Compiling a Dictionary of Irish with the Aid of an Information Retrieval System

Kieran DEVINE and F. Jack SMITH

Résumé. Un système de recherche automatique de l'information, qui comporte des possibilités linguistiques très sophistiquées, a été adapté pour aider à la compilation d'un dictionnaire de l'irlandais moderne. Les *corpora* et les *indices* sont enregistrées dans la mémoire du serveur d'un réseau de PCs et sont accessibles par le logiciel *Quill* dont des copies ont été installées sur les postes de travail PC qu'utilisent les lexicographes. Le système de Quill et un *éditeur* utilisé pour la compilation des articles de dictionnaire fonctionnent simultanément, dans des *fenêtres* différentes grâce au système *DesqView*, qui permet de transférer les citations directement d'un corpus au dictionnaire.

Keywords: information retrieval, dictionary compilation, Irish.

Mots-clés : recherche de l'information, compilation de dictionnaire, irlandais.

1. Introduction

Work on a dictionary of modern Irish by the Royal Irish Academy has been transformed by the recent installation of a new computer system based round an information retrieval system. This will greatly widen the scope of the computerised text corpora used in the preparation of the dictionary and at the same time speed up its publication.

For well over a decade no-one would have undertaken the compilation of a dictionary without the aid of a computer. Already in the 1970s and early 1980s several computerized dictionary projects were reported in the literature; for example, papers by Imbs (1971); Lara (1976); Healey (1985) amongst others are covered in a review of the use of computers for dictionary compilation by Devine, Harvey and Smith (1987). Computers can help in almost every aspect of the

✉ Department of Computer Science; The Queen's University of Belfast; Belfast BT7 1NN (Northern Ireland).
Fax: + 44 232 33 12 32

task of a lexicographer including lemmatization, the separation of homographs, the selection of examples and the classification and annotation of samples. In addition they are used for the typesetting of the final entry as well as helping with management and word processing. However, their greatest and indispensable help to a lexicographer today is in the searching for citations of particular words in selected corpora, possibly in very large corpora. Using techniques developed for on-line information retrieval systems all of the citations for a particular word can be displayed one after the other in seconds, even from a text corpus consisting of millions, or tens of millions of words and even a relatively inexpensive PC is sufficiently powerful today for this task (Salton, 1989). The use of information retrieval searching techniques makes possible exhaustive searches for word types or lemmas which would otherwise be unthinkable. The Irish dictionary project described in this paper is based on the adaptation of one such information retrieval system, called *Quill*.

2. *Quill*

The information retrieval system *Quill* (Devine and Smith, 1984) was originally designed for the storage and retrieval of legal documents (hence the Acronym: Queen's University Interrogation of Legal Literature). It has a few linguistic features which make it suited to the task of dictionary compilation, and to the compilation of an Irish dictionary in particular. It can search for any word type, like any other information retrieval system, but it also has the ability to store an index to whole lemmas for a word at the same time as it stores a more conventional index to individual word types. This is a powerful feature as dictionaries are normally based primarily on word lemmas rather than on word types and this feature facilitates searching for all words of a lemma simultaneously.

For many languages, such as English, French or Latin, the lack of this feature is not critical if the information retrieval system has the simple ability to perform a stem search, since an appropriately chosen stem will automatically find all of the word types in a lemma. For example, the stem `comput*` will find all occurrences of the word types of the lemma *compute*, *i.e.*

```
compute, computes, computed, computing
```

It will also find occurrences of the words *computer*, *computable*, *computational* and others; but this is not a major problem as they can be discarded. However, in Irish, both the beginning and end of some of the grammatical forms of each regular word may be inflected *e.g. an tábla* (the table), *ar thábla* (on a table), *na*

dtáblaí (of the tables). Therefore, the stem facility cannot pick up occurrences of all words in a lemma and a different structure is needed.

