Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4121–4129
Marseille, 11–16 May 2020

CanVEC – the Canberra Vietnamese-English Code-switching Natural Speech Corpus

Li Nguyen† and Christopher Bryant‡
Theoretical and Applied Linguistics†
Department of Computer Science and Technology‡
University of Cambridge, Cambridge, United Kingdom
{nhbn2, cjb255}@cam.ac.uk

Abstract

This paper introduces the Canberra Vietnamese-English Code-switching corpus (CanVEC), an original corpus of natural mixed speech that we semi-automatically annotated with language information, part of speech (POS) tags and Vietnamese translations. The corpus, which was built to inform a sociolinguistic study on language variation and code-switching, consists of 10 hours of recorded speech (87k tokens) between 45 Vietnamese-English bilinguals living in Canberra, Australia. We describe how we collected and annotated the corpus by pipelining several monolingual toolkits to considerably speed up the annotation process. We also describe how we evaluated the automatic annotations to ensure corpus reliability. We make the corpus available for research purposes.

Keywords: Codeswitching, Vietnamese, Natural Speech, Automatic Annotation

1. Introduction

Code-switching is the linguistic phenomenon where a speaker uses two or more languages in a single conversation or utterance; for example:

(1) *mỗi* group *phải có* a different focus
each must have
'Each group must have a different focus'

Although this language-mixing phenomenon has been studied extensively in linguistics (Poplack, 1980; Myers-Scotton, 1997; Muysken, 2000), it has received much less attention in the Natural Language Processing (NLP) community, where it remains a serious challenge (Solorio and Liu, 2008; Nguyen and Doğruöz, 2013; Li and Fung, 2014; Molina et al., 2016; Çetinoğlu et al., 2016). This is largely because of a lack of availability of large, high-quality code-switching corpora, and the reality that monolingual techniques often struggle with input from multiple languages (Hamed et al., 2017; Soto and Hirschberg, 2018).

A related issue is that, of the code-switching corpora that are available, the majority are textual and consist only of web documents (Hamed et al., 2017) or social media posts from platforms such as Twitter (Maharjan et al., 2015; Jurgens et al., 2014; Mave et al., 2018) and Facebook (Bali et al., 2014; Barman et al., 2014). Given that code-switching occurs most frequently and naturally in informal speech however (Labov, 2004; Cacoullos and Travis, 2018; Deuchar et al., 2018; Nguyen, 2018), the faithfulness of such code-switching utterances is somewhat questionable. In particular, although text is often considered a more canonical form of language use than speech, Adouane et al. (2018) points out that a heavy reliance on written input as training data is highly problematic, as it implicitly assumes all forms of languages are uniform and monolingual. This is especially not the case for code-switching however, where language use encompasses many forms of variation and speakers may use different phonological, lexical and syntactic inputs from multiple languages (Cacoullos and Travis, 2018; Deuchar et al., 2018).

Although the amount of time and effort required to build a speech corpus is significantly greater than that of a text corpus (cf. Caines et al. (2016)), we believe sufficient advances in monolingual processing have been made such that the time is right to start developing newer, more faithful, high-quality code-switching resources that can help facilitate research into more sophisticated multilingual processing. We hence introduce the Canberra Vietnamese-English Code-switching corpus (CanVEC), a corpus of 87k tokens that was designed to: i) capture the vernacular of a migrant community in its most natural form, and ii) be the first available corpus of Vietnamese-English code-switching. Information from the corpus will enable us to start addressing some of the most important questions in code-switching research, such as 'How do speakers code-switch?', and 'Are code-switching patterns universal?', which might ultimately benefit research into other NLP tasks such as multilingual speech recognition, parsing, POS tagging and machine translation in low-resourced languages.

