Automatic thesaurus generation for minority languages: an Irish example

Kevin P. Scannell
Department of Mathematics and Computer Science
Saint Louis University
St. Louis, Missouri, USA
scannell@slu.edu

Mots-clefs – Keywords

Génération automatique de thésaurus, Irlandais
Automatic thesaurus generation, Irish language

Résumé - Abstract

Nous présentons des techniques pour la génération automatique d'un thésaurus irlandais monolingue. Ces résultats ont été réalisés en dépit des ressources limitées et, comme le plupart des autres langues minoritaires, de l'absence d'outils pour le traitement de langage naturel.

Techniques are presented for the automatic construction of a monolingual Irish language thesaurus. Our results were obtained despite limited resources, including, as is the case for most other minority languages, the lack of sophisticated software tools for natural language processing.

1 Project Description

The goal of this project, taken broadly, is to provide a full suite of Irish language software tools, on par in quality with what is available in English, for everyday use by speakers of Irish. The portion of this work described in the present paper may be of some interest to researchers in computational linguistics, since some of the software that I have developed may be useful in broader contexts and is possibly portable to minority languages other than Irish. I will emphasize the practical versus the theoretical in what follows; such an emphasis is especially important in light of the precarious position of Irish as a spoken language. Given the constant pressure from English (particularly in technical domains) I believe it is essential to focus on producing software that delivers some *immediate benefit* to Irish speakers. The hope, of course, is that providing high-quality Irish software will strengthen the language by reducing (by one) the number of domains in which one is forced to use English. From a sociolinguistic perspective, the technical sphere represents a key battleground in the fight to halt or reverse language shift, particularly in light of Ireland's swiftly developing reliance on technology and the common negative associations of the language with a (real or imagined) backward, rural past.

More specifically, this paper will focus on the development of a hypertext, monolingual Irish thesaurus. In §2-§4 I will provide a detailed description of how the thesaurus was generated, with the hope that the overall process (or indeed some of the specific tools) might be applicable to other minority languages. If nothing else, it should serve as a case study of what can be achieved in this area with severely limited resources.

1.1 Thesauri and automatic thesaurus generation

Roget's English language thesaurus, first published in 1852, is the exemplar of what we will call a *classical thesaurus*: a print or electronic database of quasi-synonyms used most often by writers who are looking for a broad choice of potential synonyms to fit a given context. The basic structure of classical thesauri has remained essentially unchanged over the years; we expect Roget would easily recognize the kernel of his handiwork in the latest editions, despite their abandonment of his original classification scheme for a more convenient alphabetical arrangement, e.g. (Laird, 1999). Classical thesauri usually offer broad coverage of the lexicon and are potentially quite useful tools for the preservation of the rich linguistic heritage of endangered languages like Irish. People with a limited command of the language (as acquired, say, in the national schools) are able to use a thesaurus to expand their vocabulary and improve their writing.

It is convenient to distinguish classical thesauri from *electronic thesauri* in the modern sense: software components used in many document retrieval or indexing systems, usually for the selection of a preferred form of a given search term. The underlying data in classical and electronic thesauri are quite similar (raw lists of terms organized according to some kind of semantic hierarchy) and our goal in this project is to generate a common database of semantic relationships in Irish from which, initially, a classical thesaurus can be generated, but with the flexibility that in the future more sophisticated information retrieval tools can be developed.

There is a rich literature covering techniques for automatic thesaurus generation, but most of the work has been restricted to global languages. The best references for the elements of thesaurus construction are (Aitchison & Gilchrist, 1987), (Grefenstette, 1994), and the ANSI/NISO standard *Guidelines for the Construction, Format, and Management of Monolingual Thesauri* (ANSI/NISO, 1993). Typical systems parse a large corpus and apply some form of cluster analysis either to measurements of similarity in grammatical context or to raw counts of co-occurrence. In particular, all approaches of which we are aware rely on a sophisticated pre-existing NLP infrastructure (large corpora, parsing tools, etc.), taken for granted in languages like English but not available in Irish or most other minority languages.