Quill provides such a structure which makes the indexing and retrieval of lemmas possible. This is the characteristic that indexing is based on a division hashing algorithm rather than on a B-tree or similar structure which requires an alphabetic ordering of the words in the index. A B-tree facilitates stem searches on alphabetic lists, but as we have explained this will not pick out all of the types within a lemma in Irish. However, a hashing algorithm, in association with a ring structure, linking all words of a lemma together using pointers, allows the rapid retrieval of any word type followed immediately by the occurrences of all words in the ring, *i.e.* by all of the words in a lemma.

This works equally well for irregular words. However, the data entry of all of the grammatical forms for each noun or verb, regular or irregular, would take a considerable time and the regular words follow well defined rules, *e.g.* the genitive plural of a noun beginning with "t" after the article has the "t" changed to "dt". As we explain later this has enabled us to automatically index together word types which differ only in mutation, *i.e. dtáblaí* is indexed along with *táblaí*.

3. The FNG dictionary

The Dictionary of Modern Irish—*Foclóir na Nua-Ghaeilge* (FNG)—is to be a comprehensive monolingual dictionary of Irish, covering the period from the beginning of the 17th century until the present day. Each citation in the dictionary will include an ordered set of selected examples of use, drawn from a representative range of geographical and temporal provenances. The project is based at the Royal Irish Academy, Dublin, under the general editorship of Professor Tomás de Bhaldraithe, currently assisted by four other full-time editors and two secretarial staff. At any time there may also be one or two part-time staff, or staff on short-term contract. The first volume of FNG, covering the alphabetic range A-AL, is at an advanced stage of drafting.

The sources for the dictionary are large, and consist mainly of printed books, manuscript texts and printed and manuscript dictionaries. The collection of citations began at the end of the 1970s and at first took the form of written slips. Computer utilisation began in the early 1980s, when the project obtained an *Altos 4* microcomputer, which was used to key texts and run Basic programs (with the help of Michael Doherty from the Department of Computer Science at Trinity College Dublin) to generate concordances of words in A-AL. The present computer system arrived in 1991.

4. The PC network

Versions of *Quill* were available for both a *VAX* computer and for an *IBM* compatible PC. So the system could either have been mounted on a *VAX* mini-computer (3100) using terminals for access, or on a network of PCs linked to a central file server (also a PC) holding the text databases. The flexibility of PCs for other applications, and cost of future enhancements, both software and hardware, determined that the PC network was the one chosen for the project.

The network consists of 8 *Wang* 286 PCs networked with the Novell software, with a *Wang* 386 PC used as a file server. The network also includes 3 printers; one is a laser printer and two others are inexpensive dot-matrix printers which can print code page 850. It is interesting that in drafting and day-to-day work the lexicographers use the faster dot-matrix printers frequently and the slower but higher quality laser printer infrequently. There are also two non-networked PCs, a 386 and a 286; they are used for *MS-Windows* applications and are kept apart to reduce the complexity of the network. Data are transferred to the network workstations on diskette.

The networked PCs function under MS-DOS 3.3, with DOS code page 850 installed to provide the required character-set (Irish has acute-accented vowels). Software which is code-page transparent is therefore preferred. A keyboard utility (due to Dr Damien McKeever) is used to provide dead-key vowel accents under MS-DOS. A program has also been written to sort records, placing accented vowels correctly. The standalone PCs have *MS-Windows* 3.0.

Text entry and dictionary drafting both use *PC-Write*, a simple but effective word processor, which has the advantage of creating files which are almost pure text files (completely so if formatting features are avoided, as in text entry). Text entry is by two full time typists who use 2 networked PCs. Data are collected locally and transferred to the file server at the end of each day. Text is also entered using an optical character recognition system—a *Microtek* scanner with *Recognita Plus International* OCR software—on the standalone 286 PC. This works well when the print is of high quality.

When *PC-Write* is used to draft dictionary articles, formatting features are freely used (bold, italic, etc.). Customisation to allow the representation of phonetics is under consideration. The DTP package, *PageMaker* (on the standalone 386 PC), to which these files will ultimately be passed, has an option for reading *PC-Write* format.

The *Quill* software is resident on each networked PC and can retrieve citations from the text databases created from the texts and held on the file server. In order to paste retrieved contexts from *Quill* directly into dictionary