Improving full-text search results on *dúchas.ie* using language technology

Brian Ó Raghallaigh
Fiontar & Scoil na Gaeilge
Dublin City University
Drumcondra, Dublin 9
Ireland
brian.oraghallaigh@dcu.ie

Kevin Scannell
Department of Computer Science
Saint Louis University
220 N. Grand Blvd.
Saint Louis, Missouri 63103-2007
kscanne@gmail.com

Meghan Dowling
ADAPT Centre
Dublin City University
Glasnevin, Dublin 9
Ireland
meghan.dowling@adaptcentre.ie

[1]**Abstract**

In this paper, we measure the effectiveness of using language standardisation, lemmatisation, and machine translation to improve full-text search results on *dúchas.ie*, the web interface to the Irish National Folklore Collection. Our focus is the Schools' Collection, a scanned manuscript collection which is being transcribed by members of the public via a crowdsourcing initiative. We show that by applying these technologies to the manuscript page transcriptions, we obtain substantial improvements in search engine recall over a test set of actual user queries, with no appreciable drop in precision. Our results motivate the inclusion of this language technology in the search infrastructure of this folklore resource.

1 Background

This research is motivated by an objective to improve access to the Irish *National Folklore Collection*, one of the largest collections of folklore in Europe, and a collection that contains material in both official languages of Ireland, Irish and English. Our proposition is that the full-text search facility available on the collection's website, *dúchas.ie*,[2] can be enhanced by introducing language technology options. By demonstrating that language standardisation, demutation, lemmatisation, and machine translation technologies can improve information retrieval on the website, this paper supports this proposition and motivates the inclusion of these technologies in the search infrastructure.

1.1 The National Folklore Collection

The Irish *National Folklore Collection* (NFC) is a large archive of folkloristic material collected in Ireland mostly during the 20th century, to which material is still being added. Part of the collection was inscribed into the UNESCO *Memory of the World Register* in September 2017.[3] The NFC, which is located in University College Dublin (UCD), aims to collect, preserve and disseminate the oral tradition of Ireland.[4]

In 2012, the *Dúchas* project was established to digitise the collections of the NFC and publish them online. This project is a partnership

[2] https://www.duchas.ie/en/
[3] https://en.unesco.org/programme/mow/register
[4] http://www.ucd.ie/irishfolklore/en/

between Dublin City University (DCU) and UCD (Ó Cleircín et al., 2014). The first major NFC collection digitised under the Dúchas project was the *Schools' Collection*. The Schools' Collection is a large collection of folklore stories collected from school children throughout Ireland between 1937 and 1939, as part of a state-sponsored scheme (Ó Catháin, 1988). The collection comprises approximately 740,000 manuscript pages.

Approximately 440,000 pages of the collection were digitised, manually indexed, and made available online on *dúchas.ie* between 2013–16. About 79% of the stories on these pages are in English (348,822 stories) and about 21% are in Irish (95,511 stories). These stories are enriched with various browsable metadata, e.g. title/excerpt, collector, informant, location, language.

Stories in the collection are also indexed by topic. The first part of this work was done during the initial field work. The *Irish Folklore Commission* prepared a guide for teachers participating in the scheme, and this guide was published as a handbook entitled *Irish Folklore and Tradition* (Irish Folklore Commission, 1937). This handbook contained 55 topic headings (e.g. *a collection of riddles*, *local cures*, *the potato-crop*, *festival customs*), and teachers were instructed to collect material under these headings. Following the initial field work, researchers in the Irish Folklore Commission produced a topic-based index of the stories collected as part of the scheme. This paper-based index was based primarily on the 55 general topics from the handbook, but a large number of more specific topic headings (e.g. *Fionn Mac Cumhaill*, *1798*, *warts*) were added, bringing the total number of topics up to *c.*1,700. DCU digitised this index in 2014.

In 2016, using the digitised index and the story titles/excerpts indexed under the Dúchas project, and guided by the *MoTIF Pilot Thesaurus of Irish Folklore* (Ryan, 2015) developed by the Digital Repository of Ireland and the National Library of Ireland, DCU produced a shorter list of 208 standardised topic headings (e.g. *riddles*, *folk medicine*, *potatoes*, *events*), and mapped them to the Schools' Collection stories, resulting in a new index for the Schools' Collection on *dúchas.ie*. This index is a mixture of broad headings (e.g. *supernatural and legendary beings*, *events*, *folk medicine*) and narrow headings (e.g. *banshees*, *Halloween*, *whooping-cough*).