To the best of our knowledge, CanVEC is the first natural language Vietnamese-English code-switching speech corpus that is freely accessible.[1] It is also the first that showcases contemporary migrant repertoire in English-speaking communities, where there is ongoing tension between speakers' heritage language and the majority language.[2]

2. Background

Given the costs involved in building a speech corpus, it is unsurprising that most spoken code-switching corpora consist of either scripted speech (Chan et al., 2005; Shen et al., 2011; Modipa et al., 2013; Yilmaz et al., 2017) or are

[1] `https://github.com/Bak3rLi/CanVEC`

[2] A heritage language is the language of a speaker's indigenous origin, while a majority language is the dominant language in the location where they live. For example, a Chinese migrant family living in Spain will have Chinese as a heritage language, and Spanish as a majority language.

extremely limited in size (Solorio and Liu, 2008; Dey and Fung, 2014). For example Solorio and Liu (2008) released a bilingual code-switching corpus of English and Spanish, but the corpus itself comprises just 39 minutes of conversation between 3 colleagues at a Southwestern University in the United States. In total, the transcribed corpus contains just 922 sentences with 239 language switches, 129 of which are intra-sentential. This remained the largest annotated spoken corpus of code-switching available to NLP researchers for several years.

More recently, Lyu et al. (2015) released the South East Asia Mandarin-English corpus (SEAME), a large-scale corpus of mixed speech, containing data from 157 bilingual Mandarin-English speakers. The corpus was reported to be 63 hours long, about 11.5 hours of which were conversational. However, while the SEAME sample size is large, data is constrained to a particularly well-educated group of young speakers aged 18-34. Since most SEAME speakers were also college students accustomed to mainly speaking English on campus, only a limited number of intra-sentential code-switching utterances were collected. This is unsurprising however, as research in sociolinguistics has shown that code-switching occurs most frequently in informal, relaxed settings among bilinguals (Poplack, 1980; Poplack, 1993; Cacoullos and Travis, 2018). While unscripted interviews and conversations in a recording studio might facilitate spontaneous speech, the setting for this corpus was still contrived in that speakers were given specific topics to discuss. Research in language contact has so far concurred that utterances produced under these conditions are likely to be heavily influenced by the interviewer's language use (Cacoullos and Travis, 2018), or the psychological effects of the unnatural situations (Hofweber et al., 2016).

It is also important to recognise that, while scattered efforts have been made in NLP to investigate somewhat well-resourced language pairs such as English-Spanish (Solorio and Liu, 2008) or English-Mandarin (Lyu et al., 2015), work examining code-switching involving low-resourced, or less-described languages is still largely neglected. This means very few resources are available to automatically process this kind of data. Although one toolkit was developed for a large Welsh-English bilingual corpus released last year (Deuchar et al., 2018), the tool mainly helped with auto-glossing and translation. As Deuchar et al. (2018) also note, these resources were also largely possible thanks to a substantial government grant, which enabled them to engage a full team of people over the course of several years, up until when the corpus was finally completed.

Before we introduce our own bilingual Vietnamese-English corpus, it is also important to note that previous work has been done to create a similar corpus for the same language pair. In particular, Tuc (2003) collected a corpus of 60 hours of speech, comprising both sociolinguistic interviews and speakers' self-recorded speech, over twenty years ago in Victoria, Australia. Unfortunately however, the recordings only existed in the form of physical cassettes, and can no longer be found.[3]

In what follows, we present our original Vietnamese-English bilingual corpus, CanVEC, and introduce the procedure for semi-automatically annotating the corpus using existing monolingual toolkits. The motivation for taking advantage of NLP monolingual resources is based on evidence in code-switching research that the grammars of both languages tend to be respected when switches are made (Poplack, 1980).

3. Building the corpus

3.1. Recording procedure

Data was collected over the course of three months, from June to September 2017 in Canberra, Australia. Although Canberra is still largely English-dominated (72.7% of locals speak only English at home), the latest census shows that Vietnamese is the second most popular heritage language (N=4216) after Mandarin Chinese (ABS, 2017).

