Distilling useful linguistic information
from digitized dictionaries

John Lunney
B.A. (Mod.) Computer Science, Linguistics and a Language
Final Year Project, May 2008
Supervisor: Elaine Uí Dhonnchadha
Declaration

I hereby declare that this thesis is entirely my own work and that it has not been submitted as an exercise for a degree at any other university.

________________________________________ May 6, 2008

John Lunney
Permission to Lend

I agree that the Library and other agents of the College may lend or copy this thesis upon request.

__________________________ May 6, 2008

John Lunney
Acknowledgements

Firstly, I would like to thank my supervisor Elaine Uí Dhonnchadha for all her help and advice in the preparation of this project, and for agreeing to supervise it in the first place.

I would like to thank my family for their support during this and my previous years of study, and for putting up with my late nights throughout.

I would also like to thank Christoph Wendler and Geraldine Herbert who have both helped me a lot during the project. Christoph gave me advice on technical matters while Geraldine pointed me to relevant material in the literature.

I also have to thank my classmate Niamh for putting up with me for this year, and all her advice and support.

Thanks are also due to those at the Royal Irish Academy working on
the Foclóir Stairiúil na Nua-Ghaeilge project, and to Liam Mac Amhlaigh
for his kind correspondance and excellent book.

Support for this project was generously provided by Cathal Convery
and Foras na Gaeilge.

Finally, I would like to thank Carl Vogel. Carl has been course director
and teacher to me for 4 years, and I have greatly appreciated his support
and teaching throughout my time in university. His commitment to the
Computer Science, Linguistics and a Language course is unfailing.
Contents

1 Introduction ................................................................. 1
   1.1 The aim of this paper .................................................. 1
   1.2 Typographical notes ................................................... 2
   1.3 Miscellanea ............................................................. 3

2 The Irish language .......................................................... 4
   2.1 Names ................................................................. 4
   2.2 Celtic languages ....................................................... 5
      2.2.1 P-Celtic/Q-Celtic opposition .................................. 5
      2.2.2 Insular/Continental opposition ................................ 6
   2.3 Orthography of Irish .................................................. 8
   2.4 Syntax of Irish ........................................................ 11
      2.4.1 Word order ....................................................... 11
   2.5 Morphology of Irish .................................................. 11
   2.6 Standardization: An Caighdeán Oifigiúil .......................... 12
   2.7 Legal standing of Irish ................................................ 13
## Contents

2.7.1 Irish perspective ........................................ 13
2.7.2 European perspective ................................. 15

3 The Ó Dónaill dictionary ................................. 17
  3.1 History .................................................. 17
  3.2 Niall Ó Dónaill ........................................ 18
    3.2.1 Early life .......................................... 18
    3.2.2 Later life and death .............................. 19
    3.2.3 Name ................................................ 19
  3.3 Development .......................................... 20
    3.3.1 Objectives of the *Focloir Gaeilge-Béarla* project .. 20
    3.3.2 Terms .............................................. 20
    3.3.3 Dictionary structure ............................. 21
  3.4 Coverage ............................................. 22
  3.5 Publication ........................................... 22
  3.6 *An Focloir Póca* .................................... 22
  3.7 Electronic version ...................................... 23

4 Related projects ......................................... 25
  4.1 Overview ............................................. 25
  4.2 eDIL project .......................................... 26
    4.2.1 Dictionary of the Irish Language ............... 27
    4.2.2 Digitization process ............................ 30
    4.2.3 Launch ............................................ 32
CONTENTS

5.2.4 Future work ........................................ 32
5.3 Digital Dinneen ........................................ 33
  5.3.1 CELT ........................................... 33
  5.3.2 Why the Dinneen dictionary ................. 33
  5.3.3 Dinneen dictionary ......................... 37
4.4 Computational Lexicography ......................... 37
4.5 Web User Interface to Irish Dictionary Database ... 38
4.6 An Analyser and Generator for Irish Inflectional Morphology using Finite-State Transducers ........... 39

5 Project planning .................................. 40
  5.1 Programming language ......................... 40
    5.1.1 Perl ....................................... 41
    5.1.2 Python ................................... 42
    5.1.3 Prolog ................................... 42
    5.1.4 C/C++ ................................... 43
    5.1.5 Java ..................................... 44
    5.1.6 Decision .................................. 44
  5.2 Database ..................................... 45
    5.2.1 XML ..................................... 45
    5.2.2 Oracle .................................. 45
    5.2.3 MySQL .................................. 46
    5.2.4 PostgreSQL ............................... 46
CONTENTS

7.1 Conclusion .......................................................... 62

A Abbreviations used ............................................. 66

B Mark-up legend ................................................... 68

C Database structures ............................................. 71

C.1 Initial database structure ................................. 71

C.2 Revised database structure .............................. 74
Abstract

In this thesis, the Irish-English dictionary of Niall Ó Dónaill is analyzed with a view to extracting useful linguistic information from it, in a machine-readable form, such that it might form the basis of further lexicographical undertakings or act as a lexical back-end for other linguistic projects. We examine the particular issues associated with Irish, place the Ó Dónaill dictionary within the context in which it was first published and consider what can be learned from similar projects. One possible approach to the automatic mining of linguistic data from the Ó Dónaill dictionary is described and documented. The results are encouraging and open the possibility for much further work in the area.
Chapter 1

Introduction

“I was reading the dictionary. I thought it was a poem about everything.”
Steven Wright

1.1 The aim of this paper

The project described in this thesis was initiated when it was realized that one of the most important dictionaries of the Irish language, that by Ó Dónaill, could be accessed electronically, using a typesetting file in existence since the late 1970s, but never before used to gain access to its highly significant quantity of lexicographical data electronically. The thesis will set Ó Dónaill’s work in the linguistic and historical context of the period in which it was first published, and will then examine other projects which
have similarly exploited or created electronic versions of other important
dictionaries, such as that of the Royal Irish Academy or Dinneen. The
groundbreaking eDIL project is described in detail in Chapter 4.2.

After these introductory sections, the thesis will then discuss the choices
that were made in developing the methodologies of the project to han-
dle O’Donnell’s data electronically; the suitability of various programming
languages and infra-structural tools is assessed in Chapter 5. In the con-
cluding sections, the results of these are analysed and some conclusions
drawn. The methods developed and the experience gained, it is hoped,
will have some relevance to further use of the methodology and the lexi-
cographical resources it can be used to create, by linguists, computer sci-
entists and perhaps even language learners.

1.2 Typographical notes

Where text is in the Irish language, it will appear in italics. If a translation
or explanation is necessary in English, it will follow in brackets.

For example: Conas atá tú? (How are you?)

In some cases, it may be useful to show a snippet of code. Generally
this will not be a full program, rather a short chunk of code which serves to
demonstrate the principle. Where program code is included in the running
text, it will appear as such:
Code Listing 1.1: Example code snippet

```python
elif re.match(r'\K', chunk):
    chunk = chunk[2:]  # bite off the *K
out.write("Colon cross-reference: ")
if re.match(r'\:',chunk):
    chunk = chunk[1:]  # bite off the colon
```

1.3 Miscellanea

Given the propensity of both linguists and computer scientists for the use of abbreviations and acronyms, an annotated list of those used in the document is given in Appendix A.

Due to the nature of the project, certain sources are in the Irish language. If translation has been undertaken by this author, then the responsibility for any errors lies thereupon.

The author notes and regrets the irony of another piece of scholarship on the Irish language written in English.
Chapter 2

The Irish language

“The greatest masterpiece in literature is only a dictionary out of order.”
Jean Cocteau

2.1 Names

The Irish language is an Indo-European language. In the English language, it is known as Irish. In Irish, it is known as Gaeilge. It is often incorrectly referred to as Gaelic, which more properly refers to the Celtic language of Scotland, Scots Gaelic.

Under the ISO 639-1 standard for the representation of names of languages, Irish was given the two-letter code of ga. Under ISO 639-2, which mandates three-letter codes, this became gle, now representing only Mod-
ern Irish. Middle Irish (Gaoidhealg) and Old Irish (Goídelc) are afforded their own codes, $mga$ and $sga$ respectively. These codes are preserved in ISO 639-3.

