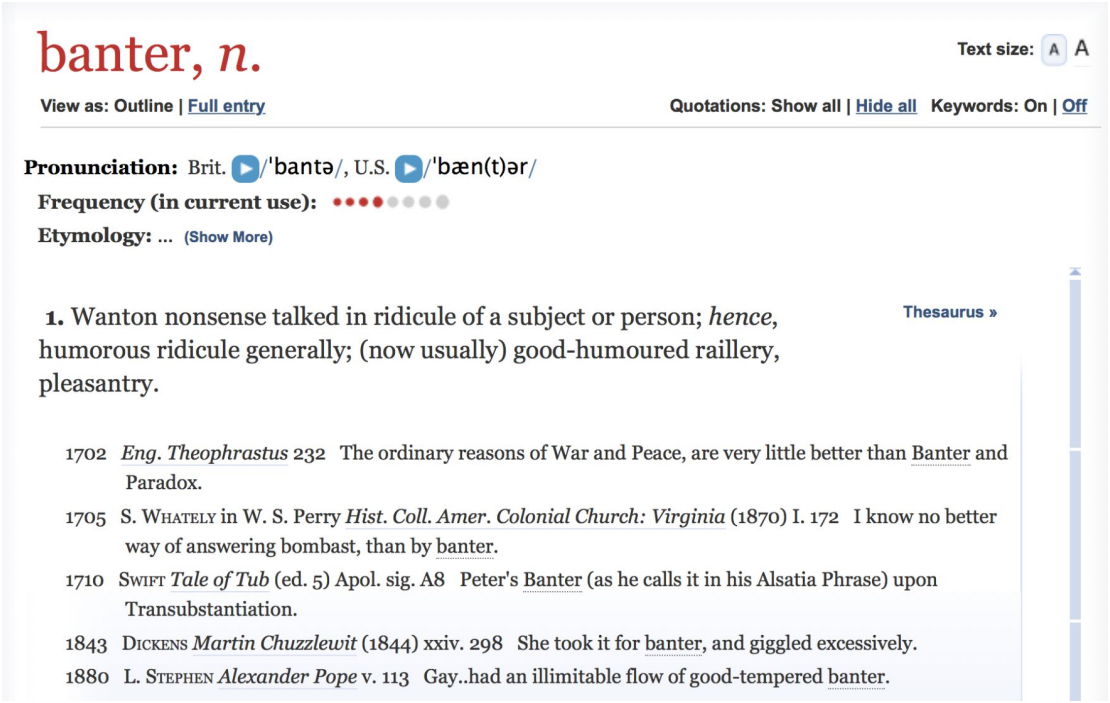
NLP tools for historical lexicography

Kevin Scannell
Cadhan Aonair
20 March 2025

What is historical lexicography?



banter, *n.*

Text size: A A

View as: Outline | Full entry

Quotations: Show all | Hide all Keywords: On | Off

Pronunciation: Brit. /ˈbantə/, U.S. /ˈbæn(t)ər/

Frequency (in current use):

Etymology: ... (Show More)

1. Wanton nonsense talked in ridicule of a subject or person; *hence*, humorous ridicule generally; (now usually) good-humoured raillery, pleasantry.

Thesaurus »

1702 *Eng. Theophrastus* 232 The ordinary reasons of War and Peace, are very little better than Banter and Paradox.

1705 S. Whately in W. S. Perry *Hist. Coll. Amer. Colonial Church: Virginia* (1870) I. 172 I know no better way of answering bombast, than by banter.

1710 Swift *Tale of Tub* (ed. 5) Apol. sig. A8 Peter's Banter (as he calls it in his Alsatia Phrase) upon Transubstantiation.

