THE BANGOR AUTOGLOSSER: A MULTILINGUAL TAGGER FOR CONVERSATIONAL TEXT

Kevin Donnelly and Margaret Deuchar

ESRC Centre for Research on Bilingualism, Bangor
{k.donnelly|m.deuchar}@bangor.ac.uk

ABSTRACT

We describe software which tags multilingual transcriptions of spoken texts in Welsh, Spanish and English to a high degree of accuracy. This is believed to be the first application to handle the tagging of Welsh text. The tagger is easily extensible to other languages, and may be of interest to researchers in natural language processing in minority languages, as well as to those working on the informal language used in speech-to-text conversion, voice recognition software, and so on.

KEYWORDS

bilingual, speech, tagging, corpora

1. INTRODUCTION

Over the last few years, Bangor University's ESRC Centre for Research on Bilingualism [1] has collected and transcribed extensive spoken bilingual corpora in three language pairs, as outlined in Table 1.

Table 1. ESRC Centre corpora.

Corpus	Languages	Publication	Length (hours)	Conversations	Words
Siarad	Welsh-English	March 2009	40	69	456k
Miami	Spanish-English	July 2011	20	32	161k
Patagonia	Welsh-Spanish	July 2011	20	31	126k

Each corpus is licensed under the GNU GPL [2], ensuring that it can be freely used by researchers. Each file in the corpora consists of a high-quality transcription of the recording of the conversation [3] in the widely-used CLAN format [4], along with a gloss - lexeme plus part-of-speech (POS) tag - for each word in the file, broadly following the Leipzig glossing schema [5], and a free translation in English of the spoken words.

2. THE BANGOR AUTOGLOSSER

The Siarad corpus was glossed manually, but this was time-consuming, and in order to save valuable researcher time with the Miami and Patagonia corpora it was decided to investigate the possibility of using software to do the majority of glossing automatically. Development of the software began in April 2010. The overall aims included:

- *Use of existing GPL or open-source software and resources where possible.* This would leverage existing work in this area, as well as lower the threshold for contributions in future.

- *Tagging across at least two languages simultaneously, preferably using a single application infrastructure.* This would simplify the tagging process, since there would be no need to use different tools for utterances in different languages (for example, to use two different taggers, perhaps with different requirements as regards input format or wordlist structure).
- *Ability to handle conversational language.* The transcripts include repetitions, restarts, overlaps, and so on, which do not occur to the same extent in formal written text.

3. How the Autoglosser Works

3.1 Text import

After some initial text preparation, the conversation text is imported via a PHP script into a PostgreSQL database table, with each utterance a separate record. Thus the following utterance by the speaker **LOR** in Siarad's **Fusser32** will form one record, with other items like manually prepared glosses or translations also attached to it:

> **xxx <oedd@1 gen@1 i@1 (ddi)m@1 prin@1 ddim@1> [?] bwyd@1 ar_ôl@1 yn@1 fridge@2 &=laughs .**
>
> *I had almost no food left in the fridge.*

The human transcriber has marked each word with a language tag (**@1** for Welsh and **@2** for English), as well as noting such speech artefacts as inaudible words (**xxx**), backtracking (text inside angle brackets), and non-linguistic cues (**&=laughs**). The constituents of multi-element "words" such as **ar ôl** are joined by an underscore, so that they can be represented by a single gloss. The details of the marking are described in [4], but since the marking as used by the Centre has changed slightly over the years (CLAN was not originally developed to handle multilingual text), the autoglosser handles four different marking systems.

Each of these utterances is then split into its constituent words, discarding artefact markers, and this segmented set is stored in another table - the utterance above would therefore have four words stored:

> **bwyd / ar_ôl / yn / fridge**
>
> *food left in [the] fridge*

The language tags are stripped off, and stored as another field against a word's record.

