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

Brian Ó Raghallaigh, Kevin Scannell, Meghan Dowling (CLTW 2019)

Abstract

In this paper, we measure the effectiveness of using language standardisation, demutation, 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.

The National Folklore Collection (dúchas.ie)

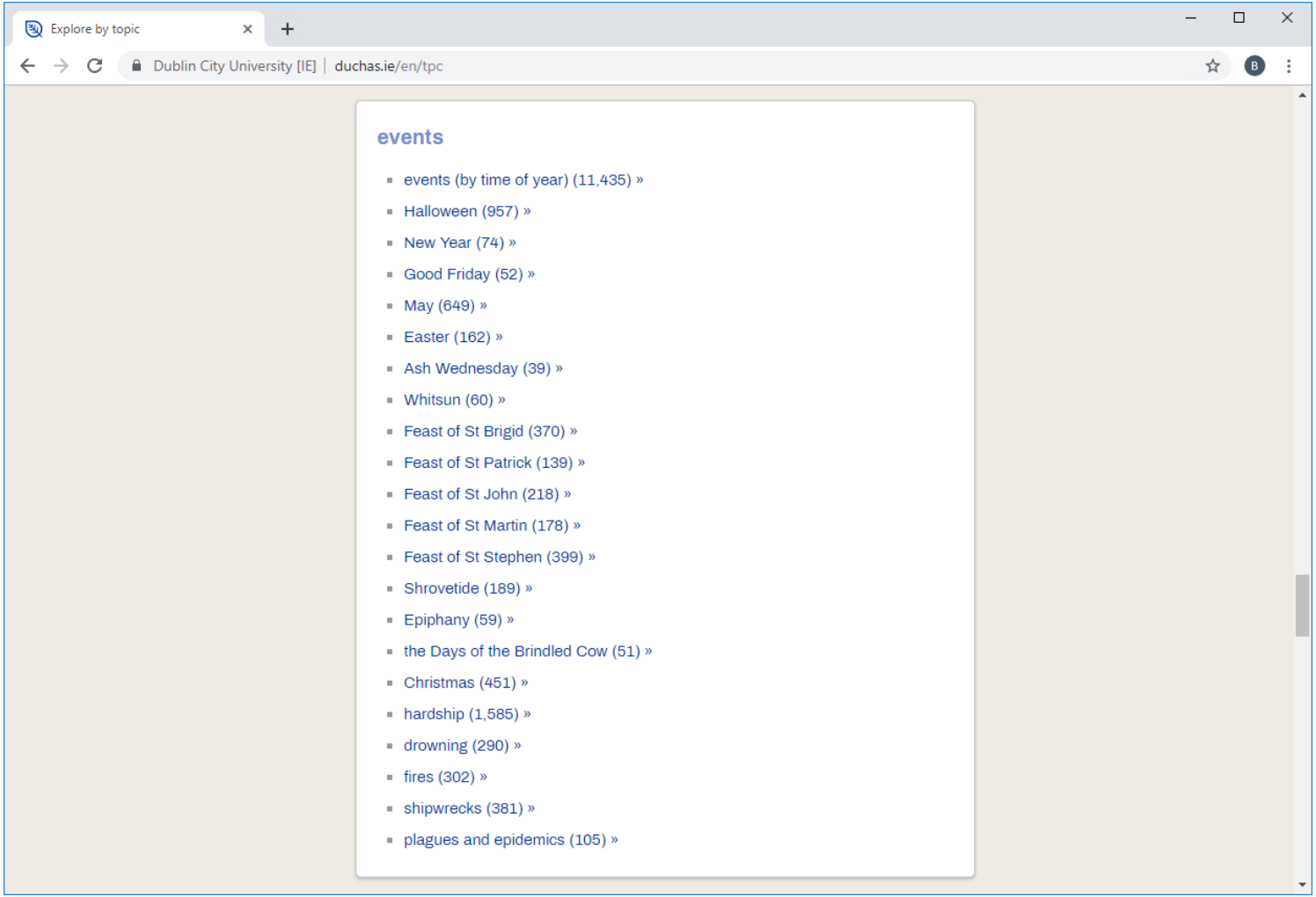
- One of the largest collections of folklore in Europe.
- Material in Irish and English.
- Collected in Ireland mostly during the 20th century.
- Part of the collection was inscribed into the UNESCO *Memory of the World Register* in 2017.
- Located in University College Dublin (UCD).
- Its aim is to collect, preserve and disseminate the oral tradition of Ireland.
- The *Dúchas* project was established in 2012 to digitise the collections and publish them online.
- Dúchas is a collaboration between UCD and DCU.

