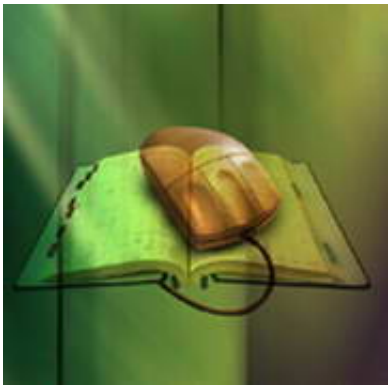
Modern tools to unlock Ancient Texts

December 1 2005



Tools for ancient texts have been successfully created that will open up rare texts and manuscripts locked away in museums, libraries and archives, and promote new kinds of scholarship while also preserving large swathes of European history and culture for the future.

With funding from the IST programme and the US, the CHLT project developed morphological analysers, citation databases, visualisation and clustering tools, and combined them with dictionaries to aid experienced scholars, students and the general public alike.

It also unified several important digital library collections – such as Isaac Newton's manuscripts in the Newton project - and early modern scientific texts, as well as creating new digital library collections of Old Norse sagas. It's a vast achievement.

In addition to providing access to primary source materials that are often rare and fragile, the project developed a series of tools and applications that will make it possible for people with no knowledge of the language to translate ancient manuscripts, albeit painstakingly.

CHLT used leading edge techniques from computational linguistics, natural language processing, and information retrieval that enables researchers to conduct new types of scholarship.

"It was a remarkably successful project between the National Science Foundation in the US and EU institutions. It generated results beyond expectations, and illustrated how essential it is to work together to create an integrated global infrastructure for scholarly research," says CHLT's European coordinator Dolores Iorizzo from the Newton Project and the London e-Science Centre.

Successful collaboration responding to users' needs

"This collaboration meant that we were able to integrate new technologies into our results that could not have been foreseen at the beginning of the project."

The team wanted to find the most effective ways to use technology to interpret digitised, historic manuscripts. CHLT responds to the challenges faced by teachers, students and scholars who are working with texts written in Ancient Greek, Mediaeval and Early-Modern Latin, and Old Norse.

The number of primary texts – arguably the most important resource for historians and linguists – is staggering. Hundreds of important texts and manuscripts, consisting of millions of words have been integrated into the CHLT open access repository that can also be viewed within the oldest and largest cultural heritage database in the world at the Perseus

Project in Tufts University, Boston.

CHLT created new text collections written in Early-Modern Latin and Old Norse. It integrated those new books and manuscripts with well-established digital texts, and it created a digital library environment that allows for high-resolution images of pages from rare and fragile printed books and manuscripts. These are presented alongside transcriptions so that the originals can be viewed alongside diplomatic and normalised versions of the material.

"The early modern printed texts can be scanned to create automatically generated hypertext, but manuscripts such as those in the Newton Project must be transcribed and XML text encoded by hand which makes it very slow and painstaking," says Iorizzo.

Powerful tools to analyse text

The project successfully developed a host of powerful language analysis tools that will help readers to understand texts written in these difficult languages by offering parsers, which automatically determine the grammatical identity of a word.

This is important because these ancient languages are highly inflected. The meaning of a word does not depend on its position in the sentence, but its grammatical case, which indicates which words are the subject or object of the sentence. Parsers analyse the underlying grammatical context to tease out the meaning.

What's more, these parsers were integrated into a digital library reading environment that automatically generates hypertext links. So a user can click on a word, register its identity and look it up in a dictionary. CHLT also built a multilingual information retrieval tool that allows users to enter queries in English and search texts written in Greek and Latin.

Experienced scholars can use the parser to check an unfamiliar word, or a word used in an unusual context. Students and scholars without Greek, Latin or Old Norse can painstakingly translate ancient texts word-by-word. The tool will provide an enormous boost to the study of these ancient languages and culture, while scholars from other fields will have access to texts even if they don't speak the language.

"CHLT is a political statement. We've lowered the barrier for access to primary texts, so now it's no longer the academic elite who have access and can read these historically important manuscripts," says Iorizzo. Users can even upload their own texts for parsing and analysis. Those texts will then be added to the library so the collection will grow organically over time.

Opening up new kinds of research

CHLT also allows scholars to engage in new kinds of research. For example, a visualisation tool clusters search results into conceptual categories, while a word-profile tool integrates statistical data about how often a particular word is used in a set of collections. It can also fetch information from different reference works, like the Liddell, Scott, and Jones' Greek-English Lexicon.

The word profile tool uses a single interface to link words to full citations of the texts in which they appear. Right now, scholars are using this to write the first new Greek-English lexicon to be created in more than one hundred years.

"We quickly realised that there must be a paradigm shift in the way lexicons are created since we now have the capacity to tailor lexicographical tools to particular texts by offering automatic analyses of words that are genre and period sensitive. This is a real breakthrough."

CHLT also created tools that allow for the computational study of writing style. This includes tools to discover common subjects and objects of Greek and Latin verbs, the relative frequency of different grammatical forms, and the distribution of grammatical forms in texts.

It has already produced a new scholarly understanding of ancient Greek culture. For example, US coordinator of CHLT, Jeffrey Rydberg-Cox, Associate Professor of the Department of English in University of Missouri, discovered that narrative descriptions of violence in Lysias are marked by high density use of the participle.

The project has revolutionised historical research by introducing new digital library architectures and protocols for resource discovery and metadata sharing in affiliated digital libraries. It represents a major step towards unifying Europe's diverse digital collections.

CHLT supports Open Access and Berlin Declaration policies, and has negotiated a free open-access agreement with Cambridge University Press for an electronic edition of the Greek-English lexicon to be published online simultaneously with the print edition; it has also explored ways that these tools can be used and shared across cooperating digital libraries.

This is another big step toward creating a global infrastructure for Cultural Heritage. The CHLT consortium now hopes to develop these technologies in a Grid-distributed network capable of linking all of Europe's 100,000-plus 'memory institutions' – libraries, archives and museums, and large-scale digital repositories.

"At present, Europe's memory is preserved in compartmentalised silos of information within separate databases and websites," says Iorizzo. "What we would like to do is to provide an infrastructure that integrates, at a metadata and data level, the rich resources of European Cultural

Heritage so that everything can be accessed, searched and preserved by anyone for generations to come."

Source: IST Results