Semi-automated construction of semantic networks using web corpora

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

Semantic Networks

- A semantic network is a database of words and semantic relationships between them
- Synonyms and antonyms, like in traditional thesauri, but also holonyms and meronyms (whole/part), hypernyms and hyponyms (broader/narrower), and more
- Princeton WordNet, for English, was the first full-scale semantic network to be developed; latest version has well over 100K “synsets”

Applications

- Created with computational applications in mind, primarily in NLP and AI
- Word sense disambiguation (gives a sense inventory and a notion of semantic distance that can be applied to raw texts)
- Information retrieval (index by synsets)
- Summarization, genre classification
- All other NLP tasks informed by “real world knowledge”, e.g. Subcategorization, attachment *(“I saw the cat with the telescope”)*
- End-user applications too; e.g. *visuwords.com*

Language Survey

- See www.globalwordnet.org
- Semantic networks listed for 46 languages
- Princeton has always made the English WordNet freely available, for any purpose; explains in part its huge impact on NLP
- Not so for “EuroWordNet” (de,es,fr,...), “BalkaNet” (bu,el,ro...) unfortunately
- Only 5 of 46 freely available (ar,en,ga,he,hi)
- Lots could be done with free, interlinked networks in many languages

Constructing Semantic Networks

- WordNet is the only one to have been constructed manually by lexicographers
- Most non-English languages induce semantic relationships by mapping words to WN and carrying English relations over
- Often the mappings are done manually
- For Irish, the mappings to English were achieved almost entirely automatically, based on statistical methods using bilingual corpora
- Encodes the knowledge of legions of Irish translators and lexicographers

Monolingual Web Corpora

- Web crawlers running right now and gathering texts written in 427 languages
- Statistical language recognition based on “character n-grams”
- Huge corpora for many “under-resourced” languages; e.g. 100M words of Welsh, 100M words of Irish gathered in 2005
- For many languages, I have everything!
- Being used widely: lexicography, proofing tools, MT, pure linguistics research

Bilingual Web Corpora

- A separate crawler looks for parallel texts for certain language pairs (usually xx/en)
- Heuristics to start, e.g. anchor text containing the word “English”, tricks based on URL
- Statistical approach to recognizing documents that are translations: learns high mutual information pairs (often dates, personal names, etc.)
- Productive domains are crawled exhaustively and all xx/en pairs checked

