

How many languages are on the web?
The Crúbadán project, 10+ years on

Kevin Scannell
Saint Louis University
14 August 2013

Felix Klein

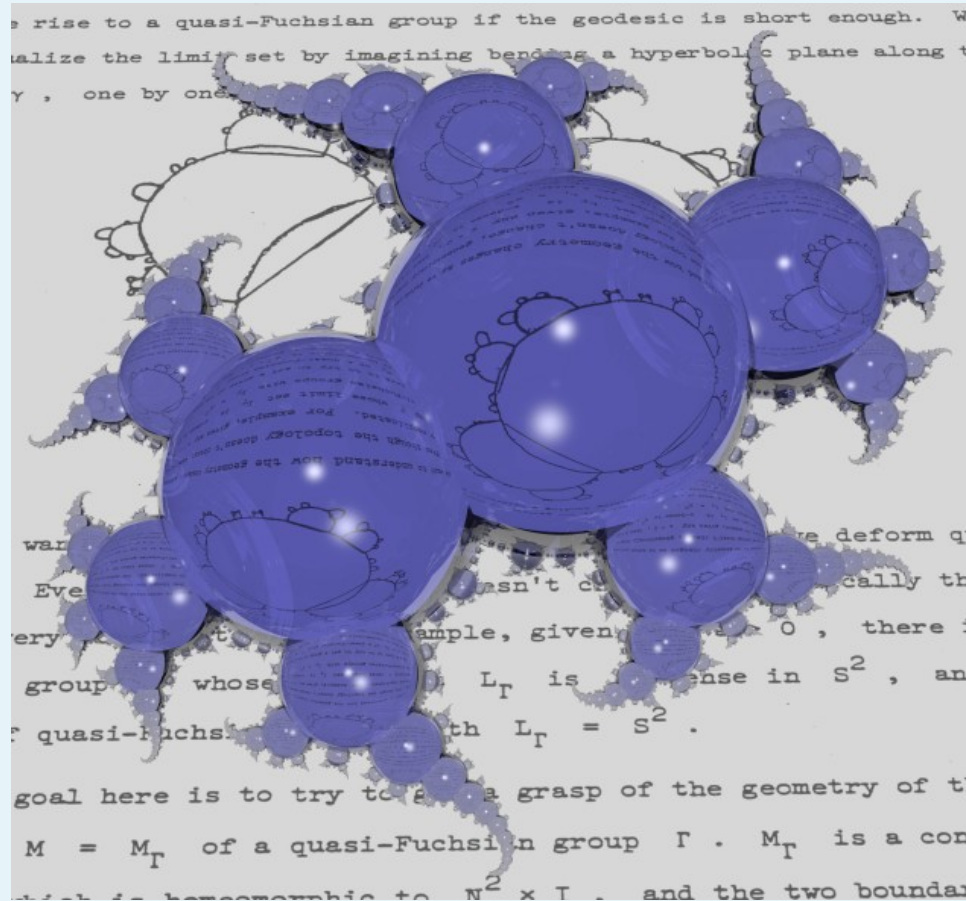


Image by David Dumas and Jeff Brock

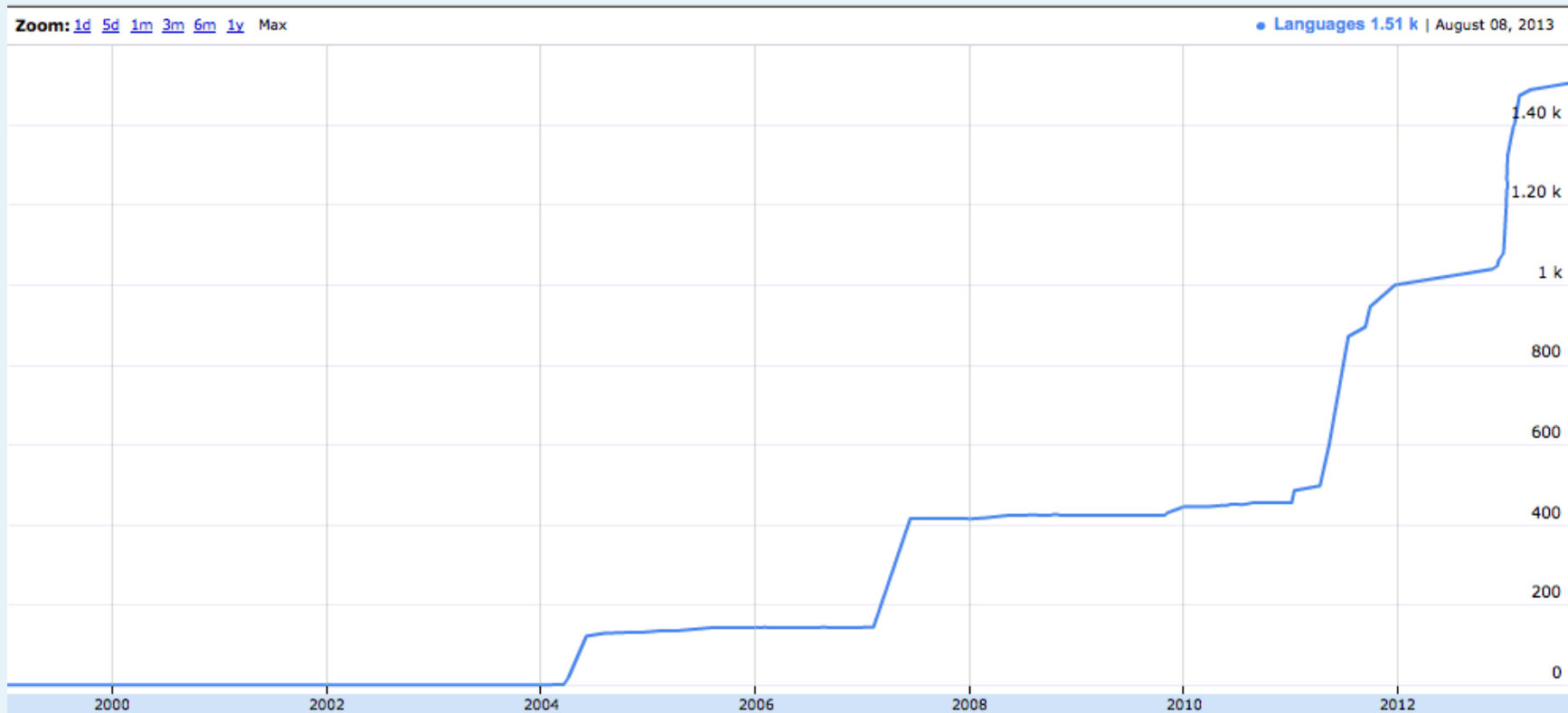
An Crúbadán: History

- 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

So, how many languages?

- Halfway through 2 year project
- Phase one: aggressively seek out new langs
- Phase two: produce free+usable resources
- Current total: 1510
- At least 100 more queued for training
- 1800? 2000?

Languages vs. time



Goals




- Interested in revitalization, first and foremost
- Building useful software for communities
- Blark: word lists, morph. analyzers
- Open data for under-resourced languages
- Linguistic typology
- Linguistic diversity of the web

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

Language ID

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

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:

 Tweet

Blog

Indigenous Blogs

Kevin Scannell



Predictive text

- T9 input
- Adaptxt
- Firefox OS



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
- Ẹ̀nì kọ̀ọ̀kan ló ní ẹ̀tọ́ sí òmì nira láti ní ìmọ̀rà̀n tí ó wù ú, kí ó sì sọ irú ìmọ̀rà̀n 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

