connection with web search (Wynne 2002).

As the issues, tools, and methods so far discussed have already shown, the emerging notion of the web as corpus can be 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 of technologies and standards in several related fields which have in common the goal of delivering linguistic content through electronic means» (Wynne 2002: 1207). It is indeed owing to such a convergence that one is tempted to argue that changes in corpus work under the impact of the web are related to the new styles and approaches to the sharing/distribution of knowledge, goods and resources which are everyday experience in contemporary society. This seems to be particularly evident in the latest development relating to the creation and exploration of large web corpora where corpus resources and corpus tools seem to be undergoing a process of transformations from products into services.

By way of conclusion is perhaps worth emphasizing again that while the web as a corpus is a promising field of research it can by no means aim at questioning the fundamental tenets of corpus linguistics. As Baroni and Ueyama (2006) suggest, it is only «a matter of research policy, time constraints and funding» that determines whether a certain project requires building a thoroughly controlled conventional corpus, or if it is better to methods that in a way or another take advantage of the web's controversial status as a linguistic corpus. What is certain is that, as any other field of human knowledge, linguistic research can only profit from the tension created between established theoretical positions and the new tools and methods devised from the needs of practicality and pragmatism. It seems more than desirable, then, that *traditional* corpus linguistics and studies on the web as corpus should coexist for time to come, providing the linguistic community with a wider spectrum of resources to choose from.

Appendix

Appendix 1 – Webcorp output for «scenery»

Sort Results on word left 1 for 'scenery'

CASE INSENSITIVE

2003103107:08:141	and buildings (67%), atmosphere (37%	scenery	(26%) and the **river** and
2002080619:29:561	well ! Here at last, a	scenery	add-on allowing you to fly
2004021402:45:031	the area - good food, accommodation	scenery	, walking, unusual and unique shops
0	boards and explore the amazing	scenery	. Fishing tackle for hire. Monster
2004021019:23:171	variety of fossils, and amazing	scenery	. Fossils are less frequent than
2003010100:00:004	and romantic resorts with amazing	scenery	and romantic sunsets, we suggest
2001010100:00:004	is a walker's paradise amidst	scenery	to die for, with some
0	the historic monuments, wildlife and	scenery	of these unique islands. Also
2002120317:06:481	for its tropical landscape and	scenery	. Located at the southern tip
2004012604:43:591	the historic monuments, wildlife and	scenery	of these unique islands. Also
0	of Orkney's landscape, moods and	scenery	. Papay Pages Images and information
2003091517:03:321	of Brecon Beacons, caves and	scenery	in South Wales. ** Black Mountains
2004012912:46:511	wealth of historic sites and	scenery	of Segovia were awarded a
2001031209:42:141	Forty-eight views of cottages and	scenery	at Sidmouth, Devon (Somers Cocks
2004010100:00:004	unrivalled variety of moods and	scenery	. © Crown Copyright 2004 Weymouth
2002072305:32:411	discover the areas history and	scenery	. If that's all a bit
2002010911:58:431	created by its situation and	scenery	. The loch is home to
2003082421:01:111	that the amenities, attractions and	scenery	, coupled with the warmth of
0	the historic monuments, wildlife and	scenery	of these unique islands. Also
2004012804:49:571	the historic monuments, wildlife and	scenery	of these unique islands. Also
2004010100:00:004	of Orkney's landscape, moods and	scenery	. http://myweb.tiscali.co.uk
2004010100:00:004	the historic monuments, wildlife and	scenery	of these unique islands. Also
2002121900:00:002	the Scottish landscape, countryside and	scenery	a lot higher than the
2002121900:00:002	surveyed, the landscape, countryside and	scenery	were by far the most
2003071611:26:051	set in the breathtaking Argyll	scenery	some seven miles from Lochgilphead
1999121413:37:111	itineraries. Nature, mostly cast as	scenery	, has been a significant attraction
2004010100:00:004	Hardy Countryside with the beautiful	scenery	as seen in the much
0	which take in some beautiful	scenery	including several woodland **walks**. The
2004020515:51:141	area. Visitors can experience beautiful	scenery	, **dramatic coastlines** and many visitor
0	**walks** through Darley Dale's beautiful	scenery	...
2000010100:00:004	or the lover of beautiful	scenery	and the accompanying wildlife, East
1999032401:16:011	the tourists. The most beautiful	scenery	is often the most fragile
2003110609:30:422	with excellent bathing and beautiful	scenery	. Visitors to Rhossili can enjoy
2003062809:28:581	natural resources, such as beautiful	scenery	and wildlife, or human-made resources
2003010100:00:004	some of the most beautiful	scenery	. Listing location, local attractions, fishing
2003100120:37:301	that we take the beautiful	scenery	in this area for granted
2001111619:30:101	areas, woodland walks and beautiful	scenery	. Sandringham House Brochure Sandringham House
0	through and across the beautiful	scenery	of the Derbyshire Dales. Walking
2003102218:03:301	courses, often set among beautiful	scenery	. The Dunes Since opening the
2002010100:00:004	some of the most beautiful	scenery	of the Dee Valley. Alternatively
2003053123:27:301	springs, mild mountain climate, beautiful	scenery	and a large well-kept park
2004010817:43:201	contains most of the beautiful	scenery	and locations immortalised in the
0	where visitors can enjoy beautiful	scenery	and excellent wine. 02/13
2003122912:37:201	Scottish fayre. Surrounded by beautiful	scenery	in a climate warmed by

0	monastery's surroundings and the beautiful	scenery . From Vecer of 15 th
2002121900:00:002	30% +37% +16% +29% Beautiful	scenery (16 +7%) +11% +15% +5%
2002121900:00:002	reality. The shift in beautiful	scenery at 7% may seem small
2003091316:05:501	many people's minds with beautiful	scenery , unspoiled landscapes, wildlife, a traditional
2004011023:36:241	through some of the best	scenery in the UK. The area
2004011420:10:441	as some of the best	scenery . Featured Castles in Scotland We
2003080612:55:411	midst some of the best	scenery in variety of carriages pulled
2004011617:14:221	through some of the best	scenery on a clearly waymarked route
2003071610:45:031	the heart of Dartmoor's breathtaking	scenery with 8 en suite bedrooms
2001111413:14:472	a wedding, boasting castles, breathtaking	scenery and some of the best
0	some of the most breathtaking	scenery anywhere on earth. All this
2003090513:11:421	to take in the breathtaking	scenery . There are numerous attractions in
2004010100:00:004	and Landseer with its breathtaking	scenery and romantic landscape. The estate
2001082912:40:041	Bradford is blessed with breathtaking	scenery including spectacular Ilkley Moor; the
2001111413:14:472	a wedding, boasting castles, breathtaking	scenery and some of the best
0	Fun Day and the breathtaking	scenery of The Great Outdoors . We
0	of Pachuca's mountains and breathtaking	scenery . And if you were hoping
2001062700:15:231	small country boasts such breathtaking	scenery . But what price do we
2003091621:49:321	some of the most breathtaking	scenery in Britain. It is the
2004011618:16:281	is renowned for its breathtaking	scenery . The Rila mountain, with its
2002091315:59:261	with panoramic views and breathtaking	scenery . Country walks and cycle routes
2002031618:52:321	walking and climbing with breathtaking	scenery throughout the Highlands and Islands
0	wide open spaces and breathtaking	scenery - come to Snowdonia, LIyn Peninsula
2102031316:16:392	of great contrast and breathtaking	scenery . On the south coast of
2003111900:00:003	spot to enjoy the breathtaking	scenery . The Park offers an abundance
2002121900:00:002	countries surveyed. In all cases	scenery was rated by at least
2003010100:00:004	Count Dracula! Enjoy the coastal	scenery on the way back via
2003122314:49:121	and miles of stunning coastal	scenery . Hotels breaks in devon cornwall
2004012804:34:371	to explore the spectacular coastal	scenery and wildlife on the Atlantic
2003051809:25:401	as marvelling at the coastal	scenery of Portmadog Bay and the
0	apart from our stunning coastal	scenery ... is the nightmare of a
0	sites, museums, countryside and coastal	scenery on offer, there is no
0	mentioned pertain to the coastal	scenery : "lovely beaches"; "rugged, craggy, coastline"
0	line in terms of coastal	scenery and topography with "smaller beaches
2002110616:38:551	why people visit the County	scenery being the single most popular
0	A stunning recipe of culture	scenery and people" Having travelled for
2003021113:35:161	and the descriptions of Devonshire	scenery in the works of Charles
2003091316:05:501	with Scotland's natural beauty, dramatic	scenery and unpolluted landscape attract both
2004021113:04:231	the north is the dramatic	scenery of the Staffordshire Moorlands and
2003101717:58:431	etc. There is more dramatic	scenery along the Yorkshire Coast , and
2004021404:50:271	west coast amidst the dramatic	scenery of the Scottish Highlands. [Visit
2004021319:41:481	an area dominated by dramatic	scenery , unrivalled countryside and an exciting
2003010100:00:004	Cruise + 7 Night Stay Enchanting	scenery and classical treasures await you
0	Clean, fresh mountain air, enchanting	scenery and eternally blissful serenity are
0	lakes to seek out Englands	scenery . Details Blueworks Minibus Tours Eight
2003010100:00:004	RESORTS Greece offers some fabulous	scenery , fantastic sea views and traditional
0	was absolutely amazing with fantastic	scenery and good opportunities for game

