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

Kevin Donnelly ***and*** **Margaret Deuchar**
ESRC Centre for Research on Bilingualism in Theory and Practice
Prifysgol Bangor University, Wales, UK
{k.donnelly|m.deuchar}@bangor.ac.uk

Abstract

We present a novel use of constraint grammar (CG) in automatic glossing software to disambiguate surface forms in connected multilingual speech. The resulting autoglosser output shows 97-99% accuracy over all three languages. We discuss the CG rules that help deliver this, noting the differences between those applying to Welsh and Spanish, and those applying to English.

1 Introduction

Bangor University's ESRC Centre for Research on Bilingualism,[1] established in Jnauary 2007, has assembled some 130 bilingual conversations in three corpora: **Siarad**[2] (Welsh-English), **Patagonia** (Welsh-Spanish), **Miami** (Spanish-English).

The conversations total some 80 hours and 750,000 words, and are all available under the GNU GPL.[3] Each recording is provided with a detailed transcription in the widely-used CLAN format[4] (MacWhinney, 2000), along with a free translation in English, and an interlinear gloss giving lexemes and part-of-speech (POS) tags for each word, so that researchers without first-hand knowledge of the languages concerned can more easily parse the utterances.

Part of a typical transcription is shown in Figure 1, in which (using CLAN terminology) three "tiers" can be discerned: the speech tier, the gloss tier, and the translation tier.

The speech tier (the words actually uttered) is marked by an initial ID to distinguish the speaker (e.g. **SER*), followed by the transcribed speech (with each word tagged for language[5] – unmarked for Welsh, *@s:eng* for English, *@s:cym&eng* for indeterminate[6]), and two numbers giving the start and end times of the utterance in the audiofile.

	Chats	*Hours*	*Words*	*Date*
Welsh-English	69	40	456k	2009
Welsh-Spanish	32	20	183k	2011
Spanish-English	31	20	126k	2011
	132	**80**	**765k**	

Table 1 – The three ESRC Centre corpora.

The gloss tier is marked by an initial *%gls*, followed by a series of lexeme+POS-tag strings.

The translation tier is marked by an initial *%eng*, and gives a free translation of the speech tier (the speaker's utterance) into English.

The corpora are valuable in examining how language is actually used: for instance, the differences between spoken language and formal written language, sociolinguistic variation (what forms of language are used where and by whom), the balance between languages in bilingual usage, and how one language handles lexical items from the other.[7]

Manual glossing of the Siarad (Welsh-English) proved to be tedious and time-consuming, so in order to save valuable specialist time it was decided to explore automating the glossing of the Miami (Spanish-English) and Patagonia (Welsh-Spanish) corpora.

Although the CLAN project provides a tag-

[1] http://bilingualism.bangor.ac.uk

[2] Siarad means "speak" in Welsh.

[3] http://www.gnu.org/licenses/gpl.html

[4] http://childes.psy.cmu.edu/clan. Note that using CLAN to record bilingual speech is an extension of its original focus on recording language development in children.

[5] The autoglosser handles 4 marking systems, which reflect changes in transcription practice in the ESRC Centre over the past 5 years, and developments in CLAN itself.

[6] Words which are used in both languages, and which therefore cannot be assigned unambiguously to one of them.

[7] For instance, Jon Stammers (Stammers, 2010) has used the Siarad corpus to show that Welsh loan-verbs such as *textio* (to text) behave more like ordinary Welsh verbs the more frequent they are.

```
*SER: dw i (y)n hopeless@s:eng efo tynnu llun . 72848_73881
%gls: be.1S.PRES PRON.1S PRT hopeless with take.NONFIN picture
%eng: I'm hopeless at drawing
*SER: dw i (y)n tynnu llun i [/] i (y)r plant <i plant> [//] <i (y)r> [//] # i er@s:cym&eng &h Helen@s:cym&eng a Susanna@s:cym&eng a +/. 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 ...
```

Figure 1 – Excerpt from the file *deuchar1* in the Siarad corpus (Welsh-English).

ging system (MOR),[8] this only caters for 11 languages, each with more than 5m speakers. Vocabulary is distributed over a number of files, and MOR requires a separate pass over the file to tag each language. Post-tagging disambiguation (using the POST program) is only available for 4 languages. Software such as Toolbox[9] offers interlinear glossing capability, but is aimed more at linguistic field researchers, and is less applicable to fully-described languages; moreover, it does not seem to be scriptable, which was essential in order to deal with the volume of data in the corpora.