In 2014, to facilitate full-text search of the collection, DCU initiated a project to crowdsource transcriptions of the Schools' Collection manuscript pages using a custom-built web-based application open to anyone (Bhreathnach et al., 2019). This project was conceived, in part, because of the problems associated with performing optical character recognition on the pages of the collection, which contain a mix of handwriting styles, a mix of scripts (i.e. Latin and Insular Celtic), and a mix of languages (i.e. Irish English and prestandard Irish).

The project has been a strong success with uptake amongst members of the public, students and folklore scholars. The transcriptions are generally of good quality and usable, in particular the Irish ones. Light editing is sometimes carried out, but bad transcriptions are rejected. Currently 49% of the English pages (*c.*170,000) and 32% of the Irish pages (*c.*31,000) have been transcribed.[5]

The *dúchas.ie* website handles around 35,000 queries per month including around 16,000 full-text searches of the Schools' Collection transcriptions. All of these queries are logged in a database.

1.2 Irish standardisation

The orthography and grammar of the Irish language was standardised in the middle of the last century with the introduction of the *Caighdeán Oifigiúil* ('Official Standard'), first published in 1958 and revised in 2012 and 2017. Today, the standard form of the language is taught in schools, is used in all modern Irish dictionaries (both print and online), and has been almost universally adopted by the Irish-speaking public when writing the language. The spelling simplifications introduced by the standard cause occasional problems for Irish language technology, such as pairs of words that were distinguished in older orthographies which collapse to the same spelling in the standard (e.g. *fiadhach* ('hunting') and *fiach* ('a raven' or 'a

[5] https://www.duchas.ie/en/meitheal/

debt'), both written *fiach* in the standard, or *bádh* ('a bay'), *báidh* ('sympathy, liking'), and *bádhadh* ('drowning'), all three written simply *bá* in the standard). Nevertheless, the overall effect of increased consistency and predictability arising from near-universal adoption of the standard has been a tremendous positive for the development of Irish language technology. It does mean, however, that taggers, parsers, and other NLP (natural language processing) tools developed for processing the standard form of the language fail badly when applied to prestandard texts. In the context of the current paper, the disconnect between the prestandard and standard orthographies makes it extremely difficult for users raised on the standard language to search corpora of prestandard texts such as the Schools' Collection.

An Caighdeánaitheoir ('The Standardiser') is an open source software package for standardising Irish texts, first developed around 2006, and detailed in (Scannell, 2014). The software treats standardisation as a machine translation (MT) problem between closely-related languages, employing a hybrid rule-based and statistical model trained on a large corpus of parallel prestandard and standardised texts (including, for example, many important novels and autobiographies first published in the 1920's and 1930's and manually standardised for a modern readership in recent years). It also attempts to correct misspellings before carrying out standardisation. The standardiser has been deployed by two important lexicographical projects in Ireland, the New English-Irish Dictionary project,[6] and the Royal Irish Academy's *Foclóir Stairiúil na Gaeilge* (Uí Dhonnchadha et al., 2014) to help both lexicographers and end-users search their corpora of prestandard texts more effectively.

1.3 Machine translation for Irish

While the 2012 META-NET study on Irish language technology resources (Judge et al., 2012) deemed Irish-language machine translation as being weak or not supported, in recent years we have witnessed some development in this field. Dowling et al. (2015) describe the development of MT for use within the translation workflow of an Irish government department, while Arcan et al. (2016) provide results of building a more general domain Irish MT system. Most recently, Dowling et al. (2018) compare the two main MT paradigms: statistical machine translation (SMT) and neural machine translation (NMT). Their results indicate that with a low-resourced language such as Irish, particularly when paired with a language that differs in terms of sentence structure and morphological richness, SMT may provide better results.