1843 Dickens *Martin Chuzzlewit* (1844) xxiv. 298 She took it for banter, and giggled excessively.

1880 L. Stephen *Alexander Pope* v. 113 Gay..had an illimitable flow of good-tempered banter.

Source: https://oed.hertford.ox.ac.uk/oed-editions/oed-online/

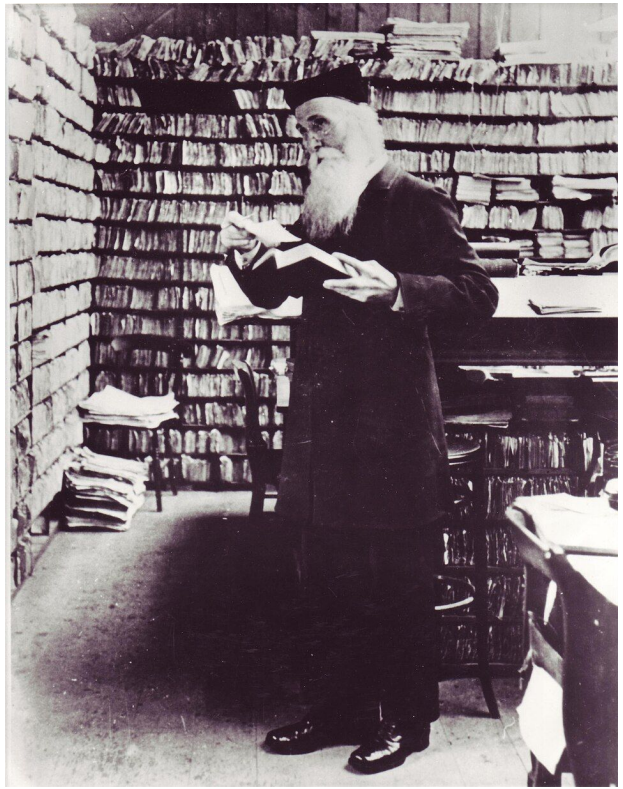
Irish language lexicography

- Virtually all dictionaries are bilingual English-Irish or Irish-English
- New English-Irish dictionary published 2020, print and online (previous 1957)
- New Irish-English dictionary hopefully by 2030 (current 1977)
- **No** full-scale monolingual dictionary
- **No** etymological dictionary
- **No** historical dictionary

Foclóir Stairiúil na Gaeilge (Historical Dictionary of Irish)

- Project of the Royal Irish Academy in Dublin
- Partial funding from the Department of the Gaeltacht in Ireland
- Intended to cover the period 1600–2000 (“modern Irish”)
- Project has existed since the 1970s

How are dictionaries compiled?



CONCORDANCE

English Web 2021 (enTenTen21)

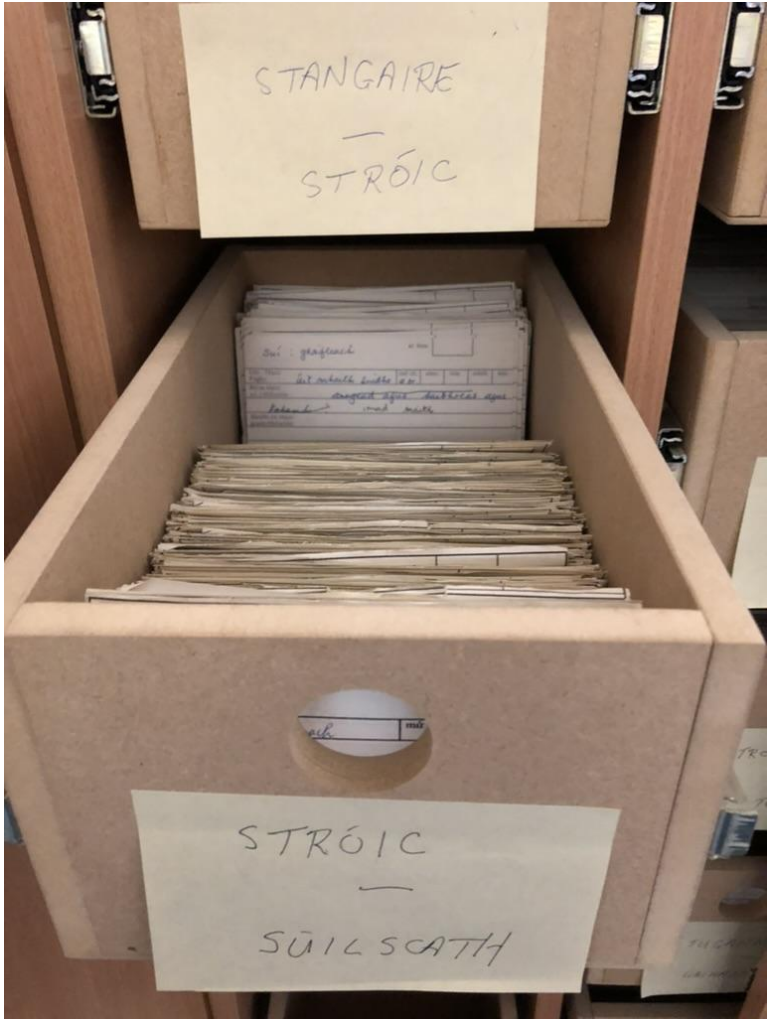
simple banter • 120,672
1.96 per million tokens • 0.0002%