2003111715:14:272	year for my holidays. Fantastic	scenery , friendly helpful people and always
0	restaurants to choose from, fantastic	scenery and eco-tourism opportunities. Monterrey is
2002010100:00:004	countryside the district offers fantastic	scenery , history and heritage and is
2003121714:54:291	Golf courses, friendly locals, fantastic	scenery and above all represents good
2003071610:45:031	The mild climate and fine	scenery make this the ideal location
2003111911:14:261	the highest, wildest and finest	scenery in England with attractive North
2002110616:44:311	some of the Roseland's finest	scenery . As it nears the coast
2004010600:00:003	the highest, wildest and finest	scenery in England. Much of the
0	as some of the finest	scenery in Scotland, visitors can marvel
0	luxury cottages amidst Scotlands finest	scenery Mains of Taymouth cottages are
0	of Kenmore amidst the finest	scenery Scotland has to offer. We
2002121900:00:002	a trip were environment focussed	scenery , nature and wildlife, wilderness). People
2003092410:15:122	but give me Swansea for	scenery and climate.â€™ However you
2004010100:00:004	famous for shopping than for	scenery . Visitors from all over the
2003122912:37:201	centrally situated in beautiful Galloway	scenery . The Bruce provides excellent food
2002070407:37:081	short river and the gentle	scenery it passes through is typical
2000010100:00:004	as a country of glorious	scenery and friendly people. You can
2001010100:00:004	some of the most glorious	scenery in Scotland can best be
3	lot to offer from great	scenery , good pubs and restaurants, and
2004010619:30:101	West Sussex) Geology: Cretaceous Great	scenery and an excellent beach. There
2003063000:00:002	said: "Scotland has always had	scenery to take your breath away
2003010100:00:004	before the 18th Century. Highland	scenery on the eastern shores of
2003111911:14:261	some of England's finest hill	scenery within range. Teesdale, Weardale and
0	hard living conditions. Stunning historical	scenery , old colonial towns that are
2003091517:03:321	and photographs. Links to history	scenery – mountains lakes and Coastline
0	even the best of Hollywood	scenery designers could provide." She said
2002030817:20:041	courses are situated in impressive	scenery and often offer unique obstacles
2004012704:23:161	Kingdom' and area rich in	scenery , ancient castles and culture. Accommodation
2003053123:27:301	for the beauty of its	scenery and a wealth of resources
2003091316:05:501	central messages about Scotland: its	scenery , history, culture. The most lucrative
0	valley and gives a lake-like	scenery . This part of the sea
2002072305:32:411	the best of the landscape	scenery , history and unusual features. The
2002080619:29:561	provide a range of local	scenery files and some special aircraft
2000021600:00:002	tourism destination. It has magnificent	scenery ; a pristine natural environment; cultural
0	in Reigate and Banstead. Magnificent	scenery and breathtaking views are available
0	remain. Many monuments and magnificent	scenery are visible from this location
0	Si-sa Wai Situated among magnificent	scenery southwest of Wat Mahathat is
2003090513:11:421	Bay is surrounded by magnificent	scenery that offers miles of cliff
2002073110:45:581	tradition, along with its magnificent	scenery , unique heritage and friendly folk
1998010100:00:004	pretty little villages and magnificent	scenery . Convenient for the Dales, the
0	Explore the wild, lonely, magnificent	scenery , perhaps see a Golden Eagle
2004021404:50:271	by Ben Nevis and majestic	scenery , less than half an hour
0	50 miles through the marvellous	scenery of the Derbyshire Dales between
2003060414:54:501	by stunning coastal and moorland	scenery , The Gurnard's Head is the
0	heightened by the awe-inspiring mountain	scenery and deeply peaceful culture, we
2004020915:51:091	to castles to spectacular mountain	scenery , Scotland has a wealth of

0	combined with the dramatic mountain	scenery , makes it a popular area
2004021404:50:271	hotel set amidst stunning mountain	scenery on the famous road to
2003071611:26:051	bunkhouse ideal for hillwalkers. Mountain	scenery . Free fishing. Ballifeary House Hotel
2000072900:00:003	beauty spots, forests and mountain	scenery . Conservation of wild animals, for
0	in stunning coastal and mountain	scenery . Award winning restaurant with coveted
2001102915:03:111	is surrounded by beautiful mountain	scenery and has several beaches along
0	their original splendour. Stunning natural	scenery , beautiful lush mountains, white sandy
2003091316:05:501	Shetland (prehistoric sites, quiet, natural	scenery). Different people come to Scotland
2003050910:52:061	walks & rambles The natural	scenery can also be explored by
2003010100:00:004	cyclists rank the importance of	scenery when choosing a destination, and
0	visitors a great diversity of	scenery , culture and leisure activities. The
2004021407:02:581	here." "My impression is of	scenery , wide open spaces, big skies
2004011023:36:241	a very agreeable variety of	scenery , with lochs, mountains, rivers and
2003052317:10:191	explored. A wide variety of	scenery awaits the visitor, with valleys
2004011617:14:221	providing a rich variety of	scenery to explore from cattle-grazed valleys
2003063000:00:002	world's best tourist attractions or	scenery . In fact they are all
2004012914:49:532	matches the beauty of our	scenery ." (Tourism in the Highalnds Towards
0	dramatic backdrop of our outstanding	scenery is one obvious draw, providing
0	with spectacular views and panoramic	scenery . A map and a good
2003021113:35:161	for romantic beauty and picturesque	scenery it cannot be surpassed. The
0	much to offer, from picturesque	scenery , history, tranquility, and rural appeal
2001031209:42:141	to the natural history, picturesque	scenery and antiquities of the western
2001031209:42:141	its situation, salubrity and picturesque	scenery . Also a account of the
2003122314:49:121	and Devon. See the picturesque	scenery and open moors of Dartmoor
0	the most dramatic and picturesque	scenery in Ireland. Rising dramatically from
2003092612:08:551	some of the most picturesque	scenery in North Yorkshire. The town
2002010100:00:004	some of the most remarkable	scenery to he found. The Park
2004010100:00:004	famous for its natural river	scenery . more details >> Lydney
2003021113:35:161	range of picturesque and romantic	scenery . The adjacent cliffs, which are
2004010117:11:091	Fort William through the rugged	scenery of Glenfinnan, over a famous
2002070407:37:081	known for its gentle rural	scenery there is still the odd
2003111715:14:272	Scottish tourism: Your views Scotland's	scenery is a a big pull
2003071611:26:051	Bay enjoying the Western Scotland	scenery . The majority of the holiday
2003071611:26:051	Croft amid the beautiful Scottish	scenery where the Red Kite, Buzzards
2003091517:03:321	photographs of Brecon Beacons showing	scenery , land use, and management problems
2003102218:03:301	challenging fairways. Combined with spectacular	scenery , these courses make Victoria an
2003111900:00:003	towns and islands with spectacular	scenery are within close proximity to
2003020713:27:001	ASHDOWN FOREST Experience the spectacular	scenery of this ancient Royal Hunting
0	road A MILD climate, spectacular	scenery , clean sandy beaches and a
0	Cruise from Ardfern through spectacular	scenery while watching birds,cetaceans,and
0	near Oban. Explore & enjoy spectacular	scenery and wildlife Mitchell's Family Amusement
0	exist within the area. Spectacular	scenery is within easy reach of
2003071611:26:051	2 bedrooms, set in spectacular	scenery between Portree and Dunvegan Castle
2004013011:07:571	of Cornwall, with its spectacular	scenery and rich heritage. The area
2003092216:04:581	some of the most spectacular	scenery in Britain. What we look
2000020322:42:051	some of the most spectacular	scenery in Wales.