3.2 Dictionary lookup

The autoglosser then runs through the words table, looking up each word against a dictionary table, and writing all the possible interpretations of the surface word into a file in a format which will allow the output to be disambiguated. For instance, using a different utterance:

> **mae@1 (y)n@1 braf@1 nice@0** (*it's fine, nice*)

we have:

```
"<mae>"
        "bod"  {25,1} [cy] v 3s pres :be:
        "bae"  {25,1} [cy] n m sg :bay: + nm
"<yn>"
        "yn"  {25,2} [cy] stat :stative:
        "yn"  {25,2} [cy] prep :in:
"<braf>"
        "braf"  {25,3} [cy] adj :fine:
"<nice>"
```

"nice" {25,4} [0] nice

"<.>"

In this example, there are two surface words (the Welsh **braf** (*fine*) and the English **nice**) which only have one entry in the dictionary. However, **mae** (*is*) and **yn** (*stative*) each have two possible entries: the former is either a part of the verb "to be", or a nasally-mutated noun meaning "bay", and the latter is either the stative particle, or the preposition "in".

The dictionaries are all heavily-adapted versions of lexicographical resources under a free license. The Welsh dictionary is based on **Eurfa**, the first free Welsh dictionary [6], while the Spanish dictionary is based on that used in the machine translation project **Apertium** [7]. The English dictionary is derived from Kevin Atkinson's part-of-speech file, which in turn combines data from Grady Ward's Moby Part-of-Speech II and the WordNet database [8].

The three languages offer an interesting spectrum of language-structure. Spanish shows extensive inflection, with different adjectival endings depending on the tense and number of the noun, and multiple verb-forms denoting tense, person and number. Historically, Welsh had extensive inflection, especially in verbs, but this is much reduced in the modern language. Modern English is an analytic language with few inflections, where a single word may share multiple grammatical roles - for instance, "back" may be a singular noun denoting a part of the body, a verb meaning "to reverse", or an adverb describing motion ("give back", "go back").

The main feature of interest in Welsh is the mutation system, where the initial sounds of a word vary based on the words beside them, or the syntax of the phrase:

du (*black*) but **cath ddu** (*a black cat*)

mae'r trên yn mynd (*the train is going*) but **cyn i'r trên fynd** (*before the train goes*)

Initial versions of the dictionaries simply listed all related forms, allowing easy lookup. So the Welsh dictionary, for instance, had separate entries for **mynd** (*to go*) and **fynd** (*to go*, with soft mutation), the Spanish dictionary separate entries for **arreglar** (*to fix*) and **arreglarlo** (*to fix it*), and the English dictionary for **walk** (the noun) and **walk** (the verb).

The major drawback of this approach, however, is that although lookup is simplified, updating the dictionaries becomes more difficult. A number of entries have to be made for each new word added, and in turn this increases the size of the dictionaries, even though many of the additional forms may be of low frequency and unlikely to occur.

More recent versions of the autoglosser therefore move towards doing at least some segmentation on-the-fly. For Welsh, this means checking for and marking mutation, while for Spanish it means checking for and marking clitic pronouns attached to verbforms. Since the dictionaries need only contain one word (in these cases, **mynd** and **arreglar**), the result is a major decrease in their size - in Spanish, 87% of the 650,000 verbforms were clitic items, while in Welsh, 49% of the 420,000 entries were mutated items. More could probably be done along these lines, for example, attempting to deconjugate verbforms on-the-fly, but this is not a top priority at present.

The English dictionary required a slightly different approach. Firstly, since words may simultaneously appear in several part-of-speech categories, it was decided to mark them as such, and allow the valid part-of-speech to be selected as part of the disambiguation process (see below). So **walk** is marked in the dictionary as *sv*, meaning that it can be either a singular noun or a verb. Secondly, some recurring affixes are checked for and marked on-the-fly - these include genitival **'s** (**my daughter's boyfriend)**, plural forms (**controversies**), agentive **-er** (**worker**), adverbial **-ly** (**happily**), and potential **-able** (**treatable**). In some cases, these rules

can lead to "noise" which requires additional disambiguation. For instance, **master** can also have the additional interpretation **master < mast** returned, which, although possible (e.g. **four-master** in describing a ship), is a lot less likely to occur in general conversation. Thirdly, some basic stemming is done on inflected verbforms, so that **walking** and **walked** get referred to the lexeme **walk**.