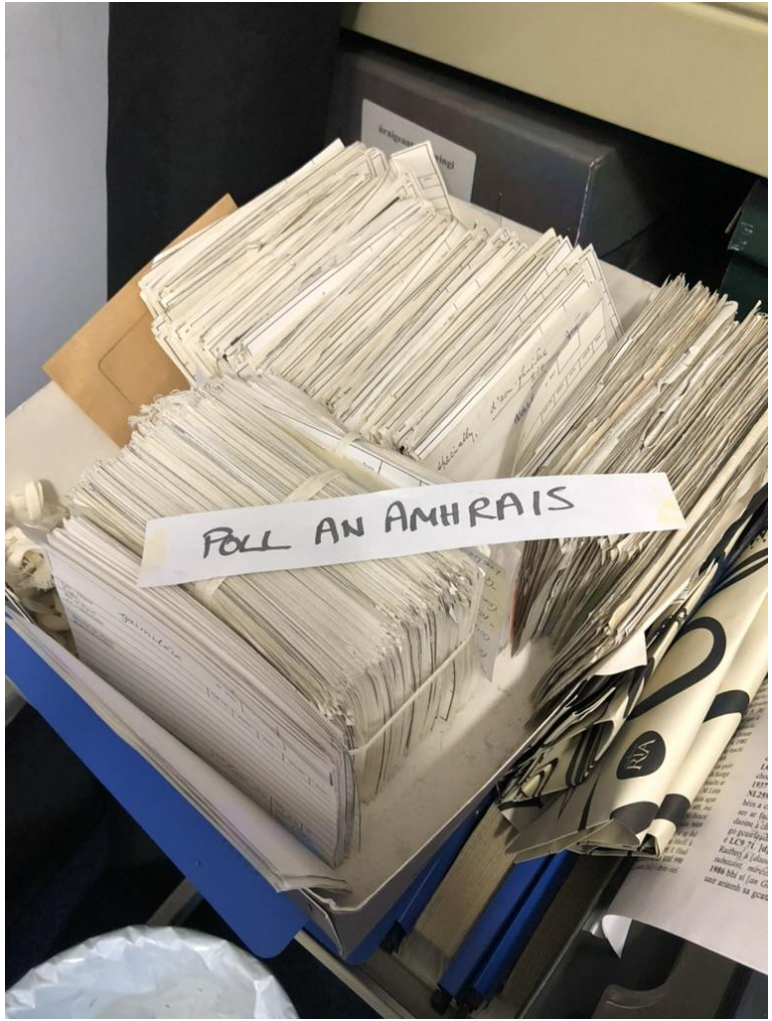
Sample 200 • 200
less than 0.01 • 3.2e-7%

Sort word

KWIC

	Details	Left context	KWIC	Right context
1	thesun.ie	ɜen them, explaining there's been "flirty	banter	" between them.</s><s>Speaking to Ol
2	newsgroove.co.u...	uded it was "friendly and good-natured	banter	".</s><s>Ballance said he felt he had "r
3	instantprint.co...	lleagues, here is our top list of "banned	banter	" in the workplace:</s><s>Hints and tip
4	redhotpawn.com	ɜ tear to my eye.</s><s>We must say '	banter	' not 'battle'. 😉</s><s>Originally posted
5	worksmart.org.u...	ɲent is sometimes dismissed as being '	banter	' or just a joke.</s><s>In fact it can hav
6	wikipedia.org	was also at this time that the 'Stand-up	Banter	' performances began, forerunners of to
7	repeatfanzine.c...	s><s>After some joyfully puerile crowd	banter	(haha- pianist sounds like penis), a firm
8	would-buy-again...	stage, they had great energy and a fun	banter	(Stone making fun of Andres' accent of
9	vdare.com	ɲers.</s><s>Most were likely just office	banter	, of the kind that was common thirty or f
10	skuds.org	ere were all the usual little in-jokes and	banter	, which I am no longer part of and I felt l
11	all-art.org	s handled with great flexibility; the light,	bantering	, somewhat ironic tone–later to become
12	thewalrus.ca	.</s><s>Still, the focus is Vic and Flo's	banter	, which reveals an easy rapport and lurk
13	gutenberg.ca	d his quaking voice into tones of gentle	banter	, forced himself to smile, to tweak her cl
14	kent.ac.uk	der children depend on proficiency with	banter	, which in turn frequently involves verba
15	casting-call.us	ɔ About Nothing - They bicker and they	banter	, they mock and charm, they are Beatric
16	splicetoday.com	eople and blogs.</s><s>There was no	banter	, no antagonizing the crowd.</s><s>Co


