Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

The Bangor Autoglosser: A Multilingual Tagger for Conversational Text

Kevin Donnelly, Margaret Deuchar

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

PRIFYSGOL CYMRU UNIVERSITY OF WALES BANGOR 1884

The Centre

Canolfan ESRC Centre
dros Ymchwil
i Ddwyieithrwydd
for Research
on Bilingualism

- ESRC Centre for Research in Bilingualism
- Established January 2007
- Five research themes
- Corpus-based research
- **bilingualism.bangor.ac.uk**

Bangor corpora

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

All available under the GPL.

A sample utterance

***SER:** dw@1 i@1 (y)n@1 tynnu@1 llun@1 i@1 [/] i@1 (y)r@1 plant@1 <i@1 plant@1> [//] <i@1 (y)r@1> [//] # i@1 er@0 &h Helen@0 a@1 Susanna@0 a@1 +/. %snd:"deuchar1"_73881_79477

%gls: be.1S.PRES PRON.1S PRT take.NONFIN picture for for DET children for children for DET for IM Helen and Susanna and

%eng: I draw a picture for . . . for the children, for, er, Helen and Susanna and . . .

(Siarad corpus, deuchar1)

Transcription format

CLAN: childes.psy.cmu.edu/clan

**SER dw@1 i@1 (y)n@1 hopeless@2 efo@1 tynnu@1 llun@1 . %snd:"deuchar1"_72848_73881*

Speaker	*SER
Utterance	dw@1 i@1 (y)n@1 hopeless@2 efo@1 tynnu@1 llun@1 .
Language tags	1=Welsh, 2=English, 0=indeterminate
Audio location	%snd:"deuchar1"_72848_73881
Manual gloss	be.1S.PRES PRON.1S PRT hopeless with take.NONFIN picture

Glossing

- Allows non-native speakers to parse the conversation
- Labour-intensive
- Tedious
- Inconsistent: *ychydig* – "a_bit"/"a_little"
- Tags difficult to revise later

Existing options

- Spanish – CLAN → MOR + POST
- Welsh – no tagger at all

Aims

- Tag across multiple languages simultaneously
- Single application infrastructure
- Handle conversational language
- Use FOSS where possible
  - speed of development
  - re-use scarce language resources
  - bootstrap new languages easily

Dictionaries

Dictionary format

- Derived from GPL or PD resources
- One database table
- Words, not morphemes
- Easily presented in a spreadsheet
- Easy to update
- Easy to get started

Welsh dictionary

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

surface	*lemma*	*enlemma*	*pos*	*gender*	*number*	*tense*
bara	bara	bread	n	m	sg	
cathod	cath	cat	n	f	pl	
mynd	mynd	go	v			infin
aeth	mynd	go	v		3s	past
hapus	hapus	happy	adj			
rhywsut	rhywsut	somehow	adv			
heb	heb	without	prep			

Spanish dictionary

surface	*lemma*	*enlemma*	*pos*	*gender*	*number*	*tense*
perro	perro	dog	n	m	sg	
canciones	canción	song	n	f	pl	
empezar	empezar	start	v			infin
empieza	empezar	start	v		23s	pres
empieza	empezar	start	v		2s	imper
rojo	rojo	red	adj	m	sg	
rojas	rojo	red	adj	f	pl	
por	por	for	prep			

English dictionary

surface	*lemma*	*pos*	*number*	*tense*
break	break	sv		infin
broke	break	av		past
broken	break	av		pastpart
car	car	n	sg	
quick	adj			
by	by	prep		
which	which	rel		

breaks, breaking, cars, quickly are derived during lookup

The autoglossing process

Stages in the autoglossing process

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- **Stage 1** – Import the unglossed file
- **Stage 2** – Look up the words it contains
- **Stage 3** – Disambiguate between alternatives for a word
- **Stage 4** – Output the glossed file

Stage 1
Import the chat file

- Read each line of the file into an utterances table
- Select the utterance and discard non-word material
- Split the resulting utterance into words
- Put them into a words table

