Irish-Language Terms for Legal Translation: Lexicon Extraction from a Parallel Legal Corpus

Fionnuala Cloke & Dr Gearóid Ó Cleircín
Fiontar, Dublin City University, Ireland

Keywords: term extraction, parallel corpus, legal terminology

Background

The aim of this project is to make Irish-language legal terminology available online to the public in general and to translators of legislation in particular. Fiontar, Dublin City University, began work on this project, which is funded by the Department of Equality, Rural and Gaeltacht Affairs, in January 2008. These terms will be published in www.focal.ie, the National Terminology Database for Irish, which was developed by Fiontar in collaboration with the Terminology Committee, Foras na Gaeilge, the statutory body responsible for terminology development for the Irish language.

In the early stages of the project, two existing sources of legal terminology were digitised — the bilingual legal dictionaries *Téarmaí Dlí* (Government of Ireland 1959) and *Focal sa Chúirt* (Ó Catháin 2000). The terms in these publications were prepared for electronic publication on www.focal.ie and are currently available to the public as auxiliary glossaries on that site.

The primary sources of the legal terminology for this project are the Irish-language versions of secondary legislation (or Statutory Instruments). The instruments cover a wide variety of domains, from agriculture to hairdressing, and many of the terms in them are 'legal' insofar as they are the chosen terms used in legislation in the Irish language rather than terms which represent purely legal concepts. There are two batches of Statutory Instruments in this project. The first batch consists of 861 relatively short Statutory Instruments from 1980 and 1981. These texts were chosen as they are the most recent translations of secondary legislation produced by the Translation Section of the Houses of the Irish Parliament. The Translation Section has responsibility for the translation to Irish of all Irish primary legislation (or Acts of Parliament) and, until the early 1980s, of all secondary legislation. Due to the linguistic and translation expertise of the Translation Section staff the translation in these texts, and so the terms contained in them, are of a high standard. The second batch consists of two larger, more recent Statutory Instruments from 1997 (District Court Rules) and 2001 (Circuit Court Rules) translated by a highly qualified translator of legal documents.

A new government Central Translation Unit was established in early 2010. This development has directed the focus of the project to meeting the translation and terminology needs of this new Unit as well as the general terminology needs of users of **www.focal.ie.**

Phase 1 (2009-2010)

Overview

Work began on the first batch of instruments in early 2009. There were 861 Statutory Instruments, or 913, 890 words of English text, produced in the years 1980 and 1981. The Irish text was only available in hard copy which had to be scanned and digitized to facilitate editing. An initial test on 79 Statutory Instruments resulted in the formulation of the following four distinct stages of work: (1) creation of the parallel corpus; (2) term extraction; (3) identification of corresponding Irish terms and bilingual concordance segments; (4) editing of the terminological entries.

(1) Creation of the parallel corpus

While the English text was available in electronic form, the Irish text was only available in hard copy. This hard copy was scanned and made editable through OCR resulting in two batches of files — PDF images of the hard copy pages for reference and Word files containing text which was editable but flawed due to the OCR process. These Word files were then cleaned up. Several changes were made to these files to make them suitable for the alignment process: errors resulting from the OCR process were corrected; publication details (page numbers, etc.) were deleted; errors not resulting from the OCR process were corrected and carefully recorded referring always to the PDF files of the published documents; and, tables, maps and forms corrupted by OCR were deleted or corrected depending on the terminological value of the content. The texts were then aligned. The resulting TMX files have two uses – as a parallel corpus for terminological work and as the basis for a translation memory for the new Central Translation Unit.

(2) Term extraction

The extraction tool used, SDL MultiTerm Extract 2007 (and later the 2009 version), works on the basis of frequency - only those words or groups of words appearing three times in the text are selected and offered as term candidates. The aim of the project at this initial stage was to record every Irish term in these texts in use as a translation for the English terms. Monolingual rather than bilingual term extraction was chosen as the volume of data to be processed was too large for the 2007 version of MultiTerm Extract meaning only simple extraction and export could be safely carried out before extraction files crashed. This meant that the bulk of the culling and editing was done after terms were exported. An empty clone of the term database behind www.focal.ie was created for this purpose. This legal database and its editorial interface were especially designed for internal terminological work and to provide a secure environment with onsite technical support for editing and managing the extracted terms.

