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, focussing on 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] – @1 for Welsh, @2 for English, @0 for indeterminate[6]), and a tag (*%snd*) giving the time-location 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, (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@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 ...
```

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.

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, and the remainder of this paper reports on how this is used in the autoglossing software developed over the past year.

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,[10] a free (GPL) platform for developing rule-based machine translation systems. The Welsh dictionary is based on Eurfa,[11] 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.[12]

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.[13]

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,[14] and refactoring of the dictionary lookup to allow a reduction in the size of the dic-

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

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

[10] http://apertium.org

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

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

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

[14] 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.

tionaries.

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 part-of-speech (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. 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.

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

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[15] (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.[16]

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 and adds tags for mutation.[17]

For Spanish, tags are added when clitic pronouns attached to verbforms are detected.[18]

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

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

[17]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.

[18]For example, **arreglarlos** ←**arreglar+los** (*to fix* + *them*)

For English, tags are added to mark the occurrence of genitives plurals and adverbs and simple verb formations.[19]

All matching entries in the dictionary are then written out to a file in the format required by the contraint grammar parser.[20] When run against the file, the parser applies grammar rules to discard invalid entries, and creates a file containing only valid, disambiguated entries.

This file is then read into the database, and the glosses (in the form of a lexeme+POS-tag string) are extracted and stored in the words table against each word of the original transcription. At this point, the words table looks like Figure 2, where the words in an utterance meaning "*And if some lorry goes in there, for example, to leave off furniture or whatever.*" have all been glossed appropriately.

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

Finally, the utterances are written out again, along with the concatenated glosses (following the Leipzig schema (Comrie et al., 2008) so far as possible), to create a final text with an interlinear gloss, similar to Figure 1.

The autoglosser produces glossed text at a rate of 900-1100 words per minute (depending on whether the original transcription file already contains a manual gloss layer). The transcription of a half-hour conversation can therefore be glossed in around 6 minutes.[21]

The grammar file currently contains about 150 rules for Spanish, about 180 for Welsh, and around 200 for English. Autoglosser accuracy is 97-99%, depending on language. [Detailed figures to be included.]

[19]For instance, **son's**←**son+'s**, **houses**←**house+s**, **quickly**←**quick+ly**, **walks**←**walk+s**, **walking**←**walk+ing**.

[20]http://visl.sdu.dk/cg3.html

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

4 Applying 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. That this was so easy is a testament to the versatility and power of the VISL-CG3 parser.

The autoglosser knows which dictionary to look up because each word is annotated by the transcriber with the language it comes from. All that needs doing is to reflect this in the lookup output. In the following example, the phrase **mewn motor newydd internacional** (*in a new international car*) oscillates between Welsh and Spanish, and this is reflected in the addition of the tags **[cy]** and **[es]** to 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:

The grammar can then use the language tag to constrain the application of its rules to the relevant language. Thus, a rule like:

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

(to choose the reading marked "noun" if the preceding word is an ordinal) will apply to all the languages covered, whereas:

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

will only apply to Spanish: **el primer viaje** (*the first journey*).

In fact, the number of cases where this is absolutely essential is very small, but at this stage of

developing the autoglosser, we are erring on the side of caution.

[We continue with further examples and discussion.]

4.2 Nature of rules

Both Spanish and Welsh are inflected languages (modern Welsh considerably less so than it was), 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 tense or the second person singular imperative of **empezar** (*to begin*). The lookup fetches these entries from the dictionary,[22] 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 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**).

[We continue with further examples and discussion.]

Finally, we discuss more general issues (on which we would welcome input from other CG practitioners) such as:

- The best way of structuring a multilingual grammar so as to prevent bleed-through of one language's rules into the others'.
- Rule types to avoid - our current view is that **remove** and **select-if-not** rules are particularly problematic unless they are carefully constrained

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

5 Further work

Although the current configuration of 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.

References

B Comrie, M Haspelmath, and B Bickel. 2008. Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, New Jersey, 3rd edition.

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

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