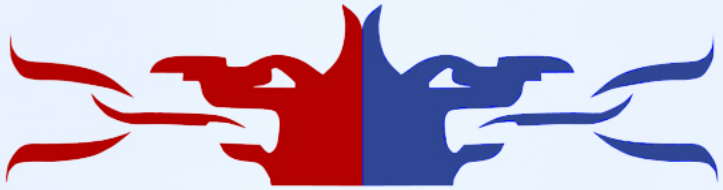


Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

Mining corpora using the Bangor Autoglosser

Kevin Donnelly

ESRC Centre for Research on Bilingualism
Bangor, Wales

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Cyngor Cyllido Addysg Uwch Cymru
Higher Education Funding Council for Wales

hefcw

Llywodraeth Cynulliad Cymru
Welsh Assembly Government

BANGOR

Outline

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Briefly introduce the autoglosser
- Give examples of using its output to mine the ESRC corpora in various ways
- Seek input re other possible uses

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

The Bangor Autoglosser

Bangor corpora

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

	Chats	*Hours*	*Words*	*Date*
Welsh-English (Siarad)	69	40	456k	2009
Welsh-Spanish (Patagonia)	32	20	161k	2011
Spanish-English (Miami)	31	20	126k	2011
	132	**80**	**743k**	

All available under the GPL.

Glosses

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Part-of-speech tags
- **walks**: walk.V.3S.PRES
- Allow non-fluent speakers to parse the conversation
- Labour-intensive, can be inconsistent
- Difficult to add further tags later
- **Solution**: use a computer to generate the glosses!

Sample file

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Demo: import and autogloss a file containing two utterances

Benefits

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Handles multiple languages simultaneously
- Handles broken text
- Highly configurable
- Reasonably fast: 1000 words a minute
- Accuracy 98%+
- First tagger for Welsh – other uses possible
- Open license (GPL) – reproducible science

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

Mining the corpora

Data analysis

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Text now in a database
- Demo: data grid interface
- Easy basic queries (eg) how many words per speaker, or per language tag
- Demo: browse columns
- Demo: query for words not in the dictionary
- Effective and flexible typo correction

Presentation of output

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Typeset files using LaTeX gloss alignment packages
- ExPex (John Frampton), Xscribe (Natalie Weber)
- Demo: sample file pdf
- Easier to read and review the text
- Easy to generate indexes and concordances
- Demo: pages from two summaries

Sequence selection

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Surface and glossed tiers in transcription file are linked at utterance level
- Database allows linking at word level
- Words can be selected based on the gloss
- Sub-select based on position (before or after the target word)
- Filter by gloss if necessary

Sequences

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Det+N+Adj (Spanish/English)
- the fair estúpido
- el tremendo tip
- About 26 sequences out of 126,000 words
- Non-sandwich (non-nested) possessive sequences (Welsh/English)
- cath fi instead of fy nghath [i]
- 340 1S instances – only 5 are sandwich-type
- Handout

Sampling

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Can be done more rigorously
- Order the utterances
- Select the utterance at the 25%, 50% and 75% points
- Sub-select ...
  - 5 utterances before and after
  - by each speaker
  - containing a finite verb

Clause delineation

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Ideal? – build dependency trees
- Surprisingly good proxy can be achieved by using
  - verbs in Welsh
  - conjunctions in Spanish
  - subject pronouns in English
- Welsh splitter has error rates of 2-5%
- Demo: clause-split the simple file