There appears to be no tagger available at all for Welsh, reflecting the dearth of linguistic tools available to many minority languages (Antonsen et al., 2010).

With no existing software meeting the purpose, a two-week test project in April 2010 looked at the viability of simply writing out entries from Spanish and Welsh dictionaries (see Section 2 below) for each word in the transcription. The results of the tests were encouraging, and the only remaining issue was how to dismbiguate between the returned entries. For this we turned to constraint grammar (Karlsson et al., 1995), and the remainder of this paper reports on how this is used in the autoglossing software developed over the past year.[10]

2 The dictionaries

A key element of any tagging or glossing system is the use of a dictionary to allow lookup of the word in the chosen language.

The Spanish dictionary used in the Autoglosser is based on the one used in Apertium,[11] a free (GPL) platform for developing rule-based machine translation systems. The Welsh dictionary is based on Eurfa,[12] developed by the first author a few years ago, and still the largest free (GPL) dictionary for Welsh. The English dictionary is based on Kevin Atkinson's Moby list.[13]

The use of material with a free or public domain license allows existing lexical resources to be easily adapted and extended for the Autoglosser without having to worry about licensing terms. This is an especially important consideration for minority languages like Welsh,(Streiter et al., 2006) where resources may be limited.

Each dictionary takes the form of one PostgreSQL database table, storing full words (not morphemes). All of the original dictionaries have undergone some refactoring to simplify and standardise their layout, and to correct errors and omissions.[14]

The dictionary table can be easily edited in place, or it can be exported to a CSV file, making it accessible via a spreadsheet for those who are unfamiliar with databases. The dictionary is therefore easy to update, since the format is a familiar glossary-style list of words. This makes expanding or editing the dictionary more accessible for people without extensive computer skills, which is again important for minority languages – no esoteric rules on word-division apply, nor are the contents distributed over several files.

In theory at least, this should simplify the addition of further languages in the future. If a simple wordlist is available, it is possible to plug it into the autoglosser, and get some useful non-disambiguated output immediately; this output can then be progressively refined by the addition of CG rules,[15] and refactoring of the dictionary

[8] http://childes.psy.cmu.edu/morgrams

[9] http://www.sil.org/computing/toolbox

[10] The Bangor Autoglosser software, licensed under the GPL, is available from http://siarad.org.uk/autoglosser.php

[11] http://apertium.org

[12] http://eurfa.org.uk

[13] http://wordlist.sourceforge.net

[14] The English dictionary is particularly prone to include non-existent "words" such as *fam*, *fath*, *gaster*, etc, and further cleaning is still required.

[15] Constraint grammar has been described as "the only grammar-based parser framework" (http://giellatekno.uit.no/cg/11/index.html), and it is indeed very easy for linguists to work with.

lookup to allow a reduction in the size of the dictionaries.

Some entries from the Welsh dictionary are in Table 2. The enlemma column gives the English lexeme for the word, and the pos column gives the part-of-speech (POS).

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			

Table 2 – Entries from the Welsh dictionary.

A similar set of entries from the Spanish dictionary is in Table 3 – it can be seen that the same columns are used in both dictionaries.

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			

Table 3 – Entries from the Spanish dictionary.

Both Spanish and Welsh are inflected languages, where the surface forms give clues about the word's POS. English, however, is an analytic language where the POS of the many homophonous words is defined by their role in the sentence. The format for the English dictionary, some entries for which are in Table 4, reflects this by having the POS reflect all of these possibilities, with the correct POS being selected during disambiguation.

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

Table 4 – Entries from the English dictionary.

For example, **walk** can be a noun (*a short walk*), an imperative verb (*walk the line!*), an infinitive verb (*to walk a mile*) and a present tense verb (*they walk everywhere*). Thus **walk** has the POS **sv**, meaning that it can be either a singular noun or a verb. The main benefit of this approach is that it minimises the number of entries which the dictionary has to include (in this case, one entry instead of four), and therefore makes maintenance of the dictionary easier.