If the ultimate goal of automatic thesaurus construction is the deduction of semantic relationships exclusively from free text corpora, systems may be viewed as more or less technically remarkable as their underlying corpora vary from free to highly-structured. According to this measure, our approach is decidedly unremarkable, as the main idea is to exploit existing English language thesauri to deduce the desired semantic relationships in Irish.

1.2 Survey of available resources

What I hope is inspiring about this case study is the fact that the end results have been achieved with virtually no financial resources, no pre-existing software infrastructure, and a limited time commitment[1]. As will become clear in a moment, though, any such inspiration must be tempered by the fact that Irish, compared with other minority languages, enjoys an embarrassment of lexicographic riches in machine-readable form. The approach I describe in §2-§4 may therefore not be feasible for the most severely marginalized languages.

A broad survey of Irish language resources on the Internet can be found at the *Gaeilge ar an Ghréasán* site maintained at Sabhal Mór Ostaig[2]. Of special interest are several online newspapers either entirely in Irish[3] or devoting special sections to Irish language articles[4]. Highly informal writing and cutting-edge usages can be gleaned from the archives of several online discussion groups[5], while the recently released CD ROM version of the Bible (Ó Fiannachta, 1981) provides a convenient source of formal literary material[6]. Most useful for lexicographic work are the resources made available by the Irish government[7] (specifically *An Coiste Téarmaíochta*, who are in charge of coining modern terminology, and *An Gúm*, the government publishing house), and by *Fiontar*, a program at Dublin City University devoted to interdisciplinary studies through the medium of Irish[8].

Irish speakers also benefit from several outstanding print resources. These include the two standard bilingual dictionaries (Ó Dónaill, 1977) and (de Bhaldraithe, 1959), and a recently

[1] My primary research areas are in pure mathematics and theoretical physics.
[2] See `http://www.smo.uhi.ac.uk/gaeilge/gaeilge.html`
[3] e.g. `http://www.beo.ie/`
[4] e.g. `http://www.ireland.com/gaeilge/teangabeo/`
[5] e.g. `http://listserv.heanet.ie/lists/gaeilge-a.html`
[6] See `http://www.fiosfeasa.com/`
[7] See `http://www.acmhainn.ie/`
[8] See `http://www.dcu.ie/fiontar/further/focloiri.html`

published monolingual thesaurus (Ó Doibhlin, 1998). Though on a much smaller scale than the present work, the latter is a finely crafted book, produced (presumably manually) by a fluent speaker and Irish language scholar. I have intentionally not incorporated its contents in the current version of the database, so that it can provide an objective "gold standard" measure of quality of the computer-generated output. Examples are discussed in §4.

While surveying the available corpus material and discussing the limitations on financial resources I should also note that there are two extent Irish corpora that I have not used; a substantial one developed as part of the European Union PAROLE project (prohibitively expensive at 250 euro) and a somewhat smaller one compiled by Ciarán Ó Duibhín[9] (free, but for use only on Windows machines).

2 Phase One: Creating a software infrastructure

The first step in the process involved the development of some simple lexicographical database software. Naturally, a great deal of the effort that went into this phase could have been avoided by using an existing package. On the other hand, starting from scratch has made it easier to integrate successive phases with the underlying database, and, where necessary, to tailor things to the specific needs of Irish. A typical record in the database stores a dictionary headword, basic grammatical information (including tags for special inflections), and a list of citations.

