Corpas na Gaeilge (1882-1926): Integrating Historical and Modern Irish Texts

Elaine Uí Dhonnchadha[3], Kevin Scannell[7], Ruairí Ó hUiginn[2], Eilís Ní Mhearraí[1], Máire Nic Mhaoláin[1], Brian Ó Raghallaigh[4], Gregory Toner[5], Séamus Mac Mathúna[6], Déirdre D'Auria[1], Eithne Ní Ghallchobhair[1], Niall O'Leary[1]

[1]Royal Irish Academy, Dublin, Ireland
[2]National University of Ireland Maynooth, Ireland
[3]Trinity College Dublin, Ireland
[4]Dublin City University, Ireland
[5]Queens University Belfast, Northern Ireland
[6]University of Ulster, Northern Ireland
[7]Saint Louis University, Missouri, USA
E-mail: uidhonne@tcd.ie, kscanne@gmail.com, ruairi.ohuiginn@may.ie, e.nimhearrai@ria.ie, nicmhaol@hotmail.com, brian.oraghallaigh@dcu.ie, g.toner@qub.ac.uk, s.macmathuna@ulster.ac.uk, d.dauria@ria.ie; e.nighallchobhair@ria.ie, nialloleary.dho@gmail.com

Abstract

This paper describes the processing of a corpus of seven million words of Irish texts from the period 1882-1926. The texts which have been captured by typing or optical character recognition are processed for the purpose of lexicography. Firstly, all historical and dialectal word forms are annotated with their modern standard equivalents using software developed for this purpose. Then, using the modern standard annotations, the texts are processed using an existing finite-state morphological analyser and part-of-speech tagger. This method enables us to retain the original historical text, and at the same time have full corpus-searching capabilities using modern lemmas and inflected forms (one can also use the historical forms). It also makes use of existing NLP tools for modern Irish, and enables integration of historical and modern Irish corpora.

Keywords: historical corpus, normalisation, standardisation, natural language processing, Irish, Gaeilge

1. Introduction

This paper describes the preparation of *Corpas na Gaeilge (1882-1926)*, a corpus of historical texts, to be used in the first instance for lexicography in the Royal Irish Academy's *Foclóir na Nua-Ghaeilge*[1] [Dictionary of Modern Irish] Project. *Corpas na Gaeilge (1882-1926)* complements *Corpas na Gaeilge 1600-1882* (2004) a corpus of earlier Irish texts, which was published in 2004. The aim of *Foclóir na Nua-Ghaeilge* is the provision of a corpus-based dictionary arranged on historical principles to cover the period from 1600 to the present. The texts found in the period 1882-1926 vary in terms of orthography, morphology and syntax; therefore the processing combines both manual and automatic elements. Manual elements include the development of specific wordlists by a panel of language experts and the automatic elements include spelling standardisation, lemmatisation and part-of-speech tagging. The work which began in 2012 (directed by a management committee), is a collaboration between staff of the Royal Irish Academy (RIA), language experts and natural language processing experts. The corpus texts have been made available online in raw text format and TEI format since December 2013[2].

[1] *Foclóir na Nua-Ghaeilge:* http://www.ria.ie/research/focloir-na-nua-ghaeilge.aspx
[2] RIA Corpus: http://research.dho.ie/fng/index.php; http://research.dho.ie/fng/cuardaigh.php

2. Background

The written history of the Irish language extends back to the seventh century, and perhaps up to two centuries earlier than that if we include the Ogham monument inscriptions, consisting of personal names written in a highly archaic form of Irish. As with any language that has such a long history, it is normal to divide it into periods, and for Irish the following are recognised: Old Irish (c.600-900), Middle Irish (c.900-1200), Early Modern Irish (c.1200-1650) and Modern Irish (c.1650-present).