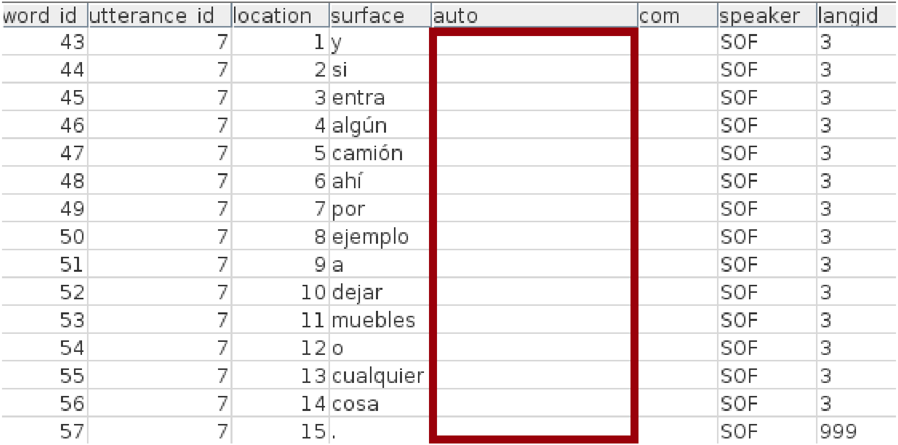
Sample import

**SOF: <y si> [/] y si entra algún camión ahí por ejemplo a dejar muebles o cualquier cosa .*

Speaker	*SOF
Utterance	y si entra algún camión ahí por ejemplo a dejar muebles o cualquier cosa .
English	And if some lorry goes in there, for example, to leave off furniture or whatever.

(Miami corpus, sastre1)

The words table



word id	utterance id	location	surface	auto	com	speaker	langid
43	7	1	y			SOF	3
44	7	2	si			SOF	3
45	7	3	entra			SOF	3
46	7	4	algún			SOF	3
47	7	5	camión			SOF	3
48	7	6	ahí			SOF	3
49	7	7	por			SOF	3
50	7	8	ejemplo			SOF	3
51	7	9	a			SOF	3
52	7	10	dejar			SOF	3
53	7	11	muebles			SOF	3
54	7	12	o			SOF	3
55	7	13	cualquier			SOF	3
56	7	14	cosa			SOF	3
57	7	15	.			SOF	999

Stage2
Dictionary lookup

- Using the language tag, look up each word against the appropriate dictionary
- Do basic segmentation (e.g clitic pronouns in Spanish, verb-tenses in English, mutation in Welsh)
- Write out all the dictionary entries (readings) for that word
- Feed these to the constraint grammar parser for disambiguation

Constraint Grammar

- Developed by Fred Karlsson in the 90s
- Third generation of the parser: **visl-cg3**
- Eckhard Bick, Tino Didriksen
- Free (GPL) license
- **beta.visl.sdu.dk/constraint_grammar.html**
- Easily-understood rules

Stage 3 Disambiguation

- **select (n) if (-1 (ord));**
- Choose the noun (**n**) reading if the first word to the left (**-1**) is an ordinal (**ord**)
- Welsh: *yr ail dro* (the second time)
- English: *the third man*
- Spanish: *el primer viaje* (the first journey)
- Verb readings for *dro*, *man* and *viaje* will be deleted

Language-specific rules

- Include that language's tag in the rule to constrain its application
- **select ([es] n) if (-1 ([es] ord));**
- Now applies only to Spanish: *el primer viaje*

Before disambiguation

```
"<ddim>"
    "dim"  {96,1} [cy] n m sg :nothing: [208789] + sm
    "dim"  {96,1} [cy] adv :not: [204176] + sm
"<yn>"
    "yn"  {96,2} [cy] stat :stative: [200654]
    "yn"  {96,2} [cy] prep :in: [204430]
    "gan"  {96,2} [cy] prep :with: [196964] + sm
"<gynnar>"
    "cynnar"  {96,3} [cy] adj :early: [209212] + sm
"<iawn>"
    "iawn"  {96,4} [cy] adv :OK: [207540]
    "iawn"  {96,4} [cy] adv :very: [203775]
```

(Patagonia corpus, patagonia1)

“not very early”

After disambiguation

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

```
"<ddim>"
    "dim" {96,1} [cy] adv :not: [204176] + sm
"<yn>"
    "yn" {96,2} [cy] stat :stative: [200654]
"<gynnar>"
    "cynnar" {96,3} [cy] adj :early: [209212] + sm
"<iawn>"
    "iawn" {96,4} [cy] adv :very: [203775]
```

