Linguistic resources from web corpora

Kevin Scannell
Saint Louis University

7 September 2013

Worst archive ever?

- Raw data only, no annotations, entirely unstructured
- Very little material in endangered languages
- Can't search archive by language
- In fact, no reliable metadata at all (no date, creator, ...)
- Resources removed from archive at random
- Resources with fix ID changed at random
- Large % use broken or non-standard encodings
- Usage rights unclear for vast majority of resources

Best archive ever?

- More than 10 billion unique resources
- Millions more added every day
- At least 1500 languages represented
- All free to download
- Full-text searchable
- (Partial) snapshots archived periodically

(I'm talking about the web)

- So it's not-so-great as an archive
- Q: Can we make it great?
- Primary goal: Language revitalization
- Subgoal #1: NLP applications
- Subgoal #2: Linguistic research
- A: Definitely yes for #1, maybe not for #2

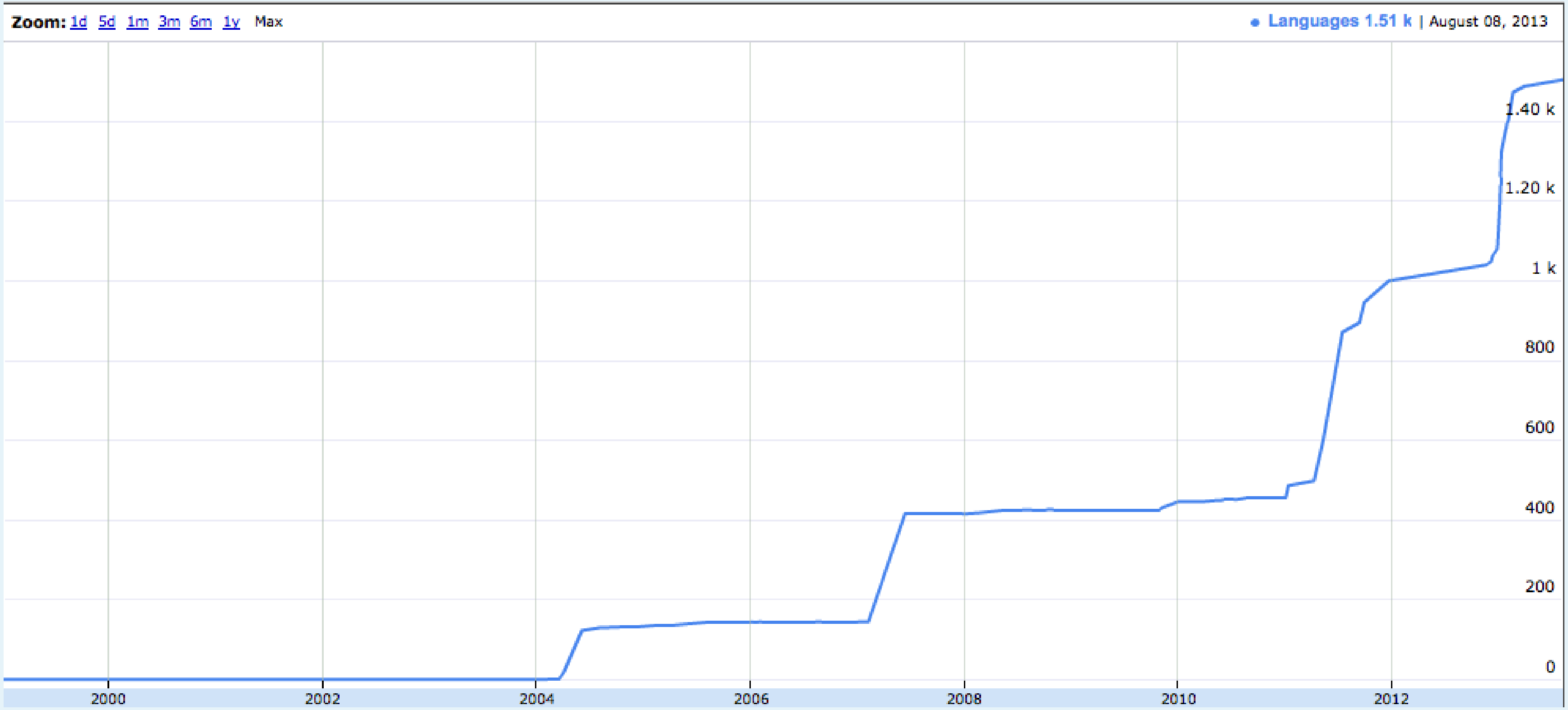
An Crúbadán: History

- 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?