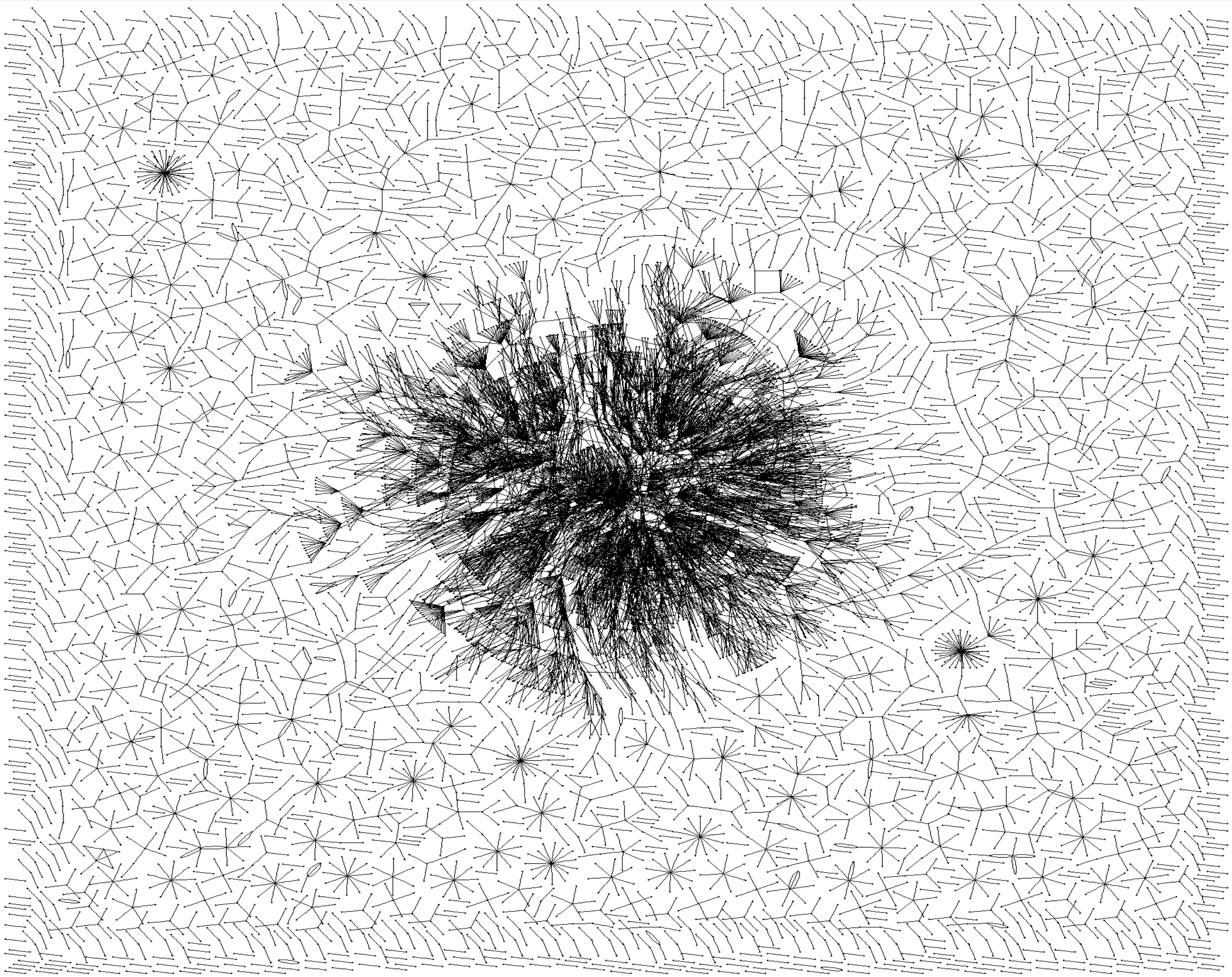
Building the Network, I

- The goal is to map each Irish word to one or more English synsets
- We start with a simple Irish-English dictionary, with one- or two-word glosses in English *e1,...,en* for each Irish word *g*.
- Need to select the correct synset for *e1,...,en*
- If *ei* is unambiguous (just one synset), done; stáplóir → stapler
- If not, use the parallel corpus; find instances where *g* appears on the Irish side and *ei* appears on the English side