The Schools' Collection

- A large collection of folklore stories collected from the school children throughout Ireland between 1937 and 1939 as part of a state-sponsored scheme.
- 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.

A topics index

- Stories in the *Schools' Collection* are also indexed by topic.
- Indexing was carried out by the collectors and by researchers in the *Irish Folklore Commission* (IFC).
- 55 general topic headings (e.g. *a collection of riddles*, *local cures*, *the potato-crop*, *festival customs*) from the IFC's handbook entitled *Irish Folklore and Tradition*.
- *c.*1,650 more specific topic headings (e.g. *Fionn Mac Cumhaill*, *1798*, *warts*).
- Digitised by DCU in 2014.
- Further refined in 2016 in conjunction with the DRI to 208 standardised topic headings (e.g. *riddles*, *folk medicine*, *potatoes*, *events*) and mapped to stories in the collection.
- 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*).

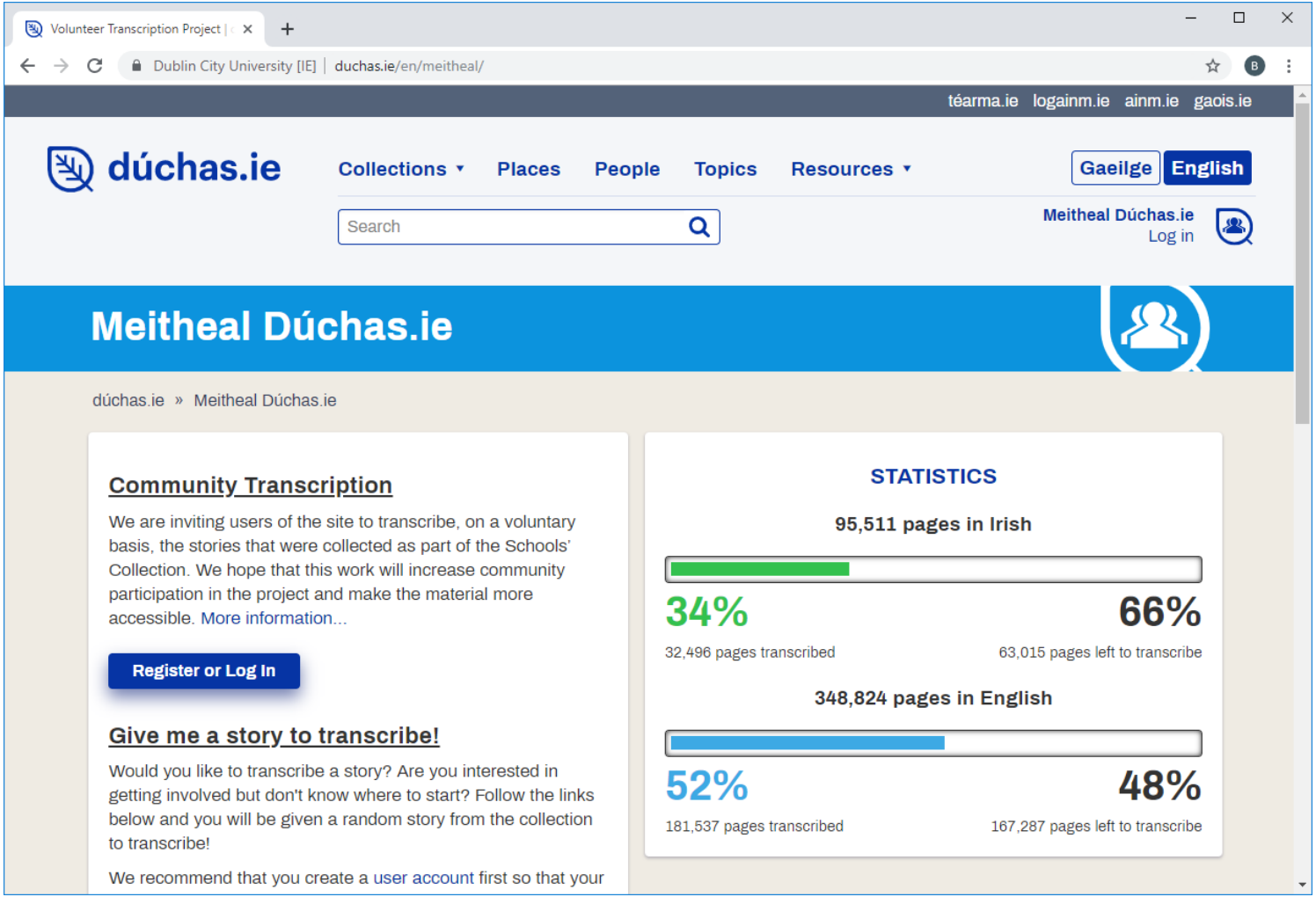


events

- events (by time of year) (11,435) »
- Halloween (957) »
- New Year (74) »
- Good Friday (52) »
- May (649) »
- Easter (162) »
- Ash Wednesday (39) »
- Whitsun (60) »
- Feast of St Brigid (370) »
- Feast of St Patrick (139) »
- Feast of St John (218) »
- Feast of St Martin (178) »
- Feast of St Stephen (399) »
- Shrovetide (189) »
- Epiphany (59) »
- the Days of the Brindled Cow (51) »
- Christmas (451) »
- hardship (1,585) »
- drowning (290) »
- fires (302) »
- shipwrecks (381) »
- plagues and epidemics (105) »

Crowdsourced transcriptions

- To facilitate full-text search of the collection, DCU initiated a project to crowdsource transcription of the Schools' Collection manuscript pages using a custom-built web-based application open to anyone.
- Conceived in part because of problems associated with 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 success (See: next slide).
- The *dúchas.ie* website now handles around 35,000 queries per month including around 16,000 full-text searches of the Schools' Collection transcriptions.


Volunteer Transcription Project