NLP Research

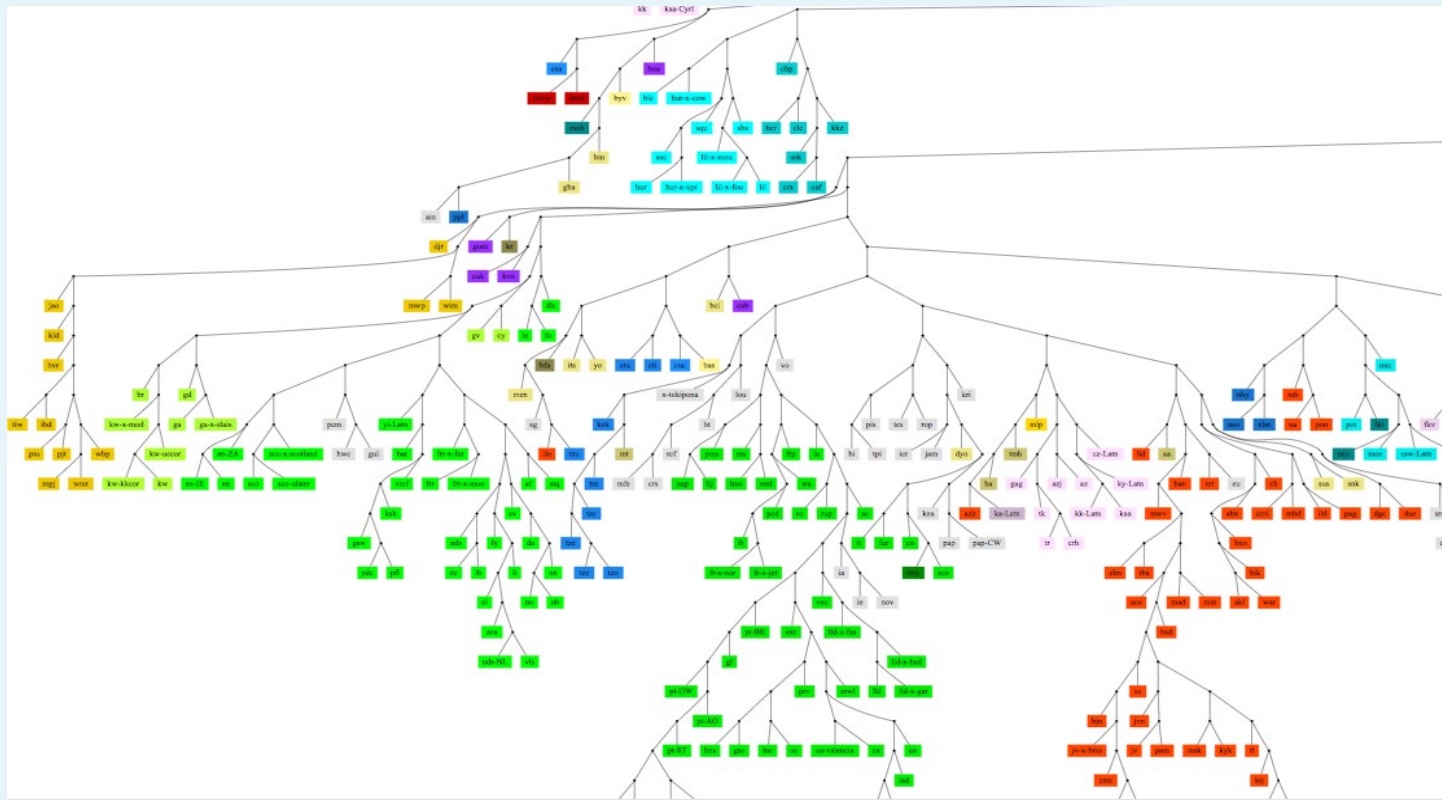
- N-gram language models for MT
- Caighdeánaitheoir, gd2ga, gv2ga
- Computational morphology
- Parsing
- OCR (e.g. Irish seanchló)
- Speech recognition/synthesis

Linguistic research

- Comparative phonology
- Syntax
- Psycholinguistics
- Selectional preferences
- ...

Orthotree

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



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.

Three modules

- Traditional web crawler
- Twitter crawler
- Blog tracker

Phase 1: Finding new languages

- Lots of web searching!
- Special code monitors WP, JW, UN, bible.is
- Typing/OCR of scanned or offline texts
- Thanks: E. Jahn, D. Joosten, J. Berlage
- NSF grant 1159174

Phase 2: Building useful resources

- Separating orthographies/dialects
- Clean boilerplate
- Convert to UTF-8 text + normalize
- Sentence segment and tokenize
- Avoid copyright issues
- Discoverability (OLAC)