3 The autoglossing process

Each line of the transcribed conversation file is read into an utterances table containing the following fields:

- utterance_id
- filename
- speaker
- surface (the utterance)
- startpoint
- endpoint
- duration
- manual gloss (if present)
- English translation (if present)
- comments (if present)
- precode[16] (if present)

Any non-lexical markers in the utterance are discarded, and it is then split into words, which are stored in a words table with the following fields:

- word_id
- utterance_id
- location of the word in the utterance
- surface (the word)
- automatic gloss (to hold the later output)
- manual gloss (if present)
- language id
- speaker
- filename

Each entry in the words table is looked up against the dictionary table for the appropriate language, using the language assigned to the word by the transcriber.[17]

The lookup includes some basic segmentation of the word. This helps to minimise the number of dictionary entries and make maintenance of the dictionary easier.

For Welsh, the lookup detects mutation[18] and adds corresponding tags:

thad →**tad** (*father*) + am (aspirate mutation)
gael →**cael** (*get*) + sm (soft mutation)

[16] This marks entire utterances in the least-frequent language of the conversation.

[17] In the absence of this, it would in principle be possible to use a brute-force lookup on each dictionary in turn.

[18] Mutation – morphophonemic alteration of initial consonants, which also marks syntactic relations at the clause level – is an important characteristic of the Celtic languages. A Welsh example is: **mae o'n marw** (*he is dying*), but **mae o'n farw** (*he is dead*), where the change **m**→**f** signifies that the mutated word is an adjective and not a verb. These mutations have to be removed in order to get to the underlying lexeme.

For Spanish, tags are added when clitic pronouns attached to verbforms are detected:

ponerle →**poner** (*put*) + **le**[pron.mf.3s]

déjanos →**déja** (*leave*) + **nos**[pron.mf.1p]

For English, tags are added for things like:

(a) elisions:

gonna →**go** # to.prep

we're →**we** # be.v.pres

(b) genitives or verb elisions:

father's →**father** # gb

(c) plural nouns or 3s present tense verbs:

breaks →**break** # pv

(d) adjectives or past tense verbs:

constructed →**construct** # av

(e) adjectives, singular nouns or present participle verbs:

thinking →**think** # asv

(f) adverbs:

quickly →**quick** # adv

All matching entries in the dictionary are then written out to a file in the format required by the constraint grammar parser.[19]

```
"<ddim>"
  "dim"  96,1 [cy] n m sg :nothing: + sm
  "dim"  96,1 [cy] adv :not: + sm
"<yn>"
  "yn"  96,2 [cy] stat :stative:
  "yn"  96,2 [cy] prep :in:
  "gan"  96,2 [cy] prep :with: + sm
"<gynnar>"
  "cynnar"  96,3 [cy] adj :early: + sm
"<iawn>"
  "iawn"  96,4 [cy] adv :OK:
  "iawn"  96,4 [cy] adv :very:
```

Figure 2 – A phrase, after lookup and before disambiguation, meaning "*not very early*", from the file *patagonia1* in the Patagonia corpus (Welsh-Spanish).

```
"<it's>"
  "it"  545,1 [en] pron.sub 3s :it: # gb
"<coming>"
  "come"  545,2 [en] sv infin :come: # asv
"<out>"
  "out"  545,3 [en] adv :out:
"<on>"
  "on"  545,4 [en] prep :on:
"<D_V_D>"
  "D_V_D"  545,5 [en] name
"<then>"
  "then"  545,6 [en] adv :then:
```

Figure 3 – A phrase, after lookup and before disambiguation, from the file *herring7* in the Miami corpus (Spanish-English).

Figures 2 and 3 show the output after lookup of a monolingual phrase in Welsh and English respectively.

The constraint grammar parser applies the rules in the grammar file to discard invalid entries and convert tags where appropriate, and creates another file containing only valid, disambiguated entries. The two phrases given above are shown after disambiguation in Figures 4 and 5.

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

Figure 4 – The Welsh phrase from Figure 2 after disambiguation.

