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

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- 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?

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- 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

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- 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			

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)

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

Import: Dictionary lookup and segmentation

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 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

Utterance-table fields

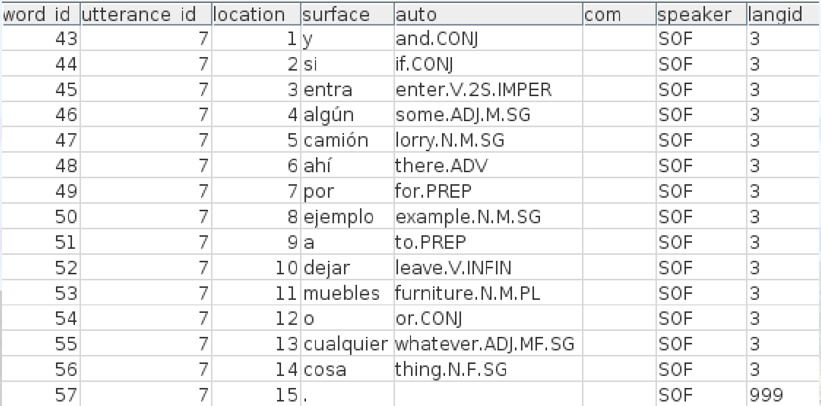
Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- 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]

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]
```

(Patagonia corpus, patagonia1)

"not very early"

Welsh after CG

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*"

Spanish before CG

```
"<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)

"*we're going to do the bathroom*"

Spanish after CG

```
"<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)

“*we’re going to do the bathroom*”

English Segmentation

- 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, singular nouns or verbs (prespart) are tagged
  - thinking →think # asv

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

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

```
"<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

- 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

Sample rule

- select (n) if (-1 (ord));
- choose the noun reading if the preceding word is an ordinal
- select ([es] n) if (-1 ([es] ord));
- applies only to Spanish (**el primer viaje**)

Welsh/Spanish

```
"<mewn>"
    "mewn" {128,4} [cy] prep :in:
"<motor>"
    "motor" {128,5} [es] n m sg :motor:
"<newydd>"
    "newydd" {128,6} [cy] adj :new:
"<internacional>"
    "internacional" {128,7} [es] adj m sg :international:
```

(Patagonia corpus, patagonia2)

"in a new international motor-car"

Spanish/English

```
"<con>"
    "con" {60,1} [es] prep :with: [132994]
"<el>"
    "el" {60,2} [es] det.def m sg :the: [45129]
"<address>"
    "address" {60,3} [en] n sg :address: [55976]
"<de>"
    "de" {60,4} [es] prep :of: [33387]
"<aquí>"
    "aquí" {60,5} [es] adv :here: [11385]
```

(Miami corpus, zeledon5)

"*with the address from here*"

Welsh/English

```
"<ac>"
    "ac" {27,1} [cy] conj :and: [209088]
"<oedd>"
    "bod" {27,2} [cy] v 3s imperf :be: [74724]
"<o>"
    "fo" {27,3} [cy] pron m 3s spoken :he: [209264]
"<gynno>"
    "gan" {27,4} [cy] prep+pron m 3s :with_him: [207424]
"<fo>"
    "fo" {27,5} [cy] pron m 3s :he: [196922]
"<background>"
    "background" {27,6} [en] n sg :background: [64983]
"<ddu>"
    "du" {27,7} [cy] adj :black: [209631] + sm
```

(Siarad corpus, deuchar1)

“and it was . . . it had a black background”

Cross-boundary rules

- Rules can apply across language boundaries
- Remove the language constraint when appropriate

Spanish adjective

```
"<es>"
    "ser"  {500,1} [es] v 23s pres :be: [51318]
"<otro>"
    "otro"  {500,2} [es] adj m sg :other: [83612]
    "otro"  {500,2} [es] pron m sg :other: [83613]
"<zip>"
    "zip"  {500,3} [en] n sg :zip: [1758]
"<code>"
    "code"  {500,4} [en] n sg :code: [81254]
```

(Miami corpus, sastre1)

"*it's another zipcode*"

Spanish adjective rule

- **otro** can be an adjective before a noun, or a pronoun
- The selection rule leaves the noun unspecified as to language:
- select ([es] adj) if (1 (n));
- *adjective* will be selected before **any** *noun* (not just Spanish)

Spanish adjective output

```
"<es>"
"ser" {500,1} [es] v 23s pres :be: [51318]
"<otro>"
"otro" {500,2} [es] adj m sg :other: [83612]
"<zip>"
"zip" {500,3} [en] n sg :zip: [1758]
"<code>"
"code" {500,4} [en] n sg :code: [81254]
```

(Miami corpus, sastre1)

“*it’s another zipcode*”

English verb

```
"<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)

"every time that we go camping"

English verb rule

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- **camping** can be an adjective, a singular noun, or a verb
- *be thinking, enjoy reading, go fishing*
- In **vamos camping**, we can get the correct end tag by specifying the meaning of the preceding verb, rather than the lemma:
- substitute (sv infin asv) (v prespart) ([en] sv infin asv) (-1 ([en] "be") or (:go:));
- The tags on **camping** are rewritten to tag it as a present participle

English verb output