0	is set within the spectacular	scenery of the Snowdonia mountain ranges
2004010117:11:091	Enjoy a day experiencing spectacular	scenery 65 miles south west of
0	some of the most spectacular	scenery on earth learn meditation techniques
0	of many tourists. The spectacular	scenery of mountain, river and gorge
0	themed carriages reflect the spectacular	scenery . Queensland Holidays - Tour Search
2004021407:09:391	a rural haven of spectacular	scenery the majority of which is
2004011316:19:221	area Crianlarich Centre of spectacular	scenery Crieff Second largest Perthshire town
2004010612:32:171	of the UK's most spectacular	scenery . Trips to the Lake District
0	its wild ponies and stunning	scenery , picture-book villages with thatched cottages
2003082021:32:571	holidays taking in Scotland's stunning	scenery and world-renowned golf courses Rampant
2003102218:03:301	is renowned for its stunning	scenery and charming coastal villages. The
2003010100:00:004	I can enjoy the stunning	scenery with a bike ride to
2002072305:32:411	show some of the stunning	scenery , mills, canals, viaducts and many
2002072305:32:411	of the Screen The stunning	scenery and timeless towns and villages
2003121611:02:421	account of regional strengths: Stunning	scenery and lakes (but not all
2003121611:02:421	of the region's appeal. Stunning	scenery and lakes (but not all
2003071610:45:031	form the door with stunning	scenery and views. Pets are also
2003010100:00:004	Enjoy 18 miles of stunning	scenery in the North Yorkshire Moors
0	play area, family friendly, stunning	scenery The Scottish Sealife Sanctuary Teeming
2003111715:14:272	that combined with Scotland's stunning	scenery would make for an irresistible
2003090513:11:421	Golf Course , with its superb	scenery overlooking the Irish sea. If
0	the major islands. The superb	scenery of the islands includes high
2003071611:26:051	and majesty of the surrounding	scenery , with the awe-inspiring backdrop of
2004020606:09:141	countryside around the town. The	scenery is wonderful and remains largely
2004010100:00:004	visitors a year enjoying the	scenery , heritage and arts in this
2004020606:09:141	near the Yunnan border. The	scenery here is very different from
0	to explore more of the	scenery and legacy of historic castles
2003100120:37:301	slow, pointing out all the	scenery which I have seen before
2003100120:37:301	outstanding natural beauty, where the	scenery is amazing. They are only
2003071611:26:051	while relaxing and enjoying the	scenery and local wildlife. The Clachan
2002010911:58:431	during summer months and the	scenery and wildlife along the banks
2004010817:43:201	explored from Pateley Bridge. The	scenery is well wooded in places
2002082011:56:551	that the enjoyment of the	scenery by the public can be
2002121900:00:002	to the holiday experience. The	scenery of Scotland was the highest
0	part of the team. The	scenery and wildlife in this remote
2002082011:56:551	Lake District and Snowdonia the	scenery is largely the result of
2004010100:00:004	so visitors can operate the	scenery . [an error occurred while processing
2001012722:29:331	nature the beauty of the	scenery and the diversity of the
2004012910:17:041	for lunch and enjoy the	scenery . Day 7. - Nha Trang - Tuy
2002011510:21:581	destroy wildlife habitats, ruin the	scenery , and increase air and noise
2004010100:00:004	recommend Malvern to anyone, the	scenery and and places to visit
2003090513:11:191	coast, whilst 35% mentioned the	scenery /landscape or peace and quiet
0	industry remains to spoil the	scenery . A mixture of mature trees
2000020322:42:051	to offer every visitor. The	scenery quite often breathtaking ; from snow
2001061510:00:481	Park just to admire the	scenery . Congestion of Villages and Beauty
2003010100:00:004	Whitby and Filey or the	scenery of the Dales and Moors
2004010100:00:004	drop down from heaven. The	scenery , the buildings. I am in
3	of the country where the	scenery is a mirror of the
2003080513:04:031	own pace, and enjoy the	scenery to the fullest. It can

0	or just take in the **scenery** . The market town of Devizes
0	staying with family, driving through **scenery** , sampling indigenous or culinary culture
2004012609:20:362	into sampling the areaâ€™s unique **scenery** and rich history. Frank Werkmeister
2000010100:00:004	our company, enjoy open, unspolied **scenery** and discover treasures of Scottish
2002062809:38:101	the quality of the upland **scenery** . The area is also outstanding
2004021407:02:581	people." "Space, lovely villages, varied **scenery** , kindly people and easy driving
0	the Black Country with varied **scenery** to tempt you to explore
0	underline } --> Water **scenery** Two hundred years ago, the
2001061510:00:421	Gradually the taste for wild **scenery** grew and Ruskin enjoyed the
2004010100:00:004	were impressed by the wild **scenery** , the prolific bird life and
0	or interest. From history, wildlife **scenery** , shopping, eating out or special
0	greatest asset is its wonderful **scenery** , whether it is the North
2004021415:38:591	visitors. You can view wonderful **scenery** , receive a warm and friendly
2002072305:32:411	never far from some wonderful **scenery** . Whether you want a gentle
2002072305:32:411	to take in the wonderful **scenery** of the Holme Valley. In
2004010100:00:004	Aviv, Eilat enjoys the wonderful **scenery** of the Edom Mountains and
2003091621:49:321	enjoy the true hospitality, wonderful **scenery** and fascinating culture and history
2004021317:58:581	they do have such wonderful **scenery** . For self-catering, there are many
2002073110:45:581	the beauty of Derwentside's wonderful **scenery** , why not take a balloon
0	holiday; there is also wonderful **scenery** , nightclubs, bars and restaurants, as
2003121012:45:371	Harrogate District Tourism Home Wonderful **scenery** , a host of attractions, excellent
2004010100:00:004	and fauna in beautiful woodland **scenery** . The majority of these walks
0	ancestry, to view the world-class **scenery** and wildlife, to explore Scotland's

Appendix 2 – Collocates of SCENERY from the BNC subcorpus *Miscellaneous:*

	WORD	# TIMES NEARBY	TOTAL IN CORPUS
1	BEAUTIFUL	54	8394
2	SPECTACULAR	40	1929
3	DRAMATIC	27	3817
4	MAGNIFICENT	20	1972
5	BREATHTAKING	19	333
6	STUNNING	16	935
7	COASTAL	15	1425
8	WILD	15	5401
9	GOOD	14	80204
10	WONDERFUL	13	4671
11	ALPINE	10	481
12	SUPERB	10	2063
13	IMPRESSIVE	9	2915
14	FANTASTIC	8	1134
15	BEST	8	34096
16	UNSPOILT	7	184
17	PASSING	7	4648
18	NATIONAL	7	37541
19	MOUNTAINOUS	6	223
20	RUGGED	6	352
21	VARIED	6	2855
22	ATTRACTIVE	6	5051
23	LOVELY	6	6038
24	SCOTTISH	6	9904
25	FULL	6	27569
26	GREAT	6	45312
27	GLORIOUS	5	1078
28	FINEST	5	1717
29	SURROUNDING	5	2990
30	CHANGING	5	6254
31	RICH	5	6705
32	FINE	5	12887
33	LOCAL	5	46003
34	OTHER	5	139504
35	ENCHANTING	4	193
36	PICTURESQUE	4	604
37	DELIGHTFUL	4	1069
38	SANDY	4	1318
39	FASCINATING	4	1643
40	OUTSTANDING	4	2996
41	EXCITING	4	3266
42	EXCELLENT	4	6620
43	FLAT	4	8158
44	OPEN	4	29268
45	BRITISH	4	35428
46	DIFFERENT	4	47607
47	OLD	4	52486
48	MAJESTIC	3	283
49	TOWERING	3	379
50	RUINED	3	1086
51	SPLENDID	3	1655
52	MARVELLOUS	3	1771
53	REMARKABLE	3	3473
54	FRIENDLY	3	3951
55	CHINESE	3	4148

56	FRESH	3	6614
57	DEEP	3	10217
58	SIMPLE	3	13711
59	NATURAL	3	14068
60	EASY	3	14414
61	SPECIAL	3	21852
62	MAIN	3	24778
63	BIG	3	24853
64	PAST	3	25393
65	VERY	3	119583
66	CREAKY	2	31
67	EVER-CHANGING	2	91
68	CRAGGY	2	116
69	PLACID	2	194
70	SILVERY	2	251
71	WILDER	2	252
72	SUBLIME	2	253
73	SCENIC	2	275
74	INSPIRING	2	340
75	WOODED	2	349
76	LUSH	2	420
77	SENSATIONAL	2	451
78	EXQUISITE	2	522
79	GLAMOROUS	2	560
80	ABUNDANT	2	598
81	VOLCANIC	2	688
82	CONTRASTING	2	699
83	AUSTRIAN	2	711
84	MAGICAL	2	831
85	ROCKY	2	1008
86	GRIM	2	1011
87	CHALLENGING	2	1227
88	SWISS	2	1424
89	GEOGRAPHICAL	2	1599
90	ACCESSIBLE	2	1625
91	BORING	2	1656
92	DULL	2	1742
93	ARTIFICIAL	2	1978
94	ROMANTIC	2	1984
95	INSTANT	2	1992
96	HISTORIC	2	2292
97	STRIKING	2	2576
98	PLEASANT	2	2580
99	GENTLE	2	2765
100	PAINTED	2	2977

Appendix 3 - Concordances for «basal» from the ORAL CANCER corpus
WordSmith Tools -- 23/07/2008 17.30.23