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

Languages vs. time



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
- Thanks: E. Jahn, D. Joosten, J. Berlage
- NSF grant 1159174

Three modules

- Traditional web crawler
- Twitter crawler
- Blog tracker

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
- “Best-guess” link to archive.org
- Small number of highly productive sites
- Manually tagging (blogs, news, bible, ...)

UTF-8 Normalization

- Fonts (Sámi, Mongolian, dozens of others)
- Lookalikes (az: ə/ә, bua: γ/ү, ro: ş/ș)
- Shortcuts (haw, mi, etc. äëïöü for āēīōū)
- Encoding issues (tn, nso: ß/š from Latin-2)
- Fun w/ apostrophes: 'ˈˊ'‘’ʽˀˊˋˎ'ˊʾ\ʼ’ˊʽ‘’ʽ′ `ˈ ' `

Tokenization

- Default tokenizer (letters in default script)
- Many exceptions: Greek in coo/hur/kab, etc.
- Word internal punctuation (ca: l•l, l·l)
- 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
- http://indigenoustweets.com/

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://indigenousblogs.com/

Deliverables

- Word frequency lists (raw data, Android, …)
- Trigram frequencies (ARPA format)
- Character n-gram frequencies (langID)
- Lists of URLs in each language
- Corpora of shuffled sentences
- OLAC repository

Spelling and grammar checkers

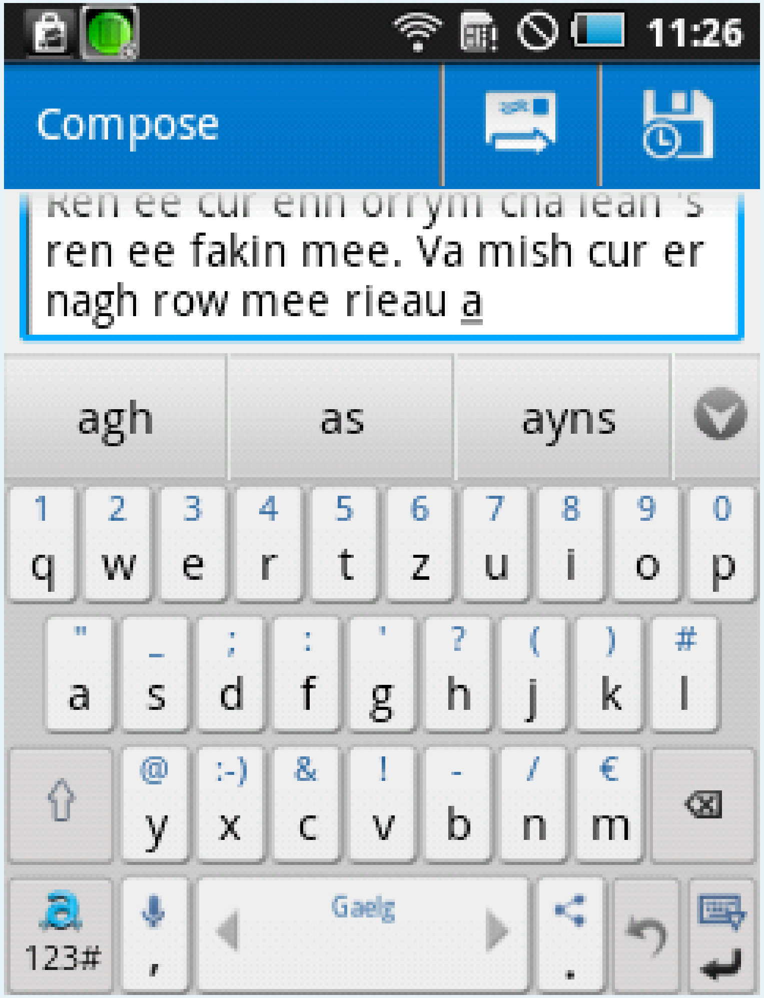
- Corpus-based Irish spellchecker, 2000
- Grammar checker, 2003
- 28 new spellcheckers since 2004
- Collaborations with native speakers
- All under open source licenses