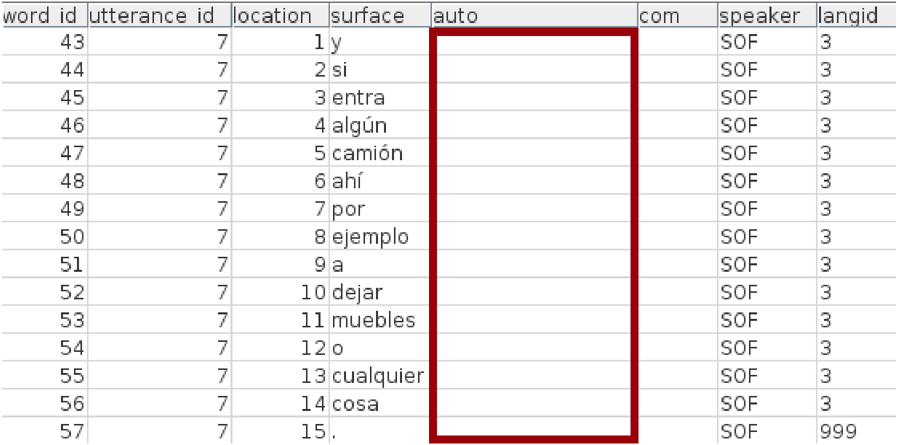
(Patagonia corpus, patagonia1)

“not very early”

Stage 4
Output the glossed file

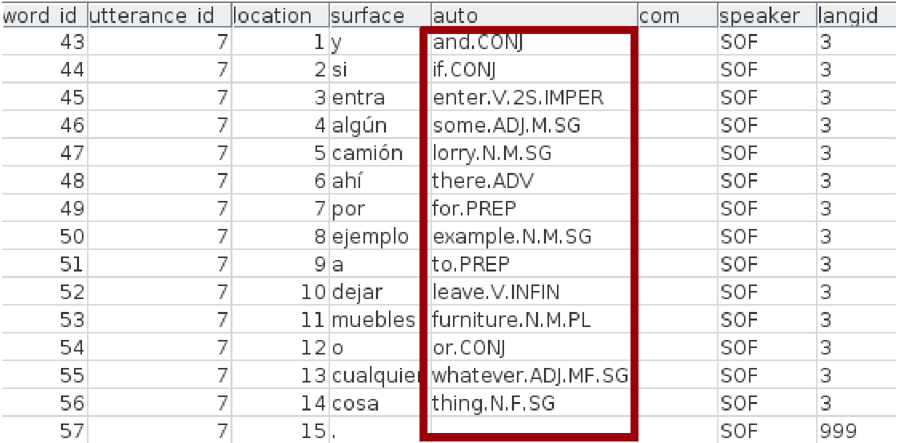
- Read the disambiguated constraint grammar output
- Insert each lexeme and its part-of-speech tags into the words table
- Use the utterances and words tables to write out an autoglossed file

The words table



word id	utterance id	location	surface	auto	com	speaker	langid
43	7	1	y			SOF	3
44	7	2	si			SOF	3
45	7	3	entra			SOF	3
46	7	4	algún			SOF	3
47	7	5	camión			SOF	3
48	7	6	ahí			SOF	3
49	7	7	por			SOF	3
50	7	8	ejemplo			SOF	3
51	7	9	a			SOF	3
52	7	10	dejar			SOF	3
53	7	11	muebles			SOF	3
54	7	12	o			SOF	3
55	7	13	cualquier			SOF	3
56	7	14	cosa			SOF	3
57	7	15	.			SOF	999

The words table – glossed



word id	utterance id	location	surface	auto	com	speaker	langid
43	7	1	y	and.CONJ		SOF	3
44	7	2	si	if.CONJ		SOF	3
45	7	3	entra	enter.V.2S.IMPER		SOF	3
46	7	4	algún	some.ADJ.M.SG		SOF	3
47	7	5	camión	lorry.N.M.SG		SOF	3
48	7	6	ahí	there.ADV		SOF	3
49	7	7	por	for.PREP		SOF	3
50	7	8	ejemplo	example.N.M.SG		SOF	3
51	7	9	a	to.PREP		SOF	3
52	7	10	dejar	leave.V.INFIN		SOF	3
53	7	11	muebles	furniture.N.M.PL		SOF	3
54	7	12	o	or.CONJ		SOF	3
55	7	13	cualquier	whatever.ADJ.MF.SG		SOF	3
56	7	14	cosa	thing.N.F.SG		SOF	3
57	7	15	.			SOF	999

Evaluation

Speed

- 900–1100 words per minute
- 1 minute to autogloss 5 minutes of speech
- Siarad: 500,000 words in 8h27m

Accuracy

	Words	*Coverage*	*MFL*	*Accuracy*
Welsh-Spanish (Patagonia[1])	15,677	100%	W:92% S:1%	99%
Welsh-English (Siarad[2])	10,411	96%	W:81% E:2%	98%
Spanish-English (Miami[3])	10,411	97%	E:54% S:42%	96%

