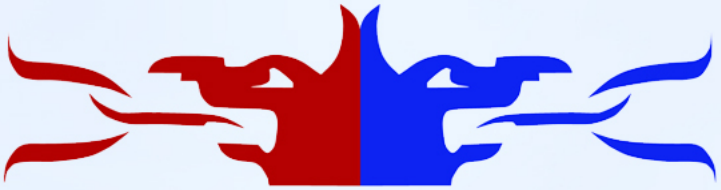


Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

Autoglossing bilingual data

Kevin Donnelly

ESRC Centre for Research on Bilingualism, Bangor

November 2010

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

Acknowledgements

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Margaret Deuchar
- ESRC Centre
- Brian MacWhinney and Leonid Spektor
- Colleagues at the ESRC Centre

Why gloss?

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Lexemes and part-of-speech (POS) tags
- Helps non-native speakers parse the conversation
- Allows morphological analysis

Example

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

***ALN:** +" oedd@1 o@1 (y)n@1 edrych@1 fath@1 â@1 cael@1 snog@2 pan@1 wnes@1 i@1 basio@1 !

%gls: be.3S.IMP PRON.3SM PRT look.NONFIN kind with have.NONFIN snog when do.1S.PAST PRON.1S pass.NONFIN

%eng: it looked like having a snog when I passed!

(Siarad corpus, stammers4)

Difficulties

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Time-consuming
- Inconsistency and errors
- Tag choice difficult to revise later
- No automatic method for small languages

Questions

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- If glossing is **A Good Thing** . . .
- Can we automate the process?
- Can we add value to the texts?

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

Can we automate the process of glossing?

A sequence . . .

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

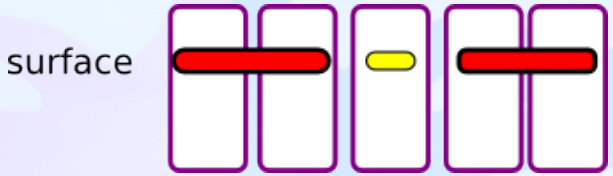
- The speech tier = a horizontal stream

- Additional tiers add vertical depth

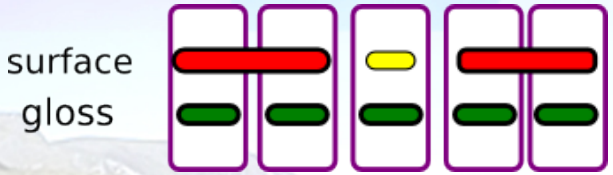
. . . segmented

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Limiting the domain to the word . . .



- . . . provides the basis for glossing automatically



Process

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Read the lines of the chat file into a database table
- Segment each line into words
- Look up the words in a digital dictionary
- Disambiguate using constraint grammar
- Write the results into a gloss tier, using Leipzig schema

Results for Welsh – manual

***ALN:** +" oedd@1 o@1 (y)n@1 edrych@1 fath@1 â@1 cael@1 snog@2 pan@1 wnes@1 i@1 basio@1 !

%gls: be.3S.IMP PRON.3SM PRT look.NONFIN kind with have.NONFIN snog when do.1S.PAST PRON.1S pass.NONFIN

%aut: be.V.3S.IMPERF he.R.M.3S.SPOKEN stative.S look.V.INFIN type.N.M.S.+SM as.C have.V.INFIN snog.V .INFIN when.C do.V.1S.PAST.SPOKEN.+SM I.R.1S pass.V .INFIN.+SM

%eng: it looked like having a snog when I passed!

(Siarad corpus, stammers4)

Results for Welsh

Canolfan ESRC Centre dros Ymchwil i Ddwyieithrwydd for Research on Bilingualism

***AVR:** neu dylai bod fi wedi mynd (be)cause@s:en mae (y)n hwyr rŵan .

%aut: or.CY.C ought.CY.V.3S.IMPERF be.CY.V.INFIN I.CY.R.1S after.CY.P go.CY.V.INFIN because.EN.C be.CY.V.3S.PRES stative.CY.S late.CY.A now.CY.B

%eng: or I ought to have gone because it's late now

(Patagonia corpus, patagonia2)

Results for Spanish – MOR

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

***LAR:** +" porque tú me apoyas en todo sabes .

%mor: conj|porque=because pro:per|tú=you pro:per|me=me vpres|apoya-2S&PRES=support prep|en=in det:indef|todo-MASC=all co|sabes=you_know^vpres|sabe-2S&PRES=know .