2.2 Celtic languages

More specifically, Irish is a Celtic language. It shares a common ancestry with other languages which descended from the supposed proto-Celtic language. There are two main hypotheses for how the different Celtic languages are related. Both hypotheses assume a common ancestor language, Proto-Celtic from which all the Celtic languages are descended, which in turn was descended from a supposed Proto-Indo-European language. The two hypotheses describe how language development branched from the common beginning of the Proto-Celtic language. The reconstruction of both this Proto-Celtic language and the Proto-Indo-European is currently in progress through historical linguistic analysis.

By both hypotheses, Irish is most closely related to Scots Gaelic and Manx.

2.2.1 P-Celtic/Q-Celtic opposition

This is the older of the two hypotheses. It suggests that there is a split between the $P$-Celtic and $Q$-Celtic. In the original formulation, the Gaulish and Brythonic languages belonged to the former group, while the Goidelic
languages alone belonged to the latter group. Eventually, it was realized that the Celtiberian languages belong to the later group, according to the hypothesis. This realization led some to abandon this hypothesis in favour of the Insular-Continental opposition hypothesis.

The difference between the groups of languages, on which the opposition is based, is the realization of the Proto-Celtic consonant *kw. In the P-Celtic languages, this is realized as *p, while it is realized as *k in the Q-Celtic languages.

By way of an example, we will take the Proto-Celtic verb root *kwrin—, meaning “to buy”. In Welsh, a P-Celtic language, the Proto-Celtic *kw is realized as *p, yielding pryn. In Old Irish, a Q-Celtic language, the Proto-Celtic *kw is realized as *k, yielding cren (compare Modern Irish ceann-aigh, meaning “to buy”).

While this appears quite compelling, any classification which relies on a single phonological feature as its determining criterion is somewhat weak, particularly if, as in this case, the sound change in question also occurs in other language groups.

### 2.2.2 Insular/Continental opposition

The Insular/Continental hypothesis denies that the *p/*k opposition is as important as previously thought. It suggests that these shifts from the Proto-Celtic language occurred independently. Rather, proponents of the Insular/Continental opposition hypothesis take a somewhat geographical
P-Celtic

- Gaulish
  - Lepontic
  - Noric
  - Galatian

- Brythonic
  - Cumbric
  - Pictish
  - Old Welsh
  - Middle Welsh
    * Welsh (Modern)
  - Southwestern Brythonic
    * Breton
    * Cornish

Figure 2.1: The hierarchy of the P-Celtic languages

Q-Celtic

- Celtiberian

- Goidelic
  - Primitive Irish
  - Old Irish
  - Middle Irish
    * Irish
    * Scottish Gaelic
    * Manx

Figure 2.2: The hierarchy of the Q-Celtic languages
stance on the matter. The Insular (island) languages are all those that are or were spoken on the islands of Britain and Ireland (including the Isle of Man), while the Continental languages are all those that are or were spoken on the European continental mainland. This geographical distribution of the languages is only one of the justifications to which proponents of this classification. Insular Celtic also share verb-subject-object word order, inflected prepositions and some other sound shifts, particularly in relation to lenition of word-initial consonants.

Continental Celtic

- Gaulish
  - Lepontic
  - Noric
  - Galatian
- Celtiberian

Figure 2.3: The hierarchy of the Insular Celtic languages

2.3 Orthography of Irish

The orthography of Irish has greatly changed over its long history.

Originally, Primitive Irish was written in the Ogham script. Up until the middle of the 20th century, Irish was written down in the uncial Gaelic script, known in Irish as the cló Gaelach (Gaelach print). The uncial alphabet
CHAPTER 2. THE IRISH LANGUAGE

Insular Celtic

- Goidelic
  - Primitive Irish
  - Old Irish
  - Middle Irish
  - Irish (Modern)
  - Scottish Gaelic
  - Manx

- Brythonic
  - Pictish
  - British
    * Cumbric
      - Old Welsh
      - Middle Welsh
      - Welsh (Modern)
    * Southwestern Brythonic
      - Breton
      - Cornish

Figure 2.4: The hierarchy of the Continental Celtic languages

An acute accent known as the síneadh fada or fada shows a lengthening of a vowel (fada meaning ‘long’).

Figure 2.5: Oham-inscribed spindle-whorl from Orkney, Scotland

The Modern Irish alphabet is the same as that shown above, but instead of a dot over a letter showering lenition, a h follows the letter which is to be lenited. Again, the acute accent known as the síneadh fada or fada shows a lengthening of a vowel

Figure 2.6: Example of cló Gaelach script in a font by Michael Everson of Evertype
2.4 Syntax of Irish

2.4.1 Word order

Irish is a VSO language. That is to say, the normal word order in Irish is verb-subject-object. For example:

\[ \textit{D’ith an fear an ceapaire.} \] (The man ate the sandwich)

\[ D’ith \] is the verb, \textit{an fear} is the subject and \textit{an ceapaire} is object.

2.5 Morphology of Irish

Irish has a far richer morphology than the English language. English is a relatively uninflected language, when compared to other Indo-European languages. Irish is fairly typical for an Indo-European language.

Nouns in Irish are inflected for number and case. Nouns are classified as either of masculine or feminine grammatical gender.

There are four cases in Modern Irish:

\begin{itemize}
\item \textit{An Tuiseal Ainmneach} (Common case, usually known as nominative, but also plays the role of accusative)
\item \textit{An Tuiseal Gairmeach} (Vocative case)
\item \textit{An Tuiseal Ginteacht} (Genitive case)
\item \textit{An Tuiseal Tabharthach} (Dative case)
\end{itemize}
Verbs are inflected for person and number. There are 11 irregular verbs in Modern Irish. Verbs can be formed either synthetically (with a “pronoun as part of the verb”) or analytically (with a following pronoun). Irish has no infinitive, the verbal noun is used instead. When combined with ag, a progressive can be expressed which is roughly equal to the English present participle.

Like all of the modern Celtic languages, Irish is notable for its initial constant mutations. There are two mutations in modern Irish, séimhiú (lenition) and úrú (eclipsis). These generally serve to show the relationship between two words. The letters t and h also interpose in certain contexts (e.g. words beginning with vowels.

2.6 Standardization: An Caighdeán Oifigiúil

The word-forms which are prescribed in the dictionary are those forms which adhere to An Caighdeán Oifigiúil (The Official Standard) [Rannóg an Aistriúcháin 1958]. This document sets out a “neutral” standardized official variant of the language. Three dialects of Irish existed at the time, Munster, Connacht and Ulster. The Caighdeán Oifigiúil was largely based on the Connacht dialect, but included aspects of the Munster and Ulster dialects.

This top-down language-planning initiative had a very strong effect on the language, but was also deeply divisive, and indeed remains so to this
2.7  Legal standing of Irish

In the recent collection of essays about the Irish language, “A New View of the Irish Language”, [Nic Pháidín and Ó Cearnaigh 2008], there is an article entitled “Irish and the Legislative Perspective”, [Ó Laighin 2008]. In the article, Ó Laighin states that “legislative protection for the Irish language has never been stronger”. This is true on both national and international levels.

2.7.1  Irish perspective

The position of the Irish language within the modern state of Ireland is enshrined on several levels. The most important of these is almost certainly the constitution. It is to this document which we will look first. We shall then consider the impact of the Official Languages Act 2003.

Constitution of Ireland

Article 8 refers to the languages of the State:

8.1 The Irish language as the national language is the first official language.
8.2 The English language is recognised as a second official language.

8.3 Provision may, however, be made by law for the exclusive use of either of the said languages for any one or more official purposes, either throughout the State or in any part thereof.

Legal precedent and court rulings

The primacy of Irish has been held in several legal cases. In a court case in 1983, the late Justice Ó hAnnluain included the following in his ruling:

“It will be noted that Irish is accorded a higher level of recognition in the Constitution of Ireland than it had in the first Constitution, since it is referred to for the first time as ‘the first official language’. At the same time greater scope is given to the Oireachtas to give priority to one language over the other in accordance with the law insofar as relates to official matters in any part of the country. Until the Oireachtas exercises the function conferred on it by the provisions of the Constitution, it must always be assumed that Irish is the first official language, and that the citizen is entitled to require that it be used when the State has official matters to administer.”

The Supreme Court has equivocated somewhat in its attitude to Irish, but has in recent years asserted that
in practical terms the state, which is expressly committed to bilingualism, is required to facilitate the use of either language for official purposes, without discrimination against Irish-speakers.

