Irish text standardization with applications to lexicography

Kevin Scannell
Saint Louis University

1 April 2017

An Caighdeán Oifigiúil

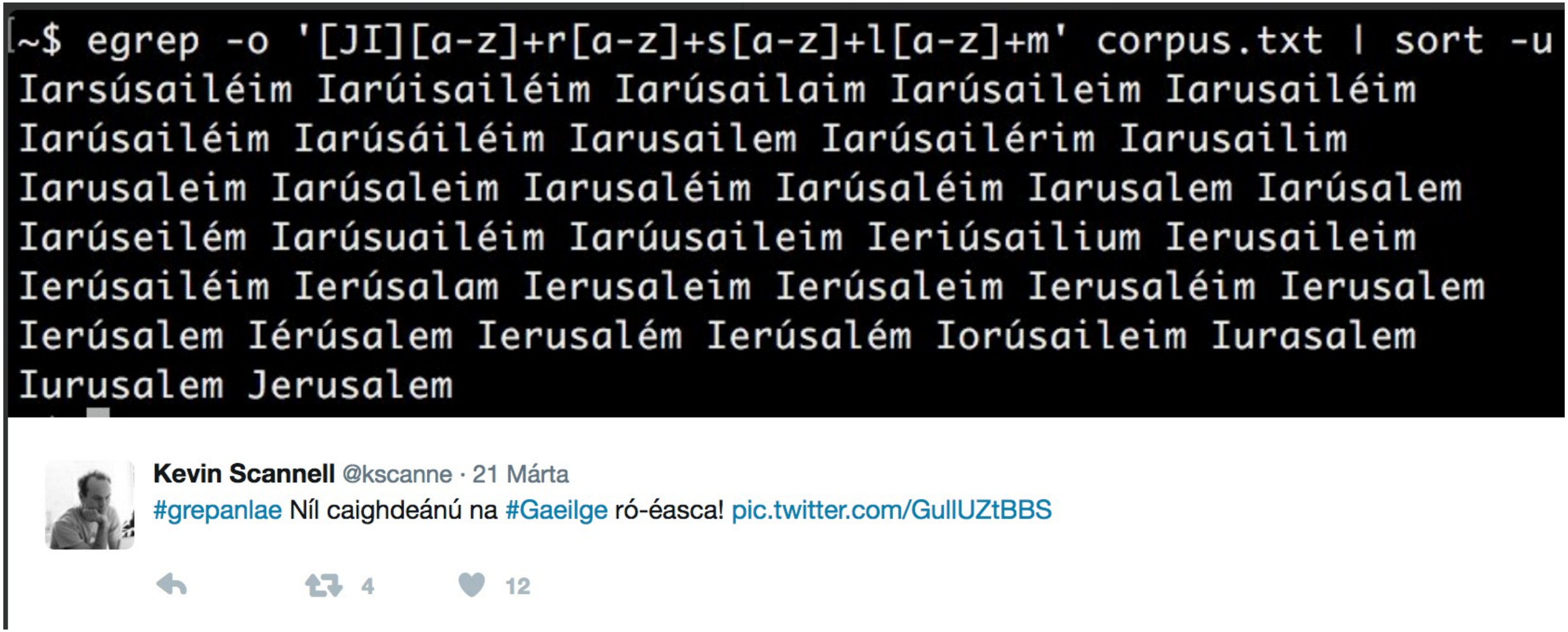
- The Official Standard
- Introduced in the 1940's and 1950's
- Simplified spelling and grammar
- Widely adopted, all domains and registers

Example

- Old: “Ní rabh 'sa dearbhughadh sin acht a chuid uchtaighe, eisean, a h-Aodh féin ag teacht na h-arraicis.”
- New: Ní raibh sa dearbhú sin ach a chuid uchtaí, eisean, a hAodh féin ag teacht ina haraicis.

