The Crúbadán Project

Kevin Scannell
Department of Mathematics
and Computer Science
Saint Louis University
Missouri, USA

Project Goals

- Creation of web-crawled corpora for many minority and “under-resourced” languages
- Development of open source language-processing tools for these languages, in collaboration with native speakers
  - Clean word lists
  - Simple morphology engines for spell checking
  - Part-of-speech tagging
  - In some cases, parsers, semantic networks, and other advanced tools

Under-Resourced NLP

- Often no public funding available and no commercial interest in this work
- Few native speakers with NLP training
- Data are too scarce for “representativeness”
- To overcome these obstacles, we use:
  - An open source development model
  - Volunteer labor by language enthusiasts
  - Free, web-crawled corpora
  - Language-independent tools (like our crawler) deployed for a large number of languages
  - Unsupervised machine learning algorithms

Project Status

- Corpora for 416 languages
- For 355 of these languages, our crawler has been unable to find additional texts: “languages with a limited web presence”
- For the remaining languages (English, French, Chinese, etc.), we have only crawled enough text to generate reliable language recognition statistics
- 278 791 documents, 320M words, 2.5GB after conversion to plain text

Some History

- Circa 2000: Original software recursively downloaded entire web sites and then distinguished English and the six Celtic languages offline
- 2003-2004: turned this into a true web crawler and trained language models for 144 under-resourced languages (Crúbadán 1.0)
- 2007: expanded coverage to include many more non-Latin-script languages, bringing the total up to 416 languages (Crúbadán 2.0)

Fundamental Algorithm: Lexicon Generator

- Algorithm takes a corpus as input and tries to output a clean word list (the “lexicon”)
- This is done with a cascade of “noise filters”:
  - Tokens with unusual characters
  - Tokens with no vowels (if appropriate)
  - Tokens with improbable three-grams
  - Tokens with late titlecase or uppercase
  - Tokens which are words in a “polluting language”
  - Tokens which may have had diacritics stripped
  - Tokens not collocated with any high-freq. word
  - Additional language-specific filters when known

Design of the Web Crawler, I

- Independent processes target one language at a time, enabling more efficient coverage of languages with a limited web presence
- Each language has one or more “stopwords” for generating queries (either from a native speaker or extracted from frequency list)
- Queries are generated by OR'ing together random words from the “lexicon” and then AND'ing a stopword; these are sent to the Google API

Design of the Web Crawler, II

- Download URLs returned by Google, convert to text using standard open source tools
- Three-gram language recognizer is applied at the document level (character 3-grams), with language-dependent threshold
- Naive Bayes classifier in problematic cases
- If document is in the target language, extract all URLs and add them to the “pending” list
- When crawling finishes, strip duplicates, flag “unproductive” domains, generate new lexicon, update 3-gram statistics, etc.

Training New Languages

- All that is needed is a sufficient amount of text to generate reliable 3-gram statistics
- Thusfar most training data have come from three sites: the Wikipedia, the Jehovah's Witnesses web site, and the UN UDHR
- More recently, also using offline training texts for languages currently lacking any web presence
- With training texts in place and tokenized, the lexicon generator is applied and 3-gram statistics generated from this “clean” word list

Native Speaker Input

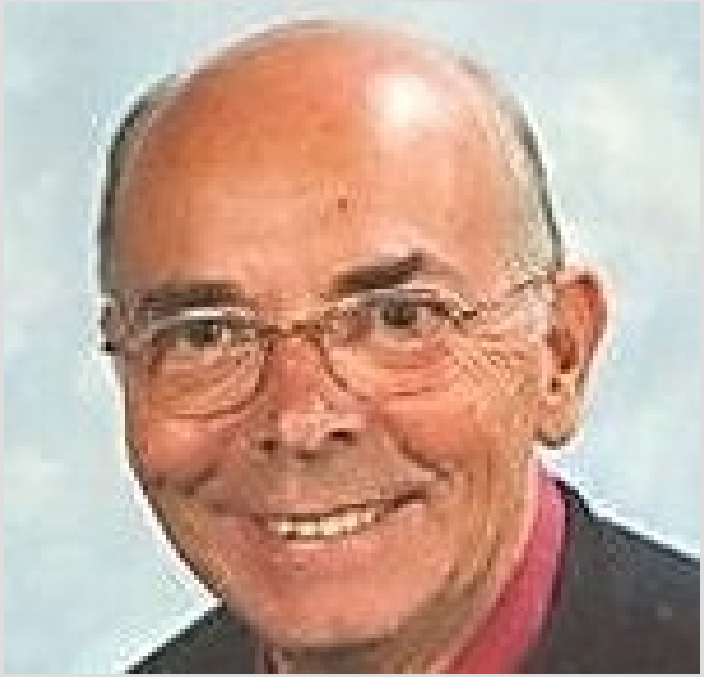
- Community-based effort is essential
- Web crawling
  - Help decipher legacy encodings (three examples in paper: Mongolian, Hawaiian, and Irish “be/al” vs. “béal”). XNLRDF (Streiter & Stuflesser)
  - Help with tokenization
  - Language-specific lexicon filters
  - Verify language recognition, separate dialects
- Applications
  - Editing and tagging word lists
  - Morphological analysis