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

Figure 5 – The English phrase from Figure 3 after disambiguation.

This file is then read into the database, and the glosses (in the form of a lexeme+POS-tag string, following the Leipzig schema (Comrie et al., 2008) so far as possible) are extracted and stored in the words table against each word of the original transcription. At this point, the words table looks like Figure 6, where the words in a Spanish utterance meaning "*And if some lorry goes in there, for example, to leave off furniture or whatever.*" have all been glossed appropriately.

Finally, a text with an interlinear gloss, as in Figure 7, is created by writing out the utterances again, along with the concatenated glosses. Comparing to Figure 1, an additional *%aut* tier has been added for each utterance, in parallel with the pre-existing *%gls* tier provided by manual glossing.

The Autoglosser produces glossed text at a rate of 900-1100 words per minute (depending on

[19] We use the visl-cg3 parser developed by Eckhard Bick and Tino Didriksen - http://beta.visl.sdu.dk/cg3.html

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

%aut: be.V.1S.PRES.SPOKEN I.PRON.1S stative.STAT hopeless.ADJ with.PREP take.V.INFIN picture.N.M.SG

%gls: be.1S.PRES PRON.1S PRT hopeless with take.NONFIN picture

%eng: I'm hopeless at drawing

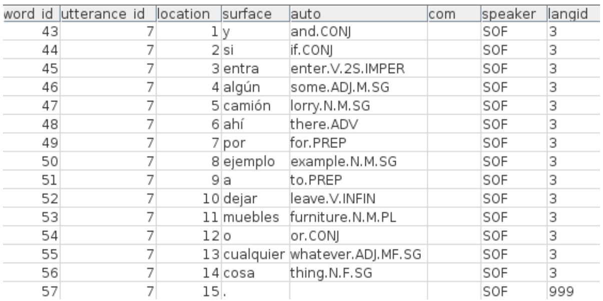
*SER: dw i (y)n tynnu llun i [/] i (y)r plant <i plant> [//] <i (y)r> [//] # i er@s:cym&eng &h Helen@s:cym&eng a Susanna@s:cym&eng a +/ . %snd:"deuchar1"_73881_79477

%aut: be.V.1S.PRES.SPOKEN I.PRON.1S stative.STAT take.V.INFIN picture.N.M.SG to.PREP to.PREP the.DET.DEF children.N.M.PL to.PREP children.N.M.PL to.PREP the.DET.DEF to.PREP er.IM name and.CONJ name and.CONJ

%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...

Figure 7 – Autoglossed excerpt from the file *deuchar1* in the Siarad corpus (Welsh-English) – compare Figure 1.



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

Figure 6 – An utterance from the words table for the file *sastre1* in the Miami corpus (Spanish-English)

whether the original transcription file already contains a manual gloss tier). The transcription of a half-hour conversation can therefore be glossed in around 6 minutes.[20]

The grammar file currently contains about 500 rules for Welsh, about 200 for English, and around 170 for Spanish. These figures reflect the fact that most work so far has been done on Welsh.

Preliminary results (see Table 5) suggest that the Autoglosser's accuracy is 97-99%, depending on the language.[21] We are confident that the accuracy rate can be further improved.

4 Using constraint grammar

We discuss here two issues:

- The addition of tags in the lookup output to specify language, and the handling of these in the grammar so as to allow one-pass disambiguation of multilingual text.
- The different approaches taken in the grammar to handle the differing nature of the languages (already reflected to some extent in the dictionary entries).

4.1 Language-specific rules

Multilingual discourse is far more common than has been assumed in classical linguistics, and it is only over the last 20 years that this important area has been given proper attention. The Autoglosser is the first attempt to apply constraint grammar to multilingual text, and in fact only two things need to be done: (1) include the language tag in the output from each word's lookup; (2) put all the rules (grouped according to language for ease of reference) into the same grammar file.

In Figure 8, the phrase oscillates between Welsh and Spanish, and this is reflected in the inclusion of the tags **[cy]** and **[es]** in the readings.

"<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 mf sg :international:

Figure 8 – A bilingual phrase (*"in a new international car"*) from the file *patagonia2* in the Patagonia corpus (Welsh-Spanish).