The lexicography of Irish has been served in an uneven manner. For its earliest stages, c.600-1650 we have the *Dictionary of the Irish Language* (1976) which appeared in a series of fascicles between 1913 and 1976. It is a dictionary compiled in broadly historical terms, giving earliest attestations, variant forms and meanings and sometimes also etymologies. Due to its long period of compilation, the standard and quality of fascicles vary and, as may be expected in a work of this nature, certain entries are incomplete or out of date. However, a digitised version of the dictionary which was published online in 2007[3] has made it possible to update certain aspects of the work through supplements and additional entries. In 2013 a second revised electronic edition containing over 4,000 amendments and additions was made available online.

[3] The electronic Dictionary of the Irish Language (eDIL): http://www.dil.ie/about.php

2.1 *Foclóir na Nua-Ghaeilge*

In the case of Modern Irish, it will be helpful to recognise two broad periods, 'revival' Irish dating from roughly 1900 to the present, and an earlier period stretching from c.1650-c.1900. Revival Irish has been relatively well served with both English-Irish and Irish-English dictionaries. A modern online English-Irish dictionary appeared in 2013[4] and an online Irish-English dictionary is scheduled to appear in 2015. Several printed Irish-English and English-Irish dictionaries appeared in the course of the twentieth century. All of these works, it should be noted, are functional dictionaries that offer English equivalents for Irish terms, or vice versa. They do not have a historical dimension and most do not give sources.

The earlier period, 1650-1900, is devoid of any modern dictionary. On the completion of the *Dictionary of the Irish Language* (600-1650) in 1976, the Royal Irish Academy established a new project, *Foclóir na Nua-Ghaeilge*, which was to provide a dictionary arranged on historical principles to cover the period from 1600 to the present, and thus continue the *Dictionary of the Irish Language*. Work on this project has been in progress since that time.

The challenges facing the compilers of this dictionary are quite daunting. In the period in question, the Irish language underwent many changes. Following the downfall of the Gaelic aristocracy in the early seventeenth century and subsequent colonisation and plantation, the language became in the course of the next centuries mainly the language of a rural peasantry. Despite its reduced status, the growth in population meant that there probably were more people speaking Irish in the seventeenth and eighteenth centuries than at any other time in its history. This, coupled with the widespread availability of paper and, to a lesser extent printing, has left us with a large body of material – devotional texts, historical tracts, songs, poems and tales from this period. The electronic corpus of material being compiled in the RIA, to be used in drafting the dictionary entries, is drawn from both written and oral sources and when complete will comprise 90+ million words, it is estimated.

The broadly standardised written language which obtained down to the seventeenth century was replaced by a written language that varied quite widely in its orthography, morphology and grammar. Dialectal forms came very much to the fore, and due to the expansion of English many words were borrowed from that language. In 1958, the spelling and orthography of Irish were again standardised and this standardised language is now used in most published works in Irish, including dictionaries, where the headwords are in a form appropriate to this standard. This standard can be at a considerable remove from forms found in our corpus of material. Processing these historical texts to enable their efficient use in lexicography presents many challenges.

[4] Foras na Gaeilge's New English Irish Dictionary: http://www.focloir.ie/

In 2004, 705 texts from the period 1600-1882 were published in CD form as *Corpas na Gaeilge 1600-1882*[5]. This 7.2 million word corpus comes with a concordance of all forms occurring in these texts, but as the texts have not been lemmatised or annotated with part-of-speech tags, the variants, inflected forms, etc., are not grouped together under one headword and may be widely dispersed.

3. Corpas na Gaeilge (1882-1926)

The corpus being described in this paper, *Corpas na Gaeilge (1882-1926)* also contains approximately 7 million words. This corpus consists of books, published by more than twenty publishers, covering a wide range of topics and genres as well as representing the three major dialects of Modern Irish. (Newspapers and periodicals are currently in preparation). A breakdown of topics and genres are given in Table 1.

