The LE-PAROLE project and the National Corpus of Irish

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

Abstract

Institiúid Teangeolaíochta Éireann/Linguistics Institute of Ireland is the Irish partner in the LE-PAROLE project - as our contribution to that project we have initiated the National Corpus of Irish, the first electronic corpus of contemporary written Irish, and we hope to illustrate some of the difficulties (and some of the joys!!) involved in that work. We would also like to demonstrate part of the corpus using the DBT corpus management tool developed in Pisa. Our corpus was used as one of the examples in the successful mid-term and final reviews of the LE-PAROLE project.

Introduction

Institiúid Teangeolaíochta Éireann (ITÉ), the Linguistics Institute of Ireland, 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 - Irish, English, French, German, Italian, Spanish. The Institute has particular responsibility in the case of Irish, where research is done not only on school syllabus but on daily language use also.

The Institute 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. LE-PAROLE has produced a large-scale harmonised set of corpora and lexica for all official European Union languages and a corpus for Irish. The resources were produced in a standard format supporting selection and customisation. The corpus part of the project has produced large (in general, at least 20 million words) monolingual harmonised corpora obeying common mark-up conventions. Part of the corpus is also morphologically tagged.

As our contribution to that project we have initiated the National Corpus of Irish, the first electronic corpus of contemporary written Irish.

In Ireland, we foresee a wide range of applications for such a resource - it is an extensive basic resources which can be used in development of basic language tools such as spell-checkers, grammar-checkers and machine translation packages. For publishers, it can form a basis for language learning material and for lexicographic work, and for academic researchers it will provide both basic resources and a basis for comparative analysis.

Status of Irish in Ireland

According to Article 8 of Bunreacht na hÉireann (Constitution of Ireland) Irish is the first official language of Ireland, with English as the second official language. In practice, however, nearly all legislation is first drafted and enacted in English, and subsequently translated in Irish. Since 1922, Irish has been a compulsory subject in Irish primary and secondary schools.

Despite official recognition, there are only about 40000 fluent speakers of Irish. In certain regions (known as Gaeltachtaí) Irish is the regular daily language. The semi-State body Údarás na Gaeltachta was established to enhance the economic prosperity and survival of these regions. Financial incentives are available to residents, and also to industries willing to establish bases in the Gaeltacht. One of the major industries, however, is tourism, which has had a negative effect on Irish language usage in the regions. In these areas, Irish is the language of instruction in schools. However, the most encouraging signs of expansion of the Irish language have been in the cities and rural areas where Irish is not the daily language, yet all-Irish schools have been founded and are flourishing. Often these schools endure poor conditions and no State funding for a number of years before official recognition is received. This movement started at primary level, but the knock-on effect was a demand for second-level all-Irish schools, and has now resulted in one of Ireland's newer universities offering a degree course (commenced in 1994) in Finance, Computing and Enterprise, all taught entirely through Irish.

The Irish language TV channel, Teilifís na Gaeilge, commenced broadcasting in 1996. There are two weekly newspapers in Irish, and radio stations broadcasting through Irish, include the State-owned Raidió na Gaeltachta, broadcasting mainly to the Gaeltacht regions, and Raidió na Life, which broadcasts to the Dublin area.

On accession to the European Union, Ireland waived its right to make Irish one of the official languages of the European Union, but all EU treaties affecting Ireland are translated into Irish.

Corpus collation/collection

In accordance with LE-PAROLE requirements, the National Corpus of Irish is composed - according to a common design - of different material taken from books,

newspapers, periodicals and other sources such as brochures and reports. It conforms to the Corpus Encoding Standard (CES) developed in collaboration with NERC and EAGLES groups and based on the Text Encoding Initiative (TEI) Guidelines.

Included in the LE-PAROLE definition of standards were certain minimum and maximum percentages for the composition of the corpus. Only texts from 1970 onwards were permitted, at least 80% of the text had to be from 1980 or later, and the medium requirements were as follows:

Medium	minimum	maximum
Newspaper	58%	72%
Book	16%	22%
Periodical	4%	10%
Miscellaneous	8%	12%

In the case of Irish special allowances were made for the fact that Ireland has no daily newspaper in Irish. At the commencement of the project there were two national weekly newspapers published in Irish, one of which (*Anois*) has since ceased publication. A new national weekly (*Foinse*) took its place in 1997. In addition, there was recognition of the weaker starting position of Irish vis-à-vis the other languages in the project - until then, Irish had no corpus whatsoever.

The starting point for corpus collection was to stop the deletion of any material that might have been held in electronic format by publishers but was, until, that point, regarded as useless. The project team identified the major potential sources of source material as follows:

- Publishers of Irish-language texts
- Typesetters of such texts
- Irish government sources (including the official parliamentary translation section)
- EU sources
- World-Wide Web pages
- In-house (ITÉ material).