3.3 Disambiguation using constraint grammar

Once all the words have been looked up in the dictionaries, the resulting file can be used as input to **VISL-CG3**, a free contraint grammar parser developed at the University of Southern Denmark [9]. Constraint grammar (CG) dates back to the early 1990s [10], and the VISL-CG3 implementation is exceptionally versatile and powerful, using subclause delineation, generalized dependency markers and semantic prototype tags [11]. It should be noted that the autoglosser currently uses CG in only the most basic form, to carry out disambiguation where there is more than one possible interpretation for a word, but in the future we may seek to extend this to clause delineation.

The main innovation in the autoglosser's use of CG is to include a language marker in the input file - the sample above, for instance, shows **[cy]** for Welsh words, and **[0]** for words like **nice** that are used in both English and Welsh. This means that rules relating to multiple languages can reside in the same grammar file, allowing multilingual text to be parsed in one iteration. It also has the side-effect that we can have rules apply across language boundaries: in

mucho speed bump (*a lot of speed bumps*)

los dry walls (*the dry walls*)

the Spanish words (**mucho**, **los**) can be assigned the correct POS tags even though the following words are in English.

An important benefit of CG is that, as a grammar-based parser, its rules are not only powerful and easy to understand, but they also "feel" right from a linguistic (rather than computer science) viewpoint. So far, we have found that basic selection or removal rules can handle most of the disambiguation in Spanish and Welsh. For Welsh:

select ("ei" a :her:) if (1 amnoun);

select ("ei" a :his:) if (1 smnoun);

says that the possessive adjective item **ei** marked "her" should be chosen to if the following word is a noun with aspirate mutation (**ei thad** < **tad**, *father*), while the one marked "his" should be chosen if the following noun is soft-mutated (**ei dad** < **tad**, *father*). For Spanish:

select (v pastpart) if (-1 ("haber") or ("estar") or ("ser");

says that the verb item marked "past participle" should be chosen if the preceding word is a form of the auxiliary verbs **haber** (*to have*), **estar** (*to be*) or **ser** (*to be*). In these cases, it is not necessary to specify the language the rule applies to, since the context makes this clear (**ei** only occurs in Welsh, and **haber/estar/ser** only in Spanish).

For English, considerably more substitution rules are included, in order to handle the multi-faceted dictionary entries. For instance, the stemmer adds the tag *asv* (adjective, singular noun, or verb) to the entry for **cooling**, and points to the lexeme **cool**, which is already tagged *sv infin* (singular noun, or verbal infinitive). The following rule converts these tags to *v prespart* (verbal present participle) when they occur together after a preceding English verb (e.g. **starts cooling**):

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

This rule:

substitute (2s123p) (3p) (v 2s123p) (-1 (pron.sub 3p) or (n pl));

converts a verb marked second person singular or first/second/third person plural to one marked third person plural if the preceding word is a third person subject pronoun or a plural noun (e.g. **they were**, or **the taxes were**).

An interesting side-effect of these conversions in English is that the tagging, because it is based on function rather than form, is correct even in cases where the speech is non-standard. For instance, in :

he	**talk**	**in**	**Spanish**	**.**
he.PRON.SUB.M.3S	*talk.V.3S.PRES*	*in.PREP*	*name*	

talk is correctly tagged as third person singular. Likewise, in:

you	**seen**	**that**	**show**	**?**
you.PRON.SUB.2SP	*seen.V.PAST*	*that.DEM.FAR*	*show.N.SG*	

seen is correctly tagged as past tense rather than past participle.

So far, there are about 150 CG rules for Spanish, about 180 for Welsh, and around 200 for English.

3.4 Output creation

Once the constraint grammar has decided which form is valid for each word, it outputs a disambiguated file, and the data for each word is then read back into the words table, and concatenated to produce a gloss string consisting of the lexeme and the POS tags. This string follows the formatting guidelines of the Leipzig glossing schema [5] as far as possible.

The original utterances in the utterances table are then written out into a final file, with the automatically-produced gloss added, somewhat similar to the example above. The autoglossed file can then be opened in the CLAN application for further work (e.g. the addition of a translation, the correction of audio placemarks, gloss checking, etc), or it can be output in different formats for gloss checking, printing, etc (see below).

The autoglosser takes 2-3 minutes to autogloss a 5,000-word file from initial import to final output.