Text classification	**Number of texts**	**% of total**
Informational works		
Folklore	75	26
Textbooks	12	4
Linguistics	20	7
Other non-fiction	65	23
Sub-Total	**172**	60%
Creative works:		
Poetry	17	6
Drama	18	6
Short story collections	57	20
Novels	20	7
Essays	2	1
Sub-Total	114	40%
Total	**286**	**100%**

Table 1: Classification of texts in Corpas na Gaeilge (1882-1926)

Modern Irish has three distinct dialects (and further sub-dialects). It is necessary for a corpus-based historical dictionary to have sufficient representation of each of the major dialects in its corpus. The dialectal composition of the corpus is given in Table 2.

Dialect	**Number of texts**	**% of total**
Connacht	54	20
Ulster	35	12
Munster	81	28
Translations from other languages	31	11
Non-dialect	85	30
Total	**286**	**100%~**

Table 2: Dialectal composition of texts in Corpas na Gaeilge (1882-1926)

[5] Corpas na Gaeilge (1600-1882): http://www.ria.ie/Research/Focloir-na-Nua-Ghaeilge/Foilseachain--Publications.aspx

As all of these texts predate computerisation, the texts had to be captured in electronic form, either by typing or by scanning. In the earlier stages of the *Foclóir na Nua-Ghaeilge* project, texts were typed, followed by a period where optical character recognition (OCR) methods were tested in parallel with typing. In recent times, the majority of texts are scanned and OCR is performed using Optopus OCR software[6] . The decision regarding whether texts are typed or scanned is based mainly on the quality of the print or the condition of the book. Close attention is paid to the accuracy of the data capture process. Through a series of experiments it emerged that the most effective method of ensuring high accuracy involves a combination of close reading and enhanced spellchecking.

4. Corpus Processing

4.1 Corpus Processing Tools for Irish

In order to process *Corpas na Gaeilge (1882-1926)*, a survey of existing natural language processing (NLP) tools for Irish was carried out. A finite-state morphological analyser and part-of-speech (POS) tagger for the modern standardised language (post 1958) were available (Uí Dhonnchadha & van Genabith, 2005). In this rule-based system, POS tagging and lemmatisation are carried out in two stages. In the first stage, each token in the text is analysed using the finite-state morphological analyser (both *xfst*[7] and *foma*[8] versions are available), and a set of possible morphological analyses and lemmas are assigned to each token. In the second stage, context specific rules (Constraint Grammar[9]) are used to determine the most likely POS for the token based on its surrounding context in the sentence. This POS tagger has 95-96% accuracy on unrestricted text.

The morphological analyser's lexicon incorporates all of the 50K headwords in the Ó Dónaill (1977) dictionary and generates all inflected forms of the headwords. It also includes a set of morphological guessers which use morphological clues (e.g. distinctive inflectional suffixes) in unrecognised words to guess the likely part of speech and features of unknown words.

These tools however could not deal with the older and varied spellings prevalent in this historical corpus. One option would be to extend the existing POS tagger's lexicon to incorporate older forms. This would probably reduce efficiency and result in increased ambiguity in this rule-based tagger. Another possibility would be to create a specialized version of the tagger for the time period in question. This would require substantial development and would have limited reusability.

A better solution would be to standardise the texts in such a way as to enable them to be processed with the modern POS tagger. Fortunately, there was a prototype standardiser available; *An Caighdeánaitheoir* (Scannell, 2009). With further development and training (described in Section 4.2) this could be used to associate a modern standard (inflected) form with pre-standard words in the texts, and thereby enable these texts to be lemmatised and POS tagged with minimal adjustments to the existing POS tagger. This solution has the advantage that sub-corpora from different historical periods can be POS-tagged in the same way and all pre-standard inflected word-forms can be united under the modern lemma. This will enable lexicographers to search the corpus for examples using the modern spelling (either lemma or inflected form) and retrieve all variant and historical forms as desired.

4.2 *An Caighdeánaitheoir* (The Standardiser)

The strategy employed by the standardiser is to treat spelling standardisation as a problem in machine translation between two very closely related languages: pre-standard and standard Irish. Indeed, our implementation uses well-known techniques in statistical machine translation and can be viewed as a variant of the word-based IBM model 1 (Brown *et al*, 1993). As such, the key elements are a *language model* for the target language (standard Irish), and a *translation model*, representing the conditional probability of a particular non-standard spelling corresponding to a particular standardised spelling. There were challenges to overcome in constructing both models.