Our principle in building CanVEC was to extract speakers' vernacular, where 'minimum attention is paid to speech' (Labov, 2004). The vernacular reflects the most natural, systematic form of the language acquired by the speaker 'before any subsequent efforts at (hyper-) correction or style shifting are made' (Poplack, 1993). Recruited speakers were thus free to choose their own interlocutors in an environment that they were most comfortable with. As Deuchar et al. (2018) point out, although this freedom limits researcher control over the environment, it optimises informality, thereby providing a better environment for language mixing.

Participants were asked to self-record a conversation or a collection of conversations totalling at least 30 minutes on their personal mobile phones, with no single conversation lasting less than 10 minutes. All participants were aware that the recording was taking place and had given written consent to participate. Participants were not required to speak both languages in the recording (nor were they asked to do so); instead, they were encouraged to converse as they normally would. The use of participants' personal phones was methodologically strategic, as it was a familiar item in everyday life, thereby substantially lessening the intrusion of a recording device such as a microphone. Two participants in their 60s did not own smart phones, so were instead given a Zoom H5–5000–2 recorder. Most recordings were of high quality, and only one sound file was considered unintelligible enough to be discarded from the corpus.

Speakers were instructed to listen to their recordings before submission and decide whether they wanted to delete any portion of the conversation that perhaps contained private or sensitive information. Once a recording was returned, it was understood that speakers consented to making all parts of their conversations available for research purposes. When transcribing the corpus however, we soon realised that several conversation segments still discussed private topics such as speakers' gambling histories or community gossip, and so, with the best interests of the speakers in mind, we eliminated these parts from the corpus. In this way, we were also able to protect the minority community

[3] This information was obtained via personal communication with the researcher, Ho-Dac Tuc.

			Gender		Education Level	
	Speakers	Age Range	Male	Female	University	School
1st Generation	28	28 - 67	15	13	19	9
2nd Generation	17	12 - 35	6	11	9	8
Total	45	12 - 67	21	24	28	17

Table 1: CanVEC demographic information

from reinforcing negative stereotypes (cf. Cacoullos and Travis (2018)).

The completed procedure generated a corpus of 10 hours and 2 minutes, comprising 23 conversations between 45 Vietnamese-English bilingual speakers. 28 of these are first generation native Vietnamese speakers who acquired English later in life, while the remaining 17 are second generation speakers who acquired both English and Vietnamese bilingually from birth or at a very young age. Various additional demographic information about the speakers in CanVEC is shown in Table 1. In particular, the corpus is also fairly well-balanced in terms of gender and education level.

3.2. Transcription via ELAN

CanVEC was transcribed using ELAN (Sloetjes and Wittenburg, 2008), a tool that makes it easy to segment utterances and organise them into linked tiers. A crucial step in transcription is segmentation, which involves splitting the stretches of utterance into consistent boundaries such as turns, clauses, or intonation units. As speech does not contain any explicit boundary markers, e.g. punctuation, this requires careful consideration; there is a trade-off between the granularity and the versatility of the transcription. For example word-level segmentation may be better for speech recognition systems, but POS-tagging and parsing work best at the sentence level. For CanVEC, this is even more challenging since spoken Vietnamese deviates significantly from the standard written form, and is naturally riddled with fragments, argument drops, disfluencies and false starts.

The first pass of segmentation roughly divided speakers into intonation units, while the second pass further divided these into clauses. Clauses were chosen as the main unit of analysis because the corpus was designed to test a specific theoretical model of code-switching in sociolinguistics: the Matrix Language Framework (Myers-Scotton, 1997). Clauses also seemed a good compromise in terms of transcription granularity and versatility.

The first author transcribed all of the recordings, but ten percent of the data (i.e. the first 10 minutes of 6 random conversations) was additionally annotated by a second transcriber to evaluate transcription reliability. Participants were assigned pseudonyms before the second transcriber was given access to this subset of the data. The assistant was a final-year undergraduate linguistics student with competence in both English (as a first language) and Vietnamese (as a second language). It was important that the assistant's primary competence was English rather than Vietnamese, because this meant the second transcriber was more likely to catch words in their native language that the first transcriber (whose native language is Vietnamese) might have missed or misheard (Cacoullos and Travis, 2018).