All sources were contacted and, in the first place, asked to store any material held in electronic format until we could assess its suitability for inclusion in the corpus. Some publishers proved very helpful in this regard, others referred us to their typesetting and design company. In this we were fortunate - a huge percentage of Irish-language typesetting is done by one company. This company had retained copies of all work done by them. They provided us with a list of their authors, and the authors were sent a letter outlining the project details and asking for a letter of consent to use their work in the corpus. The response to this was very positive.

On a regular basis this company then sent copies to us of all work done by then, once authors' permission had been received. Only a small administration fee was charged in this case.

In the case of newspapers, one of the newspapers was ceasing publication just as the project was commencing. We contacted the publishers, and offered to buy all past issues available in electronic format - in fact, we purchased the computer on which all their backups were stored! The publishers of both current newspapers have agreed to supply text on a regular basis for use in the corpus.

The largest single contribution to date has come from the State-owned publishing house which is attached to the Irish Department of Education. Its primary function is the provision of Irish-language textbooks for use in schools but it has published material on a wide range of general interest topics and has been most co-operative in its support of the project. In fact, as the publisher of the definitive Irish-English and English-Irish dictionaries, it may well prove to be one of the major users of the completed resource for lexicographic work.

The Irish parliamentary translation section and the EU have provided some material with, hopefully, more to come. Combined with ITÉ's own publications (which are usually bilingual, Irish/English) these will form the technical part of the corpus and may, in future, be used as part of a parallel bilingual corpus.

Data capture/processing

The contributed texts ranged in formats from plain ASCII, through MS-WORD and WordPerfect to Quark Xpress. A considerable amount of work was done on producing text into a single readable format suitable for tagging. Also, some text was contributed in printed format - this was scanned into a word-processor using a basic Optical Character Recognition package.

The LE-PAROLE corpus design allows for a number of levels of tagging: document wide (author, character set etc.), structural mark-up (volume, paragraph etc.), subparagraph structures (sentence, quotation, abbreviation etc.) and with linguistic annotations (morphosyntax, syntax etc.), in line with the TEI, EAGLES and preparatory PAROLE guidelines. The whole corpus first had to be tagged to paragraph level, with a detailed header on each text which included ownership details, date of text, number of words, some information relating to topic, genre etc.

In addition to this tagging level, a portion of the corpus had to be tagged to part-of-speech level. A lexical data base for use in part-of-speech tagging was created in a number of formats. A version of a standard dictionary was input to a MS-ACCESS database, as a base lexicon. In addition, work was done on the Xerox suite of lexical tools to produce the various mutated forms of Irish nouns and verbs. All of this material was then used as the initial lexicon for the Brill tagger which was trained on small portions of the corpus and extended. The contract requirement was for 150,000 words tagged to part-of-speech level, and work on the National Corpus of Irish is proceeding beyond that figure.

At the time of writing, the composition of the corpus is as follows:

Medium	% of total corpus	Words
Newspaper	43.42	6,625,000

Book	45.25	6,900,000
Periodical	4.45	680,000
Miscellaneous	6.88	1,050,000
Total	100.00	15,255,000

Newspapers

Source	Words
“Anois” (weekly)	2,525,000
“Lá” (weekly)	1,700,000
“Foinse” (weekly)	2,400,000

Books

Source	Words
An Gúm (State publishing)	4,000,000
Peanntrónaic Teo (typesetting/design company)	2,800,000
Cló Iar-Chonnachta Teo (publishers)	100,000

Periodicals

Source	Words
“Comhar”	530,000
“Teangeolas”	50,000
Other	100,000

Miscellaneous

Source	Words
ITÉ	300,000
EU	250,000
Irish Government	500,000

Corpus Management tool/corpus availability

The DBT package is a corpus management tool, for use in on-line (World-Wide Web) form or as a standalone version. Developed at the Consorzio Pisa Ricerche (CPR), the tools was used at the successful mid-term and final reviews of the LE-PAROLE project to demonstrate samples of the project corpus work. On both occasions samples of the National Corpus of Irish were demonstrated. It is hoped to show the tool in action on a sample of the Irish corpus at this LREC workshop, which will better illustrate issues such as tagging problems, character set issues etc.

As a follow-on from the corpus development work, the European Languages Activity Network (ELAN), funded by the EU's MLIS programme, will included samples from the Irish corpus.

Also, it is hoped to make at least part of the corpus available on ITÉ's own website (http://www.ite.ie) in the near future.

Future developments

It is envisaged that material will continue to be gathered for the corpus for the foreseeable future. At the time of writing, it appears that nearly all sources of existing printed material have been contacted and have indicated willingness to contribute. Major quantities of material remain to be converted and tagged, especially government and EU material. Also, ongoing supply from newspapers is regarded as crucial in maintaining the topicality and contemporaneity of the corpus.

The first major practical use of the corpus will be as resource material for an Irish spellchecker program which is being developed by Trinity College Dublin in collaboration with ITÉ. This project, scheduled for completion at end-1998, will provide the first Irish speller program compatible with the leading PC-based word-processing programs.

The corpus will also be used as resource material for Irish-language teaching material being developed in ITÉ.