Data aggregation

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

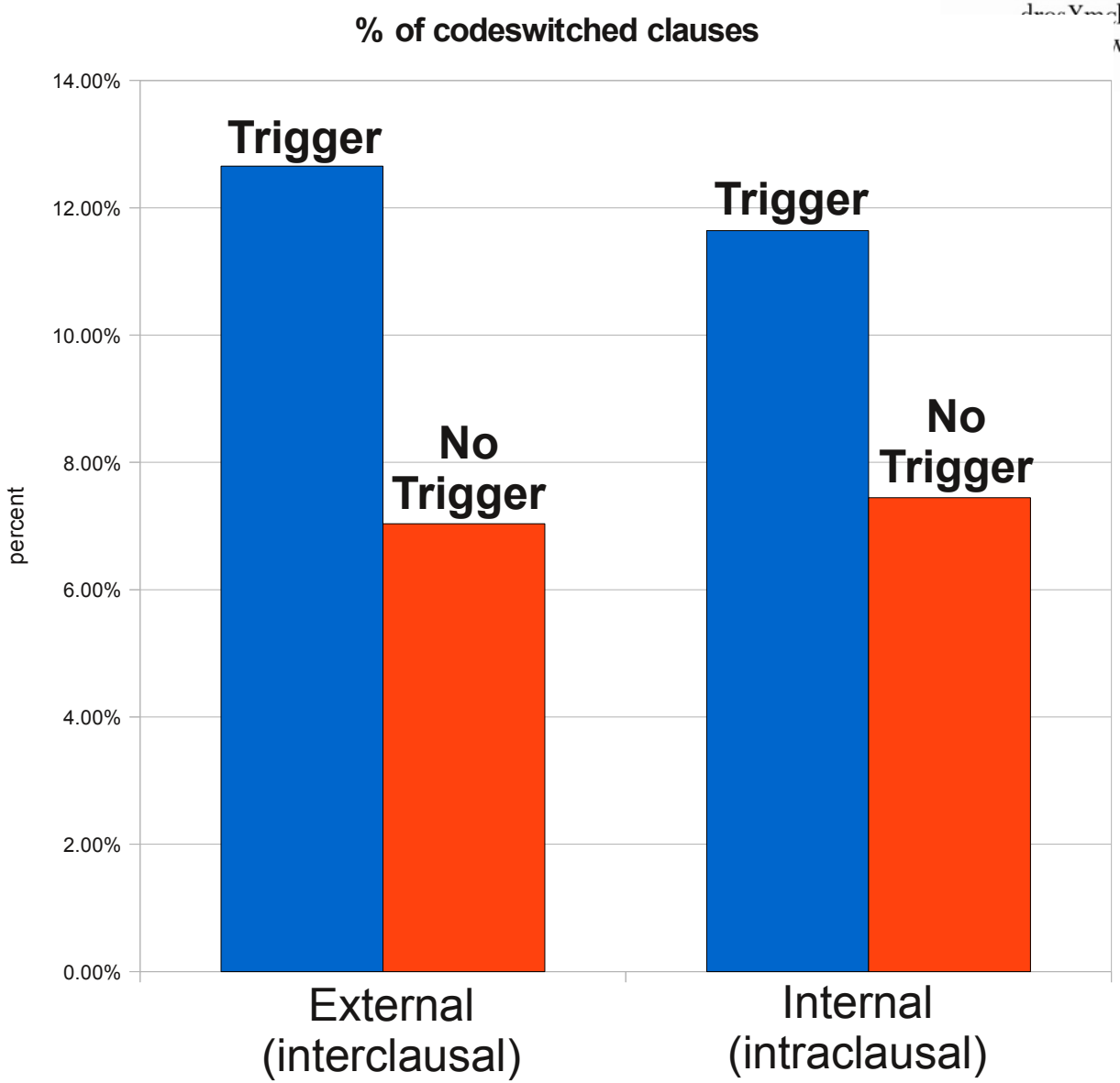
- Outputs can be combined with metadata
- Sociolinguistic analysis
- Are particular structures correlated with speakers' background?
- Demo: spreadsheet

Data reformatting

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Are codeswitches more frequent in the vicinity of triggers (cognates)?
  - names, proper nouns
  - identical words: car
  - related words: perswadio/persuade
- Split into clauses
- Group into speech-turns
- Mark triggers and count them
- Count codeswitches between and within clauses

Results


% of codeswitched clauses
14.00%
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
percent
Trigger
No Trigger
Trigger
No Trigger
External (interclausal)
Internal (intraclausal)

Timespend

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Original 2,500-word Dutch text took 50+ hours to analyse
- 456,000-word Siarad corpus analysed in 4 hours

Running text

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Input written formal text
  (instead of transcribed colloquial dialogue)
- Historical Corpus of Welsh (420,000 words)
- *Berr Hanes* (James Groniosaw) – 1779

Results

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- 1,200-word sample passage, glossed in 2 minutes
- 57% of words correctly glossed
- 3% wrongly glossed
- 40% had two or more glosses offered
- Grammar rules need to be extended to cope with this type of text.

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

Comments?

Suggestions?

http://bangortalk.org.uk