Following Deuchar et al. (2018), we used Turnitin[4], a commercial plagiarism detection service, to measure the overlap between the first author and the second transcriber. The software compared the two versions of transcriptions and calculated the overall similarity (%) between the two texts. Documents were then returned with highlighted annotations, showing where similarities and differences occurred. Turnitin ultimately reported an exceptionally high matching rate of over 95% overall, indicating a strong level of inter-annotator agreement between the two transcribers.

3.3. Semi-automatic corpus annotation

Having segmented the data into 14,047 separate clauses, we next automatically annotated each clause with additional information. The whole annotation process is illustrated in Table 2 and consists of the following steps:

1. Remove inconsistently transcribed punctuation and other artefacts from the clause.
2. Split the clause into text units on whitespace; although whitespace marks word boundaries in English, it marks syllable boundaries in Vietnamese.
3. Test each word/syllable for language membership using a Vietnamese syllable list and an English word list.
4. Send the largest contiguous sequence of Vietnamese or English text to the relevant tokeniser and POS tagger.
5. Redefine the word level language tag in terms of tokens rather than word/syllable units.
6. Assign a clause level language to the tokenised clause (explained below).
7. Translate Vietnamese to English in the monolingual Vietnamese and Mixed clauses.

When testing for language membership, we compared each whitespace separated unit against large lists of valid Vietnamese syllables[5] and English words[6]. Units that appeared in either both or neither list were held aside to be resolved manually; there were 264 of these in the data. A large number of these ambiguous units were proper nouns, interjections and fillers, such as *uhm* and *okay*, which are not exclusive to any language, and so were therefore marked as language-neutral (cf. Riehl (2005)). Similarly, units that

[4] `http://www.turnitin.com`

[5] `http://www.hieuthi.com/blog/2017/03/21/all-vietnamese-syllables.html`

[6] `https://sourceforge.net/projects/wordlist/files/speller/2018.04.16/`

<table>
<tr><th>Step</th><th>Description</th><th colspan="11">Example</th></tr>
<tr><td>1</td><td>Data cleaning</td><td colspan="11">I don’t không có really hiểu cái point of it</td></tr>
<tr><td>2</td><td>Split on whitespace</td><td>I</td><td colspan="2">don’t</td><td><i>không</i></td><td><i>có</i></td><td>really</td><td><i>hiểu</i></td><td><i>cái</i></td><td>point</td><td>of</td><td>it</td></tr>
<tr><td rowspan="2">3</td><td rowspan="2">Test language membership using word/syllable list</td><td>I</td><td colspan="2">don’t</td><td><i>không</i></td><td><i>có</i></td><td>really</td><td><i>hiểu</i></td><td><i>cái</i></td><td>point</td><td>of</td><td>it</td></tr>
<tr><td>E</td><td colspan="2">E</td><td>V</td><td>V</td><td>E</td><td>V</td><td>V</td><td>E</td><td>E</td><td>E</td></tr>
<tr><td rowspan="2">4</td><td rowspan="2">Tokenise and POS tag same-language sequences</td><td>I</td><td>do</td><td>n’t</td><td><i>không</i></td><td><i>có</i></td><td>really</td><td><i>hiểu</i></td><td><i>cái</i></td><td>point</td><td>of</td><td>it</td></tr>
<tr><td>PRON</td><td>VERB</td><td>ADV</td><td>ADV</td><td>VERB</td><td>ADV</td><td>VERB</td><td>CLS</td><td>NOUN</td><td>PREP</td><td>PRON</td></tr>
<tr><td rowspan="2">5</td><td rowspan="2">Redefine language tags in terms of tokens</td><td>I</td><td>do</td><td>n’t</td><td><i>không</i></td><td><i>có</i></td><td>really</td><td><i>hiểu</i></td><td><i>cái</i></td><td>point</td><td>of</td><td>it</td></tr>
<tr><td>E</td><td>E</td><td>E</td><td>V</td><td>V</td><td>E</td><td>V</td><td>V</td><td>E</td><td>E</td><td>E</td></tr>
<tr><td>6</td><td>Assign clause level language</td><td colspan="11">Mixed</td></tr>
<tr><td>7</td><td>Translate Vietnamese and Mixed clauses</td><td colspan="11">I don’t really understand the point of it</td></tr>
</table>

