Web Access to Corpora: the W3Corpora Project*

Doug Arnold
Department of Language and Linguistics,
University of Essex,
Wivenhoe Park,
Colchester, Essex,
CO4 3SQ, U.K.

email: doug@essex.ac.uk

1 Introduction

In this day an age, some corpus linguistics should be part of every course to do with language. But learning about corpus linguistics — its possibilities *and* limitations — is not just a matter of acquiring information. The best way to learn about corpus linguistics is to do it, and the best way to teach corpus linguistics is to put students into a position where they can do it ((Leech, 1997), (Fligelstone, 1993)). This requires corpora, and tools, in addition to teaching materials.

For a number of reasons, the World Wide Web offers a good method for delivering this (see below). This paper will present a resource that enables students to get a general introduction to corpus linguistics via the Web. The resource is currently available for general use. See Table 1 for URLs.

No very great claims will be made for the resource in terms of being highly original or visionary in style of interaction or implementation. On the contrary, the model of learning is rather traditional, and the approach taken was very simple and straightforward. However, this in itself may be interesting as providing a baseline against which more visionary approaches can be compared — this is probably the simplest way one could go about providing Internet based education. In addition, some of the design decisions and lesson learned may be of interest.

Section 2 presents the motivation for the project that produced the resource. Section 3 will give an overview of the resource. Section 4 describes and compares some similar resources that are available. Section 5 describes some problems and lessons that can be learned, and notes some open questions.

2 Motivation, Design Criteria

The motivation for the project was the observation that the up–take of corpus linguistics is not what it should be — in this day and age *some* corpus linguistics should be part of every course to do with language. The problem is that learning about corpus linguistics involves doing it, and that the overheads involved in getting started in doing or teaching corpus linguistics are considerable. Corpora in many languages are now easily available, but to use them requires a significant investment in hardware (e.g. disk space), software (tools need to be downloaded and installed), and time and effort (the tools have to be understood and techniques mastered). All this is hard enough for the individual researcher. In a teaching context, all these problems will typically be magnified by the need to deal with differences in the environment available to students (architecture, operating system, software: if something can differ, it will differ; if a difference can matter, it will matter).

Corpus linguistics should be a part of every scheme of study, and it may have a role in almost every piece of research. But it need not be a central theme, certainly not central enough to justify the effort involved. It would be nice to be able have (say) three sessions on corpus linguistics in a course with a wider focus, and in that time give students a real feeling for what can be gained, and what are the limitations. It would be nice for a researcher to be able to find out whether corpora can provide any useful data about some phenomenon without having to actually become a corpus linguist.

There are surely many areas of linguistics, even

*The project was the joint work of Ylva Berglund, Natalia Brines–Moya, Martin Rondell and the author in the period 1996–8. The results can be seen at: http://clwww.essex.ac.uk/w3c/. The project was funded by JISC (the Joint Information Systems Committee of the UK Higher Education Funding Councils), as part of JTAP, the JISC Technonology Application Programme. Thanks also to the anonymous workshop referees for valuable comments. None of this shifts responsibility for errors and other imperfections from me.

computational linguistics, that are like this: as subjects develop, it becomes impossible for students or researchers to master the full range of ideas and techniques, and there is an increasing need for the provision of knowledge of subject areas at a 'contextual' rather than specialist level. It is important to be able convey something about a wide range of areas very briefly, but (hopefully) without trivialization.

So, the goal of the project was to provide instant, and instantly usable, access to corpora, including tools to manipulate them, as well as general information and tutorial information about how and why one might manipulate them.

Of course, the World Wide Web is excellent for this. In principle, all the user needs is a Web connection and a browser. Beyond this, no investment of money, and little investment of effort should be needed: there should be no need to obtain and install corpora, or download and install software, and the interface to the corpus manipulating tools should already be familiar (since it would be based closely on their web browser). Moreover, from a teaching perspective, problems of different architecture (etc) are minimized — all that is necessary is the browser and the Web connection.