[Ó Laighin 2008]

Official Languages Act

The Official Languages Act of 2003 is legislation which aims to ensure the delivery of state or other official services in the official languages of the state. In this respect, it is a coercive act which imposes duties on the state and public bodies. This can be contrasted with legislation which expresses a more abstract desire for linguistic equality or which aims to promote bilingualism through “opt-in” measures.

The full ramifications of the Official Languages Act were not realized by many until the eve of its introduction. The market for translation into Irish exploded as public bodies realized how much of their linguistic output needed to be bilingual, or at least available bilingually. As it is more fully implemented in the future, the influence of the Act will only increase.

2.7.2 European perspective

When Ireland joined the European Union, the government requested a special status for the Irish language which had hitherto not been applied to any other language of the Union. They wanted Ireland to be a “treaty lan-
guage” of the EU. By this, only the founding Treaties of the EU and the treaty of Ireland’s accession would be translated into Irish. Irish would not be an official or working language of the EU, and the vast quantities of EU legislation would not have to be translated into Irish. Those already in the Union were very much opposed to this “two-tier” linguistic system, where one language would have less status than the others. Nonetheless, the government succeeded in its mission, and Ireland joined the EU with the Irish language as a “treaty language”.

During Ireland’s presidency of the EU in 2004, when ten new countries joined the EU and nine new official languages were recognized, a grassroots movement encouraged the government to push for official language status. It was officially made a government goal in July of 2004, was approved by the European Council in June 2005 and took effect from the 1st of January 2007. As [Ó Laighin 2008] states:

This Article enshrines the fundamental principle of equality of these languages. From a linguistic perspective, it can be said that when Irish became an official and a working language, Ireland became a full member of the European Union.
Chapter 3

The Ó Dónaill dictionary

“If a word in the dictionary were misspelled, how would we know?”

Steven Wright

3.1 History

A great deal of information about the history and development of the Ó Dónaill dictionary is contained in the recent book by [Mac Amhlaigh 2008]. The book is written in Irish, so the following paragraphs include translations into English from that publication, by this author.

Work on what would become the Ó Dónaill dictionary began in 1959, on the 19th of June. This was the same year in which the English-Irish Dic-
tionary (EID) [de Bhaldráithe 1959] was published. The project to create a new bilingual Irish-English dictionary was commenced by the Department of Education (An Roinn Oideachais), under the editorship of Niall Ó Dónaill.

### 3.2 Niall Ó Dónaill

#### 3.2.1 Early life

Niall Ó Dónaill was born in Loch an Iúir (Loughanure) in county Donegal in Ireland in 1908. He was a native speaker of Irish. He attended secondary school at Coláiste Adhamhnáin in Litir Ceanainn (Letterkenny) in Donegal from 1921 until 1925, at which point he won a scholarship to University College Dublin. He attended UCD from 1925 until 1928, but left university without a degree because he missed the day of his degree exams. After this, he spent a year teaching in Tipperary. Marrying in 1929, he moved back to Dublin and spent 2 years translating books for An Gúm. An Gúm is the state publisher of works in Irish, including dictionaries, fiction, non-fiction, school-books and music. He worked for 3 years as an assistant to Father Mac Cionnaith, who was then working on his English-Irish dictionary.

Ó Dónaill was a civil servant from 1934 until the time he took his pension and retired from the Civil Service. He worked until 1951 in the Depart-
ment of Industry and Commerce as a clerk and then in the Department of Social Welfare. He was made manager of the social welfare office on Werburgh Street. From there, he moved to An Gúm in 1959. It was here that he worked on the dictionary for which he is famous. This will be dealt with in 3.3.

3.2.2 Later life and death

He was awarded an honorary doctorate for his work by Trinity College Dublin in June 1982. Niall Ó Dónaill died on the 10th of February 1995, in Clontarf, Dublin.

3.2.3 Name

We must address the issue of the editor’s name.

Before the lexicographical work for which he was best known, he worked for Oifig an tSoláthair which published many books in Irish. At this time, great emphasis was put on the translation of “classic” books into Irish. This led to the wonderful translations of such books as Dracula, Eachtraí Sherlock Holmes (The Adventures of Sherlock Holmes), Ben-Hur. One of the translators involved in this enterprise was a young Niall Ó Dónaill. He prepared a translation of a book called “Commando” [Reitz and Ó Domhnaill 1938], describing his experiences of the Boer War. In this book, written in the older Irish script cló gaelach, the translator is credited as Niall Ó Domnaill
(Niall Ó Domhnaill).

On the eponymous dictionary, his name appears as Niall Ó Dónaill.

Furthermore, to add to the confusion, in some sources, he is referred to as Néill Úi Dhónaill, even when referring to the eponymous dictionary.

3.3 Development

3.3.1 Objectives of the Foclóir Gaeilge-Béarla project

In December 1958, the position of Eagarthóir ar Fhoclóir Gaeilge-Béarla (Editor of the Irish-English Dictionary) was offered to Ó Dónaill by the Department of Education. He was told that the Department’s aim was to make available to students and the general public a dictionary to replace the Foclóir Uí Dhuinnín (Dinneen dictionary) which was becoming out of date, from the perspective of terms included, but especially from the point of view of the typeface used (the by-then-obsolete cló Gaelach) and the spellings of words.

3.3.2 Terms

The intention of this new dictionary was to explain the most common terms in the vocabulary of Nua-Ghaeilge (Modern Irish) in English and to give examples of usage. The terms appearing in the dictionary represent a mix of caint na Ghaeltachta (words used in the Gaeltacht Irish-speaking regions
of Ireland), terms taken from the literary sources written in the language and new terms arrived at (not to say invented) by the Coiste Tearmaíochta (Terminology Committee) to describe modern ideas and objects for which words did not exist in Irish.

The word-forms which are prescribed in the dictionary adhere to An Caighdeán Oifigiúil (The Official Standard) [Rannóg an Aistriúcháin 1958]. See 2.6 for more information.

3.3.3 Dictionary structure

The dictionary is 1309 pages long in its current form. This figure includes “front matter”, the term used by [Landau 2001] to describe that text which precedes the normal A-Z content of the dictionary. There is no “back matter”, the extremely short section for the letter Z appears on page 1309 (the shortness of this section should come as no surprise, since as we saw in 2.3, the Irish alphabet should not strictly speaking contain the letter Z).

As [Landau 2001] notes, the tendency for dictionaries to ship with a large amount of front or back matter (or indeed “middle matter”, in the case of X-Y, Y-X dictionaries) is a relatively recent phenomenon. The Ó Dónaill dictionary offers 8 pages of front matter, which is no more generous than the de Bhaldratthe dictionary of 18 years before. The actual information provided is, however, more useful, explaining to some extent the structure of the dictionary entries themselves.
3.4 Coverage

According to [Mac Amhlaigh 2008] there are 50,000 headwords. Indeed, this is the most common figure cited for the Ó Dónaill dictionary.

Working from the pre-press typeset data files, we find 59,734 headwords. This includes non-unique headwords, such as acaire\textsuperscript{1} and acaire\textsuperscript{2}. Removing these different senses, leaving just one headword per orthographic form, we find 48458 headwords.

As [Landau 2001] states on many occasions, the headword count for a dictionary, while often used as a metric of quantity (not to say quality) for advertising purposes, is not particularly meaningful. It gives an idea of the ball-park scale of the dictionary, but the fluidity of the concept of the headword is such that any assessment beyond that is impossible.

3.5 Publication

The dictionary was published on the 6th of July 1978, but the copyright date in the front matter refers to the year before, [Ó Dónaill 1977].

3.6 \textit{An Foclóir Póca}

The \textit{Foclóir Póca} (Pocket Dictionary) was derived from the Ó Dónaill dictionary, by reducing the number of headwords and selecting those which would be most commonly useful. This leads to it having some strange
omissions, as anyone who has used it can attest. Interestingly, this is one of the only Irish dictionaries to include a pronunciation guide for each of the Irish words in it. Another version of the dictionary, *An Foclóir Scoile* (The School Dictionary) is the standard dictionary for school-children throughout Ireland. This version is slightly larger and has a larger print-size.