The Problem

- Searching the web
- Searching corpus texts for lexicography
- New literary editions for modern readership
- Modern tech (taggers/parsers) on old texts
- Using old texts to train modern tech (language modeling, parallel corpora for translation)

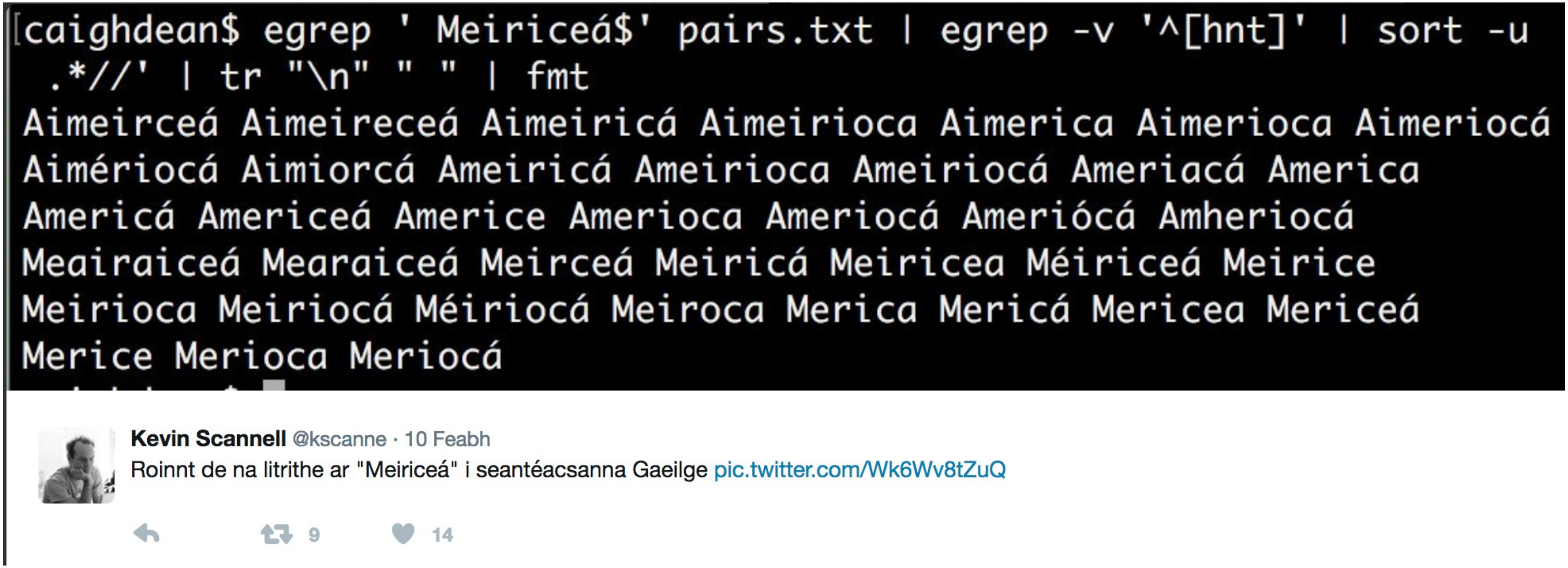


```
~$ egrep -o '[JI][a-z]+r[a-z]+s[a-z]+l[a-z]+m' corpus.txt | sort -u
Iarsúsailéim Iarúisailéim Iarúsailaim Iarúsaileim Iarusailéim
Iarúsailéim Iarúsáiléim Iarusailem Iarúsailérim Iarusailim
Iarusaleim Iarúsaleim Iarusaléim Iarúsaléim Iarusalem Iarúsalem
Iarúseilém Iarúsuailéim Iarúusaileim Ieriúsailium Ierusaileim
Ierúsailéim Ierúsalam Ierusaleim Ierúsaleim Ierusaléim Ierusalem
Ierúsalem Iérúsalem Ierusalém Ierúsalém Iorúsaileim Iurasalem
Iurusalem Jerusalem
```

Kevin Scannell @kscanne · 21 Márta

#grepanlae Níl caighdeánú na #Gaeilge ró-éasca! pic.twitter.com/GullUZtBBS

4 12



```
caighdean$ egrep ' Meiriceá$' pairs.txt | egrep -v '^[hnt]' | sort -u
 .*//' | tr "\n" " " | fmt
Aimeirceá Aimeireceá Aimeiricá Aimeirioca Aimerica Aimerioca Aimeriocá
Aimériocá Aimiorcá Ameiricá Ameirioca Ameiriocá Ameriacá America
Americá Americeá Americe Amerioca Ameriocá Ameriócá Amheriocá
Meairaiceá Mearaiceá Meirceá Meiricá Meiricea Méiriceá Meirice
Meirioca Meiriocá Méiriocá Meiroca Merica Mericá Mericea Mericeá
Merice Merioca Meriocá
```