In the following noun phrases, the last word (**dro**, **man**, **viaje**) can be both a noun and a verb.

Welsh: **yr ail dro** (*the second time*)
English: **the third man**
Spanish: **el primer viaje** (*the first journey*)

A rule such as:

select (n) if (-1 (ord));

will choose the noun (**n**) reading if the first word to the left (**-1**) is an ordinal (**ord**), meaning that the verb readings for **dro**, **man** and **viaje** will be deleted.

The language tag can be used to constrain the application of the constraint grammar rules to the

[20] The entire Siarad corpus of around 40 hours duration (456,000 words) was glossed in 8h27m.

[21] A recent comparison (Donnelly et al., 2011) suggests that accuracy is within 2% of manual glossing for Welsh, and comparable to CLAN's own MOR tagger for Spanish.

	Corpus	*Files*	*Words*	*Accuracy*	*MCL*	*Coverage*
Welsh-Spanish	Patagonia	patagonia1, 2, 3, 6	15,677	99%	W (92%)	100%
Welsh-English	Siarad	stammers4, deuchar1	10,411	98%	W (81%)	96%
Spanish-English	Miami	zeledon5	4,202	97%	S (59%)	97%

Table 5 – Autoglossing accuracy and coverage for sample files from the three ESRC Centre corpora. In the MCL (most common language) column, W=Welsh and S=Spanish. Coverage is 100% for the Patagonia files because all unknown words were added to the dictionaries before autoglossing.

relevant language.[22] Thus, if the above rule is amended to read:

select ([es] n) if (-1 ([es] ord));

it will only apply to the Spanish phrase, and not to the Welsh or English ones, meaning that the verb reading will still be available in those languages.

It is also possible to make the rules apply across language boundaries by selectively removing language constraints.

In Figure 9, the Spanish **otro** can be either an adjective before a noun, or a pronoun. If the selection rule leaves the noun unspecified as to language:

select ([es] adj) if (-1 (ord));

the adjective reading will be selected before any noun (not just Spanish nouns), as in Figure 10.

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

Figure 9 – A bilingual phrase (*"it's a different zipcode"*) from the file *sastre1* in the Miami corpus (Spanish-English).

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

Figure 10 – The bilingual phrase from Figure 9 after disambiguation.

In Figure 11, **camping** can be an adjective (*the camping ground*), a singular noun (*camping is fun*), or (as here) a verb. In **vamos camping**, the **asv** tag can be converted to the desired present participle verb tag by referring to the meaning of the preceding verb, so that the rule applies to both English (*go camping*) and Spanish (*vamos camping*):

substitute (sv infin asv) (v prespart) ([en] sv infin asv) (-1 (:go:));

as in Figure 12.

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

Figure 11 – A bilingual phrase (*"every time that we go camping"*) from the file *sastre1* in the Miami corpus (Spanish-English).

4.2 Tidying readings

The re-use of lexical resources can lead to a conflict – for many purposes, a comprehensive dictionary giving as many entries as possible for a particular word is desirable, but these multiple entries are not required for an application like the autoglosser, where one lemma will usually be sufficient for tagging purposes.

In cases where the entries are archaic or infrequent words, we use CG **select** rules to remove them from consideration. The Welsh words **huno** (*sleep*) and **pallu** (*refuse*) are low-frequency, so the following rules are applied:

remove ("huno" [cy] :sleep:);

remove ("pallu" v :refuse:);

In other cases, where a single word has different meanings we use CG **select** rules to prioritise one of the meanings. The Welsh dictionary gives two

[22] In practice, there is only a small number of cases where full constraint of the rules is essential (because only a couple of dozen words in each language overlap orthographically), but it is prudent at this stage to err on the side of over-specification.

```
"<cada>"
    "cada" 79,5 [es] adj mf sg :every:
"<vez>"
    "vez" 79,6 [es] n f sg :time:
"<que>"
    "que" 79,7 [es] pron.rel :that:
"<nos>"
    "yo" 79,8 [es] pron.obl mf 1p :us:
"<vamos>"
    "ir" 79,9 [es] v 1p pres :go:
"<camping>"
    "camp" 79,10 [en] v prespart :camp: #
```