Monolingual term extraction from the English text offered around 18,000 term candidates. Of these, 7,444 English terms were exported to the database. A further 9,907 English terms were extracted by hand - those occurring less than three times in the text. Even though such terms occur infrequently in the legislation it was important to include them in the database as they cause particular difficulties for LSP users as they may not be included in general lexical resources (Bowker and Pearson 2002:166).

(3) Identification of corresponding Irish terms and bilingual concordance segments

The TMX files were added to the database. A search function allows the editor to select an English term and generate a list of every aligned segment pair in which it appears. These segments are examined and the Irish term or terms used as translation for the English term are identified and added to the database along with the aligned segment pair in which they appear. These segment pairs have two purposes. The first is to record the format of the term as it was when extracted. All terms are edited (see below) and so in order to search for them again in the aligned text their original format must be recorded. The second purpose is to provide the raw material for usage examples when the terms are published in www.focal.ie. The layout of these segments has been designed so that they can be easily exported to the www.focal.ie database.

(4) Editing of the terminological entries

Each entry is edited according to the guidelines followed in the National Terminology Database for Irish (www.focal.ie) – for example, all nouns are in the singular form, unless the plural is more appropriate, and all verbs are in the imperative. A second check is then performed to correct or record grammatical or spelling inconsistencies. If these inconsistencies were created during the editorial process they are corrected. If the inconsistencies appear in the aligned segments they are noted and recorded but not always changed. The terminological work is still descriptive at this point.

Work is currently being carried out on a second batch of translated secondary legislation from 1997 and 2001. This contains approximately 380,000 words of English text – less than half the volume of the first batch. Again the English text is available electronically. The corresponding translated Irish text is available only in PDF. It is intended to follow the same four stages as outlined above – the creation of a parallel corpus (and TMX file), term extraction, identification of corresponding Irish terms and finally editing of the terminological entries. The main difference this time is the use of bilingual rather than monolingual term extraction. This will be performed using the more recent version of SDL MultiTerm Extract (2009) which can process larger volumes of text. These pairs will be input in the legal database and edited in the same way as the Batch 1 entries. When completed, this new batch of terms should provide an interesting opportunity to study the development of legal terminology in Irish over a period of twenty years (1981-2001).

Results so far

The terminological entries

There are currently 14,407 entries in the database in which there are 14,114 English terms, 15,943 Irish terms, 12,513 concordance segments and 688 definitions. The editing work is ongoing.

The terminology work in this phase is based entirely on the contextual data in the parallel corpus. Other Irish-language terminology sources are not routinely searched as the current aim is simply to record the terms as produced in the Statutory Instruments. This involves the basic matching of extracted English terms with every Irish-language term used as a translation for it and the editing of these entries to correspond to the www.focal.ie format. Information on usage in the aligned text is routinely recorded as editorial notes. While definitions are not routinely included in each entry, they are used in cases of homonymy, where it is clear that more than one concept is represented by the same English and/or Irish term – ward (in hospital), ward (a person who is under the protection or in the custody of another). Editorial questions are routinely recorded in the database when the context or meaning of the term is unclear to the editor or where there is a possible grammar discrepancy. These questions are then reviewed by the in-house terminologist and if necessary will be forwarded to outside subject experts before or during Phase 2.

TMX files

Another product and essential ingredient of this project are the TMX files which contain the aligned texts of 861 Statutory Instruments

(913,890 words of English text and the corresponding Irish) from 1980 and 1981 which are used internally as a parallel corpus. These TMX files have been made available to the Central Translation Unit and will later be made available to the public. Using the analysis function in SDL Trados, the English Statutory Instruments from random years between 1982 and 2008 were compared batch by batch with a translation memory created from the aligned text of 1980 and 1981. There is a high match percentage in the material from the years immediately after 1981 dropping progressively in the more recently produced legislation – 47% of the English material from 1982 had a 50-99% match in the aligned material of 1980 and 1981 compared to 4% of the material from 2008. This suggests that the closer together in time the Statutory Instruments were produced the more similar the text in them. The work on the second batch (1997 & 2001) will result in another TMX file containing more recently translated Statutory Instruments.

