The MELIN project

Donncha Ó Cróinín
Institiúid Teangeolaíochta Éireann/Linguistics Institute of Ireland
31 Fitzwilliam Place, Dublin 2, Ireland.
[doc@ite.ie]

Abstract

The MELIN (Minority European Languages Information Network) project, partly-funded by the EU's MLIS programme, will develop a WWW site for minority language dictionaries, terminology lists and other resources. This project, which commenced in January 1998, will have an initial duration of 15 months. Participating languages are Irish, Welsh, Catalan and Basque (with the aim of increasing language coverage in the near future).

Introduction

The MELIN (Minority European Languages Information Network) project is currently developing a centralised World Wide Web presence for the provision of language resources to users of minority and lesser-used languages in the EU. Existing language resources - initially, terminology lists and monolingual/bilingual dictionaries (owned by the partners or used by permission) will be converted to a standardised storage format. A standardised multilingual interface will be designed to access all resources. These standards will be readily available to encourage other minority language service providers to join the network.

The goal of this project is the provision of up-to-date language information for minority language users in an efficient, technologically-advanced manner, enabling full participation in the multilingual information society, and demonstrating the ability of such languages to thrive and make full use of such technology in their drive for equality of opportunity in Europe.

The project concentrates on the efficient delivery of dictionary and terminology material to users via a standardised WWW presence which aims to become a single focal point for such languages. In this way we can attain a higher profile for these languages in the information society.

The project will serve as an encouragement to users of these languages (the first phase of the project includes Irish, Welsh, Basque and Catalan) and to users of other lesser-used languages who may participate in the future.

In addition, the project aids the propagation of standards in text storage and internet delivery for language resources. Initial designs have encompassed SGML, HTML, XML and Unicode issues.

The key to the success of this project is ease of use by general users. In order to ensure that this is adhered to, the technology is that of the commonly-used Web browsers.

In keeping with the principle of flexibility mentioned above, the first task to be tackled is the definition of an acceptable standardised form of data storage, taking into account the partners' experience of data storage, telematic delivery of resources, language processing etc., as well as the different IT systems installed at partners' sites, and the existing electronic format of resources to be used in the project. Following a number of technical meetings, questionnaires and sample data formats were compiled. The initial version of the database will be stored in a common database in Wales, with mirroring of this site to follow later. All data will be coverted for entry into this database, but partners will retain the existing structures for their own text databases.

At the time of writing an outline database design has been written, and it is hoped that at least one language version will be demonstrated at the LREC workshop, with data conversion proceeding at the various sites. Software will be designed to perform effective storage and delivery of data, to provide efficient network management over the Web, and to allow user identification, verification, feedback etc.

This will entail the development of a package to be installed at each site containing (i) the database technology for data storage (ii) a query program to interrogate the database (iii) a standard set of multilingual web pages containing a simple and user-friendly method of entering queries, with transparent URL links to all sites participating. These programs will be written in a mixture of C++ and Java.

In this way, each site will have exactly the same appearance, with the same multilingual interface appearing to the user, and with linkage being transparent. Assuming that most users will require a simple look-up of words, definitions, synonyms and translations, the interface screens will be clear and easy to use.

Target users

This network is intended for use by anybody working, learning or communicating in a multilingual environment which involves one of the minority languages; professional linguists and translators; language engineers; schools; public authorities and semi-state bodies who are obliged to use minority languages; private companies doing business in those areas of the European Union where minority languages are in use;

cultural organisations concerned with the propagation of such languages. Basically, anybody who uses any of the languages on a regular basis is a target user.

In general, minority language speakers have been provided with printed dictionaries and terminologies (frequently state-funded) but lack of funds make constant revision and re-publication impossible. While some excellent electronic versions of such material may be available, their presence is not widely-known. Users need up-to-date dictionaries and terminologies. They also need an efficient mechanism for keeping aware of the constantly-changing state of a living language, and need to be reassured that new terminology being introduced is in fact acceptable and will have a wide currency.

These users also need the recognition from the wider world that their languages are indeed relevant and usable in the modern world. The existence of an internationally-recognised Web presence can only enhance this profile. Users also need access with the minimum of investment, with the minimum requirement for technical knowledge, with ease of use and an inherent user-friendliness. The transparency of access and multilingual design will ensure that this is the case, and familiarity with any popular Web browser will be the sole prerequisite.

Budget

The total budget for the project is 139900 ECU (with a substantial portion funded by the EU's MLIS project). Access to the site will initially be free of charge, in order to build up interest and a list of regular users.