Figure 12 – The bilingual phrase from Figure 11 after disambiguation.

meanings for **cyfeiriad** (*direction* and *address*) – the following rule ensures that the *address* meaning is ignored:

select ("cyfeiriad" [cy] :direction:);

The lookup process can generate readings which are invalid, and these need to be removed. The cohort of readings for the Welsh word **nos** (*night*) will include an incorrect one interpreting it as a nasally-mutated form of the imperative (**dos**) of the verb **mynd** (*go*), which is linguistically impossible. This sort of entry can be removed with a rule like:

remove ([cy] "mynd" v 2s imper nm);

A similar issue arises when indeterminate words are being looked up. It will be recalled that indeterminate words are those which appear in dictionaries of both languages, so it is impossible to state unequivocally which language they belong to.[23] Since the practice in the transcriptions is to use English spelling for indeterminate words, lookup for these words uses the English dictionary. The interaction with Welsh mutation can lead to invalid readings, such as the interpretation of the hesitation marker **um** as a soft-mutated form of the word gum, which is extremely unlikely. This can be removed with a rule like:

remove ([in] "gum" n sg sm);

4.3 Nature of rules

Spanish and Welsh are inflected languages,[24] while English is an analytic language with few inflections (mainly in "strong" verbs). This is reflected in the nature of the rules that have proved most efficient in the autoglosser.

For Spanish and Welsh, surface forms are fairly well-defined by their shape – **empieza**, for instance, can only be the second/third person singular present or the second person singular imperative of **empezar** (*to begin*). The lookup fetches these entries from the dictionary,[25] and so the rules consist mainly of **select** rules (with a few **remove**s and **substitute**s).

For English, on the other hand, the surface form gives us few clues about the part-of-speech a word belongs to, which is largely defined by its role in the sentence – **break** can be a singular noun, or a verb infinitive, or the non-third person singular present tense. Instead of giving **break** three entries in the English dictionary, we have chosen, as noted in Section 2, to assign it one entry, with a tag (*sv*) which reflects this diversity of role.

The result is that the the vast majority of rules for English are **substitute**s, converting one set of tags into another. For example, the surface word **miniature** can be either an adjective or a singular noun, so it is tagged *as* in the dictionary. Rules such as the following then handle its correct tagging based on context:

substitute (as) (adj) ([en] as) (1 ([en] n) or ([en] pron));

This says that an English *as* tag should be converted to an *adjective* tag when the word is followed by a noun or pronoun (e.g. **a miniature rabbit**, **miniature ones**).

Similar refinement rules can be applied to other parts-of-speech such as pronouns:

substitute (pron.sub) (pron.obj) ([en] pron.sub) (-1 ([en] v infin));

which will correctly tag **it** in **and open it** as an object pronoun, or verbs:

substitute (av past) (v past) ([en] av past) (-1 ([en] pron.sub)) (not -1 (have.v.pres)) (not -2 ("have"));

Here, **bought**, which can either be an adjective (**bought goods**) or a past verb, has the latter selected provided it is not preceded by enclitic or self-standing instances of the auxiliary verb **have**. This correctly tags **we bought**, but passes over **you've bought**, or **we have bought**. These latter examples can be handled by an additional rule converting the tag to a past participle:

[23] This is meant synchronically rather than diachronically, in terms of current usage in both languages – historically, the word may be considered a loanword.

[24] Though it should be noted that in Welsh, particularly spoken Welsh, inflected verbforms are now widely replaced by periphrastic forms.

[25] The possibility of de-conjugating inflected verbs on-the-fly is attractive, but may be too complex to attempt at this stage.

substitute (av past) (v pastpart) ([en] av past) (-1 (have.v.pres) or ("have") or ("be"));
which will also correctly tag **it was bought**.

It can be said that in general these **substitute** rules are more dependent on rule order than **select** or **remove** rules, since the output of a substitution earlier in the stack needs to be taken into account by a rule later in the stack.

4.4 Rule scope

Our current view is that **remove** and **select-if-not** rules are particularly problematic unless they are carefully constrained. A select rule is exclusive in what it applies to, and it might be considered possible to frame a select-if-not rule to be equally exclusive. By its nature, however, the set of negatives is larger than the set of positives, so it is easy to miss something obvious, particularly when dealing with rules that can apply across languages.