Given these aims and motivation, a number of design decisions are rather natural:

- The system should be immediately usable by anyone with WWW access and a Web Browser, for example:
  - it should be usable without the need to install or download any programs;
  - it should be usable with essentially any generally available browser;
  - it should be usable without the need to register and get authorization.
- The interface should be as 'friendly' and easy to use as possible; it should be supported by extensive on–line help, and backed up by detailed information about corpus linguistics in general, and how to 'do' corpus linguistics in a practical way, using a tool such as this.
- It is typical of novice users that they make mistakes with queries; thus, there should be some method for users to correct and 'refine' their queries very easily (this lead to the idea of an editable 'search history').
- It should be possible for a user to search their own Corpora — in this way a user can explore not only what is possible in general, but what is possible in relation to the kinds of material they are interested in or have to deal with.
- A major problem with Web delivery is the network overhead. Thus the source code should be freely available (in GNU 'Copyleft' style), which should allow the system to be installed and run locally over the Web at other sites.

3 Implementation, Overview of the W3Corpora Web–Site

This section will give an overview of the W3Corpora web–site. See Table 1 for URLs.

The site is divided into three main parts:

General Information where the user can learn about corpus linguistics in general (e.g. general discussions: 'What is a corpus?' issues of corpus design and annotation, research areas, bibliography, etc). This is the kind and level of information one might expect in an introductory text book, e.g. (Barnbrook, 1996) or (Kennedy, 1998).

Tutorial where the user can find out how to use the tools provided, and where some areas are described where corpus techniques are useful. A variety of tasks are described in some detail with practical examples (e.g. how to investigate the meaning of word, compare two similar words, how a word is used in different contexts, investigating spelling, and choice of preposition in a context like *an explanation of/for something*). Here an elsewhere, the emphasis is on classical corpus linguistics, neglecting e.g. statistical techniques that can be built on top.

Here the key aim is to answer, as quickly and easily as possible the two questions: 'How can I use this thing?' and 'What can I use it for?' It does not pretend to be a complete, stand-alone tutorial in Corpus Linguistics; it does not go to the length of (say) (Aston and Burnard, 1997), nor does it go into the same level of detail. The primary aim is to take the user to the point where they can answer the question 'Is Corpus Linguistics useful in my study and research?', and in case of an affirmative answer, give a basis for proceeding (perhaps, in fact most likely, with other resources and tools, installed locally to avoid network overheads).

Search Engine where the user can carry out corpus searches.

'Top level':	http://clwww.essex.ac.uk/w3c/
General Information:	http://clwww.essex.ac.uk/w3c/corpus_ling/content/introduction.html
Tutorial:	http://clwww.essex.ac.uk/w3c/help/intro/start_page.html
Search Engine:	http://clwww.essex.ac.uk/w3c/corpus_ling/content/search_engine.html

Table 1: Web Addresses for W3Corpora Resources

Apart from the Search Engine, the implementation is rather straightforward: text marked up as html, there is extensive use of frames so that users are able to maintain an overview of documents as well pursuing detail.

The implementation of the Search Engine merits more discussion, but it is also based on rather standard techniques, using cgi–bin scripts written in Perl, and fairly standard indexing techniques to speed up searching.

When a user arrives at the top–level search page, she is invited to select a corpus and from a menu, and to specify a search string and search type (e.g. regular expression, exact match, whole words, etc). Confirming these selection generates a 'session file' which records the selections. Also generated is a file recording various default values for options dictating *inter alia* what sort of results should be displayed first (frequency, or Key Word In Context — KWIC), for KWIC results, how many results should be displayed at one time, how much context should be displayed, etc. etc. The user can modify these options interactively via a form, which is generated in response to clicking the 'Options' button at the top of the screen. Currently, some 19,000,000 words (321) texts from the Gutenberg Project corpora can be searched.[1]

A flavour of the interface to the Search Engine can be gained from Figure 1, which shows the results of searching for the regular expression `/[Nn]ice/` over a subset of the Gutenberg texts, and clicking on one of the results to view the wider context. An early stage in the project defined a list of properties that a corpus searching interface should have (Brines-Moya and Hartill, 1998). This interface satisfies almost all.

A large amount of on–line help is available (via the 'Help') button (the information supplied is somewhat sensitive to the particular screen being viewed).

'Frequency' and 'Display' buttons generate different views of the search results:

- The 'Frequency' button generates frequency information for a search (total number of hits, hits per–subcorpus, and lexical information — e.g. how many of the hits for `/[Nn]ice/` arise from the the word *nice*, how many from *nicer*, *nicest*, *Venice*, *cornice*, etc).
- The 'Display' button generates a KWIC display of search results (see Figure 1). KWIC results are editable — the user can delete certain results, and can also call up wider context by clicking on a key word.

The 'Search' button allows the user to (i) carry out a totally new search, (ii) 'refine' the existing search, or (iii) view, and modify the search history.