New Open Source Spellcheckers

- Azerbaijani
- Chichewa
- Cornish
- Hiligaynon
- Irish
- Kashubian
- Kinyarwanda
- Kurdish
- Malagasy
- Manx Gaelic
- Mongolian
- Scottish Gaelic
- Setswana
- Tagalog
- Tetum
- West Frisian
- Coming Soon: Guarani, Hawaiian, Somali, ...

Case Study: West Frisian

- Germanic language with about 500 000 speakers, most in the Netherlands
- Done over three weeks in Feb. 2007 in collaboration with Eeltje de Vries, a retiree with a background in theoretical physics



Morphological Description

- Root words with one or two prefixes and one or two suffixes; not a full transducer
- This simplified description is easily encoded by novices and well-supported in open source tools (OpenOffice.org, Mozilla FF/TB)

```
# Affix file syntax:
# [PS]FX name strip add match

# moai->moaie, kreas->kreaze
SFX S 0 e [^esh]
SFX S ch ge ch
SFX S s ze s

# moai->moaier, kreas->kreazer
SFX T 0 er [^es]
SFX T 0 r e
SFX T s zer s

# moai->moaist, kreas->kreast
SFX U 0 st [^es]
SFX U 0 t s

...
```

Extract root words from corpus

```
wurdearje/V (5/5): wurdearje(18), wurdearrest(1), wurdearret(1),
wurdearre(26), wurdearren(3), wurdearjend(1)

reagearje/V (5/5): reagearje(15), reagearrest(1), reagearret(13),
reagearre(17), reagearren(3), reagearjend(1)

ynspirearje/V (4/5): ynspirearje(11), ynspirearrest(0), ynspirearret(2),
ynspirearre(23), ynspirearren(1), ynspirearjend(12)

studearje/V (4/5): studearje(27), studearrest(0), studearret(17),
studearre(34), studearren(4), studearjend(1)

konsumearje/V (4/5): konsumearje(1), konsumearrest(0), konsumearret(1),
konsumearre(2), konsumearren(1), konsumearjend(1)

funksjonearje/V (4/5): funksjonearje(7), funksjonearrest(0),
funksjonearret(9), funksjonearre(5), funksjonearren(1), funksjonearjend(1)

tramtearje/V (4/5): tramtearje(2), tramtearrest(0), tramtearret(1),
tramtearre(1), tramtearren(1), tramtearjend(1)

presintearje/V (3/5): presintearje(11), presintearrest(0),
presintearret(5), presintearre(34), presintearren(3), presintearjend(0)

komponearje/V (3/5): komponearje(1), komponearrest(0), komponearret(1),
komponearre(2), komponearren(1), komponearjend(0)

.........
```