An example of this was the results of combining a set of grammar rules for Welsh with a previously working set of rules for Spanish - the result was 304 regressions in the Spanish output. This was traced to a Welsh rule selecting an imperative if the particle **ni** did not appear in first position in the sentence:

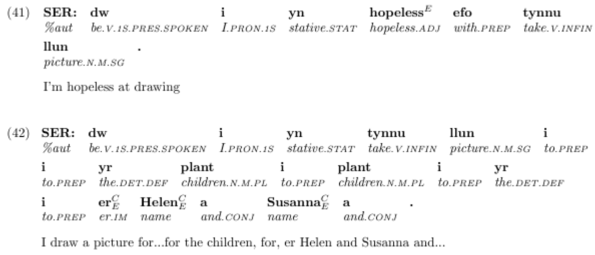
select (imper) if (not @1 ("ni"));
Since **ni** did not appear in this context in Spanish, all instances where an imperative reading was possible were selected, giving the regressions. In this case, the rule can easily be amended by adding a **[cy]** tag, but the point is that the impact of this type of rule can be subtle.

5 Spin-off benefits

The Autoglosser was primarily intended to provide reasonably accurate glosses automatically, thus saving researcher time, but it has also had a number of spin-off benefits which contribute to easier handling of the corpora.

Perhaps the most useful is the ability to use the file contents in the database tables to print out a typeset copy of the transcription in LaTeX, using John Frampton's *ExPex* package.[26] Figure 13 shows the results of this process on the file excerpt previously shown in Figure 1. This greatly facilitates checking for errors in the glossing.

Being able to access the file contents via database queries adds another tool for correcting typos. Selecting all unglossed words in the words table gives a list of words which are either unknown because they are not in the dictionaries, or could not be found in the dictionaries because they were mis-spelt (i.e. typos). It is interesting to note that even after two rounds of detailed manual proofreading such typos account for about 0.5% of the words in a file on average, and this technique provides a method of eliminating them.



(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)

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...

Figure 13 – The text in Figure 1 typeset to show alignment of the surface words and their POS-tags.

Autoglossing enforces consistency across the corpus (so that, for instance, Welsh **ychydig** does not appear in some places as *a bit*, and in other places as *a little*), and makes it much easier to change or enrich tags globally. This sort of consistency facilitates data-mining, in that queries can be correspondingly simpler.

6 Further work

Although the current configuration of CG rules is working well, we hope to explore further refinement of the grammar. This would include not only conflating similar rules within a language, but also seeking to use the grammar to mark clause relationships. The latter would be of value in the further linguistic analysis of the influence of clause structure on language switching in bilingual discourse.

Acknowledgments

The support of the Arts and Humanities Research Council (AHRC), the Economic and Social Research Council (ESRC), the Higher Education Funding Council for Wales and the Welsh Assembly Government is gratefully acknowledged. The work presented in this paper was part of the programme of the ESRC Centre for Research on Bilingualism in Theory and Practice at Bangor University.

[26] http://www.math.neu.edu/ling/tex/

References

Lene Antonsen, Trond Trosterud, and Linda Wiechetek. 2010. Reusing grammatical resources for new languages. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC-10)*. European Language Resources Association.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses. *http://eva.mpg.de/lingua/resources/glossing-rules.php*.

Kevin Donnelly, Sarah Cooper, and Margaret Deuchar. 2011. Glossing chat files using the Bangor Autoglosser. Paper presented at ISB8, Oslo, May 2011.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila. 1995. *Constraint grammar: a language-independent system for parsing unrestricted text*. Mouton de Gruyter.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates.

Jonathan Stammers. 2010. *The Integration of English-origin Verbs into Welsh: A Contribution to the Debate over Distinguishing between Code-switching and Lexical Borrowing*. Verlag Dr. Műller.

Oliver Streiter, Kevin P. Scannell, and M Stuflesser. 2006. Implementing NLP projects for non-central languages: instructions for funding bodies, strategies for developers. *Machine Translation*, 20(4).