Table 2: Table showing each step of automatic annotation using an example clause.

Underthesea	Universal
A	ADJ
ADP	PREP
C	CONJ
CCONJ	CONJ
E	PREP
I	INTJ
L	DET
M	NUM
N	NOUN
Nc	CLS
Nu	NOUN
Ny	PROPN
P	PRON
R	ADV
T	VERB
V	VERB
X	X
Z	Z

Table 3: The mapping function for Underthesea POS tags to Universal POS tags. The CLS tag (classifiers) is not a universal tag, but was considered important to preserve for Vietnamese.

were unintelligible were marked as <V> if they were considered more likely to be Vietnamese, <E> if they were considered more likely to be English, and <X> if it was impossible to decide. The remaining units, such as ‘me’, which means ‘mother’ in Vietnamese, but is an object pronoun in English, were otherwise fairly rare, and so were defined according to whichever language we considered more likely.

Having assigned language membership, we next sent the largest contiguous sequence of same-language units to a Vietnamese or English tokeniser and POS tagger as appropriate. Note that language-neutral tokens were ignored when defining the same-language sequence boundaries. We used Underthesea[7] v1.1.11 to tokenize and POS tag Vietnamese sequences, and spaCy[8] v1.9.0 to tokenise and POS tag English sequences. These resources were chosen mainly for their versatility and high performance; spaCy reports a POS tagging accuracy of 96.6% for English, while Underthesea reports a POS tagging accuracy of 92.3% for Vietnamese.

Since each POS tagger uses a different tagset however, a tag map was defined to convert all POS tags to the Universal Tagset (Petrov et al., 2012; de Marneffe et al., 2014). Although spaCy includes a function to do this automatically, Underthesea does not, so we instead defined our own mapping function (Table 3). This mapping ensured POS tag consistency across the whole corpus.

After tokenisation, we were also able to update the language tags in terms of tokens rather than units. This could not be done sooner because we previously did not know which monolingual tokeniser a clause or sequence should be processed by. The clause level language tags were then defined based on the token level language tags as follows:

1. Language-neutral tokens were excluded from the analysis.

2. If all remaining tokens were Vietnamese, the clause is monolingual Vietnamese.

3. If all remaining tokens were English, the clause is monolingual English.

4. If there is a mix of tokens from both languages, the clause is mixed.

5. Otherwise the clause consists entirely of language neutral tokens.

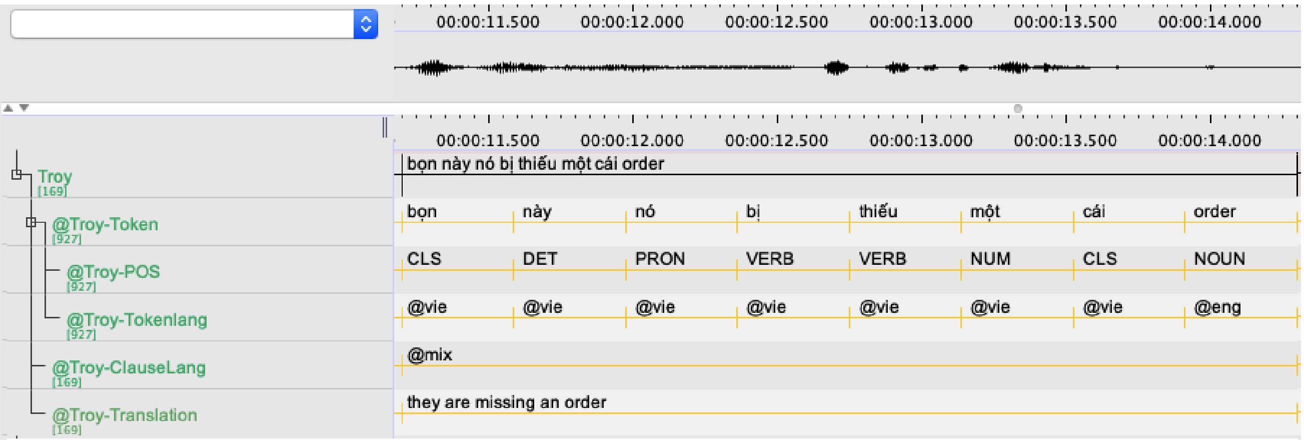
Having defined clause level language tags, we next automatically translated all the Vietnamese and mixed clauses using the Google Translate API[9]. Although we could have segmented and translated only the Vietnamese subsequences in the mixed clauses (as we did for tokenising and POS tagging), this time, we instead sent the entire bilingual clause to the translation API. This is because machine translation systems are usually designed to handle unknown words and also tend to perform better on longer sequences of input; we consequently expected better results at the clause level rather than the sub-clause level.

All the output was then imported back into ELAN and distributed across various tiers. Figure 1 hence shows how a

[7] `https://github.com/undertheseanlp/underthesea`

[8] `https://spacy.io/`

[9] `https://cloud.google.com/translate/`



Troy [169]	bọn này nó bị thiếu một cái order						
@Troy-Token [927]	bọn	này	nó	bị	thiếu	một	cái	order
@Troy-POS [927]	CLS	DET	PRON	VERB	VERB	NUM	CLS	NOUN
@Troy-Tokenlang [927]	@vie	@vie	@vie	@vie	@vie	@vie	@vie	@eng
@Troy-ClauseLang [169]	@mix							
@Troy-Translation [169]	they are missing an order							

Figure 1: Sample speech-tier alignment in ELAN. Tokens are distributed evenly across the clause.

Type	Clauses	Tokens
Vietnamese	7,508	45,640
English	2,582	15,523
Mixed	2,721	22,094
Non	1,236	3,462
Total	14,047	86,719

Table 4: Basic statistics about the CanVEC corpus.

Type	Token Language	Token POS	Clause Language
Vietnamese	96%	76%	99%
English	100%	99%	100%
Mixed	97%	75%	99%

Table 5: Accuracy report for various aspects of automatic annotation.

transcribed, time-aligned clause for each speaker is associated with separate sub-tiers for tokens, token POS tags, token language tags, clause language tag and translation. This link between transcription, encoding and speech signal not only assists with data transparency, but also facilitates preliminary analysis.

Finally, Table 4 provides some basic statistics concerning the overall composition of CanVEC after automatic annotation.

4. Evaluation

4.1. Language Identification / POS Tagging

To evaluate our automatic annotation method, we randomly selected 100 clauses of each type (i.e. monolingual Vietnamese, monolingual English, and mixed) and manually checked the accuracy of the labels for token language tags, token POS tags, and clause language tags. For each of these, accuracy was calculated as follows:

$$\text{Accuracy (\%)} = \frac{\text{\# Correct labels}}{\text{\# Total labels}} \qquad (1)$$

Table 5 hence reports the results for each of these levels of annotation.