We use a trigram language model for standard Irish, and training simply requires a sufficiently large corpus. The difficulty here is a philosophical one; namely, what is "standard Irish"? A major spelling reform was introduced in the 1940s and brought to completion in 1958, and operationalized through the publication of two major bilingual dictionaries in the second half of the twentieth century[10]. A simplified grammar was published in 1958, together with the spelling recommendations, as *Gramadach na Gaeilge agus Litriú na Gaeilge: An Caighdeán Oifigiúil* [Irish Grammar and Spelling: The Official Standard]. In practice, however, the story is quite complex; the published dictionaries do not adhere completely to the standard, nor do certain widely-used grammars, e.g. New Irish Grammar (The Christian Brothers, 1994). To add to the confusion, a major revision of the official standard was published in 2012, with the goal of simplifying the rules and bringing the standard more in line with the language as spoken by native speakers in the Gaeltacht. The result is that despite having access to a large corpus of texts (Scannell, 2007) (more than 100 million words)

[6] Optopus is a trainable OCR program from Makrolog, Germany. http://www.makrolog.com/ It is currently unsupported.

[7] Xerox Finite State Tools: http://www.stanford.edu/~laurik/fsmbook/home.html

[8] Foma Finite State Compiler: http://code.google.com/p/foma/

[9] VISL Constraint Grammar: http://beta.visl.sdu.dk/constraint_grammar.html

[10] (Ó Dónaill, 1977); (de Bhaldraithe, 1959)

published since these reforms were put into place, virtually none of the texts fully complies with any variant of the standard. How does one create a language model when there are *no* non-trivial texts written in that language?

Our solution is to employ a suite of rule-based proofing tools (spelling and grammar correction) developed by the second author to create a sub-corpus of about 40 million words consisting of the texts which are most compliant with the official standard, at least as it is implemented in the proofing tools. Additionally, we applied automated standardisations to a small number of recurring non-standard forms in order to produce a training corpus which best approximates to "standard Irish".

Our approach to the translation model is somewhat unusual in that we do not train the probabilities using a "bilingual" corpus and the Expectation Maximization (EM) algorithm (Koehn *et al*, 2007), as is typical in this context. The reason is, in short, that we can do substantially better with an *ad hoc* approach using resources we have at hand. First, we only have access to a relatively small number of texts written in both pre-standard and standard Irish (about 700,000 words), and there is a tremendous amount of variation among these pre-standard texts in terms of both dialect and time period. The statistical alignments generated from this corpus are quite noisy and unsuitable for the high-precision translation task at hand. Second, through ongoing lexicographical work, we already have a large, manually-curated database of about 22,000 pre-standard lemmas mapped to their standard forms, plus an additional 10,000 mappings taken directly from the Ó Dónaill (1977) dictionary.

Finally, whereas many spelling standardisations are essentially arbitrary choices of one form over others (for example, *eileastrom, feileastar, seileastram* are all treated as variants of *feileastram* 'wild iris'), many others are consequences of a number of general context-sensitive rules. For example, a word internal *-bhth-* is standardised to *-f-*, a word final *-ghail* standardises to *-aíl*, and *sg-* always becomes *sc-*. The current version of the standardiser implements 567 hand-written rules of this type.