%aut: because.CONJ you.PRN.SUBJ.MF.2S me.PRN.OBJ .MF.1S support.V.2S.PRES on.PREP everything.PRN.M.SG know.V.2S.PRES

%eng: because you support me in everything, you know

(Miami corpus, zeledon14)

Results for Spanish

Canolfan ESRC Centre dros Ymchwil i Ddwyieithrwydd for Research on Bilingualism

***SEB:** ellos@3 mataban@3 a@3 la@3 gente@3 como@3 nosotros@3 .

%aut: they.PRN.SUBJ.M.3P kill.V.3P.IMPERF to.PREP the.DET.DEF.F.SG people.N.F.SG like.PREP we.PRN.SUBJ .M.1P

%eng: they would kill people like us

(Miami corpus, herring7)

Benefits

Canolfan ESRC Centre dros Ymchwil i Ddwyieithrwydd for Research on Bilingualism

- Speed: 2 minutes/30-minute conversation
- Consistency: *ychydig* – “a bit”/“a little”
- Handles any number of languages in one pass
- Extensible
- Re-uses existing resources and tools
- Transferable skills

Results

Canolfan ESRC Centre for Research on Bilingualism
Canolfan drosYmchwil i Ddwyieithrwydd ESRC Centre for Research on Bilingualism

	Welsh	Spanish
Coverage (all words)	88%	96%
Tokens	5224	4827
Correlation (nouns)	82%	85%
Accuracy (nouns)	93%	97%
Nouns	459	380
Files	*stammers4*	*zeledon14*

Drawbacks

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Like MOR, still needs checking!
- Dictionary cleaning can take some time
- Rules take time to write and test

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

Can we add value to the texts?

Texts

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Check on typos – proof-reading
- Consistent glosses
- More granular analysis
- Global tag changes or enrichment

Accessibility

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

123									
124	RIC	siempre	está	en	los	mismos	esa	0	
125	aut	always.ADV	be.V.3S.PRES	on.PREP	the.DET.DEF.M.PL	same.ADJ.M.PL	that.ADJ.DEM.F.SG		
126									

Accessibility

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Interactive webpages (*siarad.org.uk*)

Words with language tag cy (2240)

yn (PRT) *[132]*, **i** (PRON.1S) *[72]*, **ti** (PRON.2S) *[58]*, **mae** (be.3S.PRES) *[53]*, **yr** (DET) *[49]*, **o** (PRON.3SM) *[46]*, **yna** (there) *[44]*, **i** (to) *[43]*, **be** (what) *[40]*, **y** (DET) *[40]*, **ddim** (NEG) *[36]*, **yn** (in) *[34]*, **na** (no) *[33]*, **wedi** (PRT.PAST) *[33]*, **ydy** (be.3S.PRES) *[31]*, **dw** (be.1S.PRES) *[29]*, **mynd** (go.NONFIN) *[29]*, **oedd** (be.3S.IMP) *[29]*, **ni** (PRON.1PL) *[25]*, **ia** (yes) *[23]*, **yma** (here) *[22]*, **bod** (be.NONFIN) *[20]*, **nhw** (PRON.3PL) *[20]*, **o** (of) *[20]*, **hwnna** (that) *[19]*, **isio** (want) *[19]*, **fo** (PRON.3SM) *[18]*, **wneud** (do.NONFIN) *[15]*, **cael** (get.NONFIN) *[14]*, **fi** (PRON.1S) *[14]*, **hi** (PRON.3SF) *[14]*, **wan** (now) *[14]*, **chi** (PRON.2PL) *[13]*, **fan** (place) *[13]*, **a** (and) *[12]*, **de** (TAG) *[11]*, **do** (yes) *[11]*, **efo** (with) *[11]*, **on** (be.1S.IMP) *[11]*, **allan** (out) *[10]*, **hanner** (half) *[10]*, **neu** (or) *[10]*, **un** (one) *[10]*, **iawn** (right) *[9]*, **o** (from) *[9]*, **pan** (when) *[9]*, **di** (PRON.2S) *[8]*, **gael** (get.NONFIN) *[8]*, **mewn** (in) *[8]*, **pwy** (who) *[8]*, **dach** (be.2PL.PRES) *[7]*, **dod** (come.NONFIN) *[7]*, **heddiw** (today) *[7]*, **i** (for) *[7]*, **nag** (NEG) *[7]*, **rywbeth** (something) *[7]*, **wna** (do.1S.NONPAST) *[7]*, **â** (with) *[6]*, **am** (for) *[6]*, **dim** (NEG) *[6]*,

