Canolfan ESRC Centre
dros Ymchwil
i Ddwyieithrwydd
for Research
on Bilingualism

Using constraint grammar in the Bangor Autoglosser to disambiguate multilingual spoken text

Kevin Donnelly and 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

Background

The Centre

- 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
Welsh-Spanish (Patagonia)	32	20	161k	2011
Spanish-English (Miami)	31	20	126k	2011
	132	**80**	**743k**	

All available under the GPL.

The conversations

- Transcribed using the CLAN format
- **childes.psy.cmu.edu/clan**
- Standard orthography
  - Elisions spelled out for Welsh:
  - **mae'n fawr** (it's big) →**mae (y)n fawr**
- Gloss added
- Free translation in English added

Sample utterances

***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
***MYF:** +< &=laugh . %snd:"deuchar1"_73196_73881
***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)

Utterance format

**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=undetermined
Audio location	%snd:"deuchar1"_72848_73881
Manual gloss	be.1S.PRES PRON.1S PRT hopeless with take.NONFIN picture

Why?

- Examine how language is actually used
- Differences between spoken language and formal written language
- Sociolinguistic variation – what is used where by whom
- Balance between languages in bilingual usage
- How one language handles lexical items from the other
    - Welsh loan-verbs such as *textio* (to text) behave more like ordinary Welsh verbs the more frequent they are

Glossing

Why gloss?

- Lexemes and part-of-speech (POS) tags:
  - Help non-native speakers parse the conversation
  - Allow further analysis - morphological, syntactic, sociolinguistic
- Difficulties:
  - Time-consuming and tedious
  - Inconsistency and errors (*ychydig* – "a_bit"/"a_little")
  - Tag choice difficult to revise later

Automation

- April 2010
- Explore automation to address difficulties above
- Move towards more granular POS information
- Welsh →Spanish →English
- Accuracy reflects timespend:
  99% for Welsh, and 95% for English.
- Work in progress

Why another wheel?

- CLAN tagging system
  - For 11 languages > 5m speakers
  - Requires one pass for each language
  - Can't mix language context
  - Vocabulary stored in a number of files
  - Disambiguation for only 4 languages
- Toolbox
- No automated system for small languages

Pilot project

- Test project over two weeks:
  - No disambiguation
  - Write out entries from Spanish dictionary
  - **apertium.org**
  - Compare them with MOR output
  - Write out entries from Welsh dictionary
  - **eurfa.org.uk**
- Good results
- Needed a way to disambiguate - enter CG!

Dictionaries

Dictionaries

- 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

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

Language differences

- Spanish and Welsh
  - Inflected (Welsh less so than it was)
  - Surface forms give clues about the POS
- English
  - Analytic
  - Homophonous surface forms
  - POS defined by role in the sentence
  - **break**
    - *a clean break* (noun)
    - *break the mould!* (imperative)
    - *to break a habit* (infinitive)
    - *they break everything* (present)

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

Import

Import the chat file

- PHP script reads each line into a PostgreSQL database table
- Selects the utterance and discards markers
- Splits the cleaned utterance into words
- Puts them into another database table

Utterance-table fields

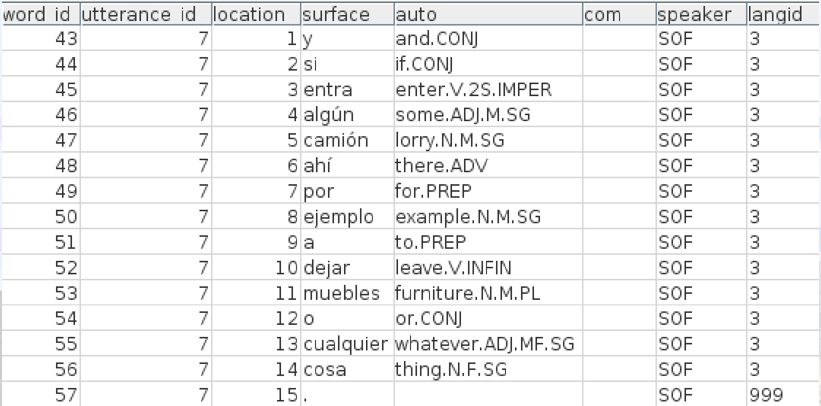
Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- utterance_id
- filename
- speaker
- surface
- startpoint
- endpoint
- duration
- manual glosses (if present)
- English translation (if present)
- comments (if present)
- precode (if present – marks entire utterances in the least-frequent language)

Word-table fields

- word_id
- utterance_id
- location of the word in the utterance
- surface
- automatic glosses
- manual glosses (if present)
- language id
- speaker
- filename

The words table



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

Lookup

- Each word is looked up against the appropriate dictionary
- Uses the language id assigned to the word
- Writes out all “hits” in the CG input format

Segmentation

- Lookup also does some basic segmentation
- Minimises number of dictionary entries (**break** above)
- Welsh: mutated words are tagged
  - thad →tad (*father*) + am
  - gael →cael (*get*) + am
- Spanish: clitic pronouns are tagged
  - ponerle →poner (*put*) + le[pron.mf.3s]
  - déjanos →dejar (*leave*)+ nos[pron.mf.1p]