Results

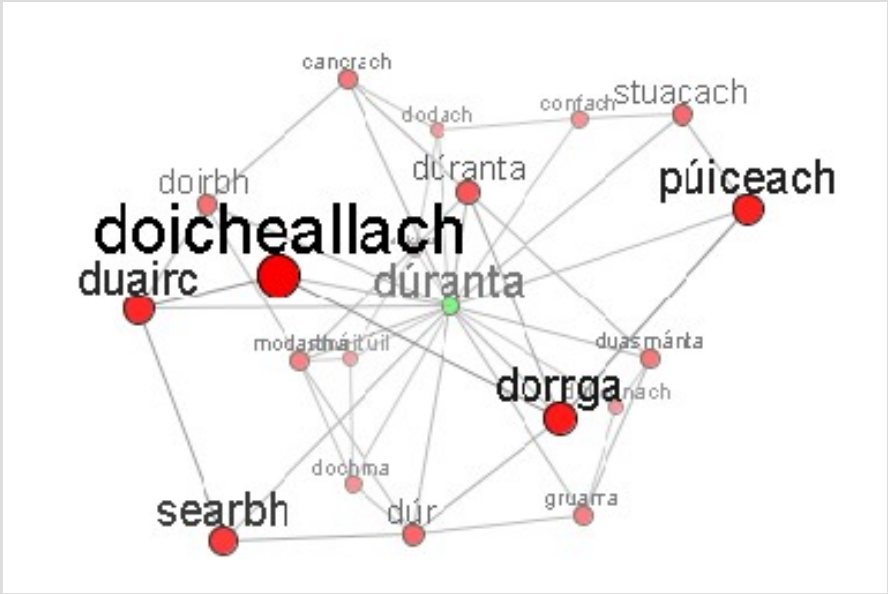
- Hand-checked lexicon with 22011 root words and 38677 derived forms
- Morphology ensures obscure inflected forms are included, unlike a pure corpus approach
- Spell checker recall is 91% on testing corpus (95% is an approximate expected upper bound on recall for uncleaned web corpora)
- It is also possible to train a part-of-speech tagger using Eric Brill's (unsupervised) transformation-based learning algorithm, but we have not evaluated this since we know of no existing tagged corpus

Semantic Networks

- Combine simple Xish-English dictionary, monolingual Crúbadán corpus for Xish, and Princeton WordNet for English
- Suffices to map Xish words over to one or more WordNet synsets
- Irish e.g.: *feileastram* defined as “flag, iris” (ambiguous English words), but collocated in the corpus with *bláth, féar* which mean (again using the dictionary) “bloom”, “grass”. These English words are closest to the correct WordNet senses of “flag” and “iris”.

3-Dimensional Graph Browser

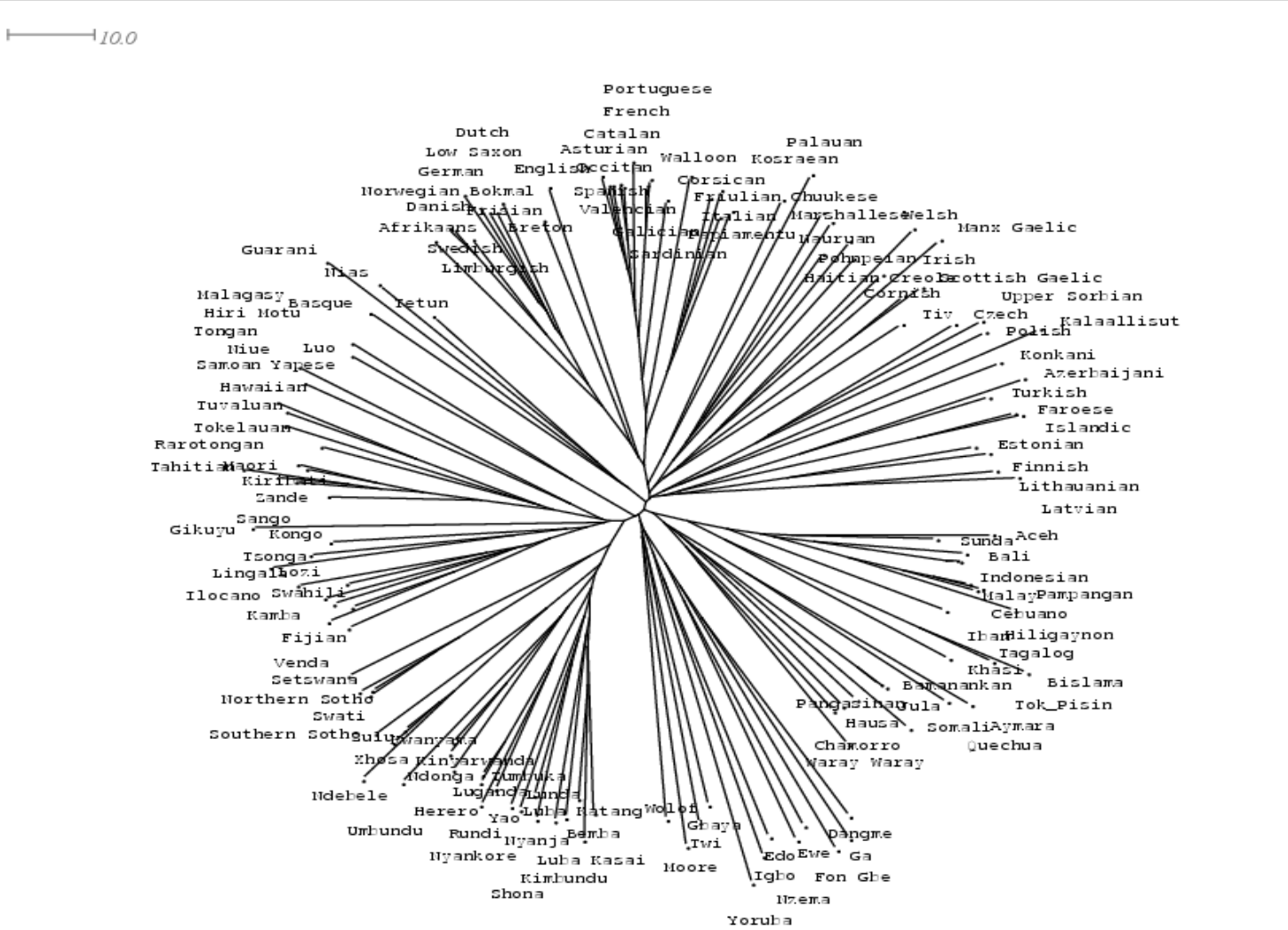
- Irish semantic network is available via an interactive three-dimensional graph browser written in Java
- Click nodes to navigate the network, click and drag to rotate in 3D