Dublin City University [IE] | duchas.ie/en/meitheal/
téarma.ie logainm.ie ainm.ie gaois.ie
dúchas.ie
Collections
Places
People
Topics
Resources
Gaeilge
English
Search
Meitheal Dúchas.ie
Log in
Meitheal Dúchas.ie
dúchas.ie » Meitheal Dúchas.ie
Community Transcription
We are inviting users of the site to transcribe, on a voluntary basis, the stories that were collected as part of the Schools' Collection. We hope that this work will increase community participation in the project and make the material more accessible. More information...
Register or Log In
Give me a story to transcribe!
Would you like to transcribe a story? Are you interested in getting involved but don't know where to start? Follow the links below and you will be given a random story from the collection to transcribe!
We recommend that you create a user account first so that your
STATISTICS
95,511 pages in Irish
34%
66%
32,496 pages transcribed
63,015 pages left to transcribe
348,824 pages in English
52%
48%
181,537 pages transcribed
167,287 pages left to transcribe

Information retrieval on dúchas.ie

- A choice between browse based on metadata added manually by trained experts and full-text search of stories hand-written in 1930s and transcribed by untrained transcribers today.
- Browse is more precise, but slower and more restrictive.
- Search is simpler and preferred by users, but limited in a number of ways.
- How to improve full-text search?
  - Language technology.
- How to test value of improvements?
  - The topics index.

Irish standardisation

- Written standard for Irish was established in the 1950's: An Caighdeán Oifigiúil (Official Standard).
- Irish standardiser originally developed in 2006 to support lexicographical work.
- Used by the New English-Irish Dictionary and Royal Irish Academy's historical dictionary.
- Views standardisation as a machine translation problem, pre-standard to standard Irish.
- Simple statistical model suffices in this case (essentially IBM model 1).
- Notable challenges in our case: errors in source texts (misplaced fadas) and transcription errors.