STANGAIRE
—
STRÓIC
STRÓIC
—
SÚILSCATH


POLL AN AMHRAIS

Irish corpora: corpas.ria.ie and corpas.ie

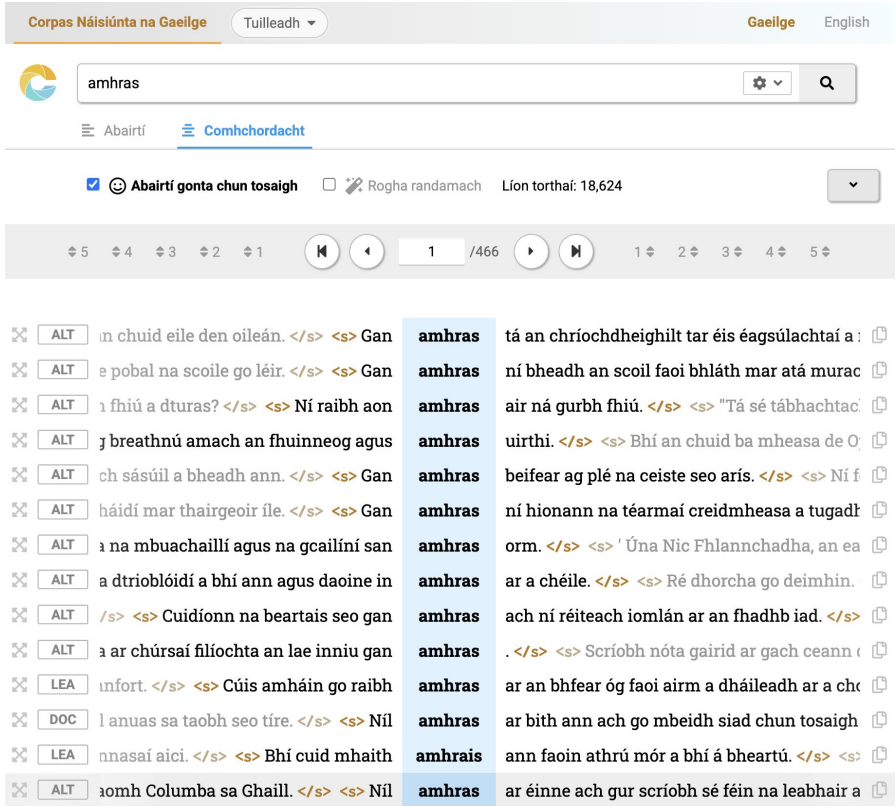
Tiomna Nuadh ár dTighearna agus ár Slánuightheora Íosa Críost (1602)
anaithnid

Search Term	Page	Line Number	Context
amharus	Mt28:17	3713	amharus ag cuid díobh air.
amharus	Lc4:23	6732	Agus adubhairt sé ríu, adértháoi rium gan amharus an seanfhocalsa,
amharus	En10:24	11661	adubhradar ris; gá fad choingéubhas tú sinn an amharus? Innis duinn
amharas	Gn2:12	13282	Agus do ghabh adhuáthmhuireachd agus amharas iád uile, ag rádh, an fear ré
amharus	Gn10:20	14434	Uime sin éiridh agus imthigh siós, agus imthigh riu, gan amharus ar
amharus	Gn11:12	14580	Agus adubhairt an spiorad riom dhul riú, gan amharus ar bioth do bheith
amharus	Gn21:22	15978	Ar anadhbharsoin créd doghénam? ní fhuil amharus gurab éigean don
amharus	Gn21:39	16047	Agus adubhairt Pól ris, gan amharus air as lúdaighe mhisi, ó
amhras	Rm7:12	17537	Ar anadhbharsin atá an reachd gan amhras náomhtha, agus anaithne
amharas	Rm8:17	17640	amharas do Dhía, agus comhoidhreagha do Chríosd: má fhuilngmíd
amhrus	Rm14:23	18217	Achd an tí ar a mbí amhrus, atá sé damanta dá nithidh sé, ar ó
amharus	1Cr7:14	18917	gan chreideamh arna náomhadh san bhfear: no no dobhiádh gan amharus ar
amharus	2Cr1:8	19961	brughadh tar mhogh sinn ós cionn ar neirt, ionnus gu raibhe amharus
amhras	Gl3:29	21148	Agus más lé Críosd sibh, as sibh siól Abrahám gan amhras, agus is
amharus	1Tm2:8	23187	tógbháil suás lámh náomhtha, gan fheirg, gan amharus.
amharus	1Tm3:16	23260	Agus gan amharus as mór seicréid na diághachda: dorinneadh Día
amhrus	Sm1:6	25153	bioth: oír an tí ar a mbí amhrus, is cosmhuil hé ré tuinn fhairge

Iomarbhágh na bhFileadh I (1604)
anaithnid

Search Term	Page	Line Number	Context
amhras	54	1495	is an amhras i dtéid sibh
n-	88	2749	bhar n-amharas orm dá mbeath
amharus	150	4494	m'amharus nocha n-iongnadh