Each record also stores, recursively, a list of records in the same format representing alternate forms. Careful handling of these alternates is essential for a language like Irish which had no standardized orthography until the middle of the 20th century (Rannóg an Aistriúcháin, 1962), and for which the standard has not taken root in the hearts of all native speakers. The majority of alternate forms in the current version of the database are either pre-standard or dialect forms, with a sprinkling of modern terminology that has been subsumed or made obsolete (e.g. a word like *glaothán* ("a pager") that appeared in (Mac Mathúna & Ó Corráin, 1995) almost ten years ago but has been supplanted by *glaoire* in usage and in the recommendations of *An Coiste Téarmaíochta*).

Next, I wrote a program in C++ called `morph-ga` that generates all inflected forms of Irish nouns, adjectives, and verbs when provided with a headword and sufficient grammatical tagging information. This piece of software is the linchpin for everything that follows, in particular providing a useful shortcut that I call "naïve stemming". Instead of taking the time to write a completely general stemmer, it suffices to implement some basic heuristics for making wild guesses at stems. Suppose, for instance, that a target word *mhantaí* appears in a corpus text. The software recognizes the ending *-aí* as (1) a common plural ending, (2) the comparative ending of an adjective ending in *-ach*, or (3) a rarely used verb ending in the subjunctive (Irish speakers may see other possibilities which must be disposed of as well). Heuristic (1) leads to a conjectural noun stem *mant* which is indeed found in the database, but `morph-ga` correctly generates its plural as *mantanna*, eliminating this case. Heuristics (2) and (3) yield *mantach* and *mantaigh* respectively, and the target word is found as a correct morphological form in each

[9] `http://www.smo.uhi.ac.uk/˜oduibhin/tobar/`

case. Probability says that possibility (2) is surely correct, but contextual markers must be used to verify this for certain.

3 Phase Two: Generating a clean list of words

The goal of this phase was, in short, to fill up the database created in phase one. Most important was the creation of an accurate list of dictionary headwords with complete grammatical information. Of secondary importance were accurate citations to print and electronic texts.

3.1 Methodology

1. **Extract the core database from a corpus**. I began by assembling a small corpus of electronic material out of the sources noted in §1.2 and wrote shell scripts that hunt for forms not already in the database, sorting by frequency. Later, improved, versions assign "editorial" weights to different texts and count an appearance in, say, the carefully edited *Oll-liosta Téarmaíochta* more heavily than one in the archives of an email discussion group. Naturally one expects that the words at the top of the list are the ones most likely to be spelled correctly; these are run through the stemmer and incorporated into the database (assuming they pass the various checks below).

2. **Add citations from print dictionaries**. A certain amount of checking by hand against print dictionaries has been performed as well, as a way of verifying the accuracy of the words being added to the database, but also as a way of fleshing out the lists of citations which are used in various ways during later phases of the project. In addition to the standard bilingual dictionaries, there are several books of terminology in print (Biology, Home Economics, Geography, etc.) representing the work of *An Coiste Téarmaíochta*. In an afternoon, one can add citations from one of these dictionaries to the database with a single keystroke per entry.

3. **Validate spelling via pattern matching**. Another powerful tool for checking the database is a shell script that uses pattern matching to look for illegal combinations of characters in a raw word list. The current version of this script implements 200 rules, varying from the trivial (only the characters 'l', 'n', and 'r' are doubled in Irish) to the subtle (a string of consonants preceded by a so-called broad vowel – 'a','o', or 'u' – is in general not allowed to be followed by a slender vowel – 'e' or 'i'). While there are many exceptions to certain rules, these exceptions can be either verified by hand or further whittled with some addition pattern matching.

4. **Validate spelling via authoritative texts**. The citation information garnered in step two is exploited to look for potential spelling problems as follows. Each source is assigned a weight that measures its "authoritativeness" from the point of view of spelling (thus modern print dictionaries get high values while materials produced before the spelling reform in the 1940's get extremely low values). Warning flags can be raised when an

alternate form has a more authoritative citation than the putatively standard form, or, similarly, if an alternate form has a greater number of citations than the standard form (authoritative or not).