Accessibility

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

Instances of "gael" in Siarad corpus: lloyd1

50	JEA	dach@1 chi@1 (y)n@1 dod@1 i@1 gael@1 paned@1 ta@1 be@1 ?
		dach chi yn dod i gael paned ta be ?
		are you coming to have a cuppa or what ?
69	JEA	troi@1 page@2 cyntaf@1 Arthur@0 # i@1 gael@1 gweld@1 pwy@1 (y)dy@1 .
		troi page cyntaf Arthur i gael gweld pwy ydy .
		turn the first page, Arthur, to see who she is
97	JEA	dach@1 chi@1 (we)di@1 bwcio@1 rywle@1 i@1 gael@1 swper@1 heno@1 ?
		dach chi wedi bwcio rywle i gael swper heno ?
		have you booked anywhere to have supper tonight ?

- Interface to CLAN queries

Data-mining

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Utterance profiling

oeddwn i yn ofnadwy er am er um darllen .	111101001	vrsaUpUUv
na mae o yn byw yn Trevelin efo Sally .	111111010	(xcca)vrsvpMpM
pan oeddwn i yn Gymru oedd lnes yn byw .	111111011	cvr(sp)(np)vMsv
ond ges i ddim gwybod bod Lea yn Chester .	111111012	cupbvvMpM
a wedyn oedden ni yn mynd wedyn er tiwy .	111111101	cbvrsvbUp
a wedyn baswn i wedi mynd i Esquel a .	111111101	cbupvvpMc
mae gen i lun o nhw chwarae yr piano .	111111110	vupn(rpp)rvtU
a wedyn mynd i gael te i le Linda .	111111110	cbvpvnpnM
a wedyn i ti mae yn tŷ modryb Christa .	111111110	cbprv(sp)nuM
a oeddwn i meddwl y byd o modryb Elsa .	111111110	cvrvtn(rpp)uM
a be oedd llong yn o_fewn i yr porth .	111111111	civn(sp)upt(nnn)
na dw i yn meddwl bydd hi yn dod .	111111111	(xcca)upsvvrsv
mae rhai yn gallu gadael ond rei eraill ddim .	111111111	vasvvcur(nb)
mae maen nhw wneud maen nhw yn stydio pen .	111111111	vvrvvr(sp)un
mi fanwodd hi pan oeddwn i cael yng ngeni .	111111111	xvrcvrvpv
a wedyn mae yr tri bachgen ddim yn smocio .	111111111	cbvtanbsv
pedwar a hanner oedd o i i Gymru ia ?	111111111	acnvrpp(np)u
mae yr adar yn gwledda arnyn nhw myfyria di .	111111111	vtnsvprvr
dw i braidd yn ddiog i sgwennu dyddiau yma .	111111111	upbsapvnb

111101001	9	1
1211111111	10	1
1111111111	10	19
1111111110	10	1
1111011121	10	1
1031110111	10	1
1111111101	10	1
1011111111	10	2
1111011111	10	2
0111111111	10	2
11111111111	11	14
11111121111	11	2
01111111111	11	2
11111110111	11	2
11101111110	11	1
11111010111	11	1
10111111111	11	1
11100011011	11	1
111111111111	12	10
111111111110	12	1
111101111111	12	1
101111111110	12	1

Data-mining

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Easier or more detailed statistical analysis
- N-gram generation (2- or 3-word collocations)
- Input to statistical machine translation

Questions

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- If glossing is **A Good Thing** . . .
- Can we automate the process?
- Can we add value to the texts?