```
"<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] pron.rel :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] v prespart :camp: [74449] #
```

(Miami corpus, sastre1)

"*every time that we go camping*"

Rule types and language type

Removal

- “Delete” items from the dictionary
- Homonym selection
- select (”cyfeiriad” [cy] :direction:);
- Archaic/infrequent words
- remove (”tasu” [cy] :stack:);

Compensate

- Remove words which are an artefact of the lookup
- remove ([cy] ”mynd” v 2s imper nm);
- *nos* $<$ *dos*
- remove ([in] ”gum” n sg sm);
- *um* $<$ *gum*

English

- substitute (n sg pv) (n pl) ([en] n sg pv);
- *house* → *houses*
- substitute (as) (adj) ([en] as) (1 ([en] n) or ([en] pron));
- *a miniature rabbit, miniature ones*

English

- substitute (pron.sub) (pron.obj) ([en] pron.sub) (-1 ([en] v infin));
- *and open it*
- substitute (sv infin av) (v past) ([en] sv infin av) (-2 ([en] pron.sub)) (-1 preverbal);
- *they closed*

English

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- substitute (av past) (v past) ([en] av past) (-1 ([en] pron.sub)) (not -1 (have.v.pres)) (not -2 ("have"));
- *we bought*, **not** *you've done, we have bought*
- substitute (av past) (v pastpart) ([en] av past) (-1 (have.v.pres) or ("have") or ("be") or (det.def) or (det.indef));
- *you've done, you have done, it was misspent, un rebuilt*

English

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Refine existing tags
- substitute (123p) (1p) ([en] v 123p) (-1 (pron.sub 1p));
- *we are*
- In general, more dependent on rule order

Default choices

- When left with an [or], we can make a “default“ choice
- select ([cy] v infin) if (0C ([cy] v infin) or ([cy] v 3s imper));
- *cerdded*
- C enforces the two conditions

Careful . . .

Canolfan ESRC Centre
dros Ymchwil i Ddwyieithrwydd
for Research on Bilingualism

- Scope of **remove** can be unexpected
- Likewise **select-if-not**
- select (imper) if (not @1 ("ni"));
- Caused 304 regressions in Spanish output!

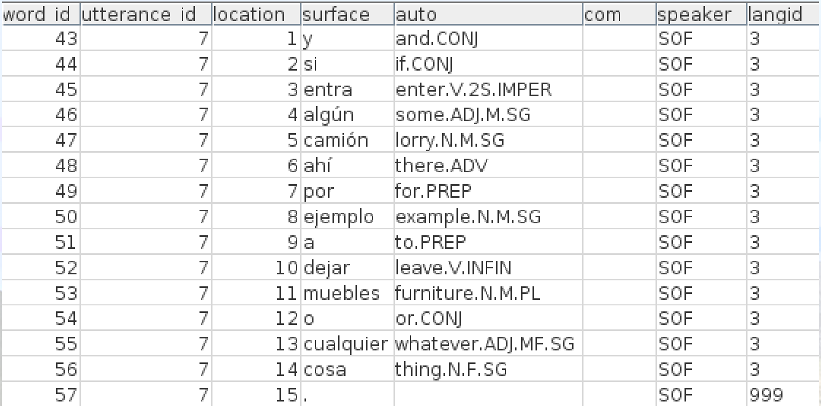
Rule numbers

- Spanish: 150
- Welsh: 180
- English: 200

Output method

- CG writes out the disambiguated text
- This file is parsed
- The glosses (lexeme + POS tag) are inserted into the words table
- The words are then written out to create the autoglossed file

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

Accuracy

Accuracy

	Words	*Coverage*	*MFL*	*Accuracy*
Welsh-Spanish (Patagonia[1])	15,677	100%	W:92%	99%
Spanish-English (Miami[2])	4,202	97%	S:59%	97%
Welsh-English (Siarad[3])	10,411	96%	W:81%	98%

[1] patagonia1,2,3,6

[2] zeledon5

[3] stammers4, deuchar1

Dictionary coverage

	Words	*Nouns*	
Welsh	209k	6k	3%
Spanish	130k	19k	15%

Speed

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

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

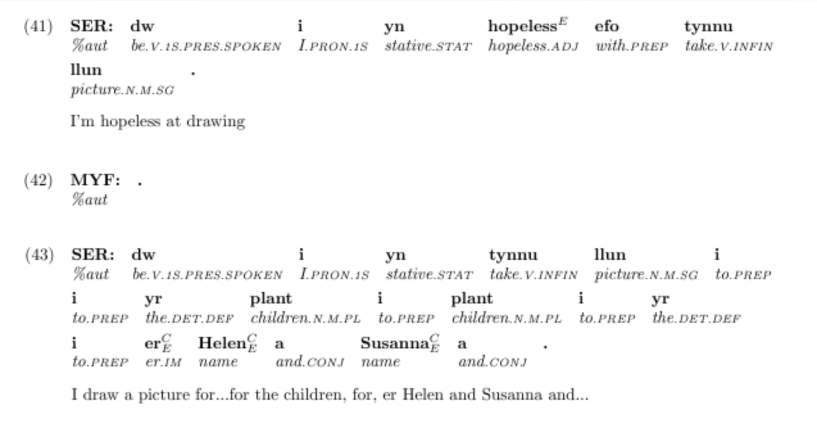
***SER:** dw@1 i@1 (y)n@1 tynnu@1 llun@1 i@1 [/] i@1 (y)r@1 plant@1 <i@1 plant@1> [//] <i@1 (y)r@1> [//] # i@1 er@0 &h Helen@0 a@1 Susanna@0 a@1 +/. %snd:”deuchar1”_73881_79477

%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)

Typesetting



(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

(42)	**MYF:**	**.**
	%aut	

(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...

Texts

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

Data-mining

- Interactive webpages (*bangortalk.org.uk*)
- Easier or more detailed statistical analysis with R
- Input to machine translation. speech-to-text, etc

thinkopen.org.uk/git