[1] patagonia1,2,3,6

[2] deuchar1, stammers4

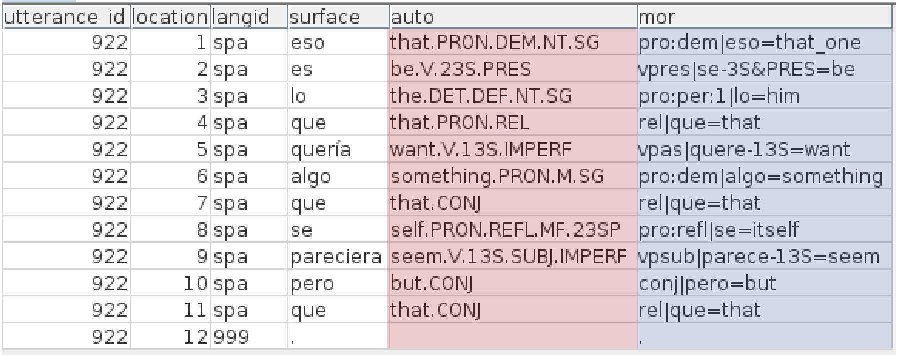
[3] herring7, sastre1, zeledon5

Comparison with other methods

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Spanish – MOR glosser (part of the CLAN suite)
- Welsh – manual (human) glossing
- Two sample files from each corpus glossed using both methods
- Aligned and then inspected manually
- Typos or missing lexemes not counted as errors
- Names omitted from consideration

Comparison between autoglosser and MOR



utterance_id	location	langid	surface	auto	mor
922	1	spa	eso	that.PRON.DEM.NT.SG	pro:dem|eso=that_one
922	2	spa	es	be.V.23S.PRES	vpres|se-3S&PRES=be
922	3	spa	lo	the.DET.DEF.NT.SG	pro:per:1|lo=him
922	4	spa	que	that.PRON.REL	rel|que=that
922	5	spa	quería	want.V.13S.IMPERF	vpas|quere-13S=want
922	6	spa	algo	something.PRON.M.SG	pro:dem|algo=something
922	7	spa	que	that.CONJ	rel|que=that
922	8	spa	se	self.PRON.REFL.MF.23SP	pro:refl|se=itself
922	9	spa	pareciera	seem.V.13S.SUBJ.IMPERF	vpsub|parece-13S=seem
922	10	spa	pero	but.CONJ	conj|pero=but
922	11	spa	que	that.CONJ	rel|que=that
922	12	999	.		.

Spanish files

- Tested on *herring11*, *sastre5*
- 8,039 tokens, 1,638 types (TTR: 0.20)

Language mix	
Spanish	88%
English	9%
indeterminate	3%

Welsh files

- Tested on *stammers7*, *stammers9*
- 9,454 tokens, 1,376 types (TTR: 0.15)

Language mix	
Welsh	87%
English	2%
indeterminate	11%

Comparison with MOR glossing (Spanish)

	Autoglosser	*MOR*
Coverage	96.9%	95.7%
Accuracy	97.4%*	97.6%†

* *wrong lexeme 0.7%, wrong POS 0.2%, ambiguous 1.7%*

† *wrong lexeme 1.6%, wrong POS 0.7%, ambiguous 0.1%*

Comparison with manual glossing (Welsh)

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

	Autoglosser	*Human*
Coverage	98.3%	99.9%
Accuracy	97.9%*	99.9%

** wrong lexeme 0.7%, wrong POS 0.1%, ambiguous 1.4%*

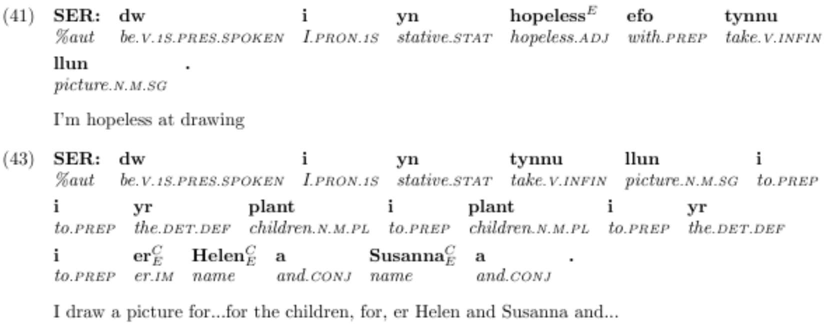
Spin-off benefits

Typesetting – before

***SER:** dw@1 i@1 (y)n@1 hopeless@2 efo@1 tynnu@1 llun@1 .
%snd:"deuchar1"_72848_73881
%gls: be.1S.PRES PRON.1S PRT hopeless with take.NONFIN picture
%eng: I'm hopeless at drawing
***SER:** dw@1 i@1 (y)n@1 tynnu@1 llun@1 i@1 [/] i@1 (y)r@1 plant@1 <i@1 plant@1> [//] <i@1 (y)r@1> [//] # i@1 er@0 &h Helen@0 a@1 Susanna@0 a@1 +/. %snd:"deuchar1"_73881_79477
%gls: be.1S.PRES PRON.1S PRT take.NONFIN picture for for DET children for children for DET for IM Helen and Susanna and
%eng: I draw a picture for . . . for the children, for, er, Helen and Susanna and . . .