While language identification was almost perfect at both the token level and the clause level, most likely because Vietnamese and English words tend to be orthographically distinct, POS tag results for Vietnamese were a lot less robust. This is likely because Vietnamese POS taggers are not only typically trained on much less data than English POS taggers, but they are also unlikely to be well-suited to speech data (Plank et al., 2016). Specifically, spoken Vietnamese is characterised by extensive use of discourse markers and lexicon variation due to regional dialects and so is significantly different to written Vietnamese. This reinforces the idea that text-trained POS taggers are not always optimal for analysing spoken discourse.

Results for mixed clause POS tags were also lower compared to English, although this is most likely for the same reason that the results for Vietnamese POS tags were low. Alternatively, since mixed clauses were split into smaller subsequences before being sent to the appropriate monolingual tagger, it might also be the case that the sequences were short enough that the tagger did not have enough context to assign a reliable tag.

Upon closer inspection, we found that the majority of Vietnamese POS tagger errors involved pronouns and classifiers. This makes sense however, because the Vietnamese personal reference system is fairly complex and uses different pronouns, kin terms, and personal names in different contexts (Nguyen, 2018). Specifically, while kin terms and personal names are frequently used as personal pronouns in spoken discourse, they are fairly unproductive in written text. Given that Vietnamese POS taggers are trained using written sentences, they understandably struggle with the spoken domain where a different set of pronouns is used. Table 6 shows the distribution and proportion of these errors across the evaluation set.

As we can see from the results, PRON and CLS were most frequently mistaken for each other (63% and 82%) rather than for something else. This is arguably a positive result however, as it shows that the confusion was systematically confined to a limited domain (i.e. PRON and CLS) and not spread out over multiple POS tags.

Above all else, it is also important to note that despite the difficulties with PRON and CLS, results for other Viet-

Correct tag	Tagged as	N	%
Pronoun (PRON)	Classifier (CLS)	59	63%
	Noun (N)	26	28%
	Particle (PRT)	7	7%
	Preposition (PREP)	2	2%
Classifier (CLS)	Pronoun (PRON)	56	82%
	Interjection (INTJ)	12	18%

Table 6: Distribution of PRON and CLS errors (N=162 errors/ 100 sample Vietnamese clauses)

namese POS tags, particularly Nouns (N)[10], Verbs (V), Adverbs (ADV), and Prepositions (PREP) remain particularly strong, with error rates ranging from 1-5% only. This means that barring PRON and CLS, other Vietnamese POS tags can be reliably extracted from the corpus.

4.2. Machine Translation

To evaluate the quality of the automatically translated clauses, we again randomly selected 100 monolingual Vietnamese and 100 mixed clauses and rated them in terms of semantic adequacy, fluency, and comprehensibility. Each of these is defined as follows:

- Semantic adequacy: Does the translation retain the intended meaning of the source clause?
- Fluency: Does the translation sound natural in the target language?
- Comprehensibility: Is the translation understandable?

Although semantic adequacy and fluency are well-known metrics in machine translation evaluation (Koehn, 2009; Dorr et al., 2011), comprehensibility is less common. We nevertheless consider comprehensibility an important aspect of code-switched speech, as speech is much more prone to idiomatic expressions or other cultural concepts that often do not literally translate into the target language.

The bilingual first author thus assigned a binary Yes/No judgement for each metric to each clause in the sample. A binary scale was used, rather than a Likert scale, because clauses were short enough to expect fewer mistakes from the translation system (Koehn, 2009, p.218). It is also worth stating that the goal of this evaluation was not to formally evaluate the Google Translate API on code-switching speech, but rather to ascertain the quality of the automatic translations for reasons of corpus reliability. We are aware that robust machine translation evaluation is an active area of research and lots of different metrics exist (Papineni et al., 2002; Snover et al., 2006; Lavie and Agarwal, 2007; Lo and Wu, 2013).

