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 greater than 95% 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] has assembled some 130 bilingual conversations totalling 80 hours and 750,000 words in three languages: Welsh, Spanish and English. The conversations are grouped into three corpora, all available under the GNU GPL.[2] Each recording is provided with a detailed transcription in the widely-used CLAN format (**?**), 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 Figures 1 and 2. ***FIX*** We need a text with the free translation. Each language utterance has an ID for the speaker (*KEV), and a tag giving the time-location of the utterance in the conversation. Below that is the gloss tier (%gls in the Welsh file, and %aut in the Spanish file). Below that is the free translation into English.

2 Autoglosser overview

The Siarad (Welsh-English) corpus (published in 2009) was glossed manually, but this was tedious and time-consuming, and in order to save valuable specialist time the Miami (Spanish-English) and Patagonia (Welsh-Spanish) corpora (to be published this summer) are instead being glossed using software developed at the Centre over the last few months.

The autoglosser uses the PHP scripting language in combination with a PostgreSQL database. Each utterance in the transcribed conversation file is imported into a database table, and then split into words, which are stored in another table. Each word is looked up against a dictionary table for the appropriate language – the transcription notes the language of each word, so this is easy.[3] The cohorts of readings are then written out to a file in the format required by the contrast grammar parser,[4] which applies the grammar rules to winnow out invalid readings and write out another file. This is then read back in to the database, with the glosses being extracted and stored in the words table. Finally, the utterances are written out again, along with the concatenated glosses (following the Leipzig schema (**?**) so far as possible), to create a final text with an interlinear gloss.

The autoglosser takes 2-3 minutes to import, gloss and output a 5,000-word text. The grammar file currently contains about 150 rules for Spanish, about 180 for Welsh, and around 200 for English. Autoglosser accuracy is more than 95% in all three languages. [Detailed figures to be included.]

3 Dictionary structure

The dictionaries used are all derived from free resources,[5] though heavily adapted. We aim, so far

---

```
*KEV: eso más bien yo creo que lo que va a hacer es como un adorno pero .
#%snd:"sastre1_b"_60356_63286#
%aut: that.PRON.DEM.NT.SG more.ADV well.ADV I.PRON.SUB.MF.1S believe.V.1S.PRES
that.CONJ the.DET.DEF.NT.SG that.CONJ
go.V.23S.PRES to.PREP do.V.INFIN be.V.23S.PRES like.CONJ one.DET.INDEF.M.SG
embellishment.N.M.SG but.CONJ
*KEV: baja la velocidad ahí ? #%snd:"sastre1_b"_63286_65895#
%aut: lower.V.2S.IMPER the.DET.DEF.F.SG velocity.N.F.SG there.ADV
*SOF: pero la calle no la van a hacer no ? #%snd:"sastre1_b"_65876_70050#
%aut: but.CONJ the.DET.DEF.F.SG street.N.F.SG not.ADV her.PRON.OBJ.F.3S
go.V.23P.PRES to.PREP do.V.INFIN not.ADV
```

Figure 1 – Excerpt from a CHAT (.cha) file in the Miami corpus (Spanish-English)

```
*KEV: eso más bien yo creo que lo que va a hacer es como un adorno pero .
#%snd:"sastre1_b"_60356_63286#
%aut: that.PRON.DEM.NT.SG more.ADV well.ADV I.PRON.SUB.MF.1S believe.V.1S.PRES
that.CONJ the.DET.DEF.NT.SG that.CONJ
go.V.23S.PRES to.PREP do.V.INFIN be.V.23S.PRES like.CONJ one.DET.INDEF.M.SG
embellishment.N.M.SG but.CONJ
*KEV: baja la velocidad ahí ? #%snd:"sastre1_b"_63286_65895#
%aut: lower.V.2S.IMPER the.DET.DEF.F.SG velocity.N.F.SG there.ADV
*SOF: pero la calle no la van a hacer no ? #%snd:"sastre1_b"_65876_70050#
%aut: but.CONJ the.DET.DEF.F.SG street.N.F.SG not.ADV her.PRON.OBJ.F.3S
go.V.23P.PRES to.PREP do.V.INFIN not.ADV
```

Figure 2 – Excerpt from a CHAT (.cha) file in the Siarad corpus (Welsh-English)

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,[6] and refactoring 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,[7] while for Spanish, tags are added when clitic pronouns attached to verbforms are detected.[8] 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.

---

[6]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.

[7]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.

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

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,[9] and so

---

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

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 ann 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

References