Lisa Biagini & Eugenio Picchi, Istituto di Linguistica Computazionale, CNR, Pisa

INTERNET and DBT

Abstract

The advent of Internet has had enormous impact on working patterns and development in many scientific sectors including Computational Linguistics. For this reason, DBT, a full-text retrieval system designed for literary and linguistic text analysis applications, will now include a set of procedures so that it can be used on the Internet circuit. Two distinct approaches are being implemented. In the first, the system adopts World Wide Web technology and the HTML document standard. The second solution is a client-server application that exploits Internet as a support: the user can access texts located on servers at geographically remote sites in the same way as when using the standard DBT system.

1. Introduction

The past two decades have seen an incredibly rapid evolution in computational technology. In particular, hardware developments regarding storage devices have increased capacity and reduced the costs of on-line data storage ten-fold in just the last few years. At the same time, there has been a proliferation of software tools (many of which are now public domain), designed for the easy management and analysis of such data. One of the results has been that lexical studies and applications are now under way which were inconceivable only a few years ago.

This near revolution has particularly affected the practice of lexicography. Lexicographic projects today frequently involve the use of the computer and the management and analysis of huge volumes of language data as a necessary prerequisite to the compilation or revision of their dictionaries. This has led to the construction of large reference corpora and textual data banks, and implies a heavy investment in terms of the resources necessary to acquire, maintain and update them and the provision of facilities for their consultation. Until a short time ago, this consultation was only possible on-site. Recently, many of the larger corpus projects have provided their users with facilities for remote consultation, using telnet utilities.

However, it is the most recent of the rapid and continuous developments to which the "Information Society" is being subjected that is

perhaps destined to have the most far-reaching effects on future studies on the lexicon. With the explosion of Internet, access to the global information highway is virtually possible for anyone; any user – with the most basic equipment (a simple PC and modem) – can access and often download enormous volumes of digitized information (text, images, etc.).

In the paper we describe DBT-NET, the latest development to DBT, the full-text retrieval system designed and developed at the *Istituto di Linguistica Computazionale*, Pisa (Picchi 1993). DBT-NET has been designed in recognition of the fact that with the advent of Internet, the behaviour and the needs of the users are changing. They are no longer ready to accept passively available resources but are stimulated to access and navigate through new sites which can provide them with new data, new knowledge. Today's users do not just want to query and analyse their own data but they want to be able to access the enormous data banks distributed world-wide. And very frequently, they want to be able to do this from the comfort of their own home, using their own simple equipment. Thus, there is a demand for procedures that allow them to access data made available by others. For linguists, of course, it is not sufficient just to be able to access this data, they must be able to analyse it with appropriate linguistic tools.

DBT-NET thus permits linguists or other kind of language scholars to perform text query and analysis operations via Internet. There are two possible operational modes. In the first and simplest, the system adopts the World Wide Web technology with the HTML document standard; this standard is well-known throughout the world and has the strong advantage of being very easy for a very wide range of users who are already acquainted with it through Mosaic and/or Netscape. The second solution is the creation of a client-server application that uses the Internet communications protocol as a support. In this case, the DBT users (clients) are able to directly access and query texts located on servers at research Institutes or other sites in exactly the same way as if they were querying their own data stored locally on their own machine. These two different approches are described in the paper.

2. The DBT System

The textual database management and query system known as the DBT has been under continuous design and development for more than ten years now. It is designed specifically for linguistic and literary text processing and analysis tasks, with particular attention being paid to the

needs of lexicography. It is now a complex system consisting of many separate modules, some very general purpose, others designed to handle the particular requirements of specific applications. Some of these modules are now fully tested and industrialised, whereas others are still in the development and testing stage. The main features of the core system are: total respect for the integrity of the source text; management of different character (Latin and non-Latin alphabets) and code sets; real time, interactive execution of all the typical functions of a text retrieval system; high performance in terms of flexibility and speed; optimisation of storage and memory requirements; management of very large text corpora; management and analysis of images in a text or associated with it (manuscripts, icons, etc.); management of structured text, in particular dictionaries.