### 3.7 Electronic version

For many years, despite the existence of an electronic version of the Ó Dónaill, it could not be accessed. The electronic version we have now is a copy of files from magnetic tape which stored the type-setting codes that were prepared by Smurfit printing in the 1980s. This would appear to be the original electronic form used to print the dictionary. [Ó Dónaill 2007]

The main files of the electronic version are in a type-setting code which seems to have been specifically designed for the creation of the Ó Dónaill dictionary. The codes and their meanings can be seen in Appendix B.

There is also a second version of the dictionary included, where all the formatting codes have been removed. This version is far more attractive because it resembles the paper dictionary much more closely in both form and layout.

At first glance, it would seem to make sense to use this cleaned-up version as the basis for any natural language-processing task. However, there is a lot of information contained in the tags of the type-set version of the
dictionary and there is nothing in the cleaned-up version which cannot be extrapolated from the type-set version, while the type-set version contains much that has been lost in the “cleaning” process.
“Words - so innocent and powerless as they are, as standing in a dictionary, how potent for good and evil they become in the hands of one who knows how to combine them.”

Nathaniel Hawthorne

4.1 Overview

The related projects which we will consider all relate to the parsing and transformation of printed dictionaries into some form of electronic form.

The first two projects we will look at involve dictionaries which have never existed in any electronic form. The first step in these projects is to get the printed word from the page to the screen.

The next two projects include earlier work on the Ó Dónaill dictionary,
from before the release of the electronic version. They relied on OCR techniques to capture the text of the dictionary, with somewhat limited success. In the light of the release of the pre-press files, these projects warrant reconsideration.

The last project we will look at is more similar to this, in that an electronic form of the dictionary was in existence, but needed reworking.

4.2 eDIL project

The Electronic Dictionary of Irish Language is a project of the Centre for Irish and Celtic Studies at the University of Ulster. The project was funded by the Arts and Humanities Research Council. [Toner et al. 2007]

The project aims to provide a digital edition of the complete contents of the Royal Irish Academy’s “Dictionary of the Irish Language: Based Mainly on Old and Middle Irish Materials”. To this end, the Royal Irish Academy allowed the digital capture of the full text of the Dictionary.

The eDIL project aims not only to capture the existing complete contents of the Dictionary, but to enrich the data and provide increased functionality and, in the future, to expand upon the dictionary itself.

The eDIL project is party to the Linking Dictionaries and Texts (LDT) project, which is a cross-border collaboration between the University of Ulster at Coleraine, where eDIL is based, and University College Cork where CELT is based. The aim is to reconnect dictionaries with the sources from
which they are drawn. In the case of the eDIL project, this means linking citations and references in the electronic text directly to CELT’s electronic copy of the text in question.

4.2.1 Dictionary of the Irish Language

While the de Bhaldraithe and Ó Dónaill together form the definitive dictionary of Modern Irish, the “Dictionary of the Irish Language: Based Mainly on Old and Middle Irish Materials” is the definitive dictionary of Old and Middle Irish [O’Donovan et al. 1976]. Among its users, which include basically every student of Old or Middle Irish, it is generally known as (the) DIL.

Overview

The dictionary project was begun in 1852 but was not completed until 1976, after an amazing 124 years of work. In this, as in other ways, the DIL is similar to the Oxford English dictionary. The approach taken, the scale and the scope of the dictionaries are comparable. The dictionary covers approximately a millennium, from circa 700 AD to 1700 AD. However, the most important and most covered periods are the Old Irish period from approximately 700 to 900 AD, and the Middle Irish period from 900 to 1200 AD. [Fomin and Toner 2006]
Irish Archaeological (and Celtic) Society

The idea for such a dictionary was first put forward in 1852 by the Irish Archaeological Society (founded in 1840). In 1853, this society merged with the Irish Celtic society to form the Irish Archaeological and Celtic Society, William Elliot Hudson being a key figure in both. This joint Society published 27 volumes of scholarly material on Irish history and antiquities, until its demise, when its publications were taken over by Sir John Gilbert. [Keenan 2006]

Work on the dictionary was begun by two of the most respected Irish scholars at the time, John O’Donovan and Eugene O’Curry. Sadly, both were dead within 10 years, with nothing even close to being published. However, O’Donovan had left behind an outline of his vision for the dictionary, which would be rigorously based in the older Irish manuscripts.

The meanings offered up for each word would be supported by citations from the manuscripts. Etymological analysis of entries would be undertaken only insofar as the derivation came from within the Irish language or from attribution of source to loan-words. These principles of punctilious citation and minimal etymological analysis were upheld throughout the dictionary’s long history. The former ensured the academic rigour of the dictionary while the latter probably ensured that the dictionary actually ever got finished.
CHAPTER 4. RELATED PROJECTS

Publication at last

From 1852, generations of scholars worked on the dictionary. It wasn’t until 1913 that the first fascicle of the dictionary, covering D–dégoir, was published by the Royal Irish Academy. At this stage, Carl Marstrander had taken over as editor of the dictionary. A period of 19 years passed before the next fascicle was published in 1932, covering the letter E. The RIA tried to speed up production by changing the format of the fascicles and changing the title to “Contributions to a Dictionary of the Irish Language”. Due to the fact that several other fascicles were already quite well-advanced, this did not have much of an effect. However, fascicles continued to appear regularly. The final fascicle appeared in 1976, the letter H, with E. G. Quin as editor. In fact, this last fascicle was only 1 page in length due to the Irish alphabet, as seen in 2.3, and the resultant paucity of terms beginning with the letter H.

Editions

The dictionary in its original form was 2525 pages long, comprising 23 fascicles for a total of about 35,000 entries. In 1983, the Academy published a Compact Edition of the dictionary, where groups of 4 pages were reduced so as to fit on one page, leading to a page-count of 640 pages, but requiring most people use a magnifying-glass to consult the dictionary.

Shortly after, the Academy came to the decision to discontinue printing
of the full-size edition, due to the difficulty of keeping all the fascicles in stock and the general awkwardness of the full-size edition.

Figure 4.1: Most recent edition of the Dictionary of the Irish Language, Compact Edition published in 2007 to coincide with the launch of the eDIL project

Since the launch of the eDIL website, the Academy has republished the Compact Edition of the dictionary in a new binding with improved reproduction of the text.

4.2.2 Digitization process

Text capture

The first stage of digitizing the Dictionary was the capture of the text. This was out-sourced to a company specializing in this field. The text was first scanned in and optical character recognition (OCR) software was used to convert the scanned images to text. At the same time, the text was “triple-keyed”. This process involves three typists each typing up a complete copy
of the dictionary. When complete, the four resulting copies of the dictionary are brought together and any conflicts resolved, leading to an error rate of less than one character in 10,000. This process only took 5 months, which is astonishing for 99.992% accuracy over more than 2200 pages.

**Formatting**

The formatting of the text in the Dictionary is quite consistent, in that bold text is always used for headwords, italics for examples etc. While the text was being scanned and keyed in, HTML (HyperText Markup Language) tags were used to represent its formatting. Also preserved were the line breaks and column breaks and the numbering of columns. This meant that the digital edition directly corresponds to the paper edition and vice versa. References with page number, column number and even line number can be used interchangeably between the two.

**HTML via XSLT to XML, pursuant to TEI**

The HTML tags used to represent the formatting were relatively easily converted into XML (eXtensible Markup Language) using XSLT (eXtensible Style Language Transformation). The resulting XML was compliant with TEI (Text Encoding Initiative) guidelines for printed dictionaries, albeit with some improvements and necessary adjustments. The authors go into great detail about their changes to the TEI DTD (Document Type Definition) in their paper [Fomin and Toner 2006] and there is no need to repeat
it here. Suffice to say that their improvements show up some of the weaknesses of the TEI guidelines.

**Interface**

With the data captured and in TEI-compliant XML format, the bulk of the work was complete. Since the resultant data is well-structured and in a well-known format, the creation of an interface to the digitized dictionary was not as onerous a task as it might have been. A web-site and search interface was developed in parallel with the data-acquisition task, and can be seen in action at the eDIL website. [Toner et al. 2007]

**4.2.3 Launch**

The electronic version of the Dictionary of the Irish Language, eDIL, was launched on the 27th of June, 2007 in the Royal Irish Academy in Dublin. It is now available for free use by anyone at [http://dil.ie/](http://dil.ie/) [Toner et al. 2007]

**4.2.4 Future work**

The eDIL team are continuing their work, actively revising the content of the DIL itself, continuing to correct errors as they are found, adding further missing parts-of-speech to the entries and otherwise improving the quality of the data.
Due to the wide-spread usage of the DIL and the length of time during which it has been in use in academia, individual scholars have amassed large collections of clarifications, amendments and additions which they have made to their personal copies of the DIL. Recognizing this, the eDIL requested the submission of as many of these extra addenda as possible and hope to include them in the process of enriching the dictionary.