3.2 Results

The data assembled in this phase enabled us to distribute the first full-scale Irish spellchecker, originally packaged for use with Geoff Kuenning's *International Ispell* and released under the GNU public license[10] in June of 2000. This initial release contained just over 13,000 dictionary headwords and some 171,000 inflected forms. Since then, I have repackaged things for use with the other widely-used spellcheckers in the open source community (`aspell` and `myspell`) and the database has grown to almost 30,000 headwords and 300,000 inflected forms[11]. Diarmaid Mac Mathúna has recently repackaged the word lists for use with Microsoft software, maintaining the open source license.

My guess is that the percentage of remaining misspellings (as of February 2003) is probably smaller than for some widely-used English spellcheckers (if so, this would be one of the rare instances in which the minority language tool outstrips the English language tool).

4 Phase Three: Generating the thesaurus

The goal of this phase was to generate a machine-readable thesaurus that can be output, for example, as a high-quality PDF document with hypertext links. Eventually we hope to refine the process described below to have the output compliant with the ANSI/NISO Z39.19 standard (ANSI/NISO, 1993). This will allow the database to be integrated more easily into information retrieval or indexing systems that rely on the standard.

The key labor-saving idea here is the introduction of English translations, allowing us to transfer semantic relationships from existing English language thesauri to Irish. While engaged in this work, I learned of a pilot study done at the University of Limerick that is akin in spirit to our approach (Sutcliffe *et al.*, 1996). They describe a prototype of a multilingual version of WordNet which, essentially, maps words from non-English languages into the existing WordNet hierarchy[12]. Modulo the ongoing port of our database to the WordNet format (discussed below), our work provides a full-scale realization of the system envisaged in their paper.

The introduction of English may raise some theoretical worries that we shall address in §4.2.

4.1 Methodology

1. **Assign raw English meanings to headwords**. This was surely the most labor inten-

[10] See `http://www.gnu.org/copyleft/gpl.html`
[11] Available from `http://borel.slu.edu/ispell/`
[12] The prototype can be found at `http://nlp01.cs.ul.ie/iwn.html`

sive phase, though it was made easier by the resources at `www.acmhainn.ie` and several other small-to-medium scale English-Irish and Irish-English electronic glossaries produced by amateur language enthusiasts[13]. Where necessary, lists of English meanings were fleshed out by reference to the standard print dictionaries (Ó Dónaill, 1977), (de Bhaldraithe, 1959), (Mac Mathúna & Ó Corráin, 1995), and (Ó Cróinin, 2000).

2. **Resolve ambiguities among English definitions**. Much of this step can be automated via standard word sense disambiguation techniques, though in doing so we relied to a certain extent on the quality of the available Irish-English dictionaries. For instance, one rarely finds a single polysemous English translation for a given headword in (Ó Dónaill, 1977), even when a human reader would surely know the correct resolution. By using a database of polysemous English words and a scheme for resolving them, as provided by a system like WordNet (Fellbaum, 1998), the software can easily decide, for instance, that the word *feileastram* with English translations "iris, flag" refers to a plant and not part of the eye or a kind of banner. When there are not sufficiently many English translations or if the translations are missing from the English database, some human intervention becomes necessary. In reality, instead of doing the sensible thing and using WordNet from the beginning, I developed my own primitive version based on the public domain Roget's Thesaurus (Roget, 1991)[14]. Were I to do it all over, I would surely use WordNet in light of the time savings, improved quality, and standardization its use would represent. I may "port" the resolutions in the database to this format at some future date.