1 sually consists of 5-7 cell layers. The basal cell layers are elongated cells ar
2 ettage is a commonly used treatment for basal cell carcinomas for tumors smaller
3 od nutrition plays a vital role in your basal cell skin cancer treatment regime.
4 5-0783, USA. AB - BACKGROUND: A case of basal cell carcinoma (BCC) developing in
5 does not abruptly stop in a palisading basal layer at a stromal interface. Rath
6 a light-activated drug that targets the basal cell skin cancer cells. The second
7 affected by these premalignant lesions. Basal cell carcinoma generally appears a
8 body. Geography can also play a role in basal cell skin cancer. If you live in a
9 00 Americans were diagnosed with either basal cell cancer or squamous cell cance
10 nant tumor, and is more aggressive than basal cell cancer, but still may be rela
11 mage or appearance issues you may have. Basal Cell Carcinoma Skin Cancer - Treat
12 his approach because we understand that basal cell skin cancer affects all of yo
13 ole-body approach. If you are exploring basal cell skin cancer treatment options
14 owing types are used to plan treatment: Basal cell cancer Basal cell cancer is t
15 K14 are expressed in the proliferating basal layer cells, whereas keratins K1 a
16 s are directly related to malnutrition. Basal cell skin cancer can cause weight
17 on, mitotic figures are the norm in the basal and parabasal layers. Mitotic figu
18 ctinic Keratosis Atypical Fibroxanthoma Basal Cell Carcinoma Keratoacanthoma Pyo
19 d to the sun such as the head and neck. Basal cell carcinoma is slow growing. It
20 1-800-422-6237); TTY at 1-800-332-8615. BASAL CELL CARCINOMA OF THE SKIN Treatme
21 Cryosurgery can be used for some small basal cell carcinomas but is not recomme
22 ctinic Keratosis Atypical Fibroxanthoma Basal Cell Carcinoma Keratoacanthoma Pyo
23 blocks, prevents, separates, or limits. basal cell Ã Ã (BAY-sul SEL) A small, ro
24 ve S., Orange, CA 92868, USA. AB - True basal cell carcinoma (BCC) involving the
25 Also called Gorlin syndrome and nevoid basal cell carcinoma syndrome. baseline
26 cer treatment regime. An individualized basal cell carcinoma skin cancer treatme
27 laces where cancer may develop. How are basal cell carcinoma dn squamous cell ca
28 on skin cancer in children, followed by basal cell and squamous cell carcinomas
29 ype of skin cancer that arises from the basal cells, small round cells found in
30 that is not cancer, but can change into basal cell or squamous cell skin cancer
31 all skin cancers. It originates in the basal cells, at the bottom of the epider
32 carcinoma and squamous cell carcinoma. Basal cell carcinoma begins in the lowes
33 blastic carcinoma that is distinct from basal cell carcinoma. 68 UI - 21336188 A
34 quamous cell cancer spreads faster than basal cell cancer , but still may be rel
35 ntiation, glandular differentiation, or basal-cell differentiation are present w
36 veral layers of cells. The cells of the basal layer are orientated vertical on t
37 s classified into five different types: Basal cell carcinoma (BCC) is the most c
38 the cells that cover or line an organ.) Basal cell carcinoma accounts for more t
39 /crr/types/skin/BasalCell.asp What are Basal Cell and Squamous Cell Carcinomas?
40 lei were detected in the well-organized basal and suprabasal layers of esophagi
41 ed under melanoma). Malignant Neoplasms Basal Cell Carcinoma A common sunlight i
42 a dn squamous cell carcinoma diagnosed? Basal cell carcinoma and squamous cell c
43 spreads, but it does so more often than basal cell carcinoma. However, it is imp
44 tivate the drug which then destroys the basal cell skin cancer cells without dam
45 Huang Z, Giovannucci E, et al. Diet and basal cell carcinoma of the skin in a pr
46 zema, infections, trauma, or psoriasis. Basal and squamous cell carcinomas are g
47 emotionally devastating a diagnosis of basal cell skin cancer can be and will h
48 squamous epithelia, the single layer of basal cells expresses K5 and K14; when t
49 exhibit multiple cell types: squamous, basal, mucoepidermoid, verrucous and jun
50 ntrol animals expressed K14 only in the basal layer. Moreover, we observed more
51 it is far more serious. Melanoma, like basal cell and squamous cell cancers, is
52 ith this syndrome have a higher risk of basal cell carcinoma. Also called Gorlin
53 skin_cancer.cfm Basal Cell Skin Cancer Basal Cell Skin Cancer To Learn More Abo
54 of America in addition to treating your basal cell skin caner, we also strive to
55 st for sun-induced lesions such as AKs, basal cell carcinoma (BCC), melanoma, or
56 y can invade and destroy nearby tissue. Basal cell carcinoma and squamous cell c
57 lack. Melanoma is much less common than basal cell and squamous cell skin cancer
58 ke a scar, and it is firm to the touch. Basal cell cancers may spread to tissues
59 ever, melanoma is much more likely than basal or squamous cell cancer to metasta
60 l excision is the usually treatment for basal cell carcinoma. This disease remai
61 low growing. It is highly unusual for a basal cell cancer to spread to distant p

eatment_of_Basal_Cell_Carcinoma_51.asp Basal cell carcinoma very rarely spreads
 recurrence rate is similar to that for basal cell cancers. Larger squamous cell
the melanocytes. It is not as common as basal cell or squamous cell skin cancer,
oot/CRI/content/CRI_2_4_4X_Treatment_of_Basal_Cell_Carcinoma_51.asp Basal cell
 the tumor out) is often used to remove basal cell carcinomas, along with a marg
ly slow-growing. It is more likely than basal cell cancer to spread (metastasize
gnant tumor. It is more aggressive than basal cell cancer , but still may be rel
ly slow-growing. It is more likely than basal cell cancer to spread (metastasize
 RN; Barr RJ; Jensen JL; Cantos KA TI - Basal cell carcinoma of the buccal mucos
ce exhibited K14 expression only in the basal layer. The expression pattern of K
wo most common kinds of skin cancer are basal cell carcinoma and squamous cell c
r C, Kurzen H, Hassfeld S: Infiltrating basal cell carcinoma of the neck 34 year
 the lowest layer of the epidermis, the basal cell layer. About 75% of all skin
he suprabasal layers in addition to the basal layer of the tongues. In contrast,
l suprabasal layers, in addition to the basal layer, in tongues from carcinogen-
tures Most are located in the mid-face. Basal cell carcinomas appear first as pe
e are three major types of skin cancer. Basal cell carcinoma: The most common fo
pical keratinocytes may be found in the basal layer and often extend deeply down
 that you may someday be diagnosed with basal cell skin cancer. Basal Cancer Cel
0% of non-melanoma skin cancers , (with basal cell carcinomas accounting for abo
ent membrane. In normal epithelium, the basal cell layer, and perhaps the one di
 it metastasizes to produce its damage. Basal-cell carcinomas are locally invasi
sible. The methods chosen to treat your basal cell skin cancer are based on your
d mitotic figures (especially above the basal layers), and atypical mitoses. In
oplasm do not occur in the oral cavity. Basal cell carcinoma is a very common sk
d to distant parts of the body than are basal cell carcinomas. Even so, very few
ayer. About 75% of all skin cancers are basal cell carcinomas. They usually begi
t Centers of America (CTCA) use various basal cancer cell skin treatment tools t
iew of treatment modalities for primary basal cell carcinoma. Arch Derm . 1999;1
nical trial of beta carotene to prevent basal-cell and squamous-cell cancers of
sk of developing pre-cancerous lesions, basal cell carcinomas, and squamous cell
ant parts of the body. After treatment, basal cell carcinoma can come back (recu
sal Cell Carcinoma of the Face >>> Both basal and squamous cell cancers are foun
n type of nonmelanoma skin cancer after basal cell carcinoma. Most SCCs occur on
e, K14 was only clearly detected in the basal layer (Fig. 3 A) . Additionally,
epidermis, the outer layer of the skin. basal cell carcinoma Â Ã (BAY-sul SEL KA
epidermis, the outer layer of the skin. basal cell nevus syndrome Â Â (BAY-sul S
e S, Garrison P, Oakleaf K, Johnson SD. Basal cell carcinoma and lifestyle chara
ined with anti-p16 antibody were in the basal and suprabasal layers of mouse ton
rcenter.com/basal_cell_skin_cancer.cfm Basal Cell Skin Cancer Basal Cell Skin C
CURRENT URL http://www.cancercenter.com/basal_cell_skin_cancer.cfm Basal Cell S
hat do not make pigment it may begin in basal cells (small, round cells in the b
ed to plan treatment: Basal cell cancer Basal cell cancer is the most common typ
ned nuclei were present not only in the basal layer, but some staining was also
therapies provides the best results for basal cancer cell skin treatment. Some o
th SCC. Pathophysiology SCC arises from basal keratinocytes of the skin. It typi
ennie G, Selwood T: The relationship of basal cell carcinomas and squamous cell
 start in the skin. The most common are basal cell cancer and squamous cell canc
ng Slides) Contact a Dermatologist Like basal cell carcinoma, squamous cell carc
lled squamous cells; round cells called basal cells; and cells called melanocyte
ice, most p16-stained cells were in the basal layer of the tongue epithelium, wh
h 4-NQO; K14 staining was found in both basal and suprabasal layers, whereas in
 for Ber-EP4 support an origin from the basal cell layer of stratified squamous
 al: Use of tanning devices and risk of basal cell and squamous cell skin cancer
kin Cancer More than 1 million cases of basal and squamous cell skin cancers wil
 the two types that are most common are basal cell carcinoma and squamous cell c
Mohs surgery has the best cure rate for basal cell carcinoma. It is especially u
 the epithelium. The cell cycle time of basal epithelial cells is approximately
s depending on the type of skin cancer: Basal cell carcinoma: generally excellen
unding tissue. Chemotherapy (Topical) - basal cell carcinoma skin cancer medicat
y (directly onto the skin) to fight the basal cell skin cancer In addition to tr
thickness atypia. If, however elongated basal cells or flattened surface cells a
tion Therapy - this type of therapy for basal cell carcinoma of the skin therapy