Other components have been designed to run on top of the core system for the management of annotated texts and data for the POS tagging of Italian texts, and for the creation, management and interrogation of bilingual parallel and comparable text archives.[1]

When using the system, the first step is to format the texts using the DBT procedures. DBT has its own encoding system and this stage is simple rapid and, once a few preliminary instructions have been given, fully automatic. The system also has an interface which permits it to acquire textual material already encoded in SGML-TEI and in HTML.

Once the texts are in DBT format, the users can use the system query procedures which provide them with a series of search functions which can be used to access a single text or a text corpus and retrieve various elements or combinations of elements. They can display all or part of the text(s) on which they are operating, search given word forms, search words containing one or a combination of given character strings or using wild cards, compute frequencies, define search functions in which words are associated in different ways and retrieve all the contexts satisfying these search conditions in the text(s), generate ordered text concordances, impose particular conditions on concordance generation, select the language of interest when several languages or language types are present in the text(s) and have been classified as such, obtain statistical data on the cooccurrences of selected words (e.g. Mutual Information Index), execute queries on annotated and lemmatised material, etc.

Another important feature of the system is that it provides procedures to manage images included in a text. An image database is associated with the text; the user can access this database dynamically and view the images. The way in which the results of the queries are displayed on the screen or printed out can be defined by the user, according to his own

particular needs and preferences. The system is extremely versatile and easy to use.

3. DBT and Internet

Over the last ten years, the standard DBT system has been widely adopted throughout the Italian academic and research world (and not only in Italy) in the course of linguistic and lexicographic studies and projects. In fact, the system procedures to put a machine readable version of a text in DBT format are so simple that a new text can be easily prepared and interfaced to capture texts from SGML and HTML documents are now available. This means that by now there are thousands of texts already structured in DBT form and this number is growing rapidly.

As is known, the importance of rendering language and text resources reusable is strongly felt in the scientific community. There is thus a strong move towards making existing resources available as widely as possible, depending on considerations of copyright and intellectual property rights. For this reason we began to study the best way of making it possible not only for local users but for scholars working anywhere in the world to consult these geographically distributed DBT archives. It became very clear that the ideal medium for this is that provided by Internet.

Our idea was to offer to the user the possibility to make the DBT system available on Internet at two levels: the first employ an easy to use approach provided by the Netscape and Mosaic standards; whereas the second offers the full functionality of the DBT environment, but requires a minimum knowledge of the system.

3.1 DBTWEB

The first approach uses the best known standards for information diffusion running on the Internet infrastructure: the addressing scheme – URL, the communication protocol – HTTP, and the description language – HTML. These standards are adopted by both Netscape and Mosaic – the systems developed for hypertext navigation on the World Wide Web (WWW). The main advantage of adopting this technology is that it facilitates navigation over the network. Information is made available in the form of pages of multimedia and hypertext data. The hypertext links point to other pages which can be located anywhere in the network.

These standards are independent and do not depend on the platform, or computing system employed by the client; this means that they are directly usable by all platforms that can communicate with Internet. This has contributed greatly to their popularity.

A version of DBT which adopts the Hypertext Transmission protocol (HTTP) and the Hypertext Markup Language (HTML) is now in an advanced stage of development. DBTWEB creates a hypertextual study environment, dynamically transforming the results of a generic query into an HTML page, which can in its turn be consulted by other queries. It offers all the main functionalities of the standard DBT system. However, despite its positive features, the adoption of the page-after-page philosophy employed by standard WWW technology tends to impose limits on the DBT system which has been developed with a very high level of user-system interaction and very fast system response times. Hopefully, this problem will be overcome to some extent in a future version of DBTWEB which runs on Netscape 2.0 which adopts a multiwindows-like philosophy.