Machine translation for Irish

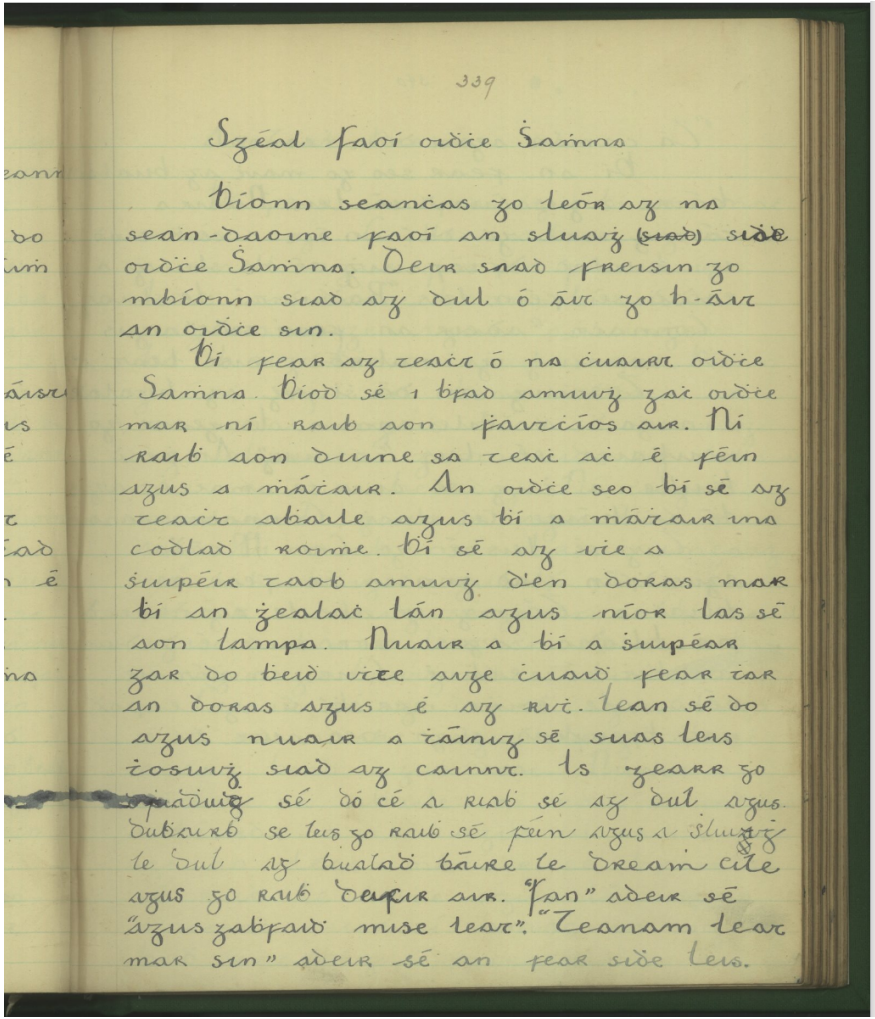
- Moses-based statistical machine translation.
- 6-gram language model.
- Hierarchical reordering tables.
- ~108,000 parallel sentences of EN-GA.
- Text preprocessed (tokenised, truecased, sentence length limited).

Methodology

- Interested in improving the search engine, so we cast this as an Information Retrieval problem.
- Used actual search engine logs to select 72 Irish language queries manually matched to 20 topics.
- 100 English and 100 Irish transcriptions randomly selected from each topic.
- Four experiments run on Irish texts, varying the preprocessing applied to texts and queries:
  - Baseline: convert to ASCII (áéíóú > aeiou) and lowercase.
  - Standardised: apply the Irish standardiser followed by the baseline.
  - Demutated: additionally, remove Irish initial mutations (bhean > bean, t-uisce > uisce).
  - Lemmatised: replace nouns, verbs, and adjectives with their lemma (aimsíonn > aimsigh, mná > bean).
- Three experiments run on English texts: translated + baseline, demutated, lemmatised.

Example

- Topic: Christmas.
- Queries: “An Nollaig”, “NOLLAIG”, “Nollag”, “nodlag”, “nodlaig”.
- Texts: “An Nodlaig”.



339

Scéal faoi oíche Samhna

Bíonn seanchas go leór ag na sean-daoine faoi an sluaigh (siad) sídhe oíche Samhna. Deir siad freisin go mbíonn siad ag dul ó áit go h-áit an oíche sin.

Bí fear ag teacht ó na cuairt oíche Samhna. Bíodh sé i bhfad amuigh gach oíche mar ní raibh aon fhaitchíos air. Ní raibh aon duine sa teach ach é féin agus a mháthair. An oíche seo bí sé ag teacht abhaile agus bí a mháthair ina codladh roimhe. Bí sé ag ithe a shuipéir taobh amuigh d'en doras mar bí an ghealach lán agus níor las sé aon lampa. Nuair a bí a shuipéar gar do bheidh ithte aige chuaidh fear thar an doras agus é ag rith. Lean sé do agus nuair a tháinig sé suas leis thosuigh siad ag cainnt. Is gearr go d'fhiadhuigh sé dó cé a riabh sé ag dúl agus dubhairbh se leis go raibh sé féin agus a shluigh le dul ag bhualadh báire le dreamh eile agus go raibh deirfir air. "Fan" adeir sé "agus gabhfaidh mise leat". "Teanam leat mar sin" adeir sé an fear sídhe leis.