gh, dry, or scaly. To see an example of Basal Cell Carcinoma of the Face >>> Bot
e following to enrich your treatment of basal cell carcinoma of the skin: Naturo
 high-energy rays to shrink or kill the basal cell skin cancer cells. There are
ll skin treatment tool is used to treat basal cell skin cancer in many instances
 skin treatment tools to help you fight basal cell skin cancer. CTCA uses both t
the aerodigestive tract or both SCC and basal cell carcinomas (BCC) of the skin.
ation. Especially, expression above the basal cell layer has been highly predict
skin treatment. Some of the traditional basal cancer cell skin treatment therapi
in tumors recur. Skin Cancer (Melanoma, Basal Cell Carcinoma, Squamous Cell Carc
the United States. Unlike some cancers, basal cell skin cancer is slow growing a
/201228.html Treatment Option Overview Basal Cell Carcinoma Of The Skin Squamou
 when viewed by light microscopy, shows basal cells which have changed shape fro
the EGFR was expressed primarily in the basal layer of the tongue epithelium, wh
nt therapies we use are: Surgery - this basal cancer cell skin treatment tool is
stained nuclei were present in both the basal and suprabasal layers of epithelia
cer can be and will help to enrich your basal cell skin cancer treatment by offe
out This Topic: Chat with Us | Email Us Basal Cell Skin Cancer is one of the mos
 diagnosed with basal cell skin cancer. Basal Cancer Cell Skin Treatment The doc
fined epithelium with a single layer of basal cells. 4-NQO-treated mouse tongues
RA; Kelly EB; Wright ST; Wagner RF TI - Basal cell carcinoma arising in a cleft
ptions. If you have been diagnosed with basal cell skin cancer, it is important
sis of your blood. Mind-body Medicine - basal cancer cell skin treatment also in

Appendix 4

7	cancer 4605	26	tumore 1014
16	may 1833	27	cellule 986
17	oral 1795	33	può 799
18	radiation 1746	35	trattamento 720
19	cell 1642	37	pazienti 684
20	treatment 1528	39	rischio 588
23	patients 1379	43	tumori 537
25	cells 1346	44	malattia 531
26	therapy 1289	46	cancro 512
27	lung 1247	47	terapia 505
28	carcinoma 1211	49	anni 483
33	can 1155	51	orale 472
53	surgery 705	52	casi 464
54	cancers 684	71	radioterapia 327
55	head 682	75	dolore 314
56	used 681	78	chirurgia 309
59	disease 646	79	tessuto 309
60	tumors 637	81	farmaci 304
65	lymph 539	82	tipo 302
66	use 539	83	caso 295
67	mouth 531	85	medico 292
71	blood 505	86	due 290
73	survival 497	89	prima 275
74	chemotherapy 494	90	lesioni 274
79	information 472	91	sintomi 270
81	body 458	96	fattori 256
84	clinical 452	98	linfonodi 247
86	should 446	100	possibile 242
88	nodes 441	102	collo 239
90	tobacco 422	106	modo 231
91	patient 417	107	cavo 230
94	called 415	108	grado 225
97	small 391	111	effetti 220
99	stage 390	112	tempo 220
100	type 389	113	chemioterapia 219
102	lesions 377	119	donne 205
103	tissue 370	120	sopravvivenza 203
106	common 365	121	test 202
111	effects 342	122	tumorali 199
112	smoking 341	124	meno 192
113	include 327	125	mammella 190
114	health 326	127	forma 186
115	medical 326	128	tessuti 186
117	node 323	129	numero 184
118	cavity 322	130	secondo 184
119	cases 321	131	risultati 182
120	new 321	132	sangue 182
121	treated 320	133	Base 181
122	tongue 319	138	metastasi 178
124	time 313	139	stadio 177
126	associated 309	141	malattie 173
127	results 309	142	persone 172
129	primary 307	144	fumo 165
130	found 306	145	vescica 165
132	human 305	152	intervento 161
133	biopsy 296	153	clinica 160
134	spread 295	154	causa 160
136	diagnosis 290	155	livello 160
138	increased 287	157	chirurgico 159
139	years 286	158	deve 159
140	bone 285	160	organi 156
141	factors 285	161	bocca 155
142	early 280	164	volte 154
144	exposure 279	165	vita 152
145	area 278	169	stomaco 150
147	side 276	170	cura 149
148	control 273	171	screening 149
149	help 273	172	corpo 148
150	normal 272	173	particolare 147
151	studies 271	174	esempio 146
153	people 270	175	età 146
154	types 265	177	piccole 145
157	two 257	179	rispetto 145
158	carcinomas 256	183	diversi 141
159	rate 256	184	fase 140
160	cause 255	185	sistema 140
163	high 253	187	esame 137
164	expression 252	190	polmonare 135
167	doctor 247	191	importante 134
168	symptoms 246	193	mucosa 132
169	radiotherapy 242	194	cute 131
172	including 237	195	mesi 129
174	malignant 233	197	tipi 128

References

Aijmer K. and Altenberg B. (eds.). 1991. *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London: Longman.

Aijmer, K. and Altenberg, B. (eds.). 2004. *Advances in Corpus Linguistics*. Amsterdam: Rodopi.

Asadi, S. and Jamali, H. R. 2004. «Shifts in search engine development: A review of past, present and future trends in research on search engines». *Webology*, 1(2),6. http://www.webology.ir/2004/v1n2/a6.html.

Aston G., Bernardini, S. and Stewart, D. (eds.). 2004. *Corpora and Language Learners*. Amsterdam: Benjamin.

Atkins S., Clear, J. and Osler, N. 1992. «Corpus Design Criteria». *Literary and Linguistic Computing*, 7 (1), 1-16.

Baayen, H.R. 2001. *Word frequency distributions*. Dordrecht: Kluwer

Banko, M. and Brill, E. 2001. «Scaling to Very Very Large Corpora for Natural Language Disambiguation». In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 26-33. http://research.microsoft.com/~brill/Pubs/ACL2001.pdf.

Baroni, M. and Bernardini, S. 2004. «BootCaT: Bootstrapping corpora and terms from the web». In *Proceedings of LREC 2004*, Lisbon: ELDA, 1313-1316. http://sslmit.unibo.it/~baroni/publications/lrec2004/bootcat_lrec_2004.pdf.

Baroni, M. and Bernardini, S. 2006. *Wacky! Working Papers on the Web as Corpus*. Bologna: Gedit. http://wackybook.sslmit.unibo.it/.

Baroni, M. and Kilgarriff, A. 2006. «Large linguistically-processed Web corpora for multiple languages». In Baroni, M. and Kilgarriff, A. (eds.), 87-90. http://www.aclweb.org/anthology-new/E/E06/E06-2001.pdf.

Baroni, M. and Kilgarriff, A. (Eds.). 2006. *Proceedings of the 2nd International Workshop on Web as Corpus (EACL06)*.Trento: Italy.

Baroni, M., Kilgarriff, A., Pomikalek, J. and Rychly, P. 2006. «WebBootCat: a Web Tool for Instant Corpora». *Proceeding of the Eu-*

raLex Conference 2006. Alessandria: Edizioni dell'Orso, 123-132. http://www.kilgarriff.co.uk/Publications/2006-BaroniKilgPomikalekRychly-ELX-WBC.doc.

Baroni, M. and Ueyama, M. 2006. «Building general- and special-purpose corpora by Web crawling». In *Proceedings of the NIJL International Symposyum, Language Corpora: Their Compilation and Application*, 31-40. http://clic.cimec.unitn.it/marco/publications/bu_wac_kokken_formatted.pdf.

Bates, M. 2002. «Toward an Integrated Model of Information Seeking and Searching». *Fourth International Conference on Information Needs, Seeking and Use in Different Contexts*, Lisbon, Portugal, September 11-13, 2002. http://www.gseis.ucla.edu/faculty/bates/articles/info_SeekSearch-i-030329.html.

Battelle, J. 2005. *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*. Nicholas Brealey Publishing.

Bergh, G. 2005. «Min(d)ing English language data on the Web. What can Google tell us?». *ICAME Journal*, 29, 25-46. http://gandalf.aksis.uib.no/icame/ij29/ij29-page25-46.pdf [Accessed 2007-06-06].