4. Results

Figures for recall (coverage) are not given here, since the texts had all their words included in the dictionary before the analysis began. Figures for precision (accuracy) are given in Table 2 - these figures count all instances of non-disambiguation, incorrect POS or incorrect lexeme as errors. (Note that the residue in the language balance column in Table 2 is due to indeterminate words, that is, words which are used in both languages, and which therefore cannot be assigned unambiguously to one of them.)

Table 2. Autoglosser precision.

Corpus	File	Words	Language balance	Precision
Siarad	stammers4	4383	Welsh 85%, English 1%	98%
Miami	herring7	4855	English 79%, Spanish 18%	94%
Miami	sastre1	5990	English 48%, Spanish 47%	96%
Patagonia	patagonia2	4709	Welsh 90%, English/Spanish 1%	99%

5. Additional Uses of the Autoglosser

Although the main purpose of the autoglosser is to generate glosses for the conversation texts, the availability of the word data in a database table allows it to be leveraged for other added-value purposes.

The most important of these is perhaps access to the conversation material through a browser. We are experimenting with a website [12] where the texts can be presented along with their audiofiles, and where the user can undertake some basic analysis simply by pointing and clicking. This means that the user does not need to download and install the CLAN application, or learn the correct syntax for its analysis commands. Much more remains to be done here, but it is a start on opening up the corpora to a less technical audience.

The website uses different fonts and colours for different aspects of the text, with the aim of making the content of the conversations easier to read. The same aim applies to printed versions. A sample of the glossed text in the default CLAN format follows:

> *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

By using the autoglosser data and John Frampton's ExPex package [13] for the LaTeX typsetting system [14], we can transform this into the much more attractive output in Figure 1.

(34)

KEV:	**eso**	**más**	**bien**	**yo**	**creo**
%aut	*that.PRON.DEM.NT.SG*	*more.ADV*	*well.ADV*	*I..PRON.SUB.MF.1S*	*believe.V.1S.PRES*
que	**lo**	**que**	**va**	**a**	**hacer**
that.CONJ	*the.DET.DEF.NT.SG*	*that.PRON.REL*	*go.V.23S.PRES*	*to.PREP*	*do.V.INFIN*
es	**como**	**un**	**adorno**	**pero**	**.**
be.V.23S.PRES	*like.CONJ*	*one.DET.INDEF.M.SG*	*embellishment.N.M.SG*	*but.CONJ*	

(35)

KEV:	**baja**	**la**	**velocidad**	**ahí**	**?**
%aut	*lower.V.2S.IMPER*	*the.DET.DEF.F.SG*	*velocity.N.F.SG*	*there.ADV*	

(36)

SOF:	**pero**	**la**	**calle**	**no**	**la**	**van**
%aut	*but.CONJ*	*the.DET.DEF.F.SG*	*street.N.F.SG*	*not.ADV*	*her.PRON.OBJ.F.3S*	*go.V.23P.PRES*
a	**hacer**	**no**	**?**			
to.PREP	*do.V.INFIN*	*not.ADV*				

Figure 1. Typeset output

Other aspects of the texts can also be probed. One such is the question of when people are more likely to switch into the other language in bilingual conversations - the autoglosser data will make it easier to examine the occurrence and context of individual words in order to shed light on this question. Another aspect is whether language switching is more likely to take place

within or between clauses - the autoglosser data will simplify the task of classifying clauses according to type and language to investigate the contribution of clause structure here.

A final aspect is using the autoglosser data to explore new ways of presenting information. Figure 2 gives an example where the "language profile" of a conversation (Miami/sastre1) has been summarised - utterances are along the X axis (one bar for each utterance), and length of utterance (in words) is along the Y axis. Black represents English words, and grey Spanish words. We can see how the language shifts from one to the other during the course of the conversation.