Subsequent funding for the maintenance and expansion of the site and its services will be obtained from the sale of advertising space on the site, other sponsorship and subscription fees for corporate users. As part of the marketing drive potential sponsors will be approached. It is not intended to charge individual users for access.

In addition, each partner will commit resources to the regular maintenance and upkeep of their sites, following completion of this project. The partners are in all cases at least partly-funded by the respective governments with a brief to provide such services.

Project Partners

InstitiúidTeangeolaíochta Éireann

ITÉ, the Linguistics Institute of Ireland, coordinator of the project, is a state-owned and state-supported research centre. Founded in 1972 as centre for research for state language policy, its principal function is the provision of research and advice services to all organizations dealing with language issues, concentrating on the main languages taught in Irish schools, with particular responsibility in the case of Irish, where research is done not only on school syllabus but on daily language use also.

ITÉ has participated in a number of European projects, including sociolingustic work, a multi-lingual terminology of teaching (EU-funded), the LE-PAROLE project and its predecessor, and is also the location of one of the SOCRATES/LINGUA offices in Ireland. ITÉ is the sole Irish partner in the LE-PAROLE project.

As a state-funded and state-owned institute, ITÉ is providing an electronic version of the definitive Irish-language dictionary, and will also provide access to its own spell-checker currently under development and scheduled for completion during 1998.

Canolfan Ymchwil Cymru

Canolfan Ymchwil Cymru, Prifysgol Cymru Bangor (Research Centre Wales, University of Wales Bangor) is the main partner responsible for technical development, and for Welsh language-specific tasks.

Its current collaborative projects include: design of telematic based training courses; network analysis to establish and utilise the strength of social networks as a learning principle; minority language planning research; linking a structured computer mediated communications environment with a database of multimedia resources for a pan-European HE and teacher development.

Its language resources include a Welsh language spell checker; a Welsh dictionary in electronic format, computer-aided language learning material and a Welsh language grammar checker.

UZEI

UZEI is a contributor to the technical work, and is responsible for all Basque language-specific tasks.

UZEI is an institute working in the area of language development in the Basque Region for more than 20 years. It has agreements with the Basque Government, the regional authorities of Araba, Bizkaia and Gipuzkoa, and with the Government of Navarre. Its main activities include: the compilation and publication of specialist terminological dictionaries; the development and maintenance of a computerised termbank with telematic consultation service; the systematic treatment of current standard Basque based on corpus material; participation in projects involving natural language processing (morphological analysers, spell-checker/corrector, automatic lemmatizers etc.); translation of technical texts; training of terminologists and lexicographers.

Its resources include: terminological data in Basque, Spanish, French and English (>600000 entries); an existing telematic query service; an Internet query service (in development).

Universitat Autonòma de Barcelona

UAB has a commendable record in its work in informatics and the provision of language and pedagogical materials via telematics. It has a section dedicated to this area in its Faculty of Education, which also has close links with the University’s Faculty of Translation and Interpretation.

High-profile projects at UAB include: creation of an integrated WWW-based multimedia environment for role-play activities in language learning; an intercontinental cooperative project involving exchange

of trainee teachers and the pooling of resources, including the establishment of electronic discussion fora, databases, etc.; provision of teaching materials via multimedia and video-conferencing.

UAB is including in the project its own electronic Catalan dictionary as well as multilingual terminology databases in the MULTITERM format.

Implications

This project combines the expertise and resources of 4 highly-respected players in the language resources field, enhancing their products and gaining a significantly larger market for all of them.

A wider implication is that this network will be a very visible, high profile implementation of cooperation by minority language users in the EU, supported by the EU, giving full recognition to the principle of cultural diversity within unity.

The project is a response by language data owners, publishers, and public organizations specialized in the production of language resources, setting up electronic networks in order to share resources and tools for their respective users.

After the initial period of free and open access, the levels of feedback received together with statistics from web counters etc. will give a reasonably accurate assessment of the impact and effectiveness of the project. Another indicator will be the willingness of other minority language service providers to join the network.

Concertation

The project will work in close cooperation with a number of other European projects, particularly in the area of corpus/lexicon delivery (e.g. the ELAN project). Initially cooperation will take place in a number of areas:

- Publicity - reference to projects in publicity material, launches etc, possible joint participation at conferences etc
- Software design - while the ELAN design is of necessity more complex than that required by the MELIN project, it has been agreed that the responsible MELIN partners will take part in the design discussions of the ELAN project, with a view to ensuring compatibility in software issues.