Presenting the Bangor Autoglosser and the Bangor Automated Clause-Splitter

D. M. Carter
The University of British Columbia, Okanagan campus, Kelowna, British Columbia, Canada; Centre for Research on Bilingualism, Bangor University, Gwynedd, Wales

M. Broersma
Centre for Language Studies, Radboud University, Nijmegen, The Netherlands; Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

K. Donnelly
Centre for Research on Bilingualism, Bangor University, Gwynedd, Wales

A. Konopka
University of Aberdeen, Aberdeen, Scotland

Abstract

Until recently, corpus studies of natural bilingual speech and, more specifically, codeswitching in bilingual speech have used a manual method of glossing, part-of-speech tagging, and clause-splitting to prepare the data for analysis. In our article, we present innovative tools developed for the first large-scale corpus study of codeswitching triggered by cognates. A study of this size was only possible due to the automation of several steps, such as morpheme-by-morpheme glossing, splitting complex clauses into simple clauses, and the analysis of internal and external codeswitching through the use of database tables, algorithms, and a scripting language.

Correspondence:
D. M. Carter, Faculty of Creative and Critical Studies, Department of Critical Studies, CCS 349, University of British Columbia, Okanagan campus, 1148 Research Road, Kelowna, BC V1V 1V7, Canada.
E-mail:
diana.carter@ubc.ca

1 Introduction

One of the main challenges faced by researchers who study natural bilingual speech is the amount of time needed to collect, transcribe, and prepare the corpus data before any type of linguistic or sociolinguistic analysis can take place. For instance, previous analyses of codeswitching patterns found specifically in the Welsh–English Siarad corpus[1] utilized in our study relied on manual morpheme-by-morpheme glossing, clause-splitting (i.e. splitting complex clauses into simple clauses), and data preparation (Carter *et al.*, 2011; Davies and Deuchar, 2010; Herring *et al.*, 2010). The manual data preparation involved processes such as determining a main language and an embedded language for each bilingual

Digital Scholarship in the Humanities, Vol. 33, No. 1, 2018.
doi:10.1093/llc/fqw065 Advance Access published on 27 February 2017

simple clause (see Section 4 for details on the Matrix Language Frame model; Myers-Scotton, 1993, 2002). The result was a slow process that limited the number of clauses included in each analysis, ranging from a few hundred to a few thousand. One of the goals of our current study was to devise more efficient automated tools and techniques that would allow us to analyze all of the 65,000 clauses in the Siarad corpus in a much shorter amount of time.

In our article, we present the methodology and innovative tools that were essential to our study of codeswitching in the Welsh–English Siarad corpus of spontaneous bilingual speech. We believe that these tools will facilitate several steps in the analysis of monolingual and bilingual corpora. For instance, the Bangor Autoglosser can be utilized to automatically gloss corpora that include languages with small speaker populations, given that tagging systems are often unavailable for languages with fewer than five million speakers. The Bangor Automated Clause-Splitter can be a helpful tool for any researcher who needs to divide complex clauses into smaller clauses for analysis and may be used for other languages in addition to Welsh, such as Spanish, for example.

Previous work has successfully used automated tools to predict codeswitching in corpora (Papalexakis *et al.*, 2014; Solorio and Liu, 2008). In the present study, to the contrary, we analyze actual occurrences of codeswitching. Specifically, our study employed automated methods with the aim of analyzing both internal codeswitches (two languages used within the same clause) and external codeswitches (switches extending over the clause boundary) triggered by cognates (Clyne 1967, 2003). Clyne (2003) defines cognates, or trigger words, as proper nouns, bilingual homophones, and lexical transfers (items from one language that have become part of the lexicon of the speaker's second language), and typically the default assumption is that cognates are nouns. However, in our study, we extended the definition to include all word types that overlap in form and meaning in the bilingual's two languages. Essentially, Clyne's triggering hypothesis proposes that cognates facilitate codeswitching, an effect that is the result of the selection of the cognate from the mental lexicon (Broersma and De Bot, 2006; Broersma, 2009). It is argued that cognates may be strongly connected in the mental lexicon and that their conceptual representations are more closely connected than those of non-cognates. Therefore, the activation of a word that is shared by two languages may lead to a change in activation of both languages at the lexical level. This in turn may 'boost' the least active language to the extent that the next time a lemma is selected it may be one from the boosted language instead of the previously spoken language.