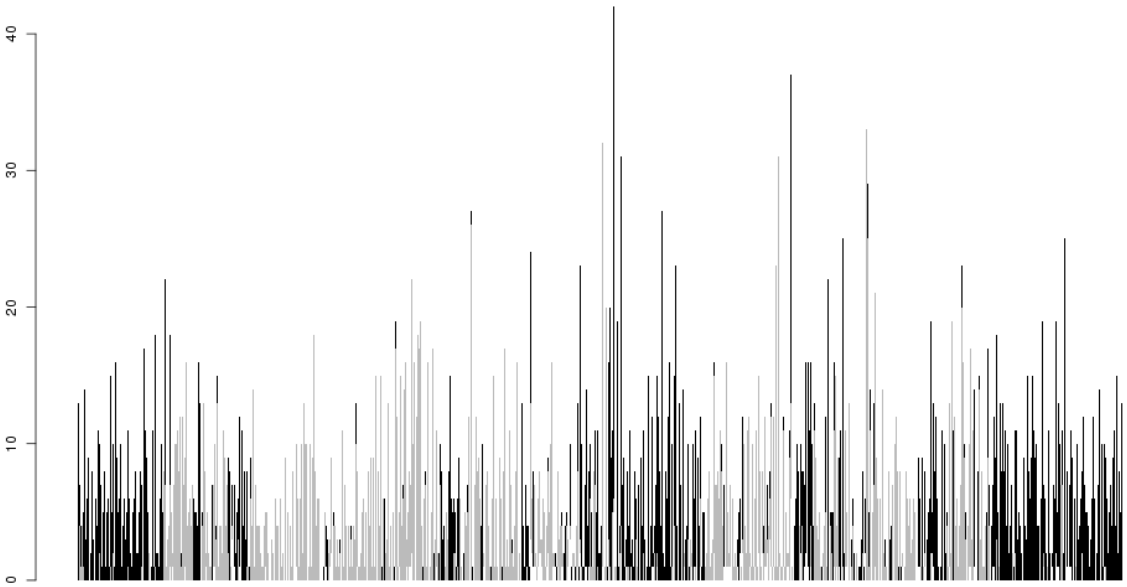

Figure 2. Language choices during a conversation

6. Conclusions

We have described a fast, extensible and accurate tagger for multilingual spoken text in Welsh, Spanish and English. Although the spoken text used is filled with quotatives, incomplete sentences, repetitions and so on, and includes multiple language switches within and between utterances, the Bangor Autoglosser is able to handle this effectively. It may therefore be of interest to those working in such non-standard language environments, for instance in speech-to-text conversion, voice recognition software, training of SMT engines, etc.

We also believe that the Bangor Autoglosser may be of interest to those trying to apply computer techniques to minority languages. These languages are often under-resourced in terms of time and skills, and it is highly desirable to be able to re-use existing materials. If a dictionary is available in a simple spreadsheet or wordlist format, the autoglosser allows it to be plugged in and used almost immediately. The user can then start adding grammar rules, which, as has been noted above, are both powerful and intuitive. Further steps might include iteratively extending the dictionary, improving the rules, and refactoring the lookup to take account of recurring inflections. The key point is that it is relatively easy to start with what you have and move quickly from that to useful output.

It is worth noting the impact of free (GPL) software and resources in allowing us to deliver the software quickly. We were able to leverage powerful tools like the constraint grammar parser, and use free dictionaries to bootstrap the lookup. This sort of re-use dramatically lessens the

timespend on getting something useable, meaning that time can be spent on fine-tuning the output and generating innovative uses of the output. The benefits of this have already been rehearsed [15, 16], and it is perhaps worth noting here that the ESRC Centre's Siarad corpus is the largest collection of Welsh data available under a free license.

Acknowledgements

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

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

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

[3] The original audiofiles are also available.

[4] MacWhinney, B. (2000) *The CHILDES Project: Tools for Analyzing Talk*, 3rd edition. Mahwah, New Jersey

[5] Comrie, B., Haspelmath, M., Bickel, B. (2008) *Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses*, http://eva.mpg.de/lingua/resources/glossing-rules.php

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

[7] http://www.apertium.org

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

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

[10] Karlsson et al. (1995), *Constraint Grammar - A Language-Independent System for Parsing Unrestricted Text,* Mouton de Gruyter

[11] http://beta.visl.sdu.dk/constraint_grammar.html

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

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

[14] http://www.latex-project.org

[15] Tyers, F. & Donnelly, K. (2009) "apertium-cy - a collaboratively-developed free RBMT sytem for English to Welsh to English", *Prague Bulletin of Mathematical Linguistics*, 91

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