English Segmentation

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Elisions are tagged
    - gonna →go # to.prep
    - we're →we # be.v.pres
- Plurals or verbs (3p sg pres) are tagged
    - breaks →break # pv
- Adjectives or verbs (past or pastpart) are tagged
    - constructed →construct # av
- Adjectives, nouns or verbs (prespart) are tagged
    - thinking →think # asv

Mutation

- **tad** (father)
  - **ei dad** (his father)
  - **ei thad** (her father)
- **marw** (die, dead)
  - **mae o'n marw** (he is dying)
  - **mae o'n farw** (he is dead)
- direct object following a verb
  - **Mi werthodd y ffermwr y mochyn** (The farmer sold the pig)
  - **Mi werthodd y ffermwr fochyn** (The farmer sold a pig)

Welsh before CG

```
"<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]
"<.>"
```

(Miami corpus, sastre1)

Welsh after CG

```
"<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)

Spanish before CG

```
"<y>"
    "y"  {122,1} [es] conj :and: [118037]
"<ahora>"
    "ahora"  {122,2} [es] adv :now: [6292]
"<vamos>"
    "ir"  {122,3} [es] v 1p pres :go: [115789]
"<a>"
    "a"  {122,4} [es] prep :to: [1]
"<hacerle>"
    "hacer"  {122,5} [es] v infin :do: [62577] + le[pron.mf.3s]
"<el>"
    "el"  {122,6} [es] det.def m sg :the: [45129]
"<baño>"
    "baño"  {122,7} [es] n m sg :bathroom: [16011]
    "bañar"  {122,7} [es] v 1s pres :bathe: [16010]
"<.>"
```

(Patagonia corpus, patagonia1)

Spanish after CG

```
‘‘<y>"
    "y" {122,1} [es] conj :and: [118037]
"<ahora>"
    "ahora" {122,2} [es] adv :now: [6292]
"<vamos>"
    "ir" {122,3} [es] v 1p pres :go: [115789]
"<a>"
    "a" {122,4} [es] prep :to: [1]
"<hacerle>"
    "hacer" {122,5} [es] v infin :do: [62577] + le[pron.mf.3s]
"<el>"
    "el" {122,6} [es] det.def m sg :the: [45129]
"<baño>"
    "baño" {122,7} [es] n m sg :bathroom: [16011]
"<.>"
```

(Miami corpus, sastre1)

English before CG

```
"<it’s>"
    "it"  {545,1} [en] pron.sub 3s :it: [130342] # gb
"<coming>"
    "come"  {545,2} [en] sv infin :come: [82193] # asv
"<out>"
    "out"  {545,3} [en] adv :out: [157287]
"<on>"
    "on"  {545,4} [en] prep :on: [156077]
"<D_V_D>"
    "D_V_D"  {545,5} [en] name
"<then>"
    "then"  {545,6} [en] adv :then: [208154]
"<.>"
```

(Miami corpus, herring7)

English after CG

```
"<it's>"
    "it" {545,1} [en] pron.sub 3s :it: [130342] # be.v.3s.pres
"<coming>"
    "come" {545,2} [en] v prespart :come: [82193] #
"<out>"
    "out" {545,3} [en] adv :out: [157287]
"<on>"
    "on" {545,4} [en] prep :on: [156077]
"<D_V_D>"
    "D_V_D" {545,5} [en] name
"<then>"
    "then" {545,6} [en] adv :then: [208154]
"<.>"
```

(Miami corpus, herring7)

Multilingual disambiguation

Multiple languages

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Previous extracts all monolingual
- But easy to use CG for multilingual speech
- Ensure that each “hit” in the input file is tagged for language
- Put all the rules into one grammar file, grouped according to language
- Constrain the rules to act only on one language by including that language’s tag in the rule

▶

▶

▶

Welsh before CG

```
"<cada>"
    "cada"  {79,5} [es] adj mf sg :every: [18541]
"<vez>"
    "vez"  {79,6} [es] n f sg :time: [116758]
"<que>"
    "que"  {79,7} [es] conj :than: [93349]
    "que"  {79,7} [es] conj :that: [93350]
"<nos>"
    "yo"  {79,8} [es] pron.obl mf 1p :us: [80717]
"<vamos>"
    "ir"  {79,9} [es] v 1p pres :go: [115789]
"<camping>"
    "camp"  {79,10} [en] sv infin :camp: [74449] # asv
```

(Miami corpus, sastre1)

- **vamos camping**
- substitute (sv infin asv) (v prespart) ([en] sv infin asv) (-1 ([en] "be") or (:go:));
- tags
-

Welsh before CG

(Miami corpus, sastre1)

Welsh before CG

(Miami corpus, sastre1)

Process

- 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

***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

- 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

	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

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

Can we add value to the texts?

Texts

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

Accessibility

Accessibility

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

Accessibility

- Interface to CLAN queries

Data-mining

- Utterance profiling

Data-mining

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