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



Ren ee cur enn orrym cha lean 's ren ee fakin mee. Va mish cur er nagh row mee rieau a

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
- Ẹnì kọ̀ọ̀kan ló ní ẹ̀tọ́ sí òmì nira láti ní ìmọ̀ràn tí ó wù ú, kí ó sì sọ irú ìmọ̀ràn bẹ́ẹ̀ 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

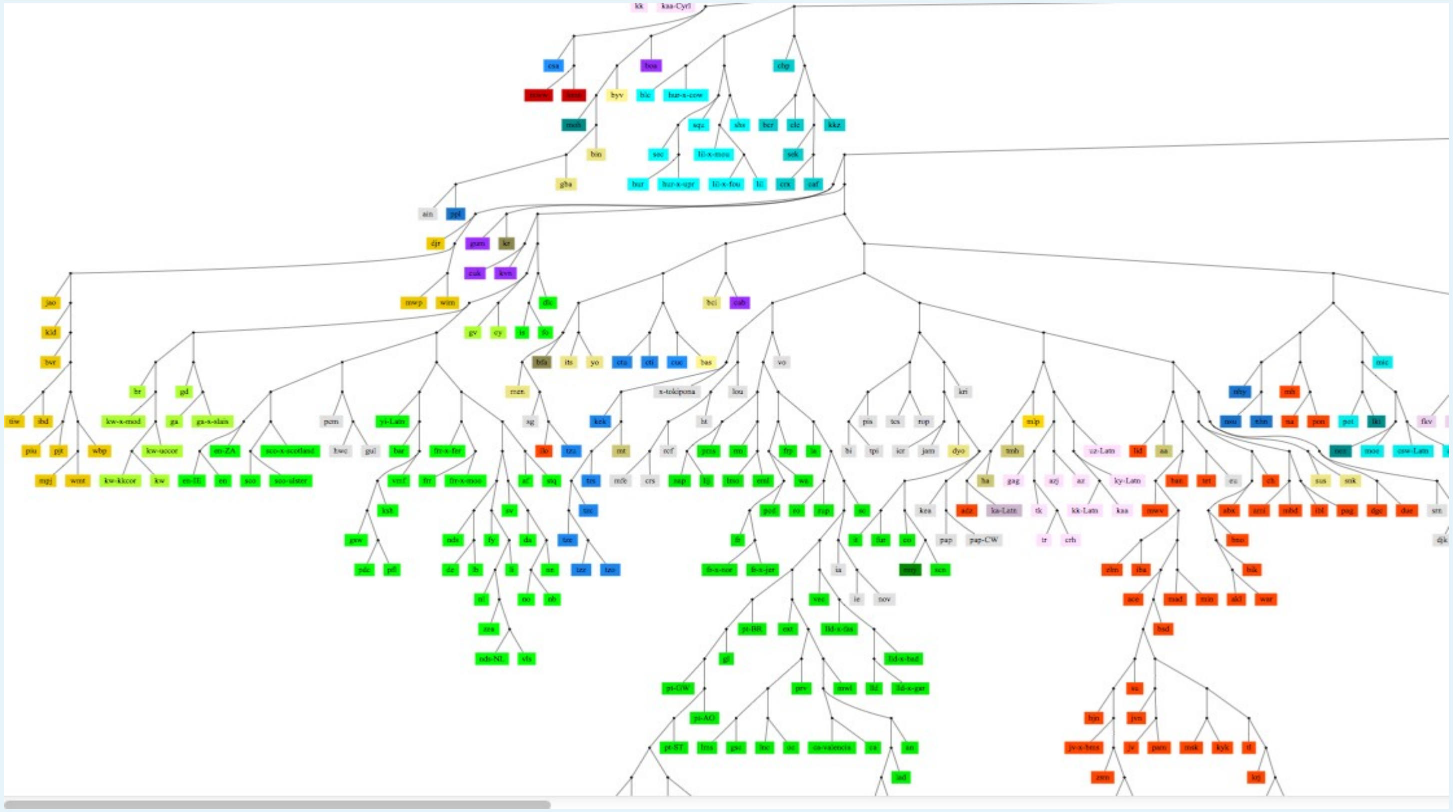
- Machine translation (gd/ga, gv/ga)
- Irish standardizer
- Speech recognition (Sphinx, KDE)
- OCR (e.g. Irish seanchló)

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.