

Corpus-based language technology for indigenous and minority languages

Kevin Scannell
Saint Louis University
5 October 2015

Background

- I was born in the US, but I speak Irish
- Trained as a mathematician
- Began developing tools for Irish in the 90s
- No corpora for doing statistical NLP
- So I built my own with a crawler
- Crúbadán = crawling thing, from crúb=paw

Web as Linguistic Archive

- Billions of unique resources, +1
- Raw data only, no annotations, -1
- > 2100 languages represented, +1
- Very little material in endangered languages, -1
- Full text searchable, +1
- Can't search archive by language, -1
- Snapshots archived periodically, +1
- Resources removed or changed at random, -1
- Free to download, +1
- Usage rights unclear for vast majority of resources, -1

An Crúbadán: History

- First attempt at crawling Irish web, Jan 1999
- 50M words of Welsh for historical dict., 2004
- ~150 minority languages, 2004-2007
- ~450 languages for WAC3, 2007
- Unfunded through 2011
- Search for “all” languages, started c. 2011

How many languages?

- Finishing a 3-year NSF project
- Phase one: aggressively seek out new langs
- Phase two: produce free+usable resources
- Current total: 2140
- At least 200 more queued for training
- 2500? 3000?

Design principles

- Orthographies, not languages
- Labelled by BCP-47 codes
- en, chr, sr-Latn, de-AT, fr-x-nor, el-Latn-x-chat
- Real, running texts (vs. word lists, GILT)
- Get “everything” for small languages
- Large samples for English, French, etc.

How we find languages

- Lots of manual web searching!
- Special code monitors WP, JW, UN, bible.is
- Typing/OCR of scanned or offline texts
- Build statistical language ID models
- Special thank to Ed Jahn, George Mason
- NSF grant 1159174

Adding Value

- Separating orthographies/dialects
- Clean boilerplate text
- Convert to UTF-8 text + normalize
- Sentence segment and tokenize
- Produce formats useful for NLP

Metadata

- URL, timestamp, page title, original format
- Most important addition: BCP-47 tag
- Language <meta> tags usually wrong
- Last-Modified header usually rubbish

Tokenization

- Default tokenizer (letters in default script)
- Many exceptions: Greek in coo/hur/kab, etc.
- Word internal punctuation (ca: |•|, |·|)
- Initial/final apostrophes or lookalikes











Three modules

- Traditional web crawler
- Twitter crawler
- Blog tracker

Twitter crawler



- Twitter's REST API
- Seed searches with words from web corpora
- Language ID particularly challenging
- Crawl social graph to find new tweets
- <http://indigenoustweets.com/>

Erabiltzailea	Euskara	Guztira	% Euskara	Jarraitzaileak	Jarraitzen	Azken tuita
1 berria 	42771	45884	93.2	18486	708	2013-08-05 15:47:30
2 txargain 	38964	53350	73.0	953	338	2013-08-05 23:17:30
3 euskalherrian 	29203	57040	51.2	4672	76	2013-08-06 04:15:58
4 toki_kom 	28922	31541	91.7	476	51	2013-08-06 01:14:02
5 eitbcomBerriak 	25525	26960	94.7	5088	188	2013-08-05 19:59:26
6 joseba01 	17330	29673	58.4	435	502	2013-08-06 00:12:08
7 theklaneh 	16242	33797	48.1	1795	355	2013-08-05 03:50:26
8 argia 	15284	16818	90.9	9598	1394	2013-08-05 21:15:57
9 euskaljakintza 	14225	24857	57.2	5176	1225	2013-08-05 07:24:48
10 joxe 	12976	22632	57.3	1578	900	2013-08-05 23:17:13
11 goiena 	12576	13932	90.3	1960	341	2013-08-05 14:07:54

Ag feitheamh le a0.twimg.com...

INDIGENOUS TWEETS.COM

Euskara

Pil-pilean:

- kariiiiis
- antraxa
- #goraligoteosana
- Carrow
- apurtzeraaaaaaaaa
- zantzlekn
- ceditutie
- #hiroshimagogoan
- erritmoen
- mortaaaaaaaaaaaaaala

Norbait falta da?

Twitter erabiltzaile-izena:

@ Bidali

Esaiozu munduari hemen zaudela:



Blog

Indigenous Blogs

Kevin Scannell



Blog tracker



- Blogger platform only
- Works hand-in-hand with traditional crawler
- Registers all blogs with an in-language post
- Tracks all past and future posts
- <http://indigenusblogs.com/>

Deliverables

- See <http://crubadan.org/>
- Word and bigram frequency lists
- Character trigram frequencies
- Lists of URLs in each language
- Discoverable as an OLAC repository

Spelling and grammar checkers

- Corpus-based Irish spellchecker, 2000
- Grammar checker, 2003
- 28 new spellcheckers since 2004
- Collaborations with native speakers
- All under open source licenses

hunspell

- Standard open source spellchecking engine
- The default in Firefox and LibreOffice
- Fast and powerful
- Good language support: ~150 languages
- Can be as simple as a word list
- But also supports complex morphology

Computational Morphology

- Finite-state transducers (Xerox FST, foma)
- Very fast + bidirectional
- Cover most morphology of human langs
- Hunspell uses “two-fold affix stripping”
- Morphological analysis only
- Not as powerful, theoretically
- BUT: simple formalism, user support FTW!

Powerful enough?