Similarly to the Welsh–English studies mentioned above, previous work on the triggering hypothesis was also performed manually and required over 250 h to tag and analyze small corpora of 2–3,000 words (Broersma and De Bot, 2006; Broersma, 2009). However, through the implementation of the Bangor Autoglosser, the Bangor Automated Clause-Splitter, as well as database tables, algorithms, and a scripting language, we were able to successfully analyze almost 450,000 words in 65,000 clauses. In the following sections, we first describe the collection and transcription of the Siarad corpus and then, crucially, the autoglossing and clause-splitting processes, and final data preparation.

2 Data collection

Here we describe the method followed to collect the large Welsh–English corpus used in our analysis of triggered codeswitching. The Welsh–English Siarad corpus consists of 447,353 words from 151 speakers across sixty-nine conversations. The corpus was collected over a 2 year period in Wales by bilingual Welsh–English researchers who were local members of the community (Deuchar *et al.*, 2014). The participants were recruited through a variety of means, such as newspaper announcements, and the 'friend of a friend' approach (Milroy, 1987). The speakers were told that the aim of the study was to record people having bilingual conversations with another bilingual friend or family member.

Conversations lasted between 19 and 64 min, with a mean length of 35 min, and were recorded using a Marantz hard disk recorder (Carter *et al.*, 2016; Deuchar *et al.*, 2014). Researchers were not

present at the time of the recording, and the participants could discuss any topic of their choice. The speakers were left alone to minimize the observer's paradox, which occurs as the result of having an interviewer or researcher present during the recordings (Labov, 1972). It was important that the participants felt comfortable and unhindered given that informal situations, rather than formal interviews, are more likely to elicit natural bilingual speech.

After the recordings were finished, the participants were asked to complete a self-assessed background questionnaire consisting of twenty questions. The questionnaire elicited a wide range of information, such as the participants' age, gender, occupation, language of education, age of exposure to each language, language of social network, language proficiency, and attitudes toward codeswitching. The anonymity of the participants was protected, which means that other researchers interested in studying social variables could also access the questionnaire data without harming anonymity.

3 Transcription

All of the recordings were transcribed by Welsh–English bilinguals using the CHAT[2] transcription system in the Computerized Language Analysis (CLAN) program (MacWhinney, 2000). Within the CHAT system, transcribers used language tags to differentiate between Welsh, English, and cognate words. Words in the most frequent language in each conversation were left untagged, while all other words were tagged according to their corresponding language using three-letter abbreviations of ISO-639-3 (i.e. @s:eng for English). Transcribers employed a 'dictionary method' to allocate language tags and ensure consistency in the transcripts. Words that occurred in the dictionaries of both English and Welsh were considered cognates and tagged as @s:cym&eng, with the language tags in alphabetical order of the abbreviation. These words included proper nouns, nouns, and verbs as well as other word classes. In the Siarad corpus, Welsh is the most frequent language in all conversations, ranging from 51 to 93% of the words, with an average of 84%.

One of the key advantages of language tagging is that it greatly facilitates the identification of cognates and codeswitching in bilingual corpora. Example (1) below illustrates a transcription tier with language tags and a translation tier.

(1) ond dw i ddim actually@s:eng isio mynd i wrando ar y stuff@s:cym&eng.

'but I don't actually want to go and listen to the stuff'.

In addition to the tiers of transcribed speech and the translation tier, another tier was included in the corpus that was essential to our study: a morpheme-by-morpheme gloss. Originally, all of these tiers were entered manually by a team of bilingual researchers. The newly developed Bangor Autoglosser provides researchers with a more efficient automatic method of glossing.