4.3 Digital Dinneen

4.3.1 CELT

The Digital Dinneen project is under the stewardship of the CELT (Corpus of Electronic Texts) project at University College Cork. CELT is a corpus of Irish and multilingual texts, predominantly internet-based. It currently has over 11 million words online in more than 960 texts, and is still growing.

4.3.2 Why the Dinneen dictionary

Since CELT and the team behind the eDIL project are working in partnership, it was decided that in order to address a period of Irish which was not covered by the Dictionary of the Irish Language [O’Donovan et al. 1976], the Dinneen dictionary should be digitized.

[Nyhan 2008b] argues that the Dinneen dictionary is important because it provides a link between the DIL (and by extension eDIL) and the Irish
written today:

Our Digital Dinneen Research Project will complement eDIL and the Lexicon by adding Irish lexicographical sources from the modern period. These are outside the scope of DIL or eDIL, but form the link to Modern Literary Irish written today. Rev. Patrick Stephen Dinneen’s Irish-English Dictionary is the most authoritative scholarly dictionary of Modern Literary Irish. It holds a rich treasure of Irish words and explanations, but for someone who wants to know Irish equivalents of English words, and is starting from the English language, it is difficult to consult. The Dictionary, published in 1934, has seen many reprints, and is still a classic. This work is not available in electronic form, nor is there any other scholarly electronic Irish-English Dictionary for the period after 1650 available at the moment.

[Nyhan 2008b]

They go on to describe the technical features of the electronic version of the Dinneen dictionary. Again, the aim is to enrich an old dictionary to make it useful to several different types of user. The ability to choose between the older spelling system and typeface and the modern one is extremely useful, especially for more casual users. The cló Gaelach, while beautiful, has often been a barrier to the use of older materials by those who have learned the language since the earlier type-face disappeared. The old
spelling system is arguably more complicated than the system used since the introduction of the Caighdeán (see 2.6).

The superior search and interrogation facilities which have become standard in other areas of research can be delivered through an electronic edition. This edition will have a number of added benefits: Many people find the old spelling difficult to deal with—we will include the modern spelling for each headword where the spelling has changed—an advantage over the printed edition. Users will also have the choice of viewing the entries in Cló Gaelach or in Cló Rómhánach. The Digital Dinneen can be used like an English-Irish dictionary, i.e. you can also enter English search terms and bring up the Irish equivalents.

[Nyhan 2008b]

Since the Digital Dinneen project is also party to the Linking Dictionaries and Texts project and is under the stewardship of CELT, it should not be surprising that it will be tightly linked with some of the other projects undertaken there:

The Digital Dinneen will be XML-encoded for integrated use on the CELT website and linked to our Text Corpus using Javascript recently developed for the Lexicon. Via this Javascript, the user will be able to highlight Old and Middle Irish words, and, at the click of a mouse, view the corresponding Lexicon headwords,
with examples, translations and usage. Preparing the Digital Dinneen in the same way will make it an ideal tool for accessing and understanding Literary Modern Irish texts. It will complement the older language periods covered by Lexicon and eDIL, giving a diachronic overview of the development of Irish and its usage. To further broaden CELT’s Text Corpus we will also make available written sources for Modern Irish (17th to 20th century) used as sources by Dinneen.

[Nyhan 2008b]

Figure 4.2: Rev. Patrick Stephen Dinneen, author of the 1904 and 1927 Foclóir Gaedhilge agus Béarla (Dictionary of English and Irish), painted by John Butler Yeats
4.3.3 Dinneen dictionary

The Rev. Patrick Stephen Dinneen was the author of a number of publications in Irish but it is for his dictionary that he is remembered today. In fact, there were two dictionaries: [Dinneen 1904] and [Dinneen 1927]. Both were called *Foclóir Gaedhilge agus Béarla* (Dictionary of English and Irish).

The plates and stock of the first, the 1904 edition, were destroyed in the 1916 Easter Rising. This led Dinneen to start work on a new, larger, edition which was published in 1934.

In some respects, Dinneen’s dictionary is superior to the Ó Dónaill dictionary which followed it, 42 years later. It provides much richer explanations of many words and is generally more literary in nature. It is a more academic dictionary than the Ó Dónaill dictionary which was designed more as a desk or reference dictionary for the general public. It is for these reasons and others that a digitization of the Dinneen is desirable. [Nyhan 2008a]

4.4 Computational Lexicography

In her thesis, [Magee 2000], Caroline Magee considered the task of parsing an OCR’d copy of the Ó Dónaill dictionary. As always with structured or formatted data, she had problems with the accuracy of the OCR. Often accents can cause serious problems for mainstream OCR software, necessitating expensive custom solutions. For the cló Gaelach, this problem is
particularly severe.

Magee considers the definition of a dictionary from a theoretical standpoint and then describes the development of the modern English dictionary. From the art and/or science of lexicography, she moves on to the rapidly expanding field of computational lexicography.

Finally, she describes her approach to the improvement of the OCR accuracy for the dictionary and the limited parsing of the dictionary which is possible with this data.

4.5 Web User Interface to Irish Dictionary Database

In her thesis, [Clear 2002], Elizabeth Clear expands on the work of Magee (above). Her idea was to create a web interface to the lexicographical data which had been created by Magee and others from the OCR’d texts. With this web interface, correcting and eventually updating the lexicographical database is made much easier.

Even with the full pre-press type-setting files for the dictionary, some sort of interface to make updating of the data easier would be very welcome. There are mistakes in the pre-press files, not to mention many inconsistencies and discrepancies. In the same way as the eDIL team are imposing a stricter order on the Dictionary of the Irish Language, the more wayward parts of the Ó Dónaill dictionary (such as most of the letter C entries!) could be brought to heel.
4.6 An Analyser and Generator for Irish Inflectional Morphology using Finite-State Transducers

In her thesis, [Úi Dhonnchadha 2002], Elaine Úi Dhonnchadha provides a comprehensive description of morphology, of the Irish language and of the overlap between the two. Irish morphology, as stated before is extremely rich and lends itself well to a finite-state morphological approach.

Úi Dhonnchadha used a series of Perl scripts to parse the *Foclóir Póca* to extract the word-forms therein and the information that the dictionary contained about them. She then created an extremely comprehensive morphological model of Irish using finite state automata techniques. Using the information from the *Foclóir Póca*, she filled out the word classes for which she had created morphological models. These she then evaluated against a list of the most common words in a corpus of Modern Irish, with very encouraging results.

She has continued to develop the morphological model and part-of-speech tagger which resulted from this work and it remains the state of the art in the field of Irish computational linguistics.
Chapter 5

Project planning

“Always remember that striving and struggle precede success, even in the dictionary.”
Sarah Ban Breathnach

5.1 Programming language

The choice of a programming language for any project of a greater than trivial size is not of insignificant importance. Certain programming languages are more suited to certain tasks than others. This should be taken into account if the luxury of choosing the tool for the job is available.
5.1.1 Perl

Perl might seem to be the “natural” choice of programming language for a project where large-scale processing of text will be required, particularly involving the use of regular expressions.

Perl’s integration of regular expression support is second to none, to the extent that regular expression operations are part of the core language and are represented with syntactic operators which allow very succinct comparison or substitution operations.

Like Python and Java, Perl is an interpreted language. Interpreters for Perl probably exist for an even larger number of platforms than do for the other two languages. Thus, portability is not a problem.

Unicode support is quite good in Perl, but as with many things in Perl, it is more complicated than it needs to be, for most purposes at least. This is shown up when you compare the Unicode support in Perl to that in Python.

However, Perl seems to encourage a certain undesirable style of programming and software development, where Python encourages neatness and careful modularity. In some senses, Perl could be seen as an older, slightly lower-level Python.
5.1.2 Python

Python is a highly-popular language at the moment. Its clean and simple syntax combined with a stable and fast virtual machine and, most importantly, an extremely comprehensive and well-structure standard library have attracted many fans.

Unlike Perl, regular expression support is not built into the syntax of the language, but it is easily accessible in the standard library and works without problems.