3. **Break word list into semantic equivalence classes**. The idea here is a completely naïve but seems to work well. To first order, we tentatively assign two words to the same equivalence class when they share a resolved English translation. This assignment is given a "confidence parameter" that increases when there are multiple shared translations. More generally, whenever two Irish words have resolved English translations that are (possibly different but) semantically close (as determined by reference to an English thesaurus) the confidence parameter is increased by an amount proportional to the semantic proximity of the English translations (equality naturally providing the largest increase). The terminology "equivalence class" is perhaps deceiving here, since transitivity of the equivalence relation fails badly. Were one to take the transitive closure by, say, further increasing the confidence parameter between two words if there is a chain of equivalences joining them, essentially unrelated words would end up marked as equivalent. For example, one might guess incorrectly that *géarchúiseach* ("shrewd") is related to *garg* ("pungent") since the polysemous Irish word *géar* shares each of these English senses. Though we have no *a priori* method for disambiguation of Irish words, there is clearly potential for some bootstrapping here. The thesaurus generated at this step (without transitivity) implicitly picks out the different senses of a word like *géar*; one could then implement transitivity as suggested above when there exists a chain of equivalences between *disambiguated* Irish words.

4. **Generate the hypertext thesaurus**. This step converts the internal database of equivalence classes into a human-readable format (namely, hyperlinked PDF). Representative

[13] See `http://www.crannog.ie/focloir.htm for a notable 14,000 word example.`

[14] Available from `http://www.promo.net/pg/`

> nouns were selected for about 1000 basic categories, similar to the classical Roget's thesaurus in English. This was done automatically, through a combination of criteria involving (1) the frequency of appearance of the representative word in the corpus, (2) a measure of its centrality in the equivalence class, and (3) its lack of ambiguity. The current PDF version displays the thesaurus in alphabetical order, each entry being followed by one or more hypertext links to the representative word(s) under which it appears. Preliminary versions are available for free download[15].

4.2 Results

As noted above, the use of English translations ought to raise some concerns about this phase in the process. The potential imposition of English language categorizations into a monolingual Irish thesaurus will surely raise some Whorfian hackles. This may be perceived as particularly dangerous ground for an endangered language; Irish readers will be reminded of Tomás Ó Rathaille's famous characterization of the then moribund Manx language as "English disguised in Manx vocabulary" (Ó Rathaille, 1932). Unfortunately, this is the sort of corner into which one is forced when working with a minority language lacking any substantial monolingual lexicographic material.

Our theoretical defense rests first on the *coarse granularity* of thesauri, that is, the semantic fuzziness inherent in a long list of quasi-synonyms. Take the canonically "untranslatable" Irish word *dúchas* in its most abstract sense of "heritage, patrimony". Though these English translations are a poor reflection of the depth of meaning in the Irish word, they are also given as translations of its nearest Irish synonym, *oidhreacht* (also meaning "inheritance" in the concrete sense). Thus, since we are not concerned with razor-sharp precision but only that these Irish words end up near each other in the thesaurus, the algorithm above suffices.

A relativist criticism is also weakened somewhat when leveled against the English-Irish language pair which has seen, for better or for worse, several centuries of heavy (mostly unilateral) lexical borrowing. One would probably need more care in trying our approach with, say, Hopi or Dyirbal.

Fundamentally, though, our strongest argument is the *a posteriori* one provided by the quality of the finished product. As noted in the introduction, an objective measure of quality can be obtained by comparing selected portions of the output with a "gold standard" thesaurus (Grefenstette, 1994) for which we use the *Foclóir Analógach* (Ó Doibhlin, 1998).

Here, for example, is the entry from (Ó Doibhlin, 1998) under the headword *anachain* ("misfortune, adversity, calamity"). It is divided into two halves, the first listing 29 general varieties of adversity and the second listing 27 more specific calamitous occurrences.

> **Cineálacha:** Angar. Mí-ádh. Míchinniúint. Mífhortún. Míshéan. Léan. Doilíos. Tubaiste. Donas. Ainnise. Léirscrios. Báine. Buaireamh. Anacair. Anbhroid. Anó. Imní. Duainéis. Crá. Clipeadh. Bearrán. Leatrom. Lionn dubh. Díomá. Drochmhisneach. Diomú. Beaguchtach. Éadóchas. Buille fill.