Kevin Scannell @kscanne · 10 Feabh

Roinnt de na litrithe ar "Meiriceá" i seantéacsanna Gaeilge pic.twitter.com/Wk6Wv8tZuQ

9 14

A Solution

- Treat as a machine translation problem
- *Very* closely-related languages
- Statistical approach; so-called “IBM model 1”
- Allow limited reordering of phrases
- Train on a parallel corpus
- Allow rule-based spelling changes
- coimh-mheasguighthe→ comh-mheasguighthe→ cóimheasguighthe→ cóimheascuighthe→ cóimheascaighthe→ cóimheascaithe→ cóimheasctha

Parallel Corpus

- Many older texts from RIA corpus
- Modern versions from a variety of sources
- New English-Irish Dictionary Project
- New Corpus for Ireland (TCD)
- Scanning/OCR of modern books
- 98,000 segments, 1.96M words on each side

Which Caighdeán?

- “Cad is brí le 'caighdeánach' san aois iar-nua-aoiseach seo?”
- Target is “standard Irish”, whatever that means
- Data-driven approach (mar is gnáth)!
- Corpus of ~150M words of Irish 1990-2017
- But, almost *no* texts conform completely
- Statistical classifier: remove “worst offenders”
- Declare “standard Irish” to be what comes out of the language model trained on the remainder

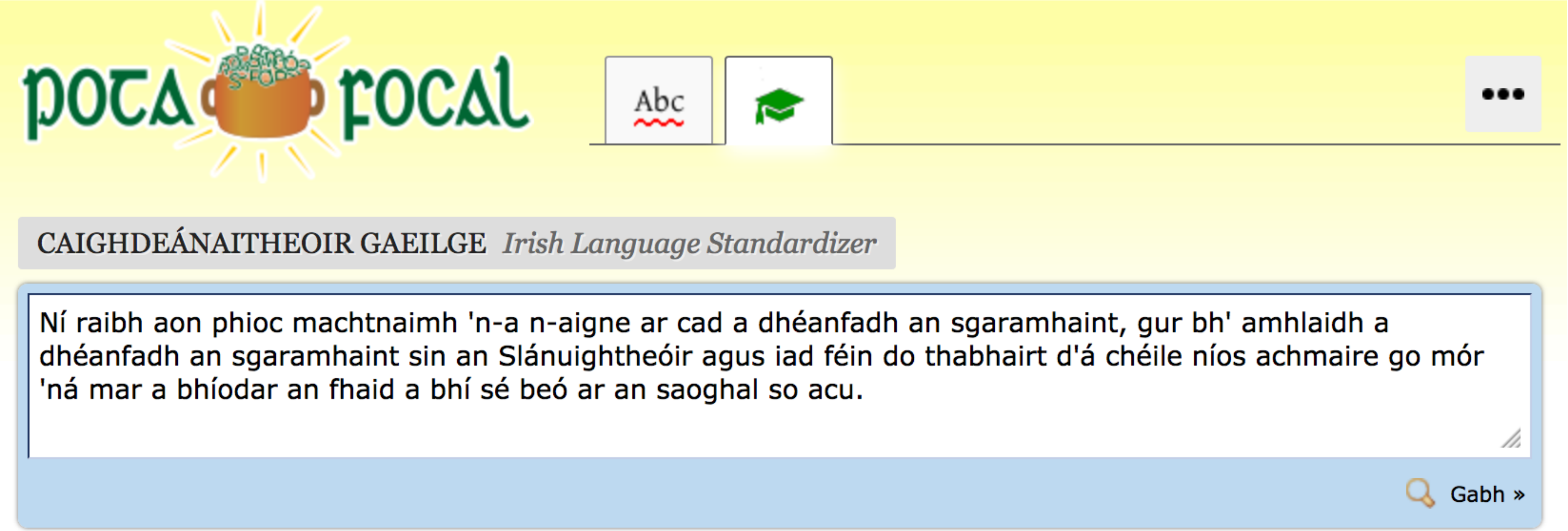
Disambiguation

- Simplified spelling => generally, easy!
- bádhadh (drowning), báidh (sympathy), bádh/bágh (bay) all map to standard “bá”
- In other cases, we need statistical modelling
- Scríobh mé leitir” (leitir=hillside, here alt. “litir”), “Bhí náire ortha” (ortha=spell, here alt. “orthu”), “Ba bhreá léithe é” (léithe=greyness, here alt. “léi”), “Chuaigh mé annsan” (annsan=emph. form ann, here alt. “ansin”), “Tá sí ar thoiseach an tslua” (toiseach=adj. dimensional, here alt. “tosach”)