Bergh, G., Seppänen, A. and Trotta, J. 1998. «Language corpora and the Internet: A joint linguistic resource». In A. Renouf (ed.), 41-54.

Bernardini, S. 2000. «Systematising serendipity: Proposals for concordancing large corpora with language learners». In L. Burnard and T. McEnery (eds.), 225-234.

Bernardini, S. 2002. «Exploring new directions for discovery learning». In B. Kettemann and G. Marko (eds.),165-182.

Bernardini, S. 2004. «Corpora in the classroom: an overview and some reflections on future developments». In J. Sinclair (ed.), 15-36.

Bernardini, S. 2006. «Corpora for translator education and translation practice. Achievements and challenges». In *Proceedings of LREC 2006*. http://mellange.eila.univ-paris-diderot.fr/bernardini_lrec06.pdf.

Biber, D. 1992. «Representativeness in Corpus Design». Reprinted in J. Sampson and D. McCarthy (eds.),174-197.

Biber, D., Conrad, S. and Reppen, R. 1998. *Corpus Linguistics. Investigating Language Structures and Use*. Cambridge: Cambridge University Press.

Biber, D. and Kurjian, J. 2007. «Towards a taxonomy of web registers and text types: A multi-dimensional analysis.» In M. Hundt *et al.* (eds.), 109-132.

Boase, E. 2005. *Stereotyping the Web: Genre Classification of Web Documents*. Unpublished M.Sc. Dissertation. Colorado State University. http://www.cs.colostate.edu/~boese/Research/masters.pdf.

Bowker, L. and Pearson, J. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge

Brandt, D. 2002. *Pagerank: Google's Original Sin*, http://www.google-watch.org/pagerank.html.

Bray, T. 1996. «Measuring the Web». *Proceedings of the fifth international World Wide Web conference on Computer networks and ISDN systems*, 993-1005, http://www.tbray.org/ongoing/When/199x/1996/05/07/OVERVIEW.HTM.

Brin, S. and Page, L. 1998. «The Anatomy of a Large-Scale Hypertextual Web Search Engine», *Computer Networks and ISDN Systems*, 30 (1), 107-117. http://infolab.stanford.edu/pub/papers/google.pdf.

Broder, A. 2002. «A taxonomy of web search». *ACM SIGIR Forum Archive*, 36(2), 3-10. http://www.acm.org/sigir/forum/F2002/broder.pdf.

Burnard, L. and McEnery, T. (eds.). 2000. *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt am Main: Peter Lang.

Calvo, H. and Gelbukh, A. 2003. «Improving Prepositional Phrase Attachment Disambiguation Using the Web as Corpus». *Progress in Pattern Recognition, Speech and Image Analysis: 8th Iberoamerican Congress on Pattern Recognition*, CIARP. http://www.gelbukh.com/CV/Publications/2003/CIARP-2003-PP-attachment.pdf.

Castagnoli, S. 2006. «Using the Web as a Source of LSP Corpora in the Terminology Classroom». In M. Baroni and S. Bernardini (eds.), 159-172.

Chakrabati, S. 1999. «Hyperserching the Web». *Scientific American*, 280. (6), 54-60. http://econ.tepper.cmu.edu/ecommerce/hypersearch.pdf.

Chakrabarti, S. 2003. *Mining the Web. Discovering Knowledge from Hypertext Data*. San Francisco: Morgan Kaufmann.

Chomsky, N. 1962. Paper given at the 3rd Texas Conference on Problems of Linguistic Analysis in English (1958). Austin: University of Texas.

Ciaramita, M. and Baroni, M. 2006. «Measuring Web-corpus randomness: A progress report». In M. Baroni and S. Bernardini (eds.), 127-158.

Connor, U. and Upton, T. (eds.). 2004. *Corpus Linguistics in North America 2002: Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics.* Amsterdam: Rodopi.

Crowston, K. and Kwasnik, B. 2004. «A Framework for Creating a Facetted Classification for Genres: Addressing Issues of Multidimensionality». In *Proceedings of the 37th Hawaii International Con-*

ference on System Sciences. http://csdl2.computer.org/comp/proceedings/hicss/2004/2056/04/205640100a.pdf.
Crowstone, K. and Williams, M. 1997. «Reproduced and Emergent Genres on the World Wide Web». In *Proceedings of the 30th Hawaii International Conference on System Sciences: Digital Documents.* http://csdl2.computer.org/comp/proceedings/hicss/1997/7734/06/7734060030.pdf.
Crystal, D. 2006. *Language and the Internet.* Cambridge: Cambridge University Press (2nd edn.).
Day, M. 2003. *Collecting and Preserving the World Wide Web. A feasibility study for the JISC and Wellcome Trust.* http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf.
De Schryver, G. 2002. «Web for / as corpus: a perspective for the African languages». *Nordic Journal of African Studies*, 11(2), 266-282. http://www.njas.helsinki.fi/pdf-files/vol11num2/schryver.pdf.
Elkiss, A. and Resnik, P. 2004. *The Linguist's Search Engine User Guide*. http://lse.umiacs.umd.edu: 8080/lseuser/lseuser.html.
Evert, S., Kilgarriff, A. and Sharoff, S. 2008. *Proceedings of the 4th Web as Corpus Workshop (WAC-4)*, Marrakech, 1 June 2008 http://webascorpus.sourceforge.net/download/WAC4_2008_Proceedings.pdf.
Fairon C., Naets H., Kilgarriff A. and De Schryver, G. 2007. *Building and Exploring Web Corpora (WAC3 - 2007).* Louvain-la-Neuve: Presses Universitaires de Louvain.
Fantinuoli, C. 2006. «Specialized corpora from the Web and term extraction for simultaneous interpreters». In M. Baroni and S. Bernardini (eds.), 173-190.
Ferraresi, A. 2007. *Building a very large corpus of English obtained by Web crawling: ukWaC*. MA thesis, University of Bologna. http: //trac.sketchengine.co.uk/wiki/Corpora/UKWaC.
Ferraresi, A., Zanchetta, E., Baroni, M. and Bernardini, S. (forthcoming), *Introducing and evaluating ukWaC, a very large web-derived corpus of English*. In S. Evert *et al.* (eds.) http://clic.cimec.unitn.it/marco/publications/lrec2008/lrec08-ukwac.pdf.
Firth, J. R. 1957. *Papers in Linguistics 1934-1951.* London: Oxford University Press.
Fletcher, W. 2001. «Concordancing the Web with KWiCFinder». *American Association for Applied Corpus Linguistics, Third North American Symposium on Corpus Linguistics and Language Teaching*, Boston, MA,23-25 March 2001. http://miniappolis.com/KWiCFinder/FletcherCLLT2001.pdf.

Fletcher, W. 2004a. «Facilitating the Compilation and Dissemination of Ad-Hoc Web Corpora». In G. Aston *et al.* (eds.), 271-300.
Fletcher, W. 2004b. «Making the Web More Useful as a Source for Linguistic Corpora». In U. Connor and T. Upton (eds.), 191-205.
Fletcher, W. 2007. «Concordancing the Web. Promise and problems, tools and techniques». In M. Hundt *et al.* (eds.), 25-46.
Francis, N. W. 1992. «Language corpora B.C.». In J. Svartvik (ed.), 17-33.
Ghani, R., Jones, R. and Mladenic, D. 2001. «Mining the Web to Create Minority Language Corpora». In *Proceedings of the 10th international conference on Information and knowledge management*, 279-286. http://widit.slis.indiana.edu/irpub/CIKM/2001/pdf36.pdf.
Ghani, R., Jones, R. and Mladenic, D. 2001. «Building Minority Language Corpora by Learning to Generate Web Search Queries». *Knowledge and Information Systems*, 1, 7. http://www.cs.cmu.edu/-afs/cs/project/theo-4/textlearning/www/corpusbuilder/papers/CMU-CALD-01-100.pdf.
Granger, S. and Petch-Tyson, S. (eds.) *Extending the scope of corpus-based research: new applications, new challenges*. Amsterdam: Rodopi.
Grefenstette, G. 1999. «The WWW as a resource for example-based MT tasks». *Translating and the Computer*, 21, ASLIB, London.
Grefenstette, G. and Nioche, J. 2000. «Estimation of English and non-English language use on the WWW». *Proceedings of the RIAO* (Recherche d'Informations Assistée par Ordinateur) Paris, 12-14 April 2000, 237-246. http:// arxiv.org/ftp/cs/papers/0006/-0006032.pdf.
Gulli, A. and Signorini, A. 2005. «The Indexable Web is more than 11.5 billion pages». *Proceedings of WWW 2005.* Chiba, Japan. http://www.di.unipi.it/~gulli/papers/f692_gulli_signorini.pdf.
Halliday, M.A.K. 1991. «Towards probabilistic interpretations». In E. Ventola (ed.), 39-62.
Halliday, M.A.K. 1991. «Corpus studies and probabilistic grammar». In K. Aijmer and B. Altenberg (eds.), 30-43.
Halliday, M.A.K. 1993. «Quantitative studies and probabilities in grammar». In M. Hoey (ed.), 1-25.
Henziger, M. and Lawrence, S. 2004. «Extracting Knowledge from the World Wide Web». *Proceedings of the National Academy of Sciences*, 101, 5186-5191. http://www.pnas.org/content/101/suppl.1/5186.full.
Hock, R. 2007. *The Extreme Searcher's Internet Handbook. A Guide for the Serious Searcher*. Medford: CyberAge Books.