Scéal faoi Oíche Shamhna

Scaip Tvuít

Sgéal faoí Oidhche Shamhna
Bíonn seanchas go leór ag na sean-daoine faoí an sluagh sidhe oidhche Shamhna. Deir siad freisin go mbíonn siad ag dul ó áit go h-áit an oidhche sin.
Bhí fear ag teacht ó na chuairt oidhche Samhna. Bhíodh sé i bhfad amuigh gach oidche mar ní raibh aon fhaitchíos air. Ní raibh aon duine sa teach ach é féin agus a mháthair. An oidhthe seo bhí sé ag teacht abhaile agus bhí a mháthair ina codladh roimhe. Bhí sé ag ithe a shuipéir taobh amuigh d'en doras mar bhí an ghealach lán agus níor las sé aon lampa. Nuair a bhí a shuipéar gar do bheidh ithte aige chuaidh fear thar an doras agus é ag rith. Lean sé do agus nuair a tháinig sé suas leis thosuigh siad ag cainnt. Is gearr go d'fhiadhuigh sé dó cé a riabh sé ag dúl agus dubhairbh se leis go raibh sé féin agus a shluigh le dul ag bhualadh báire le dreamh eile agus go raibh deirfir air. "Fan" adeir sé "agus gabhfaidh mise leat". "Teanam leat mar sin" adeir sé an fear sidhe leis.

Example: raw > baseline (lowercase ASCII)

Sgéal faoí Oidhche Shamhna Bíonn seanchas go leór ag na sean-daoine faoí an sluagh sidhe oidhche Shamhna. Deir siad freisin go mbíonn siad ag dul ó áit go h-áit an oidhche sin. Bhí fear ag teacht ó na chuairt oidhche Samhna. Bhíodh sé i bhfad amuigh gach oidche mar ní raibh aon fhaitchíos air. Ní raibh aon duine sa teach ach é féin agus a mháthair. An oidhthe seo bhí sé ag teacht abhaile agus bhí a mháthair ina codladh roimhe. Bhí sé ag ithe a shuipéir taobh amuigh d’en doras mar bhí an ghealach lán agus níor las sé aon lampa...

sgeal faoi oidhche shamhna bionn seanchas go leor ag na sean-daoine faoi an sluagh sidhe oidhche shamhna. deir siad freisin go mbionn siad ag dul o ait go h-ait an oidhche sin. bhi fear ag teacht o na chuairt oidhche samhna. bhiodh se i bhfad amuigh gach oidche mar ni raibh aon fhaitchios air. ni raibh aon duine sa teach ach e fein agus a mhathair. an oidhthe seo bhi se ag teacht abhaile agus bhi a mhathair ina codladh roimhe. bhi se ag ithe a shuipeir taobh amuigh d'en doras mar bhi an ghealach lan agus nior las se aon lampa...

Example: baseline > standard

sgeal faoi **oidhche shamhna** bionn seanchas go leor ag na sean-daoine faoi an sluagh sidhe **oidhche shamhna**. deir siad freisin go mbionn siad ag dul o ait go h-ait an oidhche sin. bhi fear ag teacht o na chuairt oidhche samhna. bhiodh se i bhfad amuigh gach oidche mar ni raibh aon fhaitchios air. ni raibh aon duine sa teach ach e fein agus a mhathair. an **oidhthe** seo bhi se ag teacht abhaile agus bhi a mhathair ina codladh roimhe. bhi se ag ithe a shuipeir taobh amuigh d'en doras mar bhi an ghealach lan agus nior las se aon lampa...

sceal faoi **oiche shamhna** bionn seanchas go leor ag na seandaoine faoi an slua si **oiche shamhna**. deir siad freisin go mbionn siad ag dul o ait go hait an oiche sin. bhi fear ag teacht o na chuairt oiche samhna. bhiodh se i bhfad amuigh gach oiche mar ni raibh aon fhaitios air. ni raibh aon duine sa teach ach e fein agus a mhathair. an **oidhthe** seo bhi se ag teacht abhaile agus bhi a mhathair ina codladh roimhe. bhi se ag ithe a shuipeir taobh amuigh den doras mar bhi an ghealach lan agus nior las se aon lampa...