Building on this knowledge, we choose to duplicate the SMT system described by Dowling et al. for use in our experiment here. We train a phrase-based SMT system using Moses (Koehn et al., 2007), incorporating hierarchical reordering tables (Galley and Manning, 2008) in an attempt to address the divergent sentence structures (verb-subject-object in Irish, subject-verb-object in English). Our translation model is trained using the same 108,796 sentences of parallel data as in Dowling et al. (2018), coming from a variety of sources such as the Department of Culture, Heritage and the Gaeltacht, Conradh na Gaeilge and Citizens Information. We build a 6-gram, rather than the traditional 3-gram, language model using KenLM (Heafield, 2011) which also aims to reduce any negative impact of the divergent word orders. Our datasets are preprocessed – sentence-tokenised, removal of blank lines, tokenisation of punctuation and truecased – before being translated by the MT system.

2 Methodology

Our principal aim in this paper is to investigate the effectiveness of standardisation and machine translation on the performance of the full-text search engine on *dúchas.ie*. We therefore cast this as an Information Retrieval (IR) problem. To this end, we set up experiments which make use of actual search queries submitted by users of the website, and aim to measure precision and recall of the search engine under various experimental conditions. Lacking a gold-standard corpus of relevant/non-relevant documents with which to measure precision and recall, we instead make use of a subset of the 208 topic labels, described in Section 1.1, attached to the stories in the

[6] https://www.focloir.ie/

Schools' Collection to create our own test corpus, as follows.

First, we manually examined the top 10,000 search queries submitted to *dúchas.ie* to date, keeping only the Irish language queries (just over 1,500 of the 10,000). Each of these was then manually compared with the list of 208 standardised topic headings, and we found 172 queries which clearly corresponded to some topic on the list; for instance, the topic *Christmas* was matched to the five queries "An Nollaig", "NOLLAIG", "Nollag", "nodlag", and "nodlaig". These 172 queries were matched to a total of 67 topics. We further restricted to those topics for which there were at least 100 transcribed stories in both Irish and English, leaving just 20 topics (*Riddles*, *Jokes*, *Fairy forts*, *The Great Famine*, *Entertainments and recreational activities*, *Folk medicine*, *Folk poetry*, *Food products*, *Religious tales*, *Clothing and accessories*, *Fianna*, *Feast of St Brigid*, *May*, *Halloween*, *Christmas*, *Prayers*, *Proverbs*, *Hardship*, *Graveyards*, *Potatoes*) and 72 of the original 172 queries. For each remaining topic, we randomly selected 100 Irish transcriptions and 100 English transcriptions from that topic in order to produce two test corpora of 2,000 documents each for our IR experiments. Finally, the English transcriptions were machine translated into Irish using the system described in Section 1.3 to allow us to evaluate the effectiveness of Irish language search queries for retrieving relevant English language documents.

The experiments differ only in the preprocessing that was applied to the documents in the test corpora before indexing and to the search queries before searching. The four experimental conditions are as follows:

- **Baseline**: For this experiment, all text was converted to lowercase, and Irish diacritics (á,é,í,ó,ú) are converted to ASCII (a,e,i,o,u). This setup is the default behavior of the existing *dúchas.ie* search engine as of May 2019.
- **Standardised**: For this experiment, the texts were standardised using the software described in Section 1.2, *An Caighdeánaitheoir*, and then lowercased and converted to ASCII as in the Baseline experiment.
- **Demutated**: Irish words are subject to so-called *initial mutations* which are triggered in certain semantic and syntactic environments and which cause the words to appear with different initial sounds. With rare exceptions, mutations in Irish can be detected and removed algorithmically in a trivial way because they are transparently reflected in the orthography (as is the case for Scottish Gaelic, but not for the other Celtic languages: Manx Gaelic, Welsh, Breton, and Cornish). For this experiment, the texts were standardised, lowercased, converted to ASCII, and finally all initial mutations were removed.
- **Lemmatised**: Irish nouns and adjectives are inflected according to their case and number, and verbs are inflected according to tense, mood, person, and number. For this experiment, the texts were standardised, lowercased, converted to ASCII, and finally lemmatised by means of a lemmatiser which is part of the open source Irish grammar checker *An Gramadóir*.[7] In the case of verbs, the lemmatised form is the singular imperative, which is the usual citation form in modern Irish dictionaries.