Irish Bardic Poetry (1609)
anaithnid

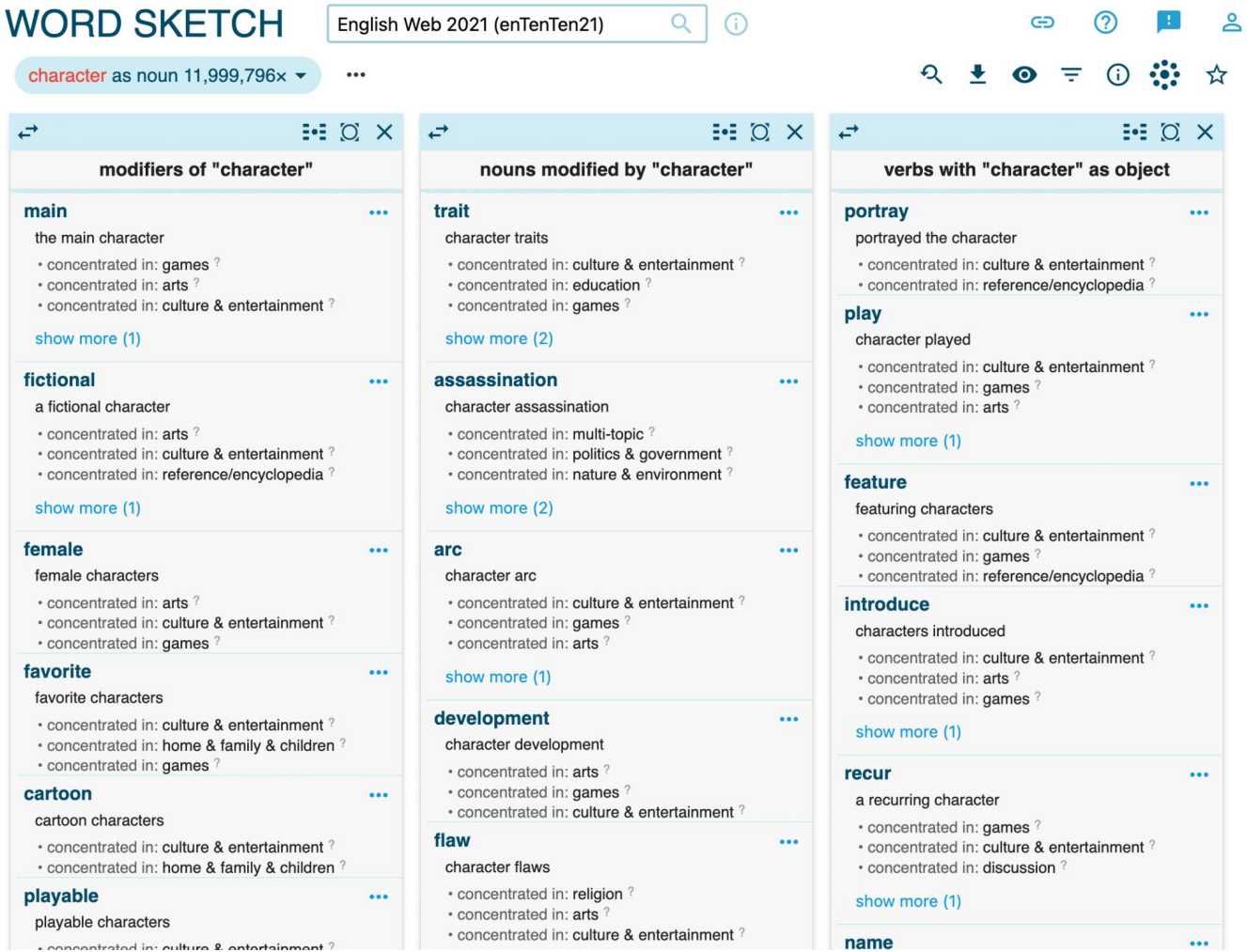


NLP tools: “One-click” dictionaries?

- Start with annotated corpus; minimally with headwords + POS, ideally parsed
- Produce headword list (frequency, language ID, noise filtering)
- For a given headword, we want to identify all senses and subsenses
- Fundamentally this is a (hierarchical) clustering problem
- Long-standing and challenging problem in NLP: “word sense induction”
- Sketch Engine implements clustering on contextual sentence embeddings
- GDEX algorithm for selecting example sentences (Kilgarriff et al 2008)
- Lots of post-editing needed! Splitting/lumping, crafting definitions etc.