Unicode causes no problems as long as the input and output streams are correctly encoded.

Programs in Python are generally succinct, clear and easy to debug due to the lack of references (Perl) and pointers (C/C++).

5.1.3 Prolog

Prolog’s excellent support for grammars and parsing cannot be disputed. However, the dictionary’s lack of a completely regular structure meant that this might not be the best match.

Despite the best efforts of the SWI-Prolog project, the Prolog ecosystem remains idiosyncratic at best.

Unicode support in SWI-Prolog is certainly better than in other Prolog implementations, but it is still far from “natural”.
5.1.4 C/C++

For sheer speed, well-written C code can generally only be beaten by hand-coded assembler. C++ code is generally marginally slower than C, but could still be an order of magnitude faster than any of the other languages being considered here. This speed advantage constitutes the main advantage to C/C++.

Disadvantages include the fact that a program in C/C++ for a task such as this would be much more complicated than necessary, and certainly much longer than a program written in one of the scripting languages (Python, Perl). This complication makes C/C++ potentially more error-prone than some of the other languages here.

Furthermore, a compiled C/C++ program is not compatible with different platforms, and there can be portability problems even if recompilation is a possibility.

Finally, regular expression support is not very good, certainly nowhere near as well-done as in Python or Perl. Unicode support is far from “natural” to C, which uses 8-bit ASCII strings as standard. In C++, the wstring type is more usable, but the fact that it is not the basic string type is somewhat bothersome.
5.1.5 Java

For the purposes of this survey, Java is roughly comparable to C/C++. Java can be many times slower than C/C++, but can be run (nearly) anywhere there is a Java interpreter available.

Programs written in Java have a tendency to be incredibly verbose as edifices of objects are built to represent the problem-set in an object-oriented manner. This increases the architectural burden for the developer. However, unlike with C/C++, it is quite difficult to write broken code. This may have something to do with the fact that each individual line in a Java program does almost nothing!

One advantage of Java over all the other options here lies in its Unicode support. Java has Unicode support built in from the ground up, whereas even Python has separate “normal” strings and Unicode strings.

5.1.6 Decision

It was decided to use Python because it appeared to offer the best set of features with the fewest disadvantages. A few years ago, Perl would have been the only reasonable choice for a program like this, and it is a measure of how far Python has come, and how little Perl has moved with the times, that Python is more suitable now.
5.2 Database

5.2.1 XML

Seeing the example of the eDIL project (see [Toner et al. 2007]), an XML data storage format was considered. However, it was decided that XML, preferably in a TEI-compliant format, should be considered as an end-product output and not as a data-store. A more traditional relational database server is much easier to work with while the project is in development. Once the data has been successfully captured and structured, exporting it to an XML format, e.g. TEI for printed dictionaries, is relatively trivial.

5.2.2 Oracle

For a database management system (DBMS), Oracle has an unnecessarily complex install process and an extremely large installer. More than one gigabyte of files should not be necessary for a database management system, even an “enterprise” one.

Furthermore, the fact that Oracle is commercial while the other two candidate DBMS packages are (relatively, in the case of MySQL) open-source and free is a mark against it.
5.2.3 MySQL

MySQL claims on its website [MySQL AB 2008] to be “The world’s most popular open source database”. Certainly, it provides the database back-end for the vast majority of web applications, particularly in the small- to medium-sized market. MySQL is the M in the LAMP software solution stack (Linux Apache MySQL PHP), and is supported by pretty much every application as a database back-end. However, it is not as robust as PostgreSQL or Oracle. In order to get reasonable speed for database operations, very careful fine-tuning of the server is required and a special database type must be used.

5.2.4 PostgreSQL

By contrast to MySQL, PostgreSQL claims to be “The world’s most advanced open source database” on its website, [PostgreSQL Global Development Group 2008]. It supports more of the SQL standard than MySQL, and approaches Oracle in its level of support. PostgreSQL is also catching up on the industry-standard enterprise Oracle in terms of speed, according to recent surveys.

5.2.5 Decision

Since we ruled out an XML database approach, preferring to see that as a goal for outputting further down the road, the choice between one be-
between the three main (semi-)SQL-compliant relational DBMs. Oracle was eliminated mainly for being so large and difficult to install. Both MySQL and PostgreSQL have decent support in Python. In the end, PostgreSQL was chosen because it provided the best balance of robustness with ease of install and support.

5.3 Server

A server in a web-hosting centre was used to provide uninterrupted hosting for the project. The project was given the sub-domain fyp of another domain to isolate the relevant server software and to make configuration easier. The open-source software Apache was used as the web-server.

5.3.1 Subversion

Subversion was used to provide version-control for the source code. This meant there was less need to back up, and no need to make multiple copies of a file before a potentially damaging change. All files used in the project were placed under Subversion version-control so that previous version could be consulted or returned to if need be.
5.3.2 Trac

Trac is a web front-end to Subversion which provides several useful tools for accessing a Subversion version-controlled source repository. It allows you to browse the source code, seeing changes and the comments made.

![Trac Interface](image)

Figure 5.1: An example of Trac’s timeline view of source changes

5.3.3 phpPgAdmin

phpPgAdmin is a web application written in PHP which acts as a frontend to a PostgreSQL database management system. It is effectively a clone of PhpMyAdmin which does the same thing for MySQL. phpPgAdmin allows the administration of the database server, creation and deletion of databases and tables, user administration and the running of SQL queries.
Figure 5.2: A demonstration of Trac’s source code browser
Chapter 6

Project implementation

“I know of no more enjoyable intellectual activity than working on a dictionary... Every day one is confronted by new problems, usually small, but absorbingly interesting; at the end of the day one feels healthily tired, but content in the thought that one has accomplished something and advanced the whole work towards its completion.” James R. Hulbert

6.1 Materials

The copy of the electronic version received was 50 megabytes in size. It contained quite a number of files, of which the relevant ones are listed below:

fgb_lip_.doc A copy of the label which would have been affixed to the CD
Two copies of the same file describing certain aspects of the electronic form of the dictionary

A more extensive version of the above file

24 files (no K or W) containing a plain text, unformatted copy of the text of the dictionary

A 6.5 megabyte file which is nothing more than the above files concatenated together

A file, included in full in Appendix B, which describes the formatting mark-up codes used in the next files

24 files (no K or W) containing a plain text pre-press copy of the dictionary with mark-up symbols

**6.2 Initial experiments**

In order to get a feel for the data, and the challenges involved, several simple programs were written to explore the sort of things that could be done. These also served to develop techniques in the Python programming language for the parsing of the data. Eventually, most of these small programs ended up being used for the main extraction process, as all the requisite functionality had already been written in the form of these little tests.
6.3 Extraction of entries

6.3.1 Simple splitting of entries

This code fragment separates the text into the entries that make up the dictionary. Each entry begins with a headword and is separated from the next entry by at least one blank line (sometimes 2 or 3).

Throughout the typesetX.txt files there are approximately 5 places where this does not hold, and an entry is apparently accidentally split by the insertion of blank lines. In these cases, the extraneous blank lines were simply edited out of the source files, since they were clearly incorrect in the context of the formatting around them and since the burden of developing some sort of heuristic to guess if this had occurred in a given case was too heavy given that there were so few cases where the problem occurred.

The code reads each line from the file and adds each to an entry. An entry is complete when a blank line is reached.

Code Listing 6.1: Contiguous blocks separated by blank lines to separate entries

```python
entry = ""
while 1:
    line = f.readline()
    if line == "": break
    if not line.strip():
        raw_input("Entry complete:")
        print entry
        entry = ""
    else:
        entry += line
```
6.3.2  Headword extraction

A headword begins each entry. The headword is separated from the rest of the entry by a comma, so we use a RegEx `split` to get just the part we need.

```python
entry = ""
while 1:
    line = f.readline()
    if line == "": break
    if not line.strip():
        lines = entry.splitlines()
        if len(lines) > 0:
            headword, rest = re.split('[, ]',lines[0],1)
            if headword == ": break
            rest = re.sub('\s\s', '',rest);
            print headword
            entry = ""
            #raw_input('Entry complete:')
        else:
            entry += line
```

6.3.3  "Equality recognition" and extraction

The Ó Dónaill dictionary has a great number of what we can call “equalities”. They show an equivalence of some degree between two words. Often, they are used to show the standardized form of pre-Caighdeán terms. They are entries of the form:
ABAIGH*X1*Y = *SAIBIGH.

