An Gramadóir

A grammar-checking framework for the Celtic languages and its applications

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

An Gramadóir

- Language-independent engine for building language processing software
- Primarily this means proofing tools (spelling and grammar), and occasionally more advanced tools (machine translation, etc.)
- Flagship application: Irish grammar checker
- Focus is on under-resourced languages: so far, the six Celtic languages, Akan, Frisian, Hiligaynon, Igbo, Kashubian, Kurdish, Malagasy, Mongolian,Tagalog, Walloon, ...

Statistical Methods

- Break the “information bottleneck”
- Web crawler gathering texts in 419 languages, running as we speak
- Texts are used initially to build up a lexical database for each language
- Word frequency counts are used by the grammar checking software in several ways
- “Machine learning” algorithms allow the program to perform tasks such as tagging words by part of speech with high precision

Free Software

- a.k.a. “open source” software
- Free as in “free speech” not “free beer”
- You are free to copy, modify, distribute or even sell my programs as long as your modified or redistributed versions preserve these same freedoms for others
- Rules, data, and code contributed by users
- Other people can do cool things with the data
- Essential for long-term viability of technology for under-resourced languages

Irish Grammar Checking

- High-quality spell checking (517 000 word database, with caveat below)
- Warnings given for correct but low-frequency words (*stáid* for *staid*, *ata* for *atá*)
- Catches many “real word” spelling errors in context: (*in ait/in áit*, *deifir idir/difear idir*)
- 2500 common spelling errors (*caisléan,* etc.)
- Checks hundreds of usage errors: e.g. *dócha* without copula, *an* vs. *na* rules, *cá bhfuair* but *cár tharla,* require genitives after *le haghaidh, trasna*, etc., singular after numbers, *beirt*

Initial Mutations

- Séimhiú missing: *bean maith, an bean, cóta an fir, cóta an fhir móir, fir móra, ba ceart go, ceithre ball, mo ceann, ...*
- Séimhiú not needed: *an fhear, fear mhaith, bean thuaithe, easpa bheathaithe, scáileáin theilifíse, an dhá theanga, ...*
- Urú missing: *deich ball, cén chaoi a oibríonn, an fuair tú, cá fuair tú, ár dhá cos, go beidh, bosca ina cuireann sé é, ...*
- Urú not needed: *do mbean, cár dtarla, don bpobal, ...*
- Prefix missing: *fichiú aois, a ocht, cá áit, cé é, chomh iontach le, an asal, den seachtain, guth an seanadóra, ...*
- Prefix not needed: *an himreoir, na t-imreoirí, an t-acmhainn, an tseanadóir, ...*
- 2310 rules in all!

Using An Gramadóir

- Enables learners to write with confidence and participate in email discussions, blogging
- Runs standalone on Mac, Linux, Windows
- With some wrestling, runs under OpenOffice
- Commercial version by Cruinneog on Mac and a Windows version coming soon
- Web interface: http://borel.slu.edu/gramadoir/

Other Languages

- Breton: word list only, 32K words (T.Vignaud)
- Cornish: 8626 words (Kemmyn), 50 mutation rules implemented (E. Werner, P. Bowden)
- Manx Gaelic: word list only, 32K words
- Scottish Gaelic: 76K words, 20 rules; Dearbhair list (500K+) to be made free soon
- Welsh: 200K words, 150 rules (K. Donnelly); available online as “Klebran”: http://www.klebran.org.uk/

Application: Standardizer

- The lexical database contains about 75 000 non-standard/standard word pairings: *abhallghort/úllord, fairrge/farraige*, etc.
- Many individual spelling rules (*sg→sc, amhail→úil, idhea→ío*, etc.) and non-standard inflections (*-eóchthá→-ófa, -aghadh→-ú, -ighinn→ínn*, etc.) implemented as well
- Automated standardization being applied to Foras na Gaeilge corpus texts for new English-Irish dictionary

Application: Text Evaluator

- Assigns a numerical score to a given text along three axes: standard vs. non-standard, high vs. low quality, and easy vs. difficult
- First two are measured by classifying the messages output by An Gramadóir into *standardizations* and *errors* (# per 1K words)
- The third measure uses word frequency counts in a more sophisticated version of the well-known Flesch-Kincaid score (years)
- Useful for evaluating potential reading material for students

Sample Output

- Book of Ruth, Ó Fiannachta Bible
- Article from An Phoblacht, 14 Dec. 2000
- S. Mac Meanman, Rácáil agus Scuabadh, 1955

Text	Non-stndness	Badness	Difficulty
Ruth	1.13	3.77	11.15
APRN	8.86	39.43	2.51
MacM55	205.94	10.68	11.51

Thanks

- Work partially supported by Foras na Gaeilge
- A big thanks to the many people around the world who have contributed to the project:

  Sanlig Badral, Anneke Bart, Paul Bowden, Thomas Cruau, Francis Dimzon, Kevin Donnelly, Paddy Dwyer, Bruno Gallart, Laurent Godard, Martin Gregory, Eugen Hoanca, Roland Illig, Paa Kwesi Imbeah, Ji ZhengYu, Petri Jooste, Andrej Kacian, Myriam Lechelt, Diarmaid Mac Mathúna, Ciarán Mac Samhráin, Alastair McKinstry, David Moreau, Tim Morley, Daniel Nylander, Andrew Ó Baoill, Séamus Ó Coileáin, Caoimhín Ó Donnaíle, Michel Robitaille, Ramil Sagum, Pablo Saratxaga, Benno Schulenberg, Clytie Siddall, Pétur Thors, Chinedu Uchechukwu, Tjaart van der Walt, Hugo Voisard, Edi Werner