Example: standard > demutated

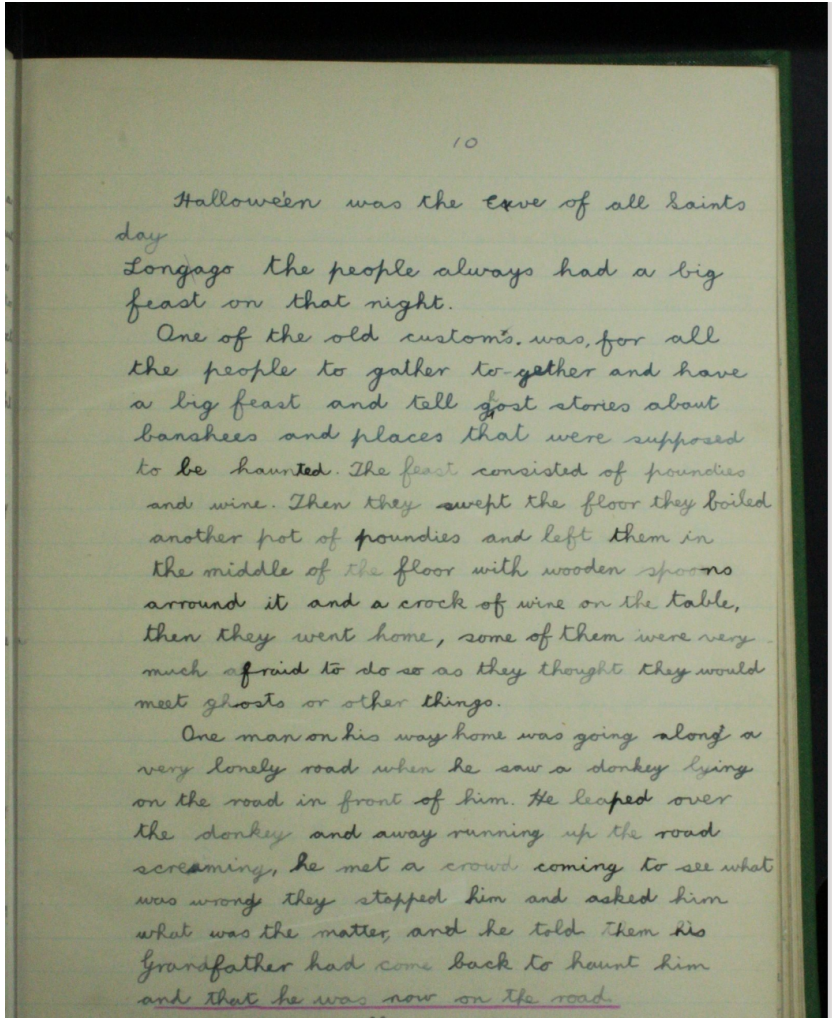
sceal faoi oiche shamhna bionn seanchas go leor ag na seandaoine faoi an slua si oiche shamhna. deir siad freisin go mbionn siad ag dul o ait go hait an oiche sin. bhi fear ag teacht o na chuairt oiche samhna. bhiodh se i bhfad amuigh gach oiche mar ni raibh aon fhaitios air. ni raibh aon duine sa teach ach e fein agus a mhathair. an oidhthe seo bhi se ag teacht abhaile agus bhi a mhathair ina codladh roimhe. bhi se ag ithe a shuipeir taobh amuigh den doras mar bhi an ghealach lan agus nior las se aon lampa...

sceal faoi oiche samhna bionn seanchas go leor ag na seandaoine faoi an slua si oiche samhna. deir siad freisin go bionn siad ag dul o ait go ait an oiche sin. bi fear ag teacht o na cuairt oiche samhna. biodh se i fad amuigh gach oiche mar ni raibh aon faitios air. ni raibh aon duine sa teach ach e fein agus a mathair. an oidhthe seo bi se ag teacht abhaile agus bi a mathair ina codladh roimhe. bi se ag ithe a suipeir taobh amuigh den doras mar bi an gealach lan agus nior las se aon lampa...