Refining a search returns a subset of the current search: the user supplies a regular expression which potential hits must satisfy in addition to the original pattern. Thus, one might refine `/[Nn]ice/` to `/^[Nn]/` to eliminate *Venice* and *cornice* as hits, a further refinement to `/e$/` would eliminate *nicer* and *nicest*. A sequence of refinements constitutes a *search history*: users can view, and edit a search history — moving backwards and forwards through the different stages of a search. The user can also delete stages (e.g. to leave just an initial and a final stage).

An aspect of the system that may be particularly useful to teachers is the ability to up–load corpora for searching. When a user up–loads a (plain text) corpus to the Web–site (by anonymous ftp), it becomes selectable for searching. When so selected, it is indexed and prepared for searching in the normal way. This may be particularly useful to teachers who want students to carry out exercises on particular corpora that are not already provided at the site.

The site has been used and positively evaluated by 'expert users' (i.e. with a background in corpus linguistics), and by students at Essex and elsewhere, but there are many open questions about how it can or should best be used in the context of different courses and learning situations (see Section 5).[2]

[1] 1 Million words of the LOB corpus, tagged and untagged can also be searched, after a user has registered and received a password.

[2] The webpage of the course at Essex which used the resources is http://privatewww.essex.ac.uk/~scholp/lg478cs.htm. It is taught by my colleague Dr. Phil Scholfield, whom I hereby thank for his advice and feedback. On this course, Corpus Linguistics is covered in

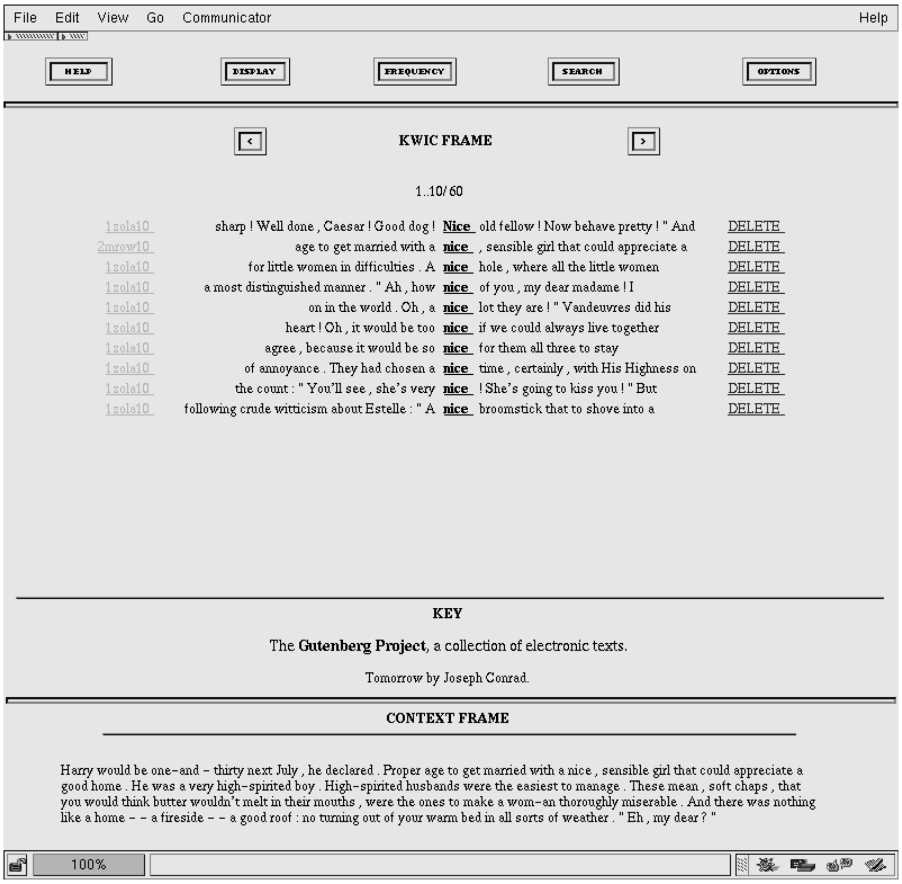

Figure 1: KWIC Display of search Results: the user has selected options which allow results to be deleted and which indicate which sub–corpus each hit comes from. At the bottom of the page the wider context of one of the hits is displayed (the user has clicked on one of the individual hits to obtain this).