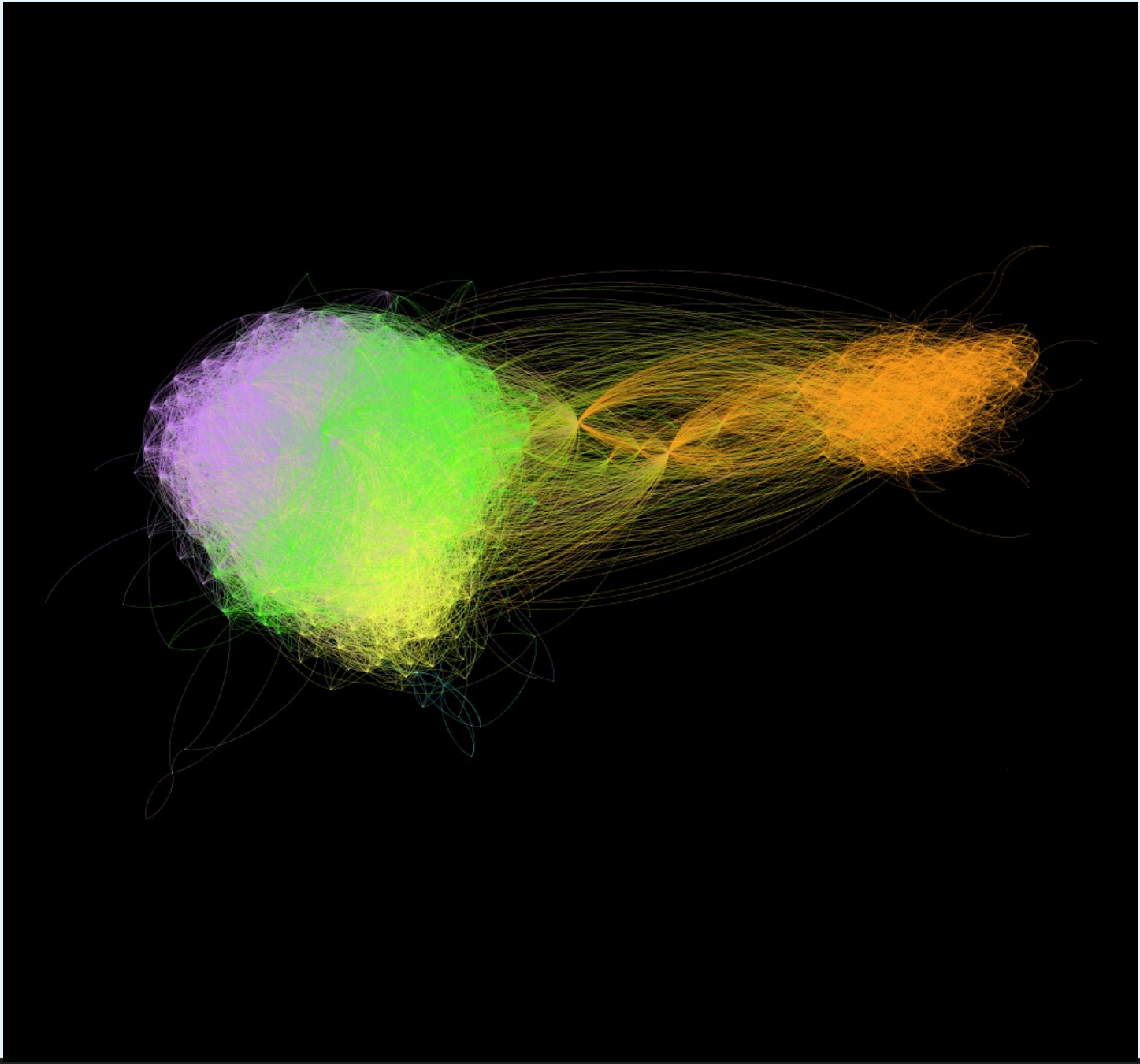
Tokenization

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

Twitter crawler

- Twitter's REST API
- Seed searches with words from web corpora
- Language ID particularly challenging
- Crawl social graph to find new tweets





Blog tracker

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

Call to action

- > 100 collaborators: speakers, linguists
- Help sort dialects, orthographies
- Tokenization and normalization
- Finding new material for training
- Help create new online material