Results

- Broad lexical coverage: 97.83% 1882-1940
- Worse on older texts, but acceptable, WIP
- Foras Feasa: 96.1% (proper names excluded)
- Being used by New English-Irish Dictionary
- Foclóir na Nua-Ghaeilge (Acadamh Ríoga)

Pota Focal Web Interface



CAIGHDEÁNAITHEOIR GAEILGE *Irish Language Standardizer*

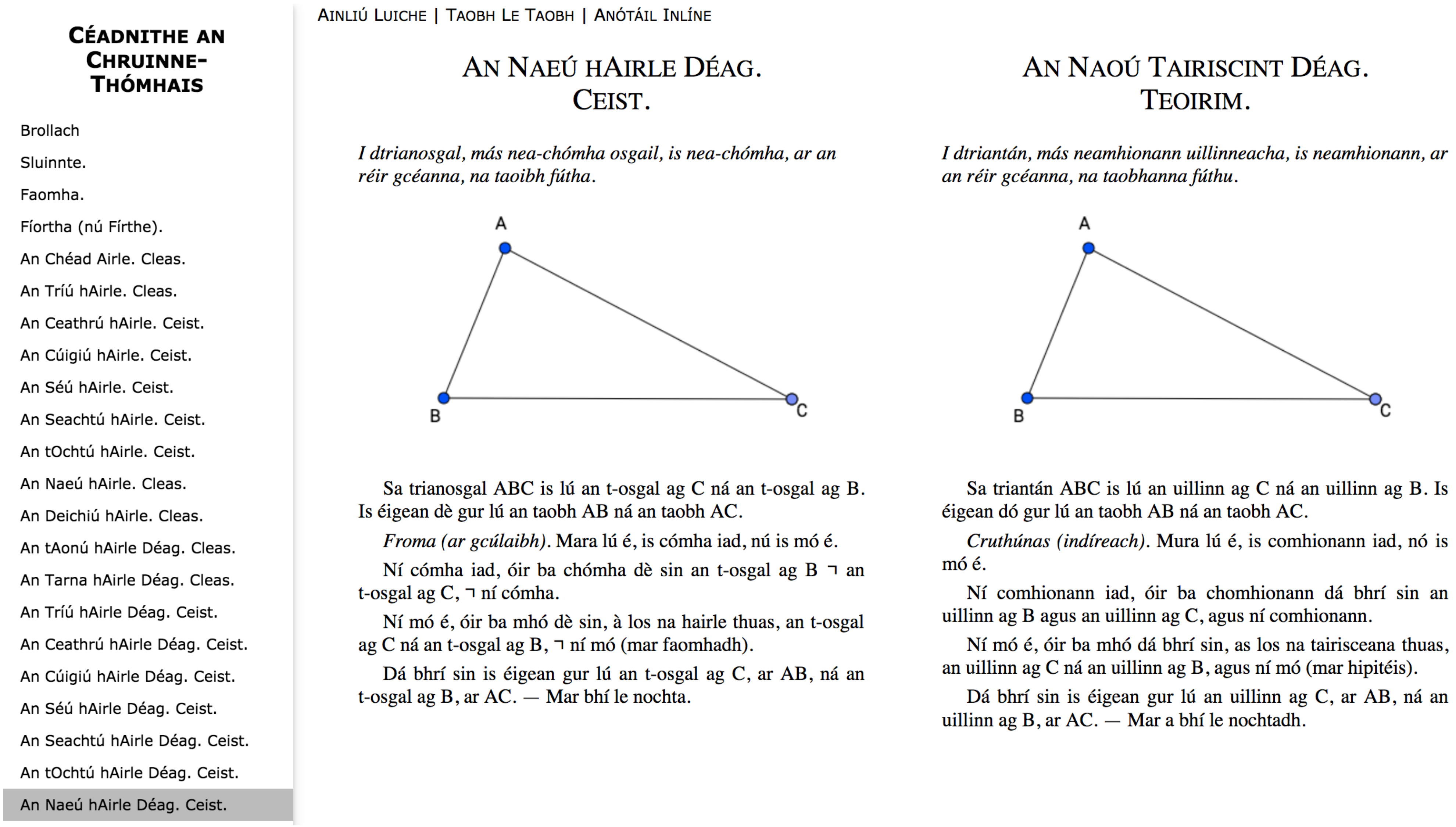
Ní raibh aon phioc machtnaimh 'n-a n-aigne ar cad a dhéanfadh an sgaramhaint, gur bh' amhlaidh a dhéanfadh an sgaramhaint sin an Slánuightheóir agus iad féin do thabhairt d'á chéile níos achmaire go mór 'ná mar a bhíodar an fhaid a bhí sé beó ar an saoghal so acu.

```
Ní raibh aon phioc machtnaimh 'n-a n-aigne ar cad a dhéanfadh an sgaramhaint, gur bh'
                   machnaimh  ina                                scaradh,     gurbh

amhlaidh a dhéanfadh an sgaramhaint sin an Slánuightheóir agus iad féin do thabhairt d'á
                        scaradh            Slánaitheoir                              dá

chéile níos achmaire go mór 'ná mar a bhíodar an fhaid a bhí sé beó ar an saoghal so acu.
            achoimre        ná                                  beo       saol    seo
```