For each experimental setup and for both test corpora, we preprocess the 2,000 documents in the corpus and the 72 search engine queries in our final list according to one of the schemes above. We then perform a full text search on the corpus for each of the 72 queries, recording the total number of returned documents and the number of those deemed to be "relevant" (here, by definition, relevant documents are those labeled with the topic corresponding to the given search query). For example, consider the query "An Nollaig". This search is typical of many others in that it conforms to the standard orthography most familiar to users of the site, and as a consequence it returns no hits at all in the Baseline experiment because the texts in the test corpus where this phrase appears all use the prestandard spelling "An Nodlaig". On the other

[7] https://www.cadhan.com/gramadoir/

hand, in the Standardised experiment, the same query returns 21 documents from the test set, with 19 of the 21 carrying the correct label "Christmas". This yields a precision of 19/21 ≈ 0.90 and a recall of 0.19 (19 of the 100 relevant documents in the corpus) for this one query.

3 Results and discussion

In Table 1 we report Precision, Recall, and F-scores for each of the experiments described in the preceding section as applied to the corpus of Irish transcriptions, totalled over all 72 search queries. Table 2 provides the analogous results for the experiments applied to the corpus of machine-translated English transcriptions. Since the MT engine produces standard Irish as its output, we do not report results for the Standardised setup separately in Table 2 since those results are the same as for the Baseline.

Experiment	P	R	F
Baseline	0.67	0.10	0.17
Standardised	0.69	0.24	0.36
Demutated	0.70	0.29	0.41
Lemmatised	0.67	0.34	0.45

Table 1. Precision/recall results – Irish transcriptions.

Experiment	P	R	F
Translated + Baseline	0.59	0.15	0.24
" + Demutated	0.59	0.17	0.26
" + Lemmatised	0.60	0.21	0.31

Table 2. Precision/recall results – English transcriptions machine-translated to Irish.

Because the experimental setup is somewhat artificial, the absolute precision and recall values are not of great importance, but we do note in Table 1 a significant increase in recall over the baseline with the introduction of standardisation, and further increases with demutation and lemmatisation. These increases occur without any large decrease in precision.

In looking more closely at the results, a few things stand out. First, even with full standardisation and lemmatisation, a recall score of 0.34 seems low. This is due in part to the quality of some of the experimental search queries. For instance, a few of the 72 queries returned no results at all under any of the experiments. In one case, a search term was misspelled beyond the ability of the standardiser to correct it ("Oiche Shanmhna" for 'Halloween', correctly spelled "Oíche Shamhna" in the standard orthography). In another case, we generously interpreted the search query "ocras mór" (lit. 'great hunger') as an attempt to retrieve documents about the Irish potato famine, which indeed is sometimes referred to in English as "the Great Hunger", but for which the correct Irish is "(an) *Gorta* Mór" (lit. '(the) great famine'). Another class of low-recall queries appear to be searches for the topics themselves as opposed to searches for terms likely to appear in the full text transcriptions; e.g. the queries "filíocht" ('poetry'), "seanfhocail" ('proverbs'), "tomhaiseanna" ('riddles') all have very low recall values under all of the experimental conditions. This could be overcome by including the topic or other metadata in the search index.

The results in Table 2 show a similar increase in recall and no corresponding decrease in precision for the machine-translated English transcriptions. It is not surprising that the scores are somewhat lower than the ones in Table 1 since the machine translation engine is far from perfect and certainly introduces some noise into the process. On the other hand, the true baseline in this case is a recall of 0.0, since the current *dúchas.ie* search engine does not support retrieval of English transcriptions via Irish queries. We are therefore encouraged by these results, especially given that they were achieved through relatively straightforward use of existing language technologies.

4 Conclusion

We set out to improve full-text search results on *dúchas.ie* using language technology, building on crowdsourced transcriptions of folklore manuscripts. We have gathered together a set of existing language technologies to achieve this goal. These include tools to standardise, demutate, lemmatise, and translate the

transcriptions of these folklore stories. We have shown that the introduction of these technologies can substantially improve search engine recall over a test set of actual user queries, with no appreciable drop in precision.

Motivated by these results, these technologies will be deployed in the search infrastructure on *dúchas.ie*. We envisage that standardisation and machine translation will be applied by default, as the query logs show that the website users tend to search using standard spellings. Demutation and lemmatisation will likely be optional, and raw searches will still be possible, however, exact implementation has yet to be specified. Search spelling suggestions using a standard lexicon and spelling distance algorithm could also be added.