4 The Bangor Autoglosser

The transcription tiers were glossed with an innovative automated tool called the Bangor Autoglosser that followed the Leipzig glossing conventions (Carter *et al.*, 2016; Donnelly and Deuchar, 2011). Given that the existing tagging system used in CLAN only handles larger languages of over five million speakers (MacWhinney, 2009), it was necessary to create a tool from scratch that could automatically gloss large multilingual Welsh–English texts. The implementation of the Bangor Autoglosser involved a combination of digital dictionaries and the application of Constraint Grammar (Karlsson, 1990; Karlsson, *et al.*, 1995). Constraint grammar assigns grammatical tags to text based on context-dependent rules written by a linguist. Each rule selects, removes, adds, or replaces the tag on any given word by taking into account surrounding words and their tags. This was the first application of Constraint Grammar to mixed-language texts. Essentially the procedure involves the separation of text into words, the lookup of each word in a dictionary that gives possible lemmas and part-of-speech (POS) for that word, and the

selection of the correct lemma and POS for the word in its current context. This is illustrated in Table 1.

The autoglossing process is as follows. First, the Bangor Autoglosser imports each utterance from a transcript into an utterance table, as seen in Fig. 1. The table facilitates the process of editing or adding items either directly to the table or to an exported spreadsheet version of the same table.

Second, the words are imported from the database into a 'words table' and tokenized (Fig. 2). Any mutations in Welsh are removed (e.g. 'gath at-cath'),[3] and any elisions or regular verb endings in English are also removed (e.g. 'gonna, I'll').

The language tags are used to decide which dictionary is consulted for the gloss. The correct dictionary accumulates all matching entries for each word and writes them in another file that is in the format required by the Constraint Grammar parser. Next, the parser applies the Constraint Grammar rules to the file.[4] For example, in the case of the English word 'dance', you would have one reading: dance, sv, infin, meaning that 'dance' can be a singular noun, or a verb (with the combined tag 'sv'), and if it is a verb, it is usually an infinitive. The Constraint Grammar rules then use context to convert the 'sv' tag into 'n.sg', or 'v.pres' (e.g. they dance). The Constraint Grammar rules for Welsh are applied by the parser in the same way it would apply rules for any language. In other words, there is no need for a special algorithm to be written specifically for Welsh. This is one of the features that allows Constraint Grammar to be used to tag multilingual text. One main difference between English and Welsh, however, is the higher number of homonyms present in English. As a result, in Welsh, each individual meaning tends to have a separate reading.

Table 1 Welsh dictionary layout

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			

The results of the application of the grammar rules are stored in a words table as a combination of a gloss and POS-tag (Fig. 3).

Finally, the entire CHAT file is written out of the database with a new autogloss tier that is generated from the glossed words. This output is illustrated in Example (2).

(2) **ond dw i ddim actually@s:eng**
but.CONJ be.v. 1S. PRES I .PRON.1S not.AD+SM actual. ADJ+ADV
isio mynd i wrando ar y
want.N.M.SG go.V.INF to.PREP listen.V.INF+SM on.PREP the .DET .DEF
stuff@s:cym&eng.
stuff.N.SG

'but I don't actually want to go and listen to the stuff'

Using this innovative method, glossed text was produced at a rate of 1,000 words per minute and the 40 h Siarad corpus was glossed in approximately 8.5 h. We performed manual checks of the complete outputs from five random transcription files which showed that the precision of the glossing was between 97 and 99%, depending on the language.

In addition to its efficiency, another advantage of the automated glossing is that it is now possible to easily access any word or attribute of texts that are available in the database. Through the use of a scripting language such as Hypertext Preprocessor (PHP) (Lerdorf, 2007) or Python (Bird *et al.*, 2009),

utterance_id	filename	speaker	surface
203	stammers4	ALN	ond # dw i (dd)im actu(ally)@s:eng [?] isio mynd i wrando ar y stuff@s:cym&eng .

eng	com	comment	durbegin	durend	duration	precode
but I don't actually want to go and listen to the stuff.	NULL	NULL	447979	451009	3030	