Building the Network, II

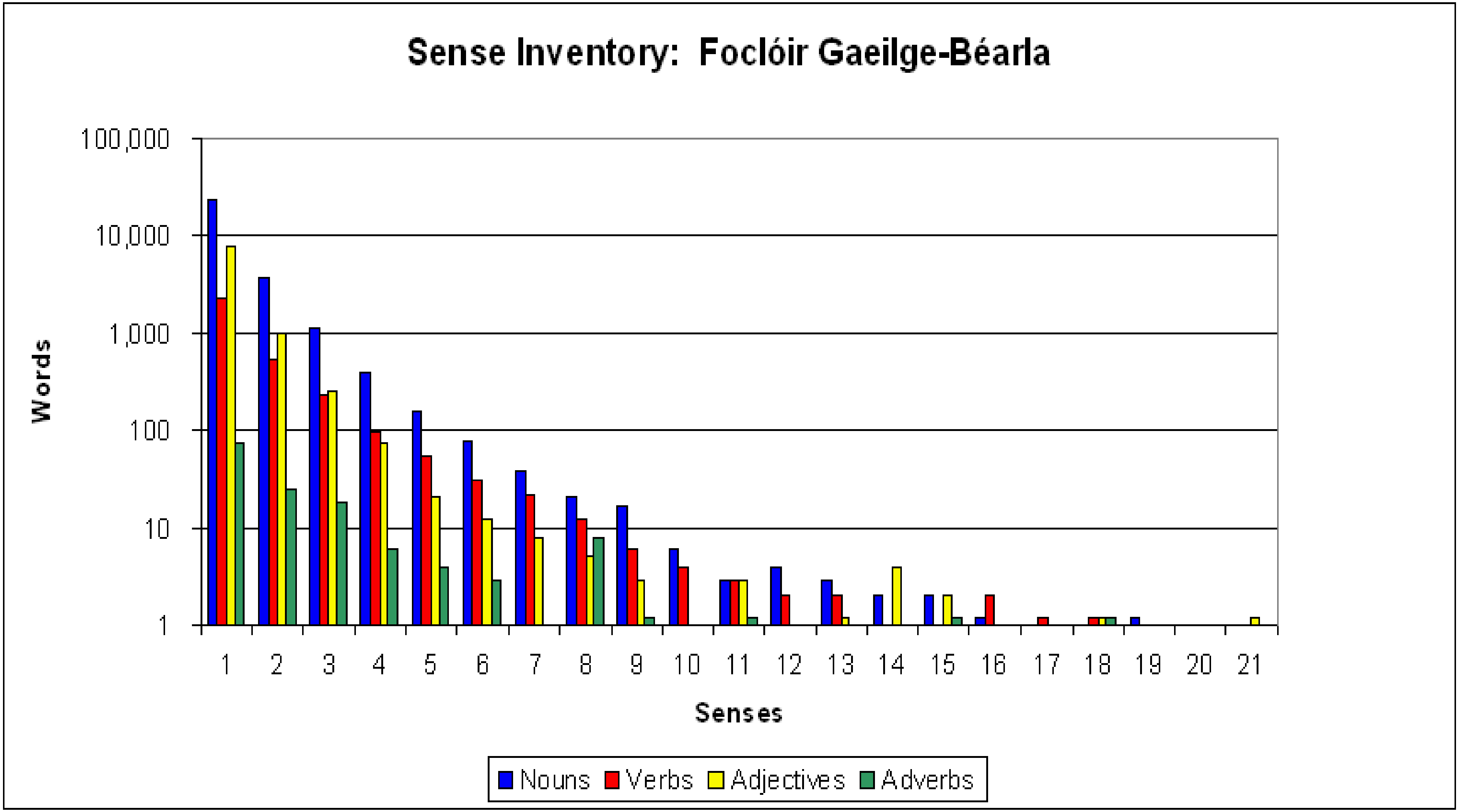
- Use contextual clues near *ei* in the English corpus, plus knowledge encoded in WordNet, to choose the best synset
- cairéal=quarry; parallel corpus citation: “the quarrying and cutting of slate at a quarry, the quarrying in rough blocks of marble or stone” = “slinn a bhaint agus a ghearradh i gcairéal, garbh-bhloic marmair nó cloiche a bhaint”
- The initial bilingual dictionary goes into the parallel corpus too, and this is enough to resolve many words: “feileastram” = “flag, iris”
- Compare Diab and Resnik, 2002, similar approach using word alignment (Giza++)