Lexicographers 🩷 recall

- If we think of corpus query software as a search engine, recall is critical!
- Especially for historical dictionaries: aim to include each sense of each word
- Better to sift through false positives than potentially miss a word/sense/citation


WORD SKETCH
English Web 2021 (enTenTen21)
character as noun 11,999,796×
modifiers of "character"
main
the main character
• concentrated in: games
• concentrated in: arts
• concentrated in: culture & entertainment
show more (1)
fictional
a fictional character
• concentrated in: arts
• concentrated in: culture & entertainment
• concentrated in: reference/encyclopedia
show more (1)
female
female characters
• concentrated in: arts
• concentrated in: culture & entertainment
• concentrated in: games
favorite
favorite characters
• concentrated in: culture & entertainment
• concentrated in: home & family & children
• concentrated in: games
cartoon
cartoon characters
• concentrated in: culture & entertainment
• concentrated in: home & family & children
playable
playable characters
nouns modified by "character"
trait
character traits
• concentrated in: culture & entertainment
• concentrated in: education
• concentrated in: games
show more (2)
assassination
character assassination
• concentrated in: multi-topic
• concentrated in: politics & government
• concentrated in: nature & environment
show more (2)
arc
character arc
• concentrated in: culture & entertainment
• concentrated in: games
• concentrated in: arts
show more (1)
development
character development
• concentrated in: arts
• concentrated in: games
• concentrated in: culture & entertainment
flaw
character flaws
• concentrated in: religion
• concentrated in: arts
• concentrated in: culture & entertainment
verbs with "character" as object
portray
portrayed the character
• concentrated in: culture & entertainment
• concentrated in: reference/encyclopedia
play
character played
• concentrated in: culture & entertainment
• concentrated in: games
• concentrated in: arts
show more (1)
feature
featuring characters
• concentrated in: culture & entertainment
• concentrated in: games
• concentrated in: reference/encyclopedia
introduce
characters introduced
• concentrated in: culture & entertainment
• concentrated in: arts
• concentrated in: games
show more (1)
recur
a recurring character
• concentrated in: games
• concentrated in: culture & entertainment
• concentrated in: discussion
show more (1)
name