Quite commonly, a standardised form is discovered through a combination of rule applications and lexical standardisations. For example, the hypothetical form *coimh-mheasguighthe* would undergo the following sequence of spelling changes; the first five represent applications of general rules, while the final change maps a non-standard second declension verb to its standard first-declension form, using a mapping found in the lexical database:

coimh-mheasguighthe → *comh-mheasguighthe* → *cóimheasguighthe* → *cóimheascuighthe* → *cóimheascaighthe* → *cóimheascaithe* → *cóimheasctha* ('coalesced')

The definition of the translation model is naive but effective in our context. All non-standard forms that are paired with a particular standard word in the lexical database are viewed as having the same conditional probability. Non-standard forms that are paired with a standard word through one or more rule applications are "penalized" for each rule that is applied; that is to say, the conditional probability is multiplied by a fixed factor $\beta < 1$ each time a rule is applied. A tuning process allows us to choose an optimal value for β; a small corpus of standard/non-standard sentence pairs was held out, and the performance of the standardiser was evaluated for different values of β.

Once the language model and translation model are in place, the decoding process is straightforward. Some standardisations map more than one word to a single word (*i mbárach* → *amárach*), but a pre-processing step treats these set phrases as a single token, so we can effectively decode wordfor word. Decoding proceeds from left to right, maintaining a data structure of all possible hypotheses and their probabilities. Since we use a trigram language model, when multiple hypotheses share the same final two words, we discard all but the highest probability candidate. At the end of each input sentence, the maximal probability hypothesis is output.

4.3 Initial Survey of *Corpas na Gaeilge (1882-1926)*

In order to establish the extent of the non-standard spelling in *Corpas na Gaeilge (1882-1926)*, the finite-state morphological analyser was run on the seven million words of raw text before standardisation. As expected, many non-standard spelling were not recognised by the morphological analyser and therefore could not be accurately assigned tags automatically. Morphological guessers were not used as they would not be able to predict the modern equivalent lemma for non-standard spellings.

Of the 7 million words, 65% of word types were not recognised, and 35% of word types were recognised, i.e. words that are the same in the modern standard. Ignoring uppercase/lowercase distinctions there are 166K (approx.) different unknown words (types) which have non-standard spelling.

The 166K types were sorted according to frequency of occurrence in the corpus, e.g. the word *chuaidh* 'went' which is near the top of the frequency list occurs 5700 times in the corpus. Therefore, by adding this word to the Standardiser's lexicon (i.e. by pairing it with its modern equivalent *chuai**g**h* 'went'), 5700 instances of this word in the corpus will be automatically identified by the POS tagger.

The 1500 most frequent unknown types account for 50% (approximately) of the unknown tokens in the corpus. By manually assigning the modern equivalent to the 1500

most frequent non-standard word types, and adding these pairing to the Standardiser's lexicon, 50% of the unrecognised word forms can immediately be assigned the correct standard form, while the remaining unrecognised word forms can be processed by using rules and probabilities.

However, as we go down the list, the types occur less and less frequently in the corpus, so that the benefit of manually adding pairings to the lexicon makes less and less of an impact on the overall recognition rates. For example adding a further 500 words pairings would only improve the recognition rates by less than 3%, (e.g. 2028 most frequent types account for 53% approximately (in this sample).

The language experts on the team took the 1500 most frequent items on the list of non-standard types and associated them with the modern standard wordform. This became the basis of a database of non-standard to standard pairings. These pairings were then used directly by *An Caighdeánaitheoir*. In addition, approximately 2500 named entities (people, places etc.) which had been manually marked up in the first corpus, *Corpas na Gaeilge 1600-1882*, were also added to this database, and used in the same way.

4.4 Processing Stages

Our processing of the historical texts goes through a series of stages: tokenisation, standardisation, lemmatisation and POS tagging. We will briefly describe each stage.

Tokenisation: Firstly, the text is segmented into units called tokens. For many languages including Irish, tokenisation is mainly based on space between words, where a word equates to a token. But sometimes we may wish to divide a word into more than one token, e.g. a contracted form such as *I'm* is separated into two tokens: *I* (pronoun) and *'m* (= am verb). And sometimes we wish to keep two or more words together as one token, e.g. names or placenames such as *Finnegan's Wake*, *Baile Átha Cliath* 'Dublin' (proper nouns) or compound prepositions, e.g. *tar éis* 'after', *os cionn* 'above', where it does not make sense to analyse the parts individually.