Fig. 1 Example (1) in the utterance table

it is possible for researchers to manipulate the database at this point and begin to analyze the corpus data. We used a scripting language in most stages of our study, including the development of the automated clause-splitter we describe next.

location	surface	langid
1	ond	cym
2	dw	cym
3	i	cym
4	ddim	cym
5	actually	eng
6	isio	cym
7	mynd	cym
8	i	cym
9	wrando	cym
10	ar	cym
11	y	cym
12	stuff	cym&eng
13	.	999

Fig. 2 Example (1) in the words list table

location	surface	auto	langid
1	ond	but.CONJ	cym
2	dw	be.V.1S.PRES	cym
3	i	I.PRON.1S	cym
4	ddim	not.ADV+SM	cym
5	actually	actual.ADJ+ADV	eng
6	isio	want.N.M.SG	cym
7	mynd	go.V.INFIN	cym
8	i	to.PREP	cym
9	wrando	listen.V.INFIN+SM	cym
10	ar	on.PREP	cym
11	y	the.DET.DEF	cym
12	stuff	stuff.N.SG	cym&eng
13	.	*NULL*	999

Fig. 3 Disambiguated words from Example (1) stored in the words table after the application of the Constraint Grammar parser

5 The Bangor Automated Clause-Splitter

Given that the Siarad corpus was not originally transcribed in simple clauses and no Welsh parser existed, we needed to devise a way of automatically splitting complex clauses into simple clauses for our codeswitching analysis. This was an essential step so that we could apply the Matrix Language Frame model (Myers-Scotton, 1993, 2002) and determine a base or matrix language for each clause. According to the model, each codeswitched clause contains a matrix language that provides the morphosyntactic frame for the clause, and an embedded language that contains inserted material, mostly consisting of content morphemes. The matrix language can usually be determined by the language of the finite verb in each clause, which was found to be true for the entire Siarad corpus. The large majority of the clauses in the Siarad corpus has Welsh as the main verb with English providing the inserted material.

As mentioned in the introduction, previous studies that involved manual clause-splitting took several weeks and many researchers to divide only a few thousand clauses (Carter *et al.*, 2011; Davies and Deuchar, 2010; Herring *et al.*, 2010). In the present study, we were able to analyze 65,000 clauses as a result of the creation of the Automated Clause-Splitter.

During the initial development phase, the first version of the clause-splitter was tested on the first 300 utterances of a single file and was checked in detail, revealing an accuracy rate of 93%. In total, twenty-one (7%) of the utterances were split incorrectly. Out of the twenty-one, eight (3%) were due to an incorrectly applied rule in the Constraint Grammar, and another three (1%) because of an error in the dictionary. The final ten (3%) were due to the splitter itself. To increase the accuracy rate of the cause-splitter, we made corrections to the Constraint Grammar application as well as the dictionaries. Additionally, we revised some of the assumptions that the splitter uses. For example, one assumption is that inflected verbs have the clause marker moved to the preceding word when the preceding word is a conjunction, a subordinator, or an adverb. The initial list of these words was increased

because it was not exhaustive, thus causing inaccurate clause-splits.

The clause-splitting procedure as applied to the Siarad corpus can be summarized as follows. First, for the purpose of the present study, we removed all conversations containing more than two speakers leaving us with fifty-two conversations and 105 speakers; this was a preemptive step that would later facilitate the statistical analysis. Note that the clause-splitter could handle conversations with more than two speakers without any problem. Second, we omitted all interactional markers, which are utterances such as 'uhhuh, mmhm' that do not fulfill any syntactic role in everyday speech. Next, we added role indicators in the 'words table' (Fig. 3) to every finite verb, which were then moved where necessary. In the following Example (3), the finite verbs are underlined, the clause-splits are marked with a forward slash/, and the word onto which the clause-split marker was moved is in bold. The example illustrates how the marker is moved from 'o'n' to 'pan' (when) because 'pan' is a conjunction, following the assumption made by the clause-splitter that the marker be moved to the word preceding an inflected verb if that word is a conjunction.