3.2 DBT-NET

For this reason, we also decided to develop a DBT client-server application operational on Internet which would maintain all the functions of the standard system. In this version, known as DBT-NET, the DBT procedures have been divided between two distinct components: the server which hosts the text archives and indices; the client which allows the user to access the text archives located at the server. Our main objective has been to offer on the one hand a user interface and a set of functions identical to the stand-alone system, while on the other to distribute the tasks, processing functions and user-system interaction, maintaining as much information as possible at the client site, so that the client-server operations are reduced to the minimum in order to optimize the system response times.

DBT-NET thus permits the user to access, consult and query richer data sources. Consisting of client and server programs, and adopting the TCP/IP protocol, it enables communication between Institutes with DBT text archives (servers) and those wishing to consult these archives (clients). The client side of the system has been developed in the first place in a WINDOWS version which runs on an ordinary PC in order to meet the needs of many scholars who prefer to work with small machines, very often from home. We are now developing a UNIX version of the client. The server (normally a research institute) has been

developed from the start in both WINDOWS and UNIX versions. The first tests have evidenced good system response times and more than satisfactory times for the transfer and loading of images (often a weak point in Internet applications). In addition, the system also provides a library of programs specifically developed for image processing; this is particularly useful for scholars analysing original texts, manuscripts, etc.

Once the data have been coded and indexed using the normal DBT procedures and made available on a server, DBT-NET is very easy to use. It is necessary to have a PC linked to Internet and the client program; this can be down-loaded directly via Internet provided the necessary authorisation has been granted. Once the program has been launched, the user is in the same environment as that of the stand-alone version of DBT and can use all the functions of the base system as described above.

All this can be done from the user's own PC and no longer on a limited quantity of data, available to a single user or single Institute, but on large data banks, located in any part of the globe. DBT-NET thus provides a user friendly system for remote text consultation, with highly powerful and flexible text analysis.

4. Conclusions

By offering two distinct ways for the DBT user to use the system on Internet, we provide solutions both for those users who want to make simple searches on already available textual material, without being obliged to learn a new query language, whereas those users who are already accustomed to using the DBT system, or those who want to make more complex searches, can use the client-server system.

For information on the DBT system and the DBT Internet servers see the WEB pages at "http://www.ilc.pi.cnr.it/dbt/pisystem.htm" or send an e-mail to "dbt@tnos.ilc.pi.cnr.it".

Note

1. For full details of the core DBT system and the different functionalities supported, see Picchi 1991; for descriptions of the version of the DBT which manages mono- and bilingual lexical data, see Marinai et al 1990; for the POS Tagger, see Picchi, 1994; for information on the DBT procedures which have been designed for bilingual text corpus construction and interrogation, see Marinai et al, 1992.

References

Marinai, E., Peters, C., Picchi, E. 1990. *The Pisa Multilingual Lexical Database System*, Esprit BRA 3030, Twelve Month Deliverable, ILC-ACQ-2-90, Pisa, 61p.

Marinai, E., Peters, C., Picchi, E. 1992. "A Project for Bilingual Reference Corpora", in: *Acta Linguistica Hungarica*, Akadémiai Kiadó, Budapest, Vol. 41 (1–2), pp. 1–15.

Picchi, E. 1983. "Textual Database", in: *Proceedings of the International Conference on Data Bases in the Humanities and Social Sciences*, Rutgers University Library, New Brunswick, New Jersey.

Picchi, E. 1991. "D.B.T.: A Textual Data Base System", in: L. Cignoni and C. Peters (eds.), *Computational Lexicology and Lexicography. Special Issue dedicated to Bernard Quemada. II, Linguistica Computazionale,* 7, pp. 177–205.

Picchi, E. 1994. "Statistical Tools for Corpus Analysis: A Tagger and Lemmatizer for Italian", in: *EURALEX '94 Proceedings*, Amsterdam, pp. 501–510.

Picchi, E., Peters, C. and Calzolari, N. 1990. "Implementing a Bilingual Lexical Database System", in: T. Magay and J. Zigány (eds.), *BUDALEX '88 Proceedings*, Budapest, pp. 317–329.