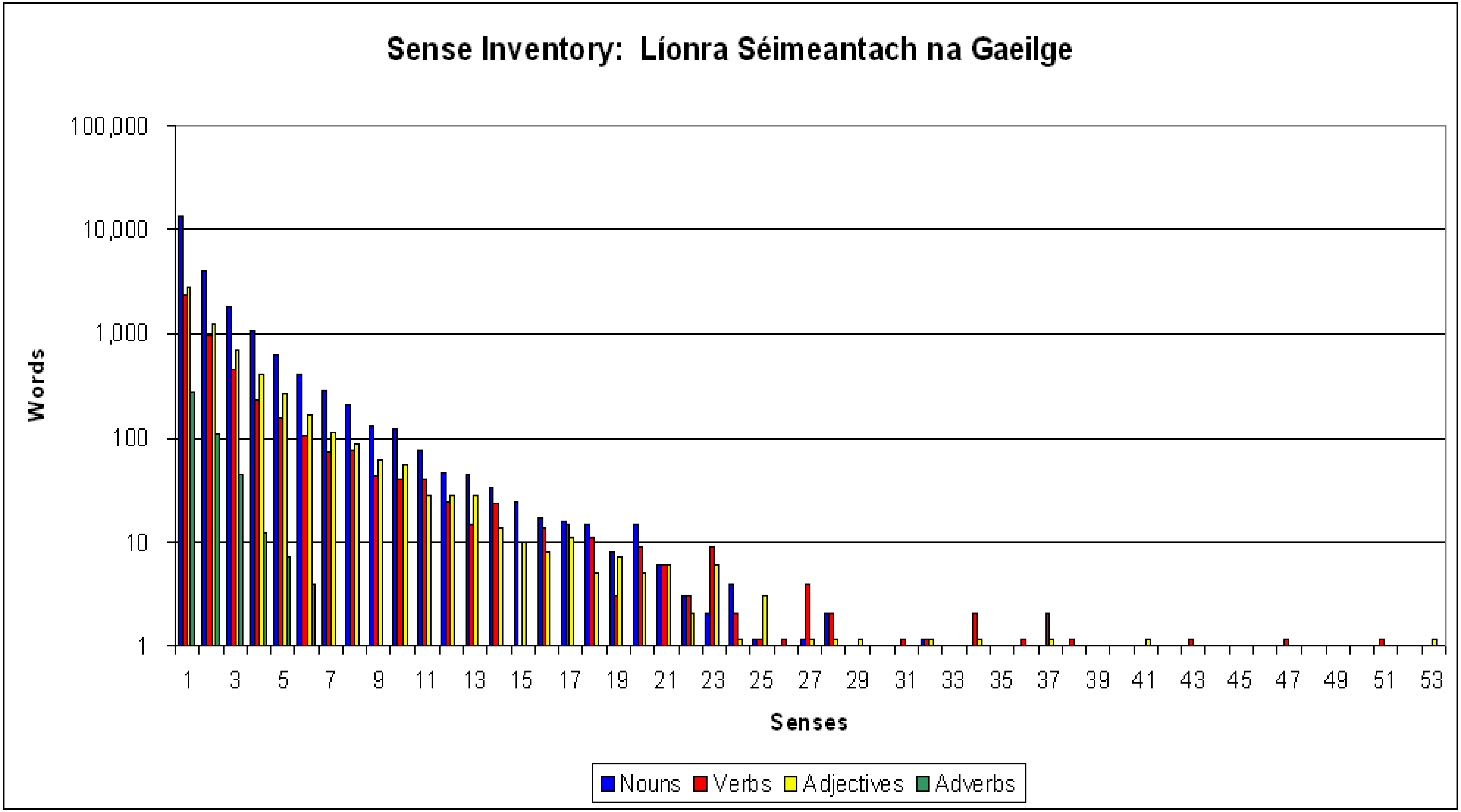
Results



Results, really

- 36262 headwords, 77596 senses organized into 32742 synsets
- Bilingual resource: linked w/ WordNet
- Open source license
- Statistical techniques → one-person job
- Approaching the scale of the standard Irish-English dictionary (Ó Dónaill Foclóir Gaeilge-Béarla) in terms of headwords (44K)
- But a much finer division of senses than one finds in Ó Dónaill as we see in the following slides...


Sense Inventory: Foclóir Gaeilge-Béarla
Words
100,000
10,000
1,000
100
10
1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
Senses
Nouns
Verbs
Adjectives
Adverbs



Sense Inventory: Líonra Séimeantach na Gaeilge

Semi-automated?

- Algorithm fails in cases where there is not enough corpus evidence to map to a single WordNet sense
- Performed some clustering of WordNet senses to alleviate this problem
- Problematic cases mapped manually
- Linguistic relativists will worry that we are imposing English semantics onto the Irish network; indeed a problem for distinctions not lexicalized in English (*dearg* vs. *rua*), but these are very much the exception