[15] http://borel.slu.edu/teasaras/

> **Saintarlú:** Tubaiste. Tionóisc, taisme. Turraing. Stoirm. Anfa. Spéirling. Tíofún. Gorta. Plá. Cogadh. Tuile. Dóiteán. Ár. Éirleach. Longbhriseadh. Coscairt. Scrios. Slad. Creach. Dochar, damáiste. Turnamh. “Tonnbhriseadh an tseanaghnáthaimh”. Teip. Titim. Briseadh.

Now, the corresponding entry from our thesaurus (138 words/phrases). The order of words is dictated by the software, which lumps together words deemed to be semantically close according the measures discussed above.

> tubaiste, drámh, donacht, uisce an cheatha, lomadh an Luain, cinniúint, caill, dochma, teann, crácamas, dua, anró, easonóir, síleáil, tónáiste, tiortáil, cruatan, splíontaíocht, mífhortún, cora crua an tsaoil, droch-chor, ciotaí, treampán, callshaoth, anchaoi, aimpléis, drochrath, caduaic, seacht gcúraimí an tsléibhe, seacht gcúraimí an tsaoil, iomard, umar na haimléise, duibheagán an éadóchais, anrath, mírath, pláinéad, saol léanmhar, saol crua, saol anróiteach, drochshaol, mí-ádh, dochraide, deacair, cacht, bráca, toirmeasc, míchonách, duais, deacracht, cruachás, smál, buairt, buaireamh, margadh éagórach, anachain, trioblóid, diachair, donas, mairg, bris, angar, dobrón, cránas, léan, dothairne, dólás, doilíos, leatrom, doghrainn, tinneas, guais, géibheann, éigeantas, éigean, duainéis, dola, dócúl, dochar, broid, anó, anbhroid, anacair, svíomh, imní, triail, cros, céasadh, cath, sciúrsáil, crá, droch-anáil, pionós, plá, sciúirse, imirt, gearradh, tuisle, tapaigean, míthapa, tionóisc, óspairt, timpiste, taisme, púir, ochlán, liach, dursan, cat mara, matalang, turraing, gátar, eirleach, meath, díomua, céim síos, dul ar gcúl, gonc, longbhriseadh, titim, turnamh, treascairt, milleadh, creachadh, faillí, cliseadh, teip, meathlú, meathlaíocht, loiceadh, feall, scrios, raic, díothú, creach, cabhóg, anás, ainriocht, aimhleas.

For reasons of space, we will restrict ourselves to a few simple observations. First, because our underlying English thesaurus tends not to give lists of specific kinds of things, our output leaves out nine of the ten calamities starting at the cognate *stoirm* and ending at *ár* “slaughter” (we picked up *plá* only because of its figurative use as “a scourge”). This seems to be a matter of taste in thesaurus construction versus a linguistic issue.

Leaving out these ten, we hit 28 of 46 ($\approx$ 60%). We missed words in places where Ó Doibhlin seems to stray farther afield from the central meaning of “adversity”: *léirscrios*, *báine* (“destruction”), *clipeadh*, *bearrán* (“teasing”), and the final seven of the first list, all variants of “sorrow” or “despair”. Undoubtedly we should have picked up *míchinniúint* which was in our database but was a bit light on English translations (“ill fate”).

The good news, of course, is the incredible richness of expression found in the expanded list. Even the most fluent speakers, we hope, will discover new idioms (*lomadh an Luain*), unusual secondary meanings (*pláinéad* as “ill luck, planetary influence”), or literary words that have fallen into disuse (*cacht, dursan*).

Finally, we do not see any “howlers” in the output (though a poor job of disambiguation of English definitions has led to some embarrassing blunders in other lists).

