Standardization of corpus texts for the NEID

Kevin P. Scannell
Saint Louis University

May 22, 2009

English-Irish Dictionaries

- Ó Beaglaoich & Mac Cruitín, 1732 (15000 words)
- Ó Coinnialláin, 1814 (8000 words)
- Foley, 1855 (22000 words)
- Fournier d'Albe, 1903 (21000 words)
- O'Neill Lane, 1904 (1st ed.)
- O'Neill Lane, 1915 (“Larger” - 33000 words)
- Mac Cionnaith, 1935 (12000 words)
- de Bhaldraithe, 1959 (42479 words)

De Bhaldraithe 1959

- Work began September 1945
- Harrap English-French dictionary as model
- Data assembled on “slips” from reading books, journals, and reversing and alphabetizing Dinneen's 1927 dictionary; roughly the first five years of the project
- Everything written, sorted by hand, on paper
- Entries drafted between Mar. '50 – Aug. '56
- Standard reference since its publication in 1959, but now sorely out-of-date

The New English-Irish Dictionary (NEID)

- First major E-I dictionary in 50+ years
- Project funded and run by Foras na Gaeilge
- Will contain about 40000 English headwords
- Began in 2003, aiming at publication in 2012
- First full-scale Irish dictionary to use modern, corpus-based lexicography
- English entry frameworks underway (through July 2010) based on 1.7 billion word corpus
- Irish entries will be drafted with the help of large Irish and bilingual (parallel) corpora

An Caighdeán Oifigiúil

- “The Official Standard” for Irish spelling and grammar, laid out in the 1940's and 1950's
- Adopted widely though some writers prefer dialect forms, not always strictly conformant
- Good: simplified and normalized spelling and grammar rules; easier for learners
- Bad: sometimes hides etymologies (muiceoil=muicfheoil, comhlacht=comhlucht, iomasach=imfhiosach, aicearra=aithghearra); breaks continuity with Scottish Gaelic, literary Irish, i.e. “lossy” conversion: bá= bàthadh (drowning), bàidh (sympathy), bàgh (bay)
- Me: neutral.

C. O. ⇒ Search Problems

- Some of the best writing in Irish is from the 1930's (translations especially); all in pre-standard orthography
- Very important material for NEID corpus
- Nearly impossible to use these texts for corpus lexicography since the pre-standard spellings effectively “hide” sample texts from the lexicographer (33 variants of “lucharachán” in my corpus for example)

Irish Standardizer

- Program that converts pre-standard texts to the Caighdeán Oifigiúil
- “Acht chan abair tú a choidhche i n-aor nó i gcúl-chainnt gur cuireadh cosg le filidheacht i nDún na nGall”
- “Ach cha   ndeir tú a choíche   in aor nó i gcúlchaint   gur cuireadh cosc le filíocht   i nDún na nGall”
- First component: “morphological transducer” encoding pre-standard word structure and C.O. spelling changes
- (coimh-mheasguighthe → comh-mheasguighthe → cóimheasguighthe → cóimheascuighthe → cóimheascaighthe → cóimheascaithe → cóimheasctha)

Irish Standardizer (cont.)

- Second component: large database of standard/non-standard pairs, from dictionaries and harvested automatically from aligning standard/non-standard texts
- Third: Techniques from statistical machine translation (n-gram statistics), among other things this handles “real word” errors:
- “Scríobh mé leitir” (leitir=hillside, here alt. “litir”), “Bhí náire ortha” (ortha=spell, here alt. “orthu”), “Ba bhreá léithe é” (léithe=greyness, here alt. “léi”), “Chuaigh mé annsan” (annsan=emph. form ann, here alt. “ansin”), “Tá sí ar thoiseach an tslua” (toiseach=adj. dimensional, here alt. “tosach”).

Applications

- Useful as an indexing tool *whether or not* you like the C.O.! Makes pre-standard texts accessible to search; standardized versions can remain invisible to user. Can also search standard texts via non-standard queries.
- Linguistic research on pre-standard texts?
- For learners: standardized versions of important books from the 1930's?
- Scottish Gaelic ↔ Irish Gaelic MT; Google Summer of Code grant with Sean Burke, undergraduate at University of Montana