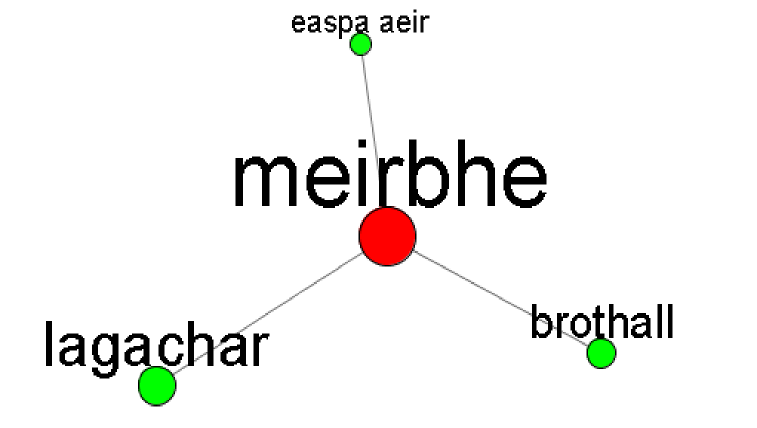
Semantic Network as Thesaurus

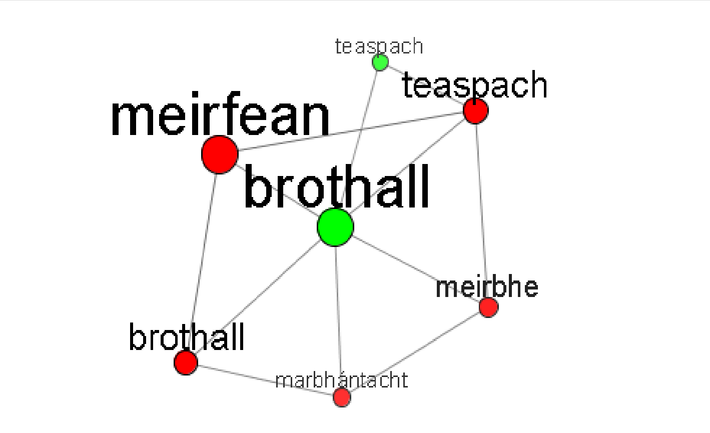
- Database exports as a hyperlinked thesaurus, but better than traditional thesauri because of the richer semantic relationships
- Available as a free PDF book with embedded cross-references, 997pp
- Available for use with the free office suite OpenOffice.org

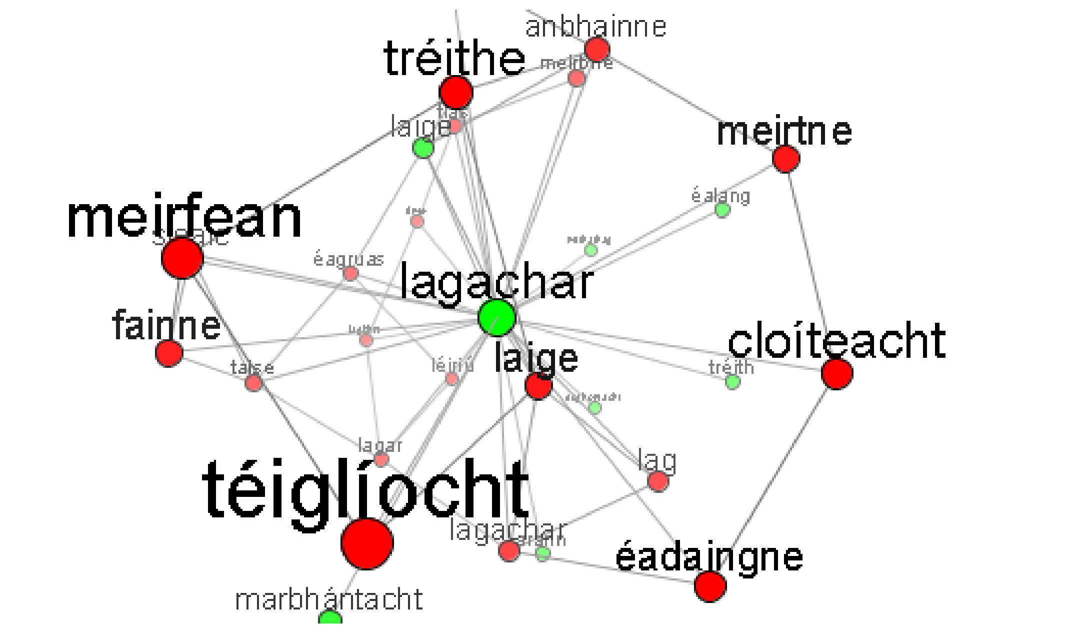
Homograph Disambiguation

- Work in progress, with Joshua Glatt, undergraduate at Washington University
- Ide and Wilks proposed homographs as the appropriate level for WSD in NLP
- No full scale systems for this, even in English
- Around 44K headwords in FGB, just 940 homographs, most (763) being nouns.
- 828 have two senses, 76 with three, 8 with four, and 1 with five (“bulla”)
- Based on looking at real corpora, less than 400 need to be considered

3D Graph Browser


easpa aeir
meirbhe
lagachar
brothall


teaspach
teaspach
meirfean
brothall
meirbhe
brothall
marbhántacht


anbhainne
tréithe
meirbhe
laige
meirtne
éalang
meirfean
éagruas
lagachar
fainne
cloíteacht
laige
léiriú
talse
tréith
lag
lagar
téiglíocht
lagachar
éadaingne
marbhántacht