Word Sketches use Dependency Parsing

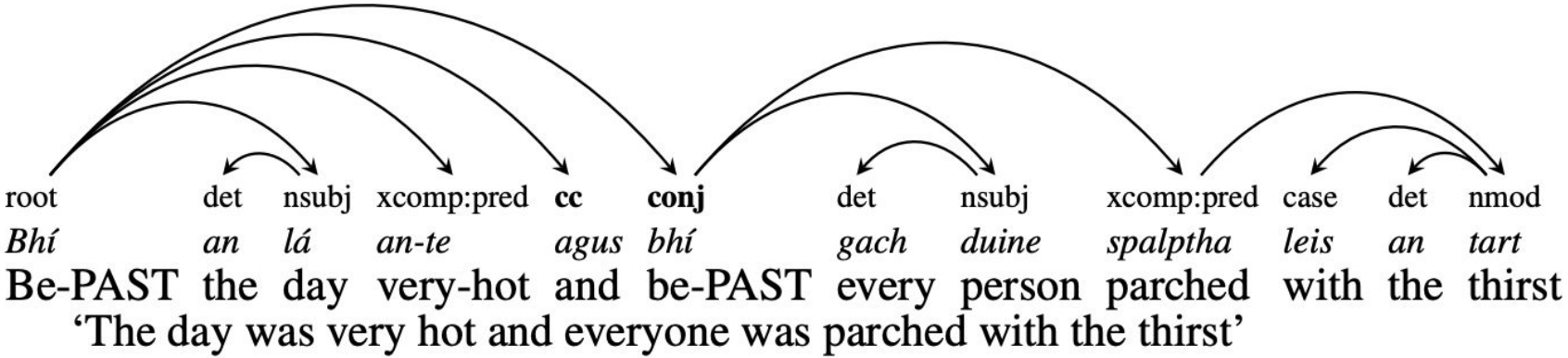


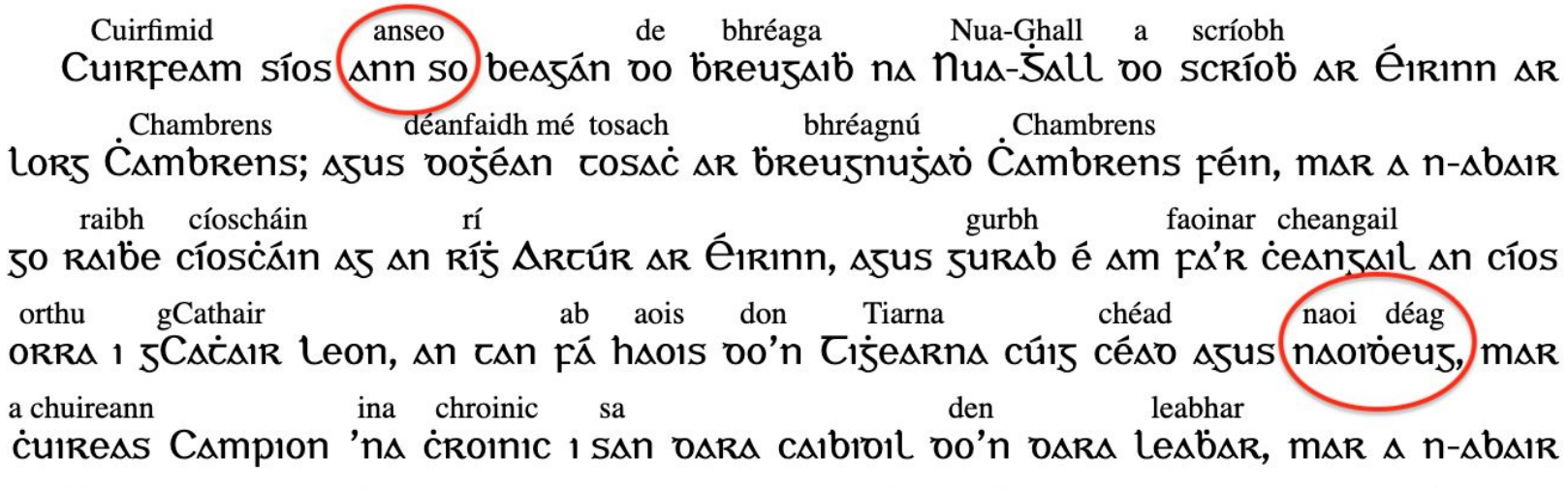
Figure 4 in *Lynn, Teresa and Foster, Jennifer (2016) Universal dependencies for Irish. In: Second Celtic Language Technology Workshop. (CLTW 2016), 4 July 2016, Paris, France.*

Standard Irish

- Official standard introduced in the 1940's and 1950's
- Significant simplifications to spelling and grammar
- Taggers and parsers we have were developed for the standard language
- These perform poorly for pre-standard texts
- Major challenge for the historical dictionary
- http://corpas.ria.ie/ has 3000 texts published between 1600 and 1926

Standardization

- I developed a tool for standardizing Irish texts, c. 2007
- Shallow statistical MT approach; *does no annotation of pre-standard text*



Gaeilge Gaedhilge Gaeilige Gaelige Gailge Gaoidhilge Gaoidheilge Gaelge
Gaidhlige Gaedheilge Gaoidhelge Gailege Gaielge Gaodhailge Gaedilge
Gaeidhilge Gaedhilige Gaoidilge Gaeilgele Gaedhlige Gaédhilge Gaoilge
Gaeillge Gaeilga Gaidhilge Gaelilge Gaodheilge Gaeiilge Gaedhilghe
Gadhilge Gaheilge Gaellge Gaoilaige Gaodhilge Gaedhilgé Gaeilege Gaaeilge
Gailige Gaeilgé Gaeghilge Gaedhailge Gaoidhlige Gaelgie Gaeiloge Gaeilgle
Gaeilghe Gaeılge Gaeidhlge Gaeidheilge Gaeeilge Gaoilige Gaóilge
Gaoilaga Gaoigheilge Gaoidhlge Gaoidelge Gaoideilge Gaodhéilge Gaieilge
Gaeulge Gaeuilge Gaeolge Gaeoidheilge Gaeilgi Gaeilgee Gaeílge Gaeidlge
Gaeidilge Gaeidhelge Gaehilge Gaeeilgee Gaedhlge Gaedhiilge Gaedhelga
Gaédhailge Gaedgilge Gadehilge Gaddhilge Gaoghailge Gaileige Gaidlhlige
Gaidhlge Gaeliage Gaelga Gaéilge Gaedilghe Gaedhulge Gaedhealg Gaedheilg
Gaédheilg Gaedhilg Gaedhilig Gaeilg Gaoidhealg Geadhilge Geailge