Phylogenic Tree Reconstruction

- Work in progress with Michael Cysouw of MPI Leipzig
- Use the huge grid of 3-gram cosines for all language pairs to reconstruct an “orthographic” family tree for languages
- The result matches up surprisingly well with the actual tree of language families, as encoded, e.g., in the Ethnologue
- Can improve this naïve approach by a transliterating non-Latin character sets, using coarse phonetics, etc.

The Orthographic Tree



Availability of Corpora

- We have provided corpora to more than 50 research groups and individuals working with under-resourced languages
- Only requirement is that results of research using the Crúbadán corpora be made freely available under an open source license
- Available as plain text without any special cleaning of boilerplate text or pollution
- Can also provide lists of URLs or raw HTML/PDF (if you have bandwidth and a preferred toolchain for converting to text)

Future Plans

- Apply high-quality cleaning algorithms to all corpora
- Continue to train new language models; goal is 1000 languages with the help of offline training texts
- Open source spell checkers for 100 languages
- Part-of-speech tagging for 25 languages

Call to Action

- Embrace an open-source approach when developing language technologies, especially for under-resourced languages or when public funding is involved (see forthcoming paper with O. Streiter and M. Stuflesser)
- Spend two or three hours a week working in support of an under-resourced language – there are thousands to choose from and many will not be around much longer

Thanks

We are grateful to 75 contributors from 40 countries for their tireless efforts of behalf of the Crúbadán project:

Metin Amiroff, Jacob Sparre Andersen, Jargal Badagarov, Sanlig Badral, Dwayne Bailey, Max Bane, Anneke Bart, Amos Batto, Kizito Birabwa, Paul Bowden, Luis Cardozo, Kenneth Rohde Christiansen, Joseph Colton, Jasmin Custic, Michael Cysouw, Eeltje de Vries, Francis Dimzon, Keola Donaghy, Roman Drzeżdżon, Alberto Escudero, Dewi Evans, Heiko Evermann, Alberto Fernández, Piotr Formella, Bruno Gallart, Biniam Gebremichael, Jason Githeko, Peter Gossner, Andrew Hawke, Paa Kwesi Imbeah, Ferli Deni Iskander, Denis Jacquerye, Orkhan Jafarov, Petri Jooste, Roger Kovacs, Gabe Lalasava, Sébastien Lanteigne, Ricardo Mones Lastra, Christin Livine, Chris Loza, Joe Maza, Mohamed I. Mursal, Rada Mihalcea, Muhirwe Jackson, Soyapi Mumba, Steve Murphy, Hirokazu Nakamura, Philibert Ndandali, Mike Nkongolo, Caoimhín Ó Donnaíle, Thapelo Otlogetswe, Iván Prieto Corvalán, Goran Rakic, Rado Ramarotafika, Brian Romanowski, Erdal Ronahi, Ramil Sagum, Pablo Saratxaga, Darrin Speegle, Oliver Streiter, Mathias Stuflesser, Andrea Tami, Toma Tasovac, Gurban Mühemmet Tewekgeli, Jim Thompson, Rêzan Tovjîn, Trond Trosterud, Ernie Turla, Chinedu Uchechukwu, Mathieu van Woerkom, Tjaart Van der Walt, Thierry Vignaud, Michel Weimerskirch, Edi Werner, and Mark Williamson.