Hoey, M. (ed.) 1993. *Data. Description. Discourse*. London: HarperCollins.
Hoey, M., Sinclair, J., Teubert, W., Stubbs, M. and Mahlberg, M. (eds.). 2007. *Text, discourse, corpora: theory and analysis*. London: Continuum.
Hundt M., Nesselhauf, N. and Biewer, C. (eds.). 2007. *Corpus Linguistics and the Web*. Amsterdam: Rodopi.
Hunston, S. 2002. *Introducing corpora in applied linguistics*. Cambridge: Cambridge University Press.
Il Ragazzini (2006). *Dizionario Inglese/Italiano-Italiano Inglese*, Bologna: Zanichelli.
Inktomi Research Institute. 2000. *Web Surpasses One Billion Documents*, http://web.archive.org/web/20011003203111/www.inktomi.com/new/press/2000/billion.html.
Jansen, B.J. «The Effect of Query Complexity on Web Searching Results», *Information Research*, Vol. 6 No. 1, October 2000, http://informationr.net/ir/6-1/paper87.html.
Joseph, B. D. 2004. «The Editor's Department: On change in Language and change in language». *Language*, 80, 3, September 2004, 381-383.
Kehoe, A. and Renouf, A. 2002. *WebCorp: Applying the Web to Linguistics and Linguistics to the Web. WWW2002 Conference*, Honolulu, Hawaii.http://www2002.org/CDROM/poster/67/.
Kehoe, A. 2006. «Diachronic linguistic analysis on the web with WebCorp», in Renouf and Kehoe (eds.), 297-308
Keller, A. and Lapata, M. 2003. «Using the web to obtain frequencies for unseen bigrams». *Computational Linguistics*, 29(3), 459-484 http://www.aclweb.org/anthology-new/J/J03/J03-3005.pdf.
Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London: Longman.
Kessler, B., Nunberg, G. and Schütze, H. 1997. «Automatic detection of text genre». http://www.aclweb.org/anthology-new/P/P97/P97-1005.pdf.
Kettemann, B. and Marko, G. (eds.). 2002. *Teaching and learning by doing corpus linguistics. Papers from the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July 2000*. Amsterdam & Atlanta, Georgia: Rodopi.
Kilgarriff, A. 2001. «Web as corpus» In *Proceedings of the Corpus Linguistics Conference (CL 2001)*. University Centre for Computer Research on Language Technical Paper Vol. 13, Special Issue, Lancaster University, 342-344. http://ucrel.lancs.ac.uk/publications/CL2003/CL2001%20conference/papers/kilgarri.pdf.

Kilgarriff, A. and Rundell, M. 2002. «Lexical Profiling Software and its Lexicographic Applications – a Case Study». In *Proceedings of the Tenth EURALEX International Congress*, Copenhagen, Denmark', August 13–17, 2002. Vol. II, 807-818. http://www.macmillandictionary.com/MED-Magazine/November2002/Word-Sketches.PDF.
Kilgarriff, A. 2003. *Linguistic Search Engine*. In *Proceedings of Corpus Linguistics 2003*, Lancaster, UK, 53-58. http://www.bultreebank.org/SProLaC/paper06.pdf.
Kilgarriff, A. and Grefenstette, G. 2003. «Introduction to the Special Issue on the Web as Corpus». *Computational Linguistics*, 29, 3, 333-347. http://acl.ldc.upenn.edu/J/J03/J03-3001.pdf.
Kilgarriff, A. 2003. *What computers can and cannot do for lexicography, or Us precision, them recall*. ftp://ftp.itri.bton.ac.uk/reports/ITRI-03-16.pdf
Kilgarriff, A., Rychly, P., Smerz, P. and Tugwell, D. 2004. *The Sketch Engine*. In *Proceedings Euralex*, Lorient, France, 105-116. http://trac.sketchengine.co.uk/attachment/wiki/SkE/DocsIndex/sketch-engine-elx04.pdf?format=raw.
Kilgariff, A. and Baroni, M. 2006. *Proceedings of the 2nd International Workshop on Web as Corpus*, 3rd April 2006, Trento.
Kilgarriff, A. 2007. «Googleology is Bad Science». *Computational Linguistics*, 33, 1, 147-151. http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.1.147.
Kübler, N. 2004. «Using WebCorp in the classroom for building specialized dictionaries». In K. Aijmer and B. Altenberg (eds.), 387-400.
Kwasnik, B., Crowston, K., Nilan, M. and Roussinov, D. 2000. «Identifying document genre to improve Web search effectiveness». *Bulletin of the American Society for Information Science & Technology*, 27(2), 23-26. http://www.asis.org/Bulletin/Dec-01/kwasnikartic.html.
Lapata, M. and Keller, A. 2004. «The Web as a Baseline: Evaluating the Performance of Unsupervised Web-based Models for a Range of NLP Tasks». In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boston, Massachusetts, 121-128. http://acl.ldc.upenn.edu/hlt-naacl2004/main/pdf/5_Paper.pdf.
Laviosa, S. 2002. *Corpus-Based Translation Studies: Theory, Findings, Applications*. Amsterdam: Rodopi
Lawrence, S. and Giles, C.L. 1999. «Accessibility of Information on the Web». *Nature*, 400, 107-109.
Lee, D. 2001. »Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path Through the BNC

Jungle». *Language Learning and Technology*, 5, 3, 37-72. http://llt.msu.edu/vol5num3/pdf/lee.pdf.
Leech, G. 1992. «Corpora and theories of linguistic performance». In J.Svartvik (ed.), 105-122.
Leech, G. 2007. «New resources or just better old ones? The Holy Grail of representativeness». In M. Hundt *et al.* (eds.), 133-150.
Leistyna, P. and Meyer, C. (eds.). 2003. *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi.
Lucchesi, M. 1987. *Dizionario Medico Inglese-Italiano Italiano-Inglese*. Milano: Garzanti.
Lüdeling, A., Evert, S. and Baroni, M. 2007. «Using Web data for linguistic purposes». In M. Hundt *et al.* (eds.), 7-24.
Mahlberg, M. 2005. *English General Nouns. A corpus theoretical approach*, Amsterdam/Phildalephia: Benjamins.
Maia, B. 1997. 'Do-it-yourself corpora... with a little bit of help from your friends'. In B. Lewandowska-Tomaszczyk and P.J. Melia (eds.). *PALC'97: practical applications in language corpora*, Łódź: Łódź University Press, 403-410.
Mair, C. 2007. «Change and variation in present-day English: integrating the analysis of closed corpora and web-based monitoring». In M. Hundt *et al.*, 233-248.
Manca, E. 2004. *Translation by Collocation: The Language of Tourism in English and Italian*. Birmingham: Tuscan Word Centre.
Manca, E. 2004. «The Language of Tourism in English and Italian: Investigating the Concept of Nature between Culture and Usage». *ESP. Across Cultures* 1, 53-65.
Manca, E. 2007. «Nominalisation vs. verbalisation: the concept of beauty and tranquillity in the British and the Italian languages of tourism». Paper presented at the XXIII AIA conference. Università di Bari. 20-22 settembre 2007.
McEnery, T. and Wilson, A. 2001. *Corpus Linguistics*, Edinburgh: Edinburgh University Press.
Martzoukou, K. 2004. «A review of Web information seeking research: considerations of method and foci of interest». *Information Research*,10(2), http://InformationR.net/ir/10-2/paper215.html.
Meyer, C. *et al.* 2003. «The World Wide Web as linguistic corpus». In P. Leistyna and C. Meyer (eds.), 241-254.
Morley, A. 2006. «WebCorp: A Tool for Online Linguistic Information Retrieval and Analysis». In A. Renouf and A. Kehoe (eds.), 283-295.
Mouhobi, S. 2005, «Bridging the North South Divide». *The New Courier* (UNESCO), November 2005. http://unesdoc.unesco.org/images/0014/001420/142021e.pdf#142065.