(3) dw i yn cofio/o'n i yn gweithio ar y nos/**pan** o'n i yn gweithio yn Beaumaris

'I remember/I was working nights/when I was working in Beaumaris'

As mentioned previously, spot checks of a random sample of the splits revealed that this method was over 97% accurate, which was deemed an acceptable rate given the speed of the process and the large number of clauses that were produced for analysis.

Next, we determined the matrix language of each clause by detecting the language of the finite verb within that clause. This step was done automatically based on the language tagging in the transcripts. Once the matrix language was assigned, we assessed whether there were any internal or external codeswitches. If two languages co-occurred within the same basic clause, it was considered an internal switch, but if the subsequent clause had a different matrix language from the previous clause, then it was an external switch. Finally, we generated additional data that characterized the clauses and the conversations. For example, we wanted to know the length of each clause in words, whether the clause contained cognates, and if yes, how many, the type of each clause, the length in letters of each cognate, and the language of the clause (Welsh, English, or bilingual). Other key information included the total number of words, clauses, cognates, and codeswitches in each conversation and per speaker.

Once the enriched data had been generated, they were exported to a comma-separated value file and could be analyzed using statistical software such as R (R Development Core Team, 2009).

7 Conclusions

In contrast to previous smaller-scale studies of codeswitching patterns in bilingual corpora, and specifically in the Welsh–English Siarad corpus, our research team was able to analyze the entire corpus of 65,000 clauses due to the development of innovative tools, namely, the Bangor Autoglosser, which applied Constraint Grammar to bilingual text for the first time, and the Bangor Automated Clause-Splitter that divided thousands of complex clauses into basic clauses at a rapid rate. All of the data were contained in database tables and were manipulated and analyzed through the use of a general-purpose scripting language, rather than a specific dedicated interface, such as the query application found in the CLAN (MacWhinney, 2009) program. The scripts were written and utilized successfully to prepare a large quantity of clauses for the analysis of several variables pertaining to our study's focus on triggered codeswitching. Although a discussion of the statistical analysis and results are outside of the scope of this current article, it should be noted that without the use of the automated tools and scripts, it would not have been possible to process the large Welsh–English Siarad corpus with such speed, efficiency, and accuracy.

Funding

This work was supported by a Small Research Grant from the British Academy awarded to the first and

second authors (grant number 101421). We also gratefully acknowledge the support of the Max Planck Institute for Psycholinguistics, the Centre for Research on Bilingualism in Wales, and the University of Calgary.

References

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. California: O'Reilly Media, Inc.

Broersma, M. (2009). Triggered codeswitching between cognate languages. *Bilingualism: Language and Cognition*, **12**: 447–62.

Broersma, M. and De Bot, K. (2006). Triggered codeswitching: a corpus-based evaluation of the original triggering hypothesis and a new alternative. *Bilingualism: Language and Cognition*, **9**: 1–13.

Carter, D., Broersma, M., and Donnelly, K. (2016). Applying computing innovations to bilingual corpus analysis. In Valenzuela, E. and de la Fuente, A. A. (eds), *Language Acquisition Beyond Parameters: Studies in honour of Juana M. Liceras*. Amsterdam: John Benjamins.

Carter, D., Deuchar, M., Davies, P., and Parafita Couto, M. C. (2011). A systematic comparison of factors affecting the choice of matrix language in three bilingual communities. *Journal of Language Contact*, **4**: 153–83.

Clyne, M. (1967). *Transference and Triggering: Observations on the Language Assimilation of Postwar German-Speaking Migrants in Australia*. The Hague: Martinus Nijhoff.

Clyne, M. (2003). *Dynamics of Language Contact: English and Immigrant Languages*. Cambridge: Cambridge University Press.

Davies, P. and Deuchar, M. (2010). Using the matrix language frame model to measure the extent of word order convergence in Welsh-English bilingual speech. In Breitbarth, A., Lucas, C., Watts, S., and Willis, D. (eds), *Continuity and Change in Grammar*. Philadelphia, PA: John Benjamins, pp.77–96.

Deuchar, M., Davies, P., and Donnelly, K. (2016). Building and using the Siarad corpus: bilingual conversations in Welsh and English. Manuscript.