Example: demutated > lemmatised

sceal faoi oiche samhna bionn seanchas go leor ag na seandaoine faoi an slua si oiche samhna. deir siad freisin go bionn siad ag dul o ait go ait an oiche sin. bi fear ag teacht o na cuairt oiche samhna. biodh se i fad amuigh gach oiche mar ni raibh aon faitios air. ni raibh aon duine sa teach ach e fein agus a mathair. an oidhthe seo bi se ag teacht abhaile agus bi a mathair ina codladh roimhe. bi se ag ithe a suipeir taobh amuigh den doras mar bi an gealach lan agus nior las se aon lampa...

sceal faoi oiche samhain bi seanchas go leor ag na seanduine faoi an slua si oiche samhain. abair siad freisin go bi siad ag dul o ait go ait an oiche sin. bi fear ag teacht o na cuairt oiche samhain. bi se i fad amuigh gach oiche mar ni bi aon faitios air. ni bi aon duine sa teach ach e fein agus a mathair. an oidhthe seo bi se ag teacht abhaile agus bi a mathair ina codladh roimhe. bi se ag ithe a suipear taobh amuigh den doras mar bi an gealach lan agus nior las se aon lampa.



10

Halloween was the Eave of all Saints day

Longago the people always had a big feast on that night.

One of the old customs was, for all the people to gather to gather and have a big feast and tell ghost stories about banshees and places that were supposed to be haunted. The feast consisted of poundies and wine. Then they swept the floor they boiled another pot of poundies and left them in the middle of the floor with wooden spoons arround it and a crock of wine on the table, then they went home, some of them were very much afraid to do so as they thought they would meet ghosts or other things.

One man on his way home was going along a very lonely road when he saw a donkey lying on the road in front of him. He leaped over the donkey and away running up the road screaming, he met a crowd coming to see what was wrong they stopped him and asked him what was the matter, and he told them his Grandfather had come back to haunt him and that he was now on the road

Halloween

Scaip Tvuít

Halloween was the eave of all Saints day.
Long ago the people always had a big feast on that night. One of the old customs was for all the people to gather together and have a big feast and tell ghost stories about banshees and places that were supposed to be haunted. The feast consisted of poundies and wine. Then they swept the floor they boiled another pot of poundies and left them in the middle of the floor with wooden spoons around it and a crock of wine on the table, then they went home, some of them were very much afraid to do so as they thought they would meet ghosts or other things.
One man on his way home was going along a very lonely road when he saw a donkey lying on the road in front of him. He leaped over the donkey and away running up the road screaming, he met a crowd coming to see what was wrong they stopped him and asked him what was the matter, and he told them his Grandfather had come back to haunt him and that he was now on the road

Tras-scríofa ag duine dár meitheal tras-scríbhneoirí deonacha.
Stair | Athraigh »

Example: English > translated

Halloween was the eave of all Saints day. Long ago the people always had a big feast on that night. One of the old customs was for all the people to gather together and have a big feast and tell ghost stories about banshees and places that were supposed to be haunted. The feast consisted of poundies and wine. Then they swept the floor they boiled another pot of poundies and left them in the middle of the floor with wooden spoons around it and a crock of wine on the table, then they went home, some of them were very much afraid to do so as they thought they would meet ghosts or other things.

Oíche Shamhna an eave de na Naomh lá . airde na daoine a bhí ag mór feast ar an oíche . ceann de na custaim agus do gach duine a bhailiú le chéile agus mór feast agus scéalta faoi banshees object name (optional agus go raibh a haunted . an feast ná poundies agus fíon . siad na han urlár siad boiled pot eile de poundies agus ar chlé orthu i lár an urlár le spoons adhmaid timpeall air agus crock fíona a chur ar an mbord , chuaigh siad baile , cuid acu nach raibh eagla ort sin a dhéanamh mar go mbeadh siad le chéile thaibhsí nó rudaí eile .

Results

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

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

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

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

Conclusion

- We have gathered together a set of existing language technologies to improve full-text search results on *dúchas.ie*.
- 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*.

Future Work

- Implement these language technologies within the full-text search on *dúchas.ie*:
  - Demutation and lemmatisation probably optional.
- Experiment with a more domain-specific MT system.
- Apply this work to future digitisation projects.

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.