(Siarad corpus, deuchar1)

Typesetting – after



(41)	**SER:**	**dw**	**i**	**yn**	**hopeless**^{E}	**efo**	**tynnu**
	%aut	*be*.V.1S.PRES.SPOKEN	*I*.PRON.1S	*stative*.STAT	*hopeless*.ADJ	*with*.PREP	*take*.V.INFIN
	llun	**.**					
	picture.N.M.SG						

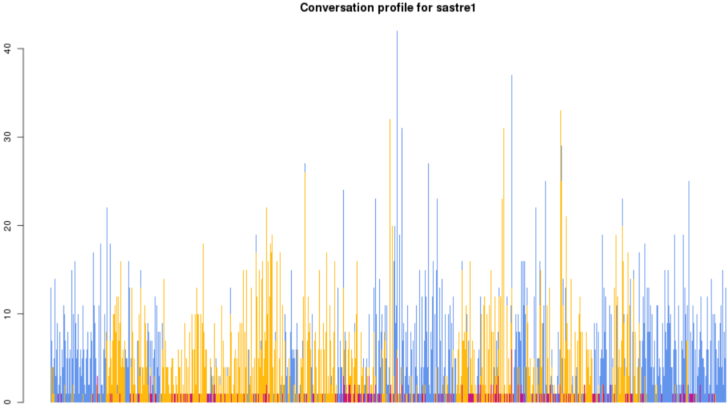
I'm hopeless at drawing

(43)	**SER:**	**dw**	**i**	**yn**	**tynnu**	**llun**	**i**
	%aut	*be*.V.1S.PRES.SPOKEN	*I*.PRON.1S	*stative*.STAT	*take*.V.INFIN	*picture*.N.M.SG	*to*.PREP
	i	**yr**	**plant**	**i**	**plant**	**i**	**yr**
	to.PREP	*the*.DET.DEF	*children*.N.M.PL	*to*.PREP	*children*.N.M.PL	*to*.PREP	*the*.DET.DEF
	i	**er**$^{C}_{E}$	**Helen**$^{C}_{E}$	**a**	**Susanna**$^{C}_{E}$	**a**	**.**
	to.PREP	*er*.IM	*name*	*and*.CONJ	*name*	*and*.CONJ	

I draw a picture for...for the children, for, er Helen and Susanna and...

Conversation profile

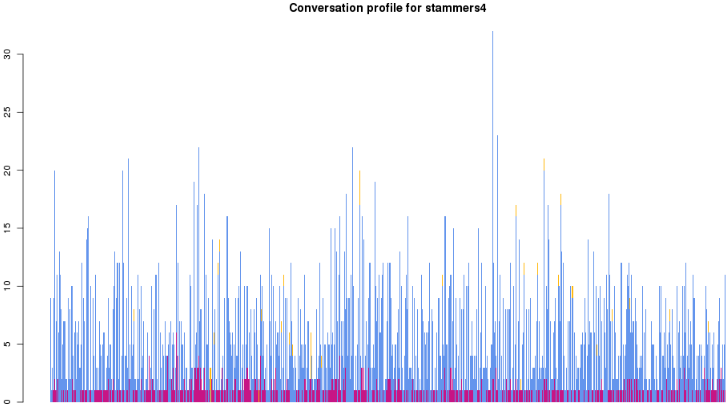
Spanish-English



Conversation profile

Welsh-English

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism



Mixed Collocations

Det+Noun+Adj

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- un healthy store (*a healthfood store*)
- the mismo papel (*the same paper*)
- the fair estúpido (*the stupid fair*)
- la cheerleader pesada (*the plump cheerleader*)
- un dealer grande (*a big dealer*)
- un pequeño pocket (*a little pocket*)

26 trigrams out of **161,000** words . . .

. . . difficult to find manually

Resources

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

bangortalk.org.uk

Web-interface to the transcripts

Transcript and audiofile download

Bangor Autoglosser code (Git repository)
Licensed under GPL v3