Deuchar, M., Davies, P., Herring, J., Parafita Couto, M.C., and Carter, D. (2014). Bilingual language use. In Thomas, E. and Mennen, I. (eds), *Advances in the Study of Bilingualism*. Bristol: Multilingual Matters, pp.93–110.

Donnelly, K. and Deuchar, M. (2011). Using Constraint Grammar in the Bangor Autoglosser to Disambiguate Multilingual Spoken Text. In *Proceedings of the NODALIDA 2011 Workshop Constraint Grammar Applications*, Riga, Latvia: NEALT Proceedings Series, Tartu.

Herring, J., Deuchar, M., Parafita Couto, M. C., and Moro Quintanilla, M. (2010). 'I saw the madre': evaluating predictions about codeswitched determiner-noun sequences using Spanish-English and Welsh-English data. *International Journal of Bilingual Education and Bilingualism*, **13**: 553–73.

Karlsson, F. (1990). Constraint Grammar as a Framework for Parsing Unrestricted Text. In *Proceedings of the 13th International Conference of Computational Linguistics*, vol. 3:168–73, Stroudsurg, PA. doi:10.3115/991146.991176.

Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila A. (1995). *Constraint Grammar: A Language-Independent System for Parsing Running Text. Natural Language Processing, 4*. Berlin and New York: Mouton de Gruyter.

Labov, W. (1972). *Some principles of linguistic methodology*. Language in Society, **1**: 97–120.

Lerdorf, R. (2007). *PHP on Hormones—history of PHP. MySQL Conference*. Santa Clara, California. http://web.archive.org/web/20130729204354id_/http://itc.conversationsnetwork.o rg/shows/detail3298.html.

MacWhinney, B. (2009). Enriching CHILDES for morphosyntactic analysis. Department of Psychology. Paper 175. http://repository.cmu.edu/psychology/17.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*, 3rd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Milroy, L. (1987). *Language and Social Networks*. Oxford: Blackwell.

Myers-Scotton, C. (2002). *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford and New York, NY: Oxford University Press.

Myers-Scotton, C. (1993). Common and uncommon ground: social and structural factors in codeswitching. *Language in Society*, **22**: 475–503.

Papalexakis, E., Nguyen, D., and Seza Doğruöz, A. (2014). Predicting Code-Switching in Multilingual Communication for Immigrant Communities. In *Proceedings of the First Workshop on Computational*

Approaches to Code Switching. Doha, Qatar, October 2014, pp. 42–50.

R Development Core Team. (2009). *R: a language and environment for statistical computing*. Vienna, Austria: R. Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org.

Solorio, T. and Liu, Y. (2008). Learning to Predict Code-Switching Points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, October 2008, pp. 973–81.

Notes

1 The Siarad corpus of Welsh–English data is available under open license at http://bangortalk.org.uk.

2 At the time the corpus was being collected, the CHAT system was one of the most suitable choices (Deuchar *et al.*, 2016). Currently, there are other options available for multilingual data, such as ELAN (https://tla.mpi.nl/tools/tla-tools/elan/), although as MacWhinney explains in the CHILDES manual (http://childes.psy.cmu.edu/manuals/chat.pdf), the CHAT data can be translated to XML which can then be used in ELAN, among other programs.

3 In Welsh, as in the other Celtic languages, some word-initial consonants change ('mutate') to reflect morphological and syntactic relationships between the words of the utterance. For example: siop llyfrau da (a shop [with] good books), but siop lyfrau dda (a good book-shop), where the change d -> dd signifies that the adjective da (good) relates to siop (shop) and not to llyfrau (books). Llyfrau is itself mutated ll -> l to show that it qualifies siop. Another example is seen here where mae o'n marw means 'he is dying', but mae o'n farw means 'he is dead'. The change m -> f signifies that marw (die, dead) is the adjective and not the verb. These mutations have to be removed to get to the underlying lemma.

4 The scripts for the Constraint Grammar rules for Welsh are available at https://github.com/donnekgit/autoglosser.