

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 January 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 format⁴ (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

	<i>Chats</i>	<i>Hours</i>	<i>Words</i>	<i>Date</i>
Welsh-English	69	40	456k	2009
Welsh-Spanish	42	20	183k	2011
Spanish-English	31	20	126k	2011
	142	80	765k	

Table 1 – The three ESRC Centre corpora

(e.g. **SER*), followed by the transcribed speech (with each word tagged for language⁵ – @1 for Welsh, @2 for English, @0 for indeterminate⁶), and a tag (%*snd*) giving the time-location of the utterance in the audiofile.

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 (for instance, using the Siarad corpus (Stammers, 2010) has shown that Welsh loan-verbs such as *textio* (to text) behave more like ordinary Welsh verbs the more frequent they are).

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>

²Siarad 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.

⁵The 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.

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

```

*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),⁷ 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 Toolbox⁸ 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.

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 for each word in the transcription. The Spanish dictionary was based on the one used in Apertium,⁹ a free (GPL) platform for developing rule-based machine translation systems, and the Welsh dictionary was based on Eurfa,¹⁰ developed by the first author a few years ago, and still the largest free (GPL) dictionary for Welsh. The results of the tests were encouraging, and the only remaining issue was how to disambiguate between the returned entries. For this we turned to constraint grammar, and the rest of this paper reports on how this is used in the software we have developed over the last year (the Bangor Autoglosser).

2 Overview of the process

Each utterance (that is, each line that is preceded by a speaker ID) in the transcribed conversation file is read into a PostgreSQL database table containing the following fields:

- utterance_id
- filename

⁷<http://chilides.psy.cmu.edu/morgrams>

⁸<http://www.sil.org/computing/toolbox>

⁹<http://apertium.org>

¹⁰<http://eurfa.org.uk>

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

Each utterance in this table is cleaned (non-lexical markers are discarded), and then split into words, which are stored in another 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 word in this table is looked up against a dictionary table for the appropriate language, using the language id assigned to the word by the transcriber.¹¹

All matching entries in the dictionary are then written out to a file in the format required by the constraint grammar parser.¹² 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

¹¹In the absence of this, it would in principle be possible to use a brute-force lookup on each dictionary in turn.

¹²<http://visl.sdu.dk/cg3.html>

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

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.

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.¹³

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.]

3 Dictionary structure

The dictionaries used are all derived from free resources,¹⁴ though heavily adapted. We aim, so far as possible, to maintain the dictionaries through a simple spreadsheet, in order to allow easy update of the current language support, and also to simplify the addition of further languages in the future. For instance, 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 refactor-

¹³The entire Siarad corpus of around 40 hours duration (456,000 words) was glossed in 8h27m.

¹⁴<http://eurfa.org.uk>, <http://apertium.org>, <http://wordlist.sourceforge.net>

¹⁵Constraint 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.

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

This process has been applied to all three languages to differing degrees, and it could be taken further. For Welsh, the lookup now detects and adds tags for mutation,¹⁶ while for Spanish, tags are added when clitic pronouns attached to verb-forms are detected.¹⁷ For English, tags are added to mark the occurrence of genitives, plurals and adverbs.

[We go on to give some examples of these, since they are relevant to the discussion of the CG rules in the main section below.]

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 CLAN transcription system was originally devised to study the development of language in children, but it has been adapted by researchers at the ESRC Centre to capture bilingual spoken interactions. Likewise, although constraint grammar has to date mainly been applied to monolingual texts, we have adapted it to handle 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**

¹⁶Mutation – 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.

¹⁷For example, **arreglarlos** ← **arreglar+los** (*to fix + them*)

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,¹⁸ and so the rules consist mainly of **select** rules (with a few **removes** and **substitutes**).

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 vast majority of rules for English are **substitutes**, 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

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

¹⁸The possibility of de-conjugating inflected verbs on-the-fly is attractive, but may be too complex to attempt at this stage.

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. Müller.