ABAIGH*X2*Y*K: *SABACH.

Code Listing 6.3: Recognizing and extracting “equalities”

```python
headword, rest = re.split('[],',lines[0],1)
if headword == ':': break
rest = re.sub('\s', '', rest);
if re.match('=\',rest):
    # what follows is a reference
    print headword + '___' + rest
```

6.3.4 Text normalization

Since all the text in the typeset files is encoded using the typesetting tags, we must interpret them. First, we make everything lower-case. Then we apply the [ tag(s), which makes the letter that follows it a capital. Then we apply the $ tag(s) which apply an acute accent to the following vowel. This is achieved using Unicode entity codes.

Code Listing 6.4: Function to normalize text, interpreting typesetting tags

```python
def normalize_text(text=""):    text = text.lower()
    while text.find('[') > -1:
        index = text.find('[')
```
if index > 0:
    before = text[0:(index)]
else:
    before = ""

after = text[(index+1):]

capped = after[0].upper() + after[1:]
text = before + capped

# fadas for lower-case
text = text.replace("$a", u'\u00e1')
text = text.replace("$e", u'\u00ef')
text = text.replace("$i", u'\u00ed')
text = text.replace("$o", u'\u00f3')
text = text.replace("$u", u'\u00fa')

# fadas for upper-case
text = text.replace("$A", u'\u00c1')
text = text.replace("$E", u'\u00c9')
text = text.replace("$I", u'\u00cd')
text = text.replace("$O", u'\u00d3')
text = text.replace("$U", u'\u00da')

return text

6.3.5 Extraction of parts of speech and grammar explanations

Once the headword has been extracted, we try to extract the part-of-speech tag which follows it. Beyond it, there usually appears a small grammatical explanation, which describes how the plural and/or the different cases are formed. These are each marked with helpful type-setting codes ("G and
*R), so we search for these. The idea of “eating” the string as we go along continues here here.

Code Listing 6.5: To extract parts of speech and grammar

```python
lines[0] = re.sub('`\s+', '', lines[0]);
try:
    headword, rest = re.split(' [ ]', lines[0], 1)
except ValueError:
    print "error with headword"
# chew space off the front
rest = re.sub('`\s+', '', rest);

headword = normalizeText(headword)
if re.match('/\*G', rest): # part of speech
    rest = rest[2:] # bite off the "\*G"
    pos, rest = re.split(' [ ]', rest, 1)

if re.match('/\*R(\)', rest): # grammatical speech
    rest = rest[2:] # bite off the "\*R"
    m = re.match('/\(.*\)', rest)
    # extract the match and chomp first and last
    # characters which will be brackets
    if m:
        grammar = m.group()[1:-1]

out.write(headword + " : " + pos + " : " + grammar + "\n")
```

6.3.6 Extraction of super-script markers from headwords

If the same word must appear as several different headwords because it can assume several parts-of-speech, these are distinguished using a small superscript after the word. In the type-setting coding, this is shown as *X1*Y, for a superscript number 1, for example. The code below searches
for this tell-tale pattern and extracts the number from the super-script. It then removes this coded super-script from the headword and normalizes the word.

Code Listing 6.6: Extraction of super-script markers from headwords

```python
m = re.search(r'\*X(.*)\*Y', headword)
if m:
    superscript = m.group(1)
    headword = headword[0:m.start(0)]
    headword = normalize_text(headword)
else:
    superscript = ""
    headword = normalize_text(headword)
```

6.4 Initial database structure

The initial design for the database structure was planned at the same time as the initial experimentation with the data set. The design was based largely on observation of the printed dictionary, which later turned out to be somewhat misleading. The SQL statements used to create the structures for the database can be seen in Appendix C.1.

6.4.1 List of normalised relations

In the following list of normalised relations, an underline indicates a primary key, while italic indicates a foreign key. These are implicit structural integrity constraints on the database.
Figure 6.1: Determinacy diagram for initial database structure
• headword (headword_id, word, part_of_speech, gender, grammar)

• sense (headword_id, sense_id)

• entry (headword_id, sense_id, entry_id, type, collocation_id, crossrefereGRANTing
privilegesnce_id, gloss_id)

• collocation (collocation_id, irish, english)

• crossreference (crossreference_id, to_sense, from_sense)

• gloss (gloss_id, text)

6.4.2 Notes on design

Because of the inherently hierarchical nature of a dictionary, the headword, sense and entry tables “stack” together. On the other hand, the collocation, crossreference and gloss are stored in vertical tables across the whole database. This make sense because they represent “solid” information, and this structure could eliminate redundancy and will allow for easier access to all data-types of a particular kind. An example of where this would be useful is in the creation of a parallel corpus for English and Irish, where the parallel texts are drawn from the collocations.

6.4.3 Problems with initial database structure

The original structure of the database was planned before much analysis of the data had taken place. It had a relatively complex structure, with a
somewhat object-oriented aesthetic. Implementing a pseudo-object-oriented structure within the confines of a relational database model is possible, but difficult.

A headword contained information about the headword as well as 1 or more senses. Each sense consisted of 1 or more “entries”, where an entry can be a gloss, translation, example etc. This multi-functional data-type was implemented using polymorphism, of a kind.

After parsing a number of dictionary entries by hand, it became clear that this rigid hierarchical structure was not flexible enough to handle the less-than systematic structure of entries in the Ó Dónaill dictionary.

6.5 Revised database structure

The new database structure is flatter. The hierarchical, object-oriented model has been left aside.

The many different “data-types” have been replaced with just two tables. This allows the data-base to represent the almost free-form structure of the entries in the Ó Dónaill dictionary.

In the following list of normalised relations, an underline indicates a primary key, while italic indicates a foreign key. These are implicit structural integrity constraints on the database.

The various references to “superscript” are necessary because as well as including multiple senses under a single head-word, in cases where the
same word can be two or more different parts-of-speech, it is split into several head-words, which are differentiated with a small superscript number. Where cross-references or “equality” references are used in the dictionary, these superscripts are replicated in order to distinguish which of the identically-named headwords is being referred to.

The SQL statements used to create the structures for the database can be seen in Appendix C.2. In the following list of normalised relations, an underline indicates a primary key, while italic indicates a foreign key. These are implicit structural integrity constraints on the database.

- headword (headword_id, word, superscript, part_of_speech, gender, grammar, equalsref, equalsref_superscript, colonref, colonref_superscript)

- definition (headword_id, sense_id, definition_id, category, irish, english, crossref, crossref_superscript)
Chapter 7

Conclusion

“Dictionaries are like watches; the worst is better than none,
and the best cannot be expected to go quite true.”
Samuel Johnson

7.1 Conclusion

In this work, the Ó Dónaill Irish-English dictionary is considered as a potential gold-mine for lexicographical data about Modern Irish. At present, it is unique as the only full-size Modern Irish dictionary to even approach the Caighdeán standard.

As such, if its words and definitions could be unpicked, the impact on the automatic processing of Irish could be immense. To this end, we take the recently-released electronic version of the dictionary to side-step the
previous stumbling-block of scanning the dictionary.

By looking at other similar projects, we can see some of the same problems cropping up again and again. Hopefully now that we are aware of the common problems such as scanning, and the shared solutions such as triple-keying, we can expedite individual projects by working together.

We found that the problem of maintaining consistency across a dictionary does not end with the lexicographer or even the reader. Those who wish to put the dictionary to yet another new use must face up to the problems of the past in order to create the linguistic tools for the future.

### 7.2 Future work

The approach taken by the eDIL team in triple-keying and scanning their chosen target volume shows us the way forward. With almost perfect accuracy compared to the original volume, they can at last ignore the issue of OCR inaccuracy. This process could be applied to several of the other important Irish dictionaries, in particular the Mac Cionnaith dictionary.

Daniel Foley’s dictionary of 1855 has been scanned by an American organization and released into the public domain, along with many other

Applying similar finite-state morphological techniques as used in [Uí Dhonnchadha 2002] to a suitably-parsed and re-structured Ó Dónaill dictionary would lead to a great improvement in the quality of part-of-speech tagging for Irish. Indeed, this task is already under way.
Using the techniques demonstrated in this thesis and those used by the similar projects which are discussed herein, hopefully we can make quicker progress toward machine-readable lexicographical database of all the major Irish dictionaries ever created.