Standardisation: The pre-standard inflected forms encountered in historical and dialectal texts are annotated with their modern standard inflected equivalents using *An Caighdeánaitheoir*.

Part-of-speech (POS) Tagging and Lemmatisation: The standard inflected word forms are processed using the POS tagger, enabling part-of-speech tag and lemma annotations to be added.

The following sentence (1), taken from the beginning of a short story printed in 1913, illustrates the process.

```
(1)  Bhí baintreabhach mhná  ann
     fad  ó shoin.
     Was widow         woman there
     long ago
     'There was a widow long ago'
```

Table 3 shows the sentence in vertical form, i.e. column 1 represents the original text showing one token per line. In column 2 we see the standard forms, in column 3 we see the PAROLE[11] POS tag and in column 4 we have the lemma.

Original	Std. form	POS	Lemma (base)
Bhí	Bhí	Vmis	bí
baintreabhach	baintreach	Ncfsc	baintreach
mhná	mná	Ncfsg	bean
ann	ann	Rl	ann
fad	fada	Rt	fada
ó	ó	Sp	ó
shoin	shin	Pd	sin
.	.	Fe	.

Table 3: Sample annotation of historical text

Formatting: The annotated corpus is formatted in both vertical form (similar to Table 3) and in XML Corpus Encoding Standard (XCES[12]) format as follows , using the `<w>` word tag with attributes `tag` for POS, `base` for lemma and `std` for standard form.

```
<p>
<s>
<w tag = "Vmis" base = "bí" std =
"Bhí">Bhí</w>
<w tag = "Ncfsc" base = "baintreach" std =
"baintreach">baintreabhach</w>
<w tag = "Ncfsg" base = "bean" std =
"mná">mhná</w>
<w tag = "Rl" base = "ann" std =
"ann">ann</w>
<w tag = "Rt" base = "fada" std =
"fada">fad</w>
<w tag = "Sp" base = "ó" std = "ó">ó</w>
<w tag = "Pd" base = "sin" std =
"shin">shoin</w>
<w tag = "Fe" base = "." std = ".">.</w>
```

This same vertical/XCES format is used for modern texts (e.g. the 30 million word *Nua-Chorpas na hÉireann*[13]) in which case the `std` value is usually the same as the original token (except for dialectal variants). In this manner historical texts and modern texts can be seamlessly integrated. Therefore, rather than normalising pre-standard forms, (e.g. as in Bollmann *et al* (2012), the

[11] Full specification of the Parole tags can be found at https://www.scss.tcd.ie/SLP/parole.htm
[12] XCES: http://www.xces.org/
[13] http://corpas.focloir.ie/ which can be queried using the SketchEngine interface (http://the.sketchengine.co.uk)

original form is kept and is annotated with modern standard form. Historical forms which do not have a modern equivalent are added to the finite-state POS tagger lexicon and associated with the modern lemma where one exists.

5. Evaluation

A detailed evaluation of three texts, (each one representative of one of the three major dialects), was carried out by the language experts. There were a number of difficulties. Firstly, the texts are in pre-standardised orthography. Secondly, there are differences, sometimes quite significant, between the dialects in terms of morphology, inflexion and syntax. Finally, there was small number of typographical errors and errors made in the scanning process. Consequently, there were problems noted at every stage of processing, including some errors in the original texts themselves.

Texts:
There were occasional typographical errors in the printed works, and some residual OCR errors.

Tokenisation:
There were many issues surrounding the non-standard use of hyphens to connect words which should be two tokens e.g. *oidche-sin* 'night-that' or even to connect suffixes to words, e.g. *táim-se* in place of *táimse* 'I am (emphatic)'. There were also difficulties regarding the widespread use of apostrophes for elided and contracted forms, particularly in stories using direct speech, again causing two tokens to be joined, e.g. *c'acu* in place of *cé acu* 'which of-them'.