Nakov, P. and Hearst, M. 2005a, «A Study of Using Search Engine Page Hits as a Proxy for n-gram Frequencies». In *Proceedings of RANLP '05.* http://biotext.berkeley.edu/papers/nakov_ranlp2005.pdf.
Nakov, P. and Hearst, M. 2005b, «Search Engine Statistics Beyond the n-gram: Application to Noun Compound Bracketing», *Proceedings of CoNLL-2005*, Ann Arbor, MI. http://biotext.berkeley.edu/papers/conll05.pdf.
Nelson, T. 1981. *Literary Machines.* Swarthmore, Pa.
Ntoulas, A., Junghoo C. and Olston, C. 2004, «What's New on the Web? The Evolution of the Web from a Search Engine Perspective». *WWW 2004*. ACM Press, 1-12. http://www2004.org/proceedings/docs/1p1.pdf.
O'Neill, E. T., Lavoie, B.F. and Bennett, R. 2002. «Trends in the Evolution of the Public Web 1998-2002». *D-Lib Magazine*, 9 / 4 (April). http://www.dlib.org/dlib/april03/lavoie/04lavoie.html.
Ohlan, M. 2004. *Introducing Corpora in Translation Studies*. London: Routledge.
Pastore, M. 2000. *Web Pages by Language*. http://www.clickz.com/stats/sectors/demographics/print.php/408521.
Pearson, J. 2000. «Surfing the Internet: Teaching Students to Choose their Texts Wisely». In L. Burnard and T. McEnery (eds.), 235-239.
Rayson, P. *et al.* (eds.) *Proceedings of the Corpus Linguistics 2001 Conference*, Lancaster: UCREL.
Renouf, A. (ed.). 1998. *Explorations inCorpus Linguistics*. Amsterdam: Rodopi.
Renouf, A. 2003. «WebCorp: providing a renewable data source for corpus linguists». In S. Granger and S. Petch-Tyson (eds.), 39-58.
Renouf, A., Kehoe, A. and Mezquiriz D. 2004, «The Accidental Corpus: issues involved in extracting linguistic information from the Web». In K.Aijmer and B.Altenberg (eds.), 403-419.
Renouf, A., Kehoe, A. and Banerjee, J. 2005. «The WebCorp Search Engine: a holistic approach to Web text Search». *Electronic Proceedings of CL2005*, University of Birmingham. http://www.corpus.bham.ac.uk/PCLC/cl2005-SE-pap-final-050705.doc.
Renouf, A. and Kehoe, A. (eds.). 2006. *The Changing Face of Corpus Linguistics*, Amsterdam: Rodopi.
Renouf, A., Kehoe, A. and Banerjee, J. 2007, «WebCorp: an integrated system for web text search». In M. Hundt *et al.* (eds.), 47-67.
Resnik, P. and Smith, N. 2003. «The Web as Parallel Corpus». *Computational Linguistics* (Special Issue on the Web as a corpus), 29 (3), 349-380.
Risvik, K.M. and Michelsen, R. 2002. «Search engines and web dy-

namics». *Computer Networks*, 39, 289-302. http://www.iconocast.com/zresearch/se-dynamicweb1.pdf.
Robb, T. 2003. «Google as a Quick 'n Dirty Corpus Tool». *TESL-EJ*, 7 (2), September 2003. http://www-writing.berkeley.edu/TESL-EJ/ej26/int.html.
Rosenbach, A. 2007. «Exploring constructions on the Web: a case study». In M. Hundt *et al.* (eds.), 167-190.
Rundell, M. 2000. *The biggest corpus of all*. http://www.hltmag.co.uk/may00 /idea.htm.
Sampson, J. and McCarthy, D. (eds.). 2005. *Corpus Linguistics. Readings in a Widening Discipline*. London: Continuum.
Santini, M. 2005. «Web pages, text types and linguistic features: some issues». *ICAME Journal*, Vol. 30, 67-86. http://icame.uib.no/ij30/ij30-page67-86.pdf.
Santini, M. 2007. «Characterizing Genres of Web Pages: Genre Hybridism and Individualization». *40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. http://csdl2.computer.org/comp/proceedings/hicss/2007/2755/00/27550071.pdf.
Santini, M. 2004. «Identification of Genres on the Web: a Multi-Faceted Approach». In M.P.Oakes (ed.). *Proceedings of the ECIR 2004* (*26th European Conference on IR Research*), University of Sunderland (UK), 5-7 April, 2004.
Scannell, K. 2007. «The Crúbadán Project: Corpus building for under-resourced languages». In C. Fairon *et al.*, 5-16, http://borel.slu.edu/pub/wac3.pdf.
Sharoff, S. 2006. «Creating general-purpose corpora using automated search engine queries». In M. Baroni and S. Bernardini (eds.), 63-98
Sharoff, S. 2007. «Classifying Web corpora into domain and genre using automatic feature identification». In C. Fairon *et al.* (eds.), 83-94.
Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.
Sinclair, J. (ed.) 2004. *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.
Sinclair, J. 2005. «Corpus and Text. Basic Principles». In M. Wynne (ed.). *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford: Oxbow Books, 1-16. http://ahds.ac.uk/linguistic-corpora/.
Spink, A. and Jansen, B.J. 2004. «A study of Web search trends». *Webology*, 1(2), Article 4. http://www.webology.ir/2004/v1n2/a4.html.
Stubbs, M. 2001. *Words and Phrases*. Oxford: Blackwell.
Stubbs, M. 2007. «On texts, corpora and models of language», in M. Hoey *et al.* (eds.), 127-162.
Sullivan, D. 2005. «Search Engine Sizes», *Search Engine Watch*, 28th January 2005. http://searchenginewatch.com/showPage.html?page=2156481.
Svartvik, J. 1992. (ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*, Berlin: Mouton de Gruyter.
Teich, E. 2003. *Cross-linguistic variation in system and text: a methodology for the investigation*. Berlin: Mouton de Gruyter.
Teubert, W. 2005. «My version of corpus linguistics». *International Journal of Corpus Linguistics*, 10 (1), 1-13.
Teubert, W. 2007 «Parole-linguistics and the diachronic dimension of the discourse», in Hoey *et al.* (eds.), 57-88.
Tognini Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.
Tognini Bonelli, E. and Manca, E. 2001. «Welcoming Children, Pets and Guests: A problem of Non-equivalence in the Languages of «Agriturismo» and «Farmhouse Holidays». *Textus*, XV, 317-334
Van Rijsbergen. C.J.K. 1979, *Information Retrieval*. http://www.dcs.gla.ac.uk/Keith/Preface.html.
Varantola, K. 2003. «Translators and disposable corpora». In Corpora in translator education. Manchester: St Jerome. 55-70.
Ventola, E. (ed.) 1991. *Functional and Systemic Linguistics: Approaches and Uses*, Berlin: Walter de Gruyter.
Véronis, J. 2005. *Google missing pages: mistery solved?*. http://aixtal.blogspot.com/2005/02/web-googles-missing-pages-mystery.html.
Volk, M. 2002. «Using the Web as Corpus for Linguistic Research». In R. Pajusalu and T. Hennoste (eds.). *Tahendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Oim*, Publications of the Department of General Linguistics 3. University of Tartu. http: //www.ifi.uzh.ch/cl/volk/papers/Oim_Festschrift_2002.pdf
Wehmeier, N. W. 2003. *Using web search for machine translation.* Unpublished BSc Computing and German. University of Leeds http://www.comp.leeds.ac.uk/fyproj/previous-titles/bsc2003.html.
Werthner, H. and Klein, S. 1999. *Information Technology and Tourism: A Challenging Relationship*, Springer, New York.
Wynne, M. 2002. *The Language Resource Archive of the 21th century*, Oxford Text Archive. http://gandalf.aksis.uib.no/lrec2002/pdf/271.pdf.
Wynne, M. (ed.) 2005. *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. http://ahds.ac.uk/linguistic-corpora/.
Yates, S.J. and Sumner, T.R. 1997. «Digital Genres and the New Bur-

den of Fixity». *Proceedings of the Hawaiian International Conference on System Sciences (HICCS 30)*, VI, 3-12. http://csdl2.computer.org/comp/proceedings/hicss/1997/7734/06/7734060003.pdf.
Zanettin, F. 2002. «DIY Corpora: the WWW and the Translator». In Maia B. *et al.* (eds), *Training the Language Services Provider for the New Millennium*, Porto, 239-248. http: // www.federicozanettin.net/DIYcorpora.htm.
Zanettin, F., Bernardini, S. and Stewart, D. (eds). 2003. *Corpora in Translator Education*. Manchester: St. Jerome.
Zipf, G. K. 1935. *The psychobiology of language*. New York: Houghton Mifflin.
Zuraw, K. 2006, «Using the Web as a Phonological Corpus: a case study from Tagalog», in Baroni, M. and Kilgarriff, A. (eds.), 59-66

WEBSITES

BootCaT, sslmit.unibo.it/~baroni/bootcat.html.
Elsevier, www.elsevier.com.
Global Reach, http://global-reach.biz/globstats/.
Internet World Stats, «Internet Top Ten Languages», http: //www.internetworldstats.com/stats7.htm (June 2007).
Multilingual Glossary of technical and popular medical term, users.ugent.be/~rvdstich/eugloss/welcome.html.
Our Search: Google Technology, http://www.google.com/technology/
Pubmed, www.ncbi.nlm.nih.gov/pubmed/.
Sketch Engine, www. sketchengine.co.uk.