4 Existing Work, Comparisons

There are a large number of tools and systems that offer something similar to what the W3Corpora site seeks to provide. They range from simple Unix command–line style programs like Ptx, to sophisticated GUI interfaces. For local installation, on a Macintosh one has **Conc 1.7**, and **ParaConc**; for DOS/Windows, one has **ICEUP** (from ICE), **LEXA** (from ICAME), **Micro–OCP**, **Multiconcord**, **LDB** (from Nijmegen), **Wordsmith Tools**, **TACT**, and **Sara** (for the BNC); for Unix, there many standard utilities, as well as **ptx**, and **Xkwic** (from Stuttgart).

As regards Web–accessible resources, the following should me mentioned:

BNC The BNC site provides access to a subset of the British National Corpus on a trial basis. This permits simple searches on–line, but with limited number of hits, and limited information about the hits. Registration for a trial account (20 days) is required. Full access requires downloading (Windows) client program (available for Windows95, and Windows3.x only), and payment of an annual registration fee. It is restricted to users within the EC.

Canadian Hansard This site permits access to the proceedings of the Canadian Parliament in English and French. These are parallel corpora (English and French), searches may be mono– or bi–lingual (in either case, the results returned are bi–lingual — i.e. the user sees both the context where the search term appears, and translation). In the mono–lingual case one can see how an expression is used and translated. The bi–lingual case allows one to see, e.g. where English *commitment* is translated as French *attachement*.

four 2-hour sessions, two of which are descriptive, two practical; in the latter two the students use the W3Corpora search engine, under supervision. A practical corpus investigation, using tools such as the W3Corpora search engine, is one of the options for course assessment.

In addition to verbatim (case independent) searches, it is also possible to perform a dictionary search, e.g. the query: `pull+ the plug` will match *pull the plug*, *pulling the plug*, *pulls the plug*, etc, and to search for words that do not appear contiguously (e.g. *make ... arrangements*). No frequency information is provided.

Cobuild This site gives limited access to the Cobuild Corpora: the "Bank of English" (over 50million words), giving an idea of the kinds of search possible with the full system. It is possible to search for regular expressions (including a special character which matches inflectional endings), combinations of words, and part of speech tags. Only 40 lines of concordance are returned, and no information about frequency, or wider context is accessible. It is also possible to search for collocates of words, based on either of two statistical scores (mutual information and T-score). The site does not provide much in the way of help pages, and there is no tutorial.

TACTWeb a pilot version of the TACTWeb software can be used on the Bergen Corpus of London Teenager Language (TACTWeb is intended to make a TACT style text database accessible over the WWW). This is close in intention to the present project. At the time of writing, it is still under development.

LDC/Brown Corpus Text Corpora, and Speech Corpora, are accessible via the Linguistic Data Consortium. After registration, it is possible to access the Brown Corpus. For individuals who are not (affiliated to) members of the LDC it is possible to register as a guest, and access corpora with the password that is sent to the user by email. Frequency information is available, and a wide variety of searches is supported, concordances can be generated, and collocational information retrieved. Access to the TIMIT Speech Corpus is similar.

It is obvious that some of these sites provide functionality that is not available at the W3Corpora site — notably (i) multi–lingual searching and searching over parallel corpora, (ii) collocational information, and (iii) 'dictionary style searching' — and several provide access to far more extensive corpus resources.

On the other hand none of these sites duplicates what is available at the W3Corpora site. In particular, none of them provides the balance of easy (immediate) access to usable quantities of corpus material, with easy, customizable functionality, and extensive user support and tutorial facilities. So far as I know, in no case is the source code freely available. Where they do provide semi–introductory access (e.g. by means of free registration and/or a guest account), there is generally very little in the way of of tutorial material.[3]

5 Conclusion: some Problems, Lessons and Open Questions

By far the most serious problem that the project faced was the difficulty of getting corpus resources that could be made freely available (i.e. without registration) over the Web.

The whole system took about two years (three person years) to complete. This is a considerable effort, and one that is only worthwhile for a relatively stable area like corpus linguistics, where there one can reasonably expect several years of use for a resource.

The finished system is very large: the search engine and interface involves over 12,000 lines of code, much of it very straightforward (Perl commands to generate the html forms that provide the interface). It is hard to resist the sense that there should be an easier way to do this.

Using html forms brings some problems. In particular, the lack of any kind of 'interactive' forms means that the interface is more complicated than it might otherwise need be (a form must be completed in *toto* and then submitted — it cannot be partially completed and updated on the fly).