Euclid’s Elements



Céadnithe an Chruinne-Thómhais

Brollach

Sluinnte.

Faomha.

Fíortha (nú Fírthe).

An Chéad Airle. Cleas.

An Tríú hAirle. Cleas.

An Ceathrú hAirle. Ceist.

An Cúigiú hAirle. Ceist.

An Séú hAirle. Ceist.

An Seachtú hAirle. Ceist.

An tOchtú hAirle. Ceist.

An Naeú hAirle. Cleas.

An Deichiú hAirle. Cleas.

An tAonú hAirle Déag. Cleas.

An Tarna hAirle Déag. Cleas.

An Tríú hAirle Déag. Ceist.

An Ceathrú hAirle Déag. Ceist.

An Cúigiú hAirle Déag. Ceist.

An Séú hAirle Déag. Ceist.

An Seachtú hAirle Déag. Ceist.

An tOchtú hAirle Déag. Ceist.

An Naeú hAirle Déag. Ceist.

Ainliú Luiche | Taobh Le Taobh | Anótáil Inlíne

An Naeú hAirle Déag. Ceist.

I dtrianosgal, más nea-chómha osgail, is nea-chómha, ar an réir gcéanna, na taoibh fútha.

Sa trianosgal ABC is lú an t-osgal ag C ná an t-osgal ag B. Is éigean dè gur lú an taobh AB ná an taobh AC.

Froma (ar gcúlaibh). Mara lú é, is cómha iad, nú is mó é.

Ní cómha iad, óir ba chómha dè sin an t-osgal ag B ⁊ an t-osgal ag C, ⁊ ní cómha.

Ní mó é, óir ba mhó dè sin, à los na hairle thuas, an t-osgal ag C ná an t-osgal ag B, ⁊ ní mó (mar faomhadh).

Dá bhrí sin is éigean gur lú an t-osgal ag C, ar AB, ná an t-osgal ag B, ar AC. — Mar bhí le nochta.

An Naoú Tairiscint Déag. Teoirim.

I dtriantán, más neamhionann uillinneacha, is neamhionann, ar an réir gcéanna, na taobhanna fúthu.

Sa triantán ABC is lú an uillinn ag C ná an uillinn ag B. Is éigean dó gur lú an taobh AB ná an taobh AC.

Cruthúnas (indíreach). Mura lú é, is comhionann iad, nó is mó é.

Ní comhionann iad, óir ba chomhionann dá bhrí sin an uillinn ag B agus an uillinn ag C, agus ní comhionann.

Ní mó é, óir ba mhó dá bhrí sin, as los na tairisceana thuas, an uillinn ag C ná an uillinn ag B, agus ní mó (mar hipitéis).

Dá bhrí sin is éigean gur lú an uillinn ag C, ar AB, ná an uillinn ag B, ar AC. — Mar a bhí le nochtadh.

Future Work

- Clean digital versions of Dinneen, O’Reilly, etc.
- Link headwords to modern spellings
- Push further into past: 1600-1882
- Annotated versions of “hard” early M. Ir. books