If we had access to the many dictionaries made over the past 400 years, in similar and compatible formats, many areas of linguistic research would be made easier. Historical linguists could follow a word through the centuries, etymologists could locate hitherto-unknown derivations or lexicographers could have instant access to all the definitions that their predecessors have ever given to a word.

The area of Irish computational linguistic research is still very young and growing fast, but approaches to lexicographical resources such as these could help it make great leaps forward to the state of the art globally.
Bibliography


Irish Materials, volume 1-23. Royal Irish Academy, 19 Dawson Street, Dublin, Ireland.


Appendix A

Abbreviations used

ASCII  American Standard Code for Information Interchange

CELT  Corpus of Electronic Texts

DBMS  Database management system

DIL  Dictionary of the Irish Language: Based Mainly on Old and Middle Irish Materials [O’Donovan et al. 1976]

DTD  Document Type Definition

eDIL  Electronic Dictionary of the Irish Language [Toner et al. 2007]

EID  English-Irish Dictionary, de Bhaldraithe dictionary [de Bhaldraithe 1959]

FGB  *Foclóir Gaeilge-Béarla* (Irish-English Dictionary), Ó Dónaill dictionary [Ó Dónaill 1977]
HTML  HyperText Markup Language

LAMP  Linux Apache MySQL PHP (or Python, Perl, Ruby etc.)

OCR  Optical character recognition

RegEx  Regular expression

RIA  Royal Irish Academy

SQL  Structured Query Language

TEI  Text Encoding Initiative

UCC  University College Cork

XML  eXtensible Markup Language

XSLT  eXtensible Style Language Transformation
Appendix B

Mark-up legend

‘Following is the content of the file legend.txt supplied with the electronic form of the Ó Dónaill dictionary, which goes some way to explaining the rather arcane formatting codes used in the mark-up of the type-set files.

This legend applies to the typesetting files called TYPESETA.TXT thru TYPESETZ.TXT.

[  = The next letter is capitalized (count = 89086).
$  = The next letter (vowel) has an acute acent ($A = a-fada) (count 76907).
*  = Indent (count 291)
*A  = See also (count = 1961).
*B  = Irish bold type (count = 85635).
*C  = Classification, e.g. ”Gym.” (Gymnastics) (count = 7368).
APPENDIX B. MARK-UP LEGEND

*D = Not used (count = 0).
*E = English in italics (count = 182).
*F = Not used (count = 0).
*G = Grammar (count = 51535).
*H = Hyphen or tilde (count = 701).
*I = Irish italics (count = 7993).
*J = Not used (count = 0).
*K = Colon (count = 3923).
*L = Letter sub-reference, e.g. "(a)" (count 6773).
*M = Not used (count = 0).
*N = Numeral reference, e.g. "1." (count = 2538).
*O = Of -- (count 1784).
*P = Period (count = 960).
*Q = Not used (count = 0).
*R = English in roman font (count = 102647).
*S = Referenced Irish word (count = 22649).
*T = Not used (count = 0).
*U = Not used (count = 0).
*V = Variant (count = 5516).
*W = Not used (count = 0).
*X = Superscript marker (18029).
*Y = Return from superscript (18029).
*Z = Not used (count = 0).
APPENDIX B. MARK-UP LEGEND

fl = blank space
Appendix C

Database structures

C.1 Initial database structure

```
CREATE TABLE headword
(
    headword_id INTEGER NOT NULL UNIQUE,
    word CHARACTER(255) NOT NULL,
    part_of_speech CHARACTER(15) NOT NULL,
    gender CHARACTER(15),
    grammar CHARACTER(255) NOT NULL,

    PRIMARY KEY (headword_id)
);
```
CREATE TABLE sense
(
    headword_id INTEGER NOT NULL,
    sense_id INTEGER NOT NULL UNIQUE,
    PRIMARY KEY (headword_id, sense_id),
    FOREIGN KEY (headword_id)
    REFERENCES headword(headword_id)
);

CREATE TABLE collocation
(
    collocation_id INTEGER NOT NULL UNIQUE,
    irish CHARACTER(255),
    english CHARACTER(255),
    PRIMARY KEY (collocation_id)
);

CREATE TABLE crossreference
(
    crossreference_id INTEGER NOT NULL UNIQUE,
    to_sense INTEGER NOT NULL,
    from_sense INTEGER NOT NULL,
APPENDIX C. DATABASE STRUCTURES

```
CREATE TABLE gloss
(
  gloss_id INTEGER NOT NULL UNIQUE,
  text CHARACTER(255) NOT NULL,
  PRIMARY KEY (gloss_id)
);

CREATE TABLE entry
(
  headword_id INTEGER NOT NULL,
  sense_id INTEGER NOT NULL,
  entry_id INTEGER NOT NULL,
  type CHARACTER(15) NOT NULL,
  collocation_id INTEGER,
  crossreference_id INTEGER,
  gloss_id INTEGER UNIQUE,
  PRIMARY KEY (headword_id),
  FOREIGN KEY (headword_id) REFERENCES sense(sense_id),
  FOREIGN KEY (crossreference_id) REFERENCES sense(sense_id)
);
```

APPENDIX C. DATABASE STRUCTURES

CHECK (type IN ('collocation', 'crossreference', 'gloss')),

PRIMARY KEY (headword_id, sense_id, entry_id),
FOREIGN KEY (headword_id)
  REFERENCES headword(headword_id),
FOREIGN KEY (sense_id)
  REFERENCES sense(sense_id),
FOREIGN KEY (collocation_id)
  REFERENCES collocation(collocation_id),
FOREIGN KEY (crossreference_id)
  REFERENCES crossreference(crossreference_id),
FOREIGN KEY (gloss_id)
  REFERENCES gloss(gloss_id)
);

C.2 Revised database structure

CREATE TABLE headword
(
  headword_id INTEGER NOT NULL UNIQUE,
  word CHARACTER(255) NOT NULL,
  superscript INTEGER,
  part_of_speech CHARACTER(15) NOT NULL,
APPENDIX C. DATABASE STRUCTURES

    gender CHARACTER(15),
    grammar CHARACTER(255),
equalsref CHARACTER(255),
equalsref_superscript INTEGER,
    colonref CHARACTER(255),
colonref_superscript INTEGER,

    PRIMARY KEY (headword_id)
    );

CREATE TABLE definition
    (  
        headword_id INTEGER NOT NULL,
        sense_id INTEGER NOT NULL,
    definition_id INTEGER NOT NULL,
category CHARACTER(15),
    english CHARACTER(255),
    irish CHARACTER(255),
crossref CHARACTER(255),
crossref_superscript INTEGER,

    PRIMARY KEY (headword_id, sense_id, definition_id),
    FOREIGN KEY (headword_id)
REFERENCES headword(headword_id),
);
## List of Figures

2.1 The hierarchy of the P-Celtic languages ........................................... 7
2.2 The hierarchy of the Q-Celtic languages ........................................... 7
2.3 The hierarchy of the Insular Celtic languages ...................................... 8
2.4 The hierarchy of the Continental Celtic languages ................................. 9
2.5 Oham-inscribed spindle-whorl from Orkney, Scotland ............................ 10
2.6 Example of cló Gaelach script in a font by Michael Everson of Evertype ................................................................. 10

4.1 Most recent edition of the Dictionary of the Irish Language, Compact Edition published in 2007 to coincide with the launch of the eDIL project ................................................................. 30
4.2 Rev. Patrick Stephen Dinneen, author of the 1904 and 1927 *Foclóir Gaedhilge agus Béarla* (Dictionary of English and Irish), painted by John Butler Yeats ................................................................. 36

5.1 An example of Trac’s timeline view of source changes ............................ 48
5.2 A demonstration of Trac’s source code browser ..................................... 49
6.1 Determinancy diagram for initial database structure . . . . 58
Code Listings

1.1 Example code snippet ........................................... 2
6.1 Contiguous blocks separated by blank lines to separate entries 52
6.2 Extracting headwords from entries .......................... 53
6.3 Recognizing and extracting “equalities” .................... 54
6.4 Function to normalize text, interpreting typesetting tags ... 54
6.5 To extract parts of speech and grammar .................... 56
6.6 Extraction of super-script markers from headwords ........ 57