Caoimhín Ó Scanaill
@kscanne

174 litriú ar fhocal amháin, in ord minicíochta 😀

```
kps@borel:~/gaeilge/crubadan/twitter$ egrep -i -o 'ch?[oó](mh?)?gh?[a-z]*r[a-z]*
d[a-z]*s' sonrai/ga-tweets.txt  | tolow | sed 's/^ch/c/' | sort | uniq -c | sort
 -r -n | sed 's/^ *[0-9]* //' | tr "\n" " " | fmt
comhghairdeas comhghairdeachas comhgháirdeachas comhgairdeas
comhgairdeachas comghairdeas comhgháirdeas comghairdeachas
comhgáirdeachas comhghardeas cómhghairdeas comgháirdeachas
comghairdeaghas cómhgháirdeas comhghairdeachais comhgháirdeachais
comhghairdis comhgheardas comhgairdeagas comhgáirdeas comhghairdheas
comhghairdes cómhgháirdeachas comhgháirdis comhghairdeagas
comhgairdegas comgháirdeas comhgairdeachais comhghdeardas comhghàirdeas
comhgardachas comhgháirdeachas comhgháirdeagas comhghairdigeas
comhgaridgeas comhghairdeacheas comhghairdachas comhgairdheas
comghgairdeas comhghirdeas coghairdeachas comhgheardais comhghardeachas
comgairdeas comgairdeachas comhghairdeachs cómhghairdeachas comhgardaghas
comhghraideas comhghiardeas comhghairdras comhghàirdeachas comhghairdeaas
comhgjairdeas comhghairdres comhghairdegas comhghairdechas comhghairdeahas
comhghairdas comghairdheas coghairdeas comhghsirdeachas comhghrds
comhghárdeachas comhghairedeas comhghairedeachas comhghairdreas
comhghairdeaches comhgardechas comhgardeachas comhgárdachas
cómhgairdeas comghghairdeachas comghardeas cóghairdeas comhgradas
comhgharideas comhghardieas comhghardeachais comhghairdwas comhghairdeads
comhgháirdeachas cómhgháirdeachais comhgards comhgárdeas comhgardeas
comhgardeachais comhgairds comhgairdgeas comhgairdeghas comhgairdeagais
comhgairdeacheas cómhgáirdeachas cómhgairdeachas cómhgairdeachais
comhgairdeacas comhgairdais comhgáiirdeachas comghardas cómghairdeachas
cógáirdeachas comhgradeas comhgjairdeachas comhgiardeachas comhghsirdeas
comhghéirdeachas comhgheardeachas comhghdeardás comhghdairdeas
comhghardis comhghárdeas comhghárdachas comhghaordeas comhghairgdeas
comhgháiredeachas comhghairdís cómhghairdheachas comhghairdheachas
comhghairdess comhghairdeghas comhghairdeeas comhgháirdechas
comhghairdeass comhghairdeasannaagus comhghairdeás comhghairdeais
comhghairdeaghas comhgháirdeacheas comhghairdeachás comhgháirdéachas
comhghairdéachas comhghâirdeachas comhghairdeacahas comhgháirdas
comhgghairdeas comhggardeas comhgeardeas comhgeardais comhgdairdeas
comhgardgas comhgardeaghas comhgardas comhgardaicheas comhgairedeachas
comhgairdeis comhgairdegeas comhgairdecais comhgáirdeagas comhgairdeachss
comhgâirdeachas comhgàirdeachas comhgairdeacahas comhgahirdeas
comghrdeas comghgháirdeas comghghairdeas comgheardeas comghairghdeas
comghairedeachas comghairdreas comghairdigeas comghairdegas comghairdecas
cómghairdeas comghairdeachais comggairdeas comgardeas comgairdheas
comgáirdeachas comgairdeachais coghárds cógardachas cogáirdeachas
cógairdeachas
```

8:57 r.n. · 10 DFómh 2020 · Twitter Web App

Traditional processing pipeline

- Run an older text through the standardizer, outputs word-level alignments
- Tag/parse the standardized text using tools for the modern language
- “Project” the annotations back to the original text
  - One-to-many standardization (“naoidheug”): adjust tokenization of source text
  - Many-to-one standardization (“ann so”): DB of 750 most common examples + annotations
- Essentially the pipeline used for the corpas.ria.ie site
- But how well does this work?
- We knew this approach was inherently limited
  - Limited by accuracy of the shallow standardizer, which is very good but will never be perfect
  - More importantly, discards grammatical features which have disappeared in standard Irish

Test corpus of pre-standard Irish texts

- 150 sentences, just under 4000 tokens
- 25 sentences from three 20th c. books, one per major dialect: “Older” corpus
- 25 sentences from three very challenging texts: “Oldest” corpus
  - 1602 Irish New Testament
  - Foras Feasa ar Éirinn (1630s)
  - Cín Lae Amhlaoibh (1820s)
- Manually tagged/parsed following the Universal Dependencies guidelines
- https://github.com/UniversalDependencies/UD_Irish-Cadhan/blob/dev/ga_cadhan-ud-test.conllu

Results (unlabeled attachment)

Model	Standard Irish	Older (1900-1950)	Oldest (1600-1900)
UD	81.8	77.6	61.2
Standardize+Project	81.1	84.8	73.0
UD+Silver training	82.0	84.0	70.6

Observations

- Modern taggers/parsers perform poorly on older texts
- Traditional pipeline using the standardizer gives good results
- But, promising results w/o gold training and w/o using the standardizer directly
- With enough gold training data, we can eliminate the standardizer

Thank you! / Go raibh maith agaibh!

- https://cadhan.com/
- https://github.com/kscanne/