The Perl–cgi-bin combination is powerful and excellent for small applications, but there is a severe lack of good debugging tools.

It had originally been hoped to make the resource both 'future proof' and 'past proof'. The former is not too problematic — the technology involved is likely to be supported for many years to come. But the latter — the intention to make the resource usable with essentially any kind of browser — quickly proved impossible, because of the need to use frames in serving the search engine interface.

The resource is now fully operational and available. While it has been evaluated by a number of different kinds of users in a number of contexts, there are still many open questions about how it can or should best be used.

In designing the resources, we had in mind a casual, novice user, either an individual student or researcher with an interest in, but no strong

[3]See (Arnold et al., 1999) for a fuller discussion of alternatives.

commitment to, Corpus Linguistics, or a student on a course where Corpus Linguistics has a minor place (in the order of, say, three two hour sessions). See (Arnold and Berglund, 1998) for a little more discussion of this. (We took the view that committed users would invest the effort in installing corpora and corpus searching tools locally, and would find the overheads of WWW access unacceptable). Similarly, the resource was intended to be 'stand-alone' — this was intended to make it as generally usable as possible. This means it does not form part of a larger suite of materials, and there are open questions about how it should best be integrated into schemes of study, and about what sorts of teaching method are appropriate. At one extreme, a teacher may simply note the resources as one among many resources available for further investigation, at another, one could imagine entire classes trying to access the resources at the same time, with similar queries, under the direct supervision of a teacher. Apart from obvious remarks about the machine and network loading implications of the latter, I have nothing to offer here. But these are important issues, and since this range of possibilities exist in principle for any Web-Based resource, quite general.

The resources and tools were designed for WWW based access. But many of the advantages (and a few other benefits) can be gained by a local area (LAN) installation. The cost is that the tools and corpora must be installed and maintained locally, the advantage is that one eliminates the WWW network overhead, and no longer has to rely on a remote site to provide the resource.[4] Again, this is a general question for WWW based teaching, but one on which it is hard to say anything general. From a users point of view, the key questions are obviously the reliability of the remote site, compared to the reliability of local systems, and the inconvenience of the network overhead. These are matters which will vary greatly from one place to another, and will depend on the resources being provided — in the case of the W3Corpora, there is still insufficient experience in practice to do much more than raise the questions.

References

Doug Arnold and Ylva Berglund. 1998. WWW access to corpora: a tool for teaching and learning about corpora. In *TALC-98 (Third International Conference on Teaching and Language Corpora)*, Keeble College, Oxford, 24–27 July. Humanities Computing Unit, Oxford University, Oxford. `http://clwww.essex.ac.uk/w3c/corpus_ling/TALC.html`.

Doug Arnold, Bas Aarts, Justin Buckley, Ylva Berglund, Gerald Nelson, and Martin Rondell. 1999. Corpora and grammars on the web: the W3Corpora–IGE Project, final report JTAP–2/247. `http://clwww.essex.ac.uk/w3c-ige/FinalReport/`, February.

D.J. Arnold. 1997. WWW–IGE: World Wide Web access to Corpora and the Internet Grammar of English. In *Proceedings of DRH-97 (Digital Resources in the Humanities)*, pages 711–716, St. Anne's College Oxford, Sept. (abstract: `http://users.ox.ac.uk/~talc98/arnold.htm`).

Guy Aston and Lou Burnard. 1997. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh.

Geoff Barnbrook. 1996. *Language and Computers: a practical introduction to the computer analysis of language*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh University Press, Edinburgh.

Natalia Brines-Moya and Julie Hartill. 1998. Criteria for user-oriented evaluation of monolingual text corpora interfaces. In *Proceedings of the First International Conference on Language Resources and Evaluation*, volume 2, pages 893–898, Granada, Spain, 28-30 May.

Steve Fligelstone. 1993. Some reflections on the question of teaching, from a corpus linguistic perspective. *ICAME Journal*, 17:97–109.

Graeme Kennedy. 1998. *An Introduction to Corpus Linguistics*. Studies in Language and Linguistics. Addison Wesley Longman Ltd, London.

Geoffrey Leech. 1997. Teaching and language corpora: a convergence. In Ann Wichmann, Steven Fligelstone, Tony McEnery, and Gerry Knowles, editors, *Teaching and Language Corpora*, pages 1–23. Addison Wesley Longman, Harlow.

[4] Of course, once one has decided to go for a local installation, there are many alternatives to the W3Corpora resources, and one is not tied to a Web browser style interface.