We emphasize that the example just given is the *unedited output* of the sequence of algorithms described above. A change to the underlying database (say, the addition of a new English definition for an Irish word) automatically propagates itself (sometimes in subtle ways) when we give the command to rebuild the thesaurus from scratch. This enables continuous updating of the thesaurus, and allows end users to make contributions or corrections in a standardized way. Such continuous maintenance is essential for any piece of software, but especially so when the primary goal of the software is the accurate reflection of the various idiosyncrasies of a living language: new terminology, shifting usages, etc.

We believe this kind of collective approach to software development and maintenance will be essential to the future provision of quality software to speakers of minority languages. In concrete terms, this approach is facilitated by releasing our thesaurus and its LaTeX sources under the GNU Free Documentation License[16] which says, in short, that everyone has the freedom to copy, modify, or even sell the thesaurus as long as redistributed versions preserve the same freedoms. This kind of license guarantees the widest possible dissemination of the materials we have developed, but more importantly, empowers speakers of minority languages by placing control of these resources directly in their hands, eliminating the generally fruitless reliance on the benevolence of large corporations for the provision of such material.

Acknowledgments

Thanks to Diarmaid Mac Mathúna, Michael Conry, Vincent Morley, and Alastair McKinstry for their interest in the spellchecking project and for helping to spread the news of its availability. Alastair, in particular, deserves credit for developing the first non-trivial Irish spellchecker (circa 1997). I benefitted from enjoyable email exchanges with Andrew Dunbar and Alan Horkan about machine translation and software localization, respectively. Many people deserve credit for producing the electronic texts upon which this work was based, but especially helpful were Caoimhín Ó Donnaíle (who provided some well-edited lexicographic material) and Antain Mac Lochlainn and the people at `www.acmhainn.ie` who have made available much of the work of *An Coiste Téarmaíochta*.

References

AITCHISON J. & GILCHRIST A. (1987). *Thesaurus construction: a practical manual*. Aslib, London, 2nd edition.

ANSI/NISO (1993). Z39.19 – 1993 Guidelines for the Construction, Format, and Management of Monolingual Thesauri.

T. DE BHALDRAITHE, Ed. (1959). *English-Irish Dictionary*. An Gúm, Baile Átha Cliath.

FELLBAUM C. D. (1998). *WordNet: an electronic lexical database*. MIT Press, Cambridge, Mass.-London.

[16] See `http://www.gnu.org/licenses/fdl.html`

GREFENSTETTE G. (1994). *Explorations in automatic thesaurus discovery*. Kluwer Acad. Publ., Dordrecht.

C. LAIRD, Ed. (1999). *Webster's New World Roget's A-Z Thesaurus*. Macmillan, New York.

S. MAC MATHÚNA & A. Ó CORRÁIN, Eds. (1995). *Collins Gem Irish Dictionary*. Harper-Collins Publishers, New York.

B. Ó CRÓININ, Ed. (2000). *Pocket Oxford Irish Dictionary*. Oxford Univ. Press, Oxford.

Ó DOIBHLIN B. (1998). *Gaoth an Fhocail*. Coiscéim, Baile Átha Cliath.

N. Ó DÓNAILL, Ed. (1977). *Foclóir Gaeilge-Béarla*. An Gúm, Baile Átha Cliath.

P. Ó FIANNACHTA, Ed. (1981). *An Bíobla Naofa*. An Sagart, Maigh Nuad.

Ó RATHAILLE T. (1932). *Irish dialects past and present*. Institiúid Árd-Léinn, Baile Átha Cliath.

RANNÓG AN AISTRIÚCHÁIN (1962). *Gramadach na Gaeilge agus Litriú na Gaeilge: An Caighdeán Oifigiúil*. Oifig an tSoláthair, Baile Átha Cliath.

ROGET P. M. (1991). Project Gutenberg Roget's Thesaurus.

SUTCLIFFE R. F. E., O'SULLIVAN D., MCELLIGOTT A. & Ó NÉILL G. (1996). Irish-English mappings in International WordNet: a pilot study. Unpublished.