A secondary outcome of this research was a list of common errors in the crowdsourced transcriptions. These terms were identified as errors by virtue of them not being present in a large corpus of texts from the period (Uí Dhonnchadha et al., 2014). We have categorised these errors, and they mostly involve (1) accented characters, (2) missing or spurious lenition (i.e. an orthographic 'h' following the initial consonant indicating phonetic weakening or deletion of the initial consonant (Welby et al., 2017)), or (3) the disordering of the letters 'iu' or 'iú'. This information will be used to improve instructions given to transcribers.

Lastly, other collections in the NFC being digitised by the Dúchas project include another large manuscript collection and a large audio collection. If these collections are made available for transcription by members of the public, and if such efforts are as successful as previous efforts, access to these collections could be improved using the language technologies tested in the paper.

Acknowledgements

The *Dúchas* project is funded by the Department of Culture, Heritage and the Gaeltacht with support from the National Lottery, University College Dublin, and the National Folklore Foundation, Ireland.

References

Arcan, Mihael, Caoilfhionn Lane, Eoin Ó Droighneáin, and Paul Buitelaar. 2016. IRIS: English-Irish Machine Translation System. *The International Conference on Language Resources and Evaluation*, Portorož, Slovenia.

Bhreathnach, Úna, Ciarán Mac Murchaidh, Gearóid Ó Cleircín, and Brian Ó Raghallaigh. 2019. Ní hualach do dhuine an léann: meithleacha pobail i ngort na Gaeilge. *Léachtaí Cholm Cille*, Maynooth, Ireland.

Dowling, Meghan, Teresa Lynn, Alberto Poncelas, and Andy Way. 2018. SMT versus NMT: Preliminary Comparisons for Irish. *Proceedings of the Workshop on Technologies for MT of Low Resource Languages*, Boston, MA, 12–20.

Dowling, Meghan, Teresa Lynn, Yvette Graham, and John Judge. 2016. English to Irish Machine Translation with Automatic Post-Editing. *2nd Celtic Language Technology Workshop*, Paris, France.

Dowling, Meghan, Lauren Cassidy, Eimear Maguire, Teresa Lynn, Ankit Srivastava, and John Judge. 2015. Tapadóir: Developing a statistical machine translation engine and associated resources for Irish. *4th LRL Workshop: Language Technologies in support of Less-Resourced Languages*, Poznan, Poland.

Galley, Michel, and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 848–856.

Heafield, Ken. 2011. KenLM: Faster and smaller language model queries. *Proceedings of the 6th Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 187–197.

Irish Folklore Commission. 1937. *Irish Folklore and Tradition*. Department of Education, Dublin, Ireland.

Judge, John, Ailbhe Ní Chasaide, Rose Ní Dhubhda, Kevin P. Scannell, and Elaine Uí Dhonnchadha. 2012. The Irish Language in the Digital Age. *META-NET White Paper Series: Europe's Languages in the Digital Age.* Springer.

Ó Cathain, Séamas. 1988. Súil siar ar Scéim na Scol 1937-1938. *Sinsear*, 5:19–30.

Ó Cleircín, Gearóid, Anna Bale, and Brian Ó Raghallaigh. 2014. Dúchas.ie: ré nua i stair Chnuasach Bhéaloideas Éireann. *Béaloideas*, 82:85–99.

Ryan, Catherine. 2015. *MoTIF: Thesaurus Construction Guidelines*. Digital Repository of Ireland, Dublin, Ireland.

Scannell, Kevin. 2014. Statistical models for text normalization and machine translation. *Proceedings of the 1st Celtic Language Technology Workshop*, Dublin, Ireland, 33–40.

Uí Dhonnchadha, Elaine, Kevin Scannell, Ruairí Ó hUiginn, Eilís Ní Mhearraí, Máire Nic Mhaoláin, Brian Ó Raghallaigh, Gregory Toner, Séamus Mac Mathúna, Déirdre D'Auria, Eithne Ní Ghallchobhair, and Niall O'Leary. 2014. Corpas na Gaeilge (1882-1926): Integrating Historical and Modern Irish Texts. *Proceedings of the Workshop on Language Resources and Technologies for Processing and Linking Historical Documents and Archives – Deploying Linked Open Data in Cultural Heritage*, Reykjavik, Iceland, 12–18.

Welby, Pauline, Máire Ní Chiosáin, and Brian Ó Raghallaigh. 2017. Total eclipse of the heart? The production of eclipsis in two speaking styles of Irish. *Journal of the International Phonetic Association*, 47(2):125–153.