Phase 2 (2011– 2012)

Overview

The main objective of Phase 1 is the provision of TMX files for the new Central Translation Unit and the extraction and recording of terms from the Statutory Instruments. Phase 2 will involve more detailed analysis of the terminological entries created in Phase 1 separate from their original context in the Statutory Instruments. The aim in the medium term is to publish the terms on the National Terminology Database for Irish, www.focal.ie.

All terms published in www.focal.ie must first be approved by the Terminology Committee. Material from the two bilingual dictionaries processed in the early stages will be input to the legal database along with the terms from both batches. There will be four main term sources in the database at this point: terms from the two dictionaries, *Focal sa Chúirt and Téarmaí Dlí;* terms from 1980-81; and terms from 1997 and 2001. In order to expedite the Terminology Committee authorisation process all the entries in the legal database will be compared with the terminology already approved in www.focal.ie. This

work will follow a model for terminology work already developed in Fiontar in which Irish terms are provided to correspond to selected entries from the EU's IATE database. This terminology work is done according to a system of research and checks carried out in another modified clone of the www.focal.ie database. It is proposed to use the same approach in order to research the legal terms and make recommendations or engage in further investigations where there are multiple equivalents. To facilitate feedback in the IATE work Fiontar created a separate extranet to which term-lists are uploaded on a monthly basis. The Irish-language translators of the various EU institutions can access this website and provide feedback on the new terms which is automatically transferred to Fiontar's editorial interface. A similar system is envisaged for Phase 2 of this legal terms project.

This stage of the project should yield four categories of entry: (1) entries which are the same as those in www.focal.ie and therefore don't need Terminology Committee approval; (2) entries in which the English terms are the same as those in www.focal.ie entries but in which the Irish terms differ; (3) legal database entries in which the Irish terms are the same as those in www.focal.ie entries but the English terms differ; (4) and, finally, entries consisting of entirely new terms not previously published in www.focal.ie. The entries in the second, third and fourth categories will require review by the Terminology Committee with potential input from the Central Translation Unit. At this point all entries should contain a substantial amount of terminological information drawn from the research carried out in both Phases.

Publication

The terms will be published in www.focal.ie when they have been approved by the Terminology Committee. Much research and planning will be needed to work out the best way of inputting these terms into www.focal.ie and of presenting them to the public. The volume of terms and the diverse uses and sources of these terms has grown hugely since the inception of www.focal.ie and so the methods of storage and presentation will need to be developed and refined. There is currently a draft-proposal under review for the design and construction of a new term management system for the various term collections. It is envisaged that this new design will overcome the present constraints.

Users

While it is intended that the TMX files will be made available to the public in general, it is likely that the main users of these files will be specialist translators of legislation and other legal material. The extracted terms will be of interest to a wide variety of users of *www.focal.ie* including translators, students, journalists, civil servants and academics. Both the TMX files, when used as the basis for translation memories, and the terms could significantly assist the work or studies of these users.

Further development and research

Another considerable source of valuable Irish-language terms could be the translated primary legislation or Acts of the Irish Parliament. Much research and planning would be needed for this due to the large volume of data involved.

When completed, this new batch of terms should provide an interesting opportunity to study the development of legal terminology in Irish over a period of twenty years. The bilingual corpus offers the potential for further research in other fields such as translation, linguistics and lexicography. More specific domain-based research could also be carried out.

References

Bowker, L. and Pearson, J. (2002) *Working with specialized language: a practical guide to using corpora,* London/New York: Routledge.

Government of Ireland (1959) *Téarmaí Dlí,* Dublin: Oifig an tSoláthair.

National Terminology Database for Irish *[http://www.focal.ie]* (accessed 13 May 2010).

Ó Catháin, L. (ed) (2000) *Focal sa Chúirt,* Dublin: Coiscéim.