- Hungarian
- Northern Sámi
- Basque
- Lingala, Kinyarwanda, Swahili, Chichewa
- Nishnaabemwin

Language ID

- Component *and* an application of Crúbadán
- Character n-grams + word models
- NLTK 3-gram data set
- Indigenous Tweets and Blogs

Predictive text

- T9 input
- Adaptxt
- Firefox OS
- Dasher



accentuate.us



- Web service for diacritic restoration
- Eni kookan lo ni eto si omi nira lati ni imoran ti o wu u, ki o si so iru imoran bee jade
- Ènì kòòkàn ló ní ẹ̀tọ́ sí òmì nira láti ní ìmọ̀ràn tí ó wù ú, kí ó sì sọ irú ìmọ̀ràn bẹ̀ẹ̀ jáde
- End-user clients for Firefox, LibreOffice
- Perl, Python, Haskell libraries
- Joint work with Michael Schade

Lexicography

- Geiriadur Prifysgol Cymru
- Foclóir Nua Béarla-Gaeilge
- Foclóir na Nua-Ghaeilge
- SketchEngine

N-gram language models

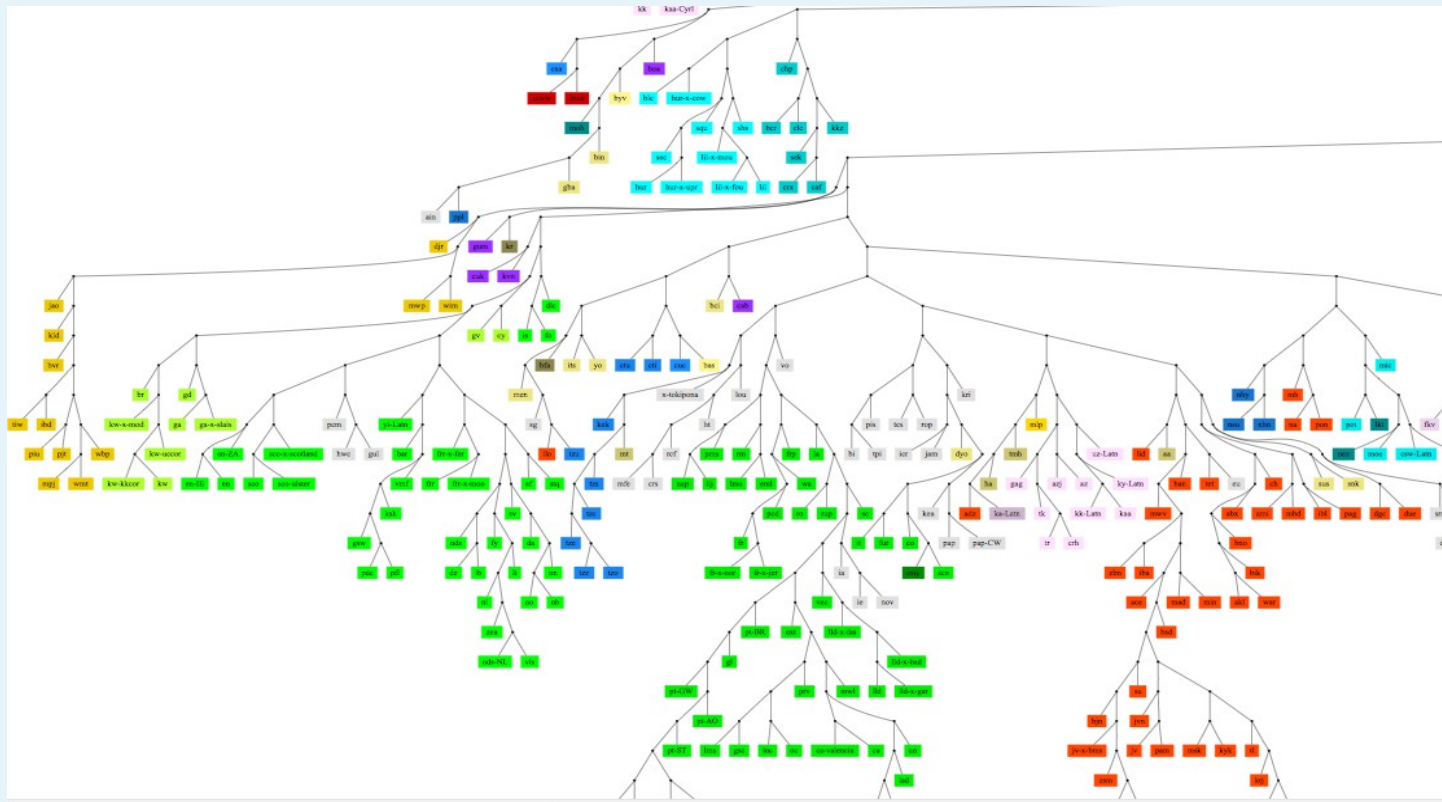
- Machine translation (gd/ga, gv/ga)
- Irish standardizer
- Speech recognition
- OCR (e.g. Irish seanchló)

Linguistic research

- Comparative phonology
- Syntax and morphology
- Psycholinguistics
- Selectional preferences

Orthotree

- <http://indigenoustweets.blogspot.com/2011/12/>
- <https://github.com/kscanne/orthotree>



Help wanted

- > 100 collaborators: speakers, linguists
- Help sort dialects, orthographies
- Tokenization and normalization
- Finding new material for training
- Help create new online material
- Collaborate on spell checkers, etc.