Standardisation:
There were problems connected to residual non-standard spellings and non-standard inflectional morphology, e.g. *dubhras* 'I said' with non-standard spelling and incorporated pronoun, rather than the standard form *dúirt mé* 'I said' which has a separate pronoun. The most difficult problem to remedy (systematically) is where a non-standard root is used e.g. *dtearn* 'did' rather than the modern standard form *ndearna* 'did'.

POS tagging:
There were also problems at the POS tagging stage; the most common problem related to older verbal forms which used to have a preverbal particle *do* with the past-tense form. This preverbal particle takes the same form as the modern preposition *do* 'to', causing the POS tagger to tag some past tense verbs as nouns. There are also a small number of nouns which had a different gender in historical texts, based on inflection and accompanying definite article (or anaphoric references). These are tagged with the modern gender which will be wrong in those instances.

After these problems were addressed in the corpus processing tools, the same three texts were re-processed and the current accuracy of the standardiser's output is calculated to be approximately 95%. Other random samples have been selected from the corpus for detailed evaluation and initial calculations show accuracy rates ranging from 91-96%. The accuracy of POS tagging and lemmatisation has not yet been evaluated.

6. Conclusions

We believe this to be a very promising method of processing and integrating historical and contemporary documents, which makes maximum use of existing tools while developing specific standardisers for specific time periods. The original texts are not changed[14], rather additional information is added. The project relies on the collaboration of many different individuals each with different skills, without which the work could not be accomplished.

7. Future Work

In the immediate future, a substantial body of material from newspapers and periodicals from the period 1882-1926 will be processed. This work is currently in hand and it is envisaged that a further 3.5 million words from over twenty-five periodicals (many of which are of great historical interest) will be added to the corpus. Following this, the possibility of processing and integrating the earlier *Corpas na Gaeilge 1600-1882* with the current corpus will be investigated.

8. References

Bollmann, M., Dipper, S., Krasselt, J., and Petran, F. (2012). Manual and Semi-automatic Normalization of Historical Spelling - Case Studies from Early New High German. In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*. Vienna, September, 2012

Brown, P., Della-Pietra, S., Della-Pietra, V. and Mercer, R. (1993). The Mathematics of Statistical Machine Translation. *Computational Linguistics*, 19(2), pp. 263-313.

Corpas na Gaeilge 1600-1882. (2004). Royal Irish Academy: Dublin.

Dictionary of the Irish Language (1976). Edited by E.G. Quin. Royal Irish Academy: Dublin.

de Bhaldraithe, T. (1959). *English-Irish Dictionary*. An Gúm.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., … & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 177-180). Association for Computational Linguistics.

Ó Dónaill, N. 1977. *Foclóir Gaeilge-Béarla.* [Irish-English Dictionary]. Baile Átha Cliath: An Gúm.

Scannell, K. (2009) *Standardization of corpus texts for the New English-Irish Dictionary*, paper presented at the 15th annual NAACLT conference, New York, 22

[14] Except for obvious typographical and OCR errors. There is also an online archive which preserves the original printed layout http://research.dho.ie/fng/index.php

May 2009. (http://borel.slu.edu/pub/naaclt09.pdf)
Scannell, K. (2007). The Crúbadán Project: Corpus building for under-resourced languages, Cahiers du Cental 4 (2007), pp. 5-15, C. Fairon, H. Naets, A. Kilgarriff, G-M de Schryver, eds., "*Building and Exploring Web Corpora*", Proceedings of the 3rd Web as Corpus Workshop in Louvain-la-Neuve, Belgium, September 2007.
The Christian Brothers. (1994). *New Irish Grammar*. Baile Átha Cliath: C. J. Fallon.
Uí Dhonnchadha, E. and van Genabith, J. (2005) Scaling an Irish FST morphology engine for use on unrestricted text. *In* editor(s)A. Yli-Jyrä, L. Karttunen, J. Karhumäki, *Lecture Notes in Artificial Intelligence (LNAI): Proceedings of the FSMNLP 2005 Finite-State Methods in Natural Language Processing*, Berlin, Springer-Verlag, 2006, pp247-58