With this in mind, results are shown in Table 7. The results suggest that the overall quality of corpus translation for monolingual Vietnamese is relatively positive, with more than half of the clauses meeting all three requirements of semantic adequacy, fluency, and comprehensibility (i.e. Total "Yes" = 3). Although the results for mixed utterances are not as robust, 43% is still a promising number given that MT systems are not usually explicitly designed to handle code-switching. In fact this result is arguably more impressive when we consider that natural speech is permeated with non-standard grammatical and discourse features such as argument drop or disfluencies (McCarthy and Carter, 2015), and that code-switching speech is especially notorious for exhibiting all sorts of different combination of grammatical features from both participating languages (Adel et al., 2015; Çetinoğlu et al., 2016).

Metric	Vietnamese	Mixed
Adequate	67%	64%
Fluent	77%	54%
Comprehensible	80%	72%

N Metrics Satisfied	Vietnamese	Mixed
0	11%	20%
1	11%	14%
2	22%	23%
3	56%	43%

Table 7: Table showing: i) the proportion of clauses meeting each criterion per metric, and ii) the distribution of clauses meeting at least N criteria in a sample of 100 Vietnamese and 100 Mixed clause translations. For example, 77% of Vietnamese clauses were considered fluent, while 22% of Vietnamese clauses received positive scores in any 2 metrics.

Additionally, it is worth noting that MT performed best at comprehensibility on both sets of data, scoring 80% and 72% on monolingual Vietnamese and mixed clauses respectively. Although MT struggled most with maintaining the fluency of the translated output in the mixed utterances (54%), the fact that most translations were still considered comprehensible suggests that the output is still probably understandable and useful to users of CanVEC.

In terms of specific errors, we also found that similar to Vietnamese POS taggers, MT seems to struggle most with Vietnamese pronouns. Table 8 thus illustrates contrasting occasions when the pronoun was translated incorrectly and correctly in a monolingual Vietnamese and mixed clause respectively. In particular, the first person subject *con* (kin term meaning 'child') was erroneously translated as a 3SG common noun in the monolingual Vietnamese clause, but accurately translated as a 1SG subject pronoun in the mixed clause. Although this is only an isolated example, it is nevertheless surprising that the correct translation is found in a mixed clause, which typically scored lower in the evaluation overall.

[10] Note that although it is apparent from Table 6 that 1/3 of PRON were incorrectly tagged as N, these only count towards PRON errors and do not count towards N error rates. This is because in Vietnamese (and many other languages), PRON is considered an open-class subset of N, and hence a PRON can be a N in essence, but not vice versa.

Input:	*con*	*đi*	*bộ*	
Gloss:	1SG.kin	go	foot	
MT:	child walking			
Human:	I walked.			
				(Penny.Marie.Rory.0912, 11:48.8 - 11:49.4)

Input:	*mà*	*giống-như*	*con*	pick up a little bit of Busan Busan dialect
Gloss:	but	like	1SG.kin	
MT:	but like I pick up a little bit of Busan Busan dialect			
Human:	but like I pick up a little bit of Busan Busan dialect.			
				(Tim.Jess.Chloe.0705, 08:03.7 - 08:10.2)

Table 8: Examples of pronoun translations in CanVEC. The pronoun *con* is translated incorrectly in the first monolingual Vietnamese clause, but correctly in the second, much longer mixed clause.

5. Conclusion

To conclude, the main contributions of this paper are threefold:

1. We introduce a new natural speech corpus, the Canberra Vietnamese-English Code-switching corpus (CanVEC), which we make available for research purposes with this paper.[11] CanVEC is also the first Vietnamese-English code-switching corpus to be collected in a maximally natural setting.

2. We describe a simple method to semi-automatically annotate Vietnamese-English intra-clausal code-switching data. We believe the method can also be adapted for other language pairs, and have already obtained promising results (Kidwai et al., 2019) on a Hindi-English code-switching corpus (Jamatia et al., 2015).

3. We identify areas of potential difficulty in automating the annotation of a spoken code-switching corpus. These findings can perhaps be further tested on other corpora and may have strong implications for improving the processing of bilingual data.

A final important point to make is that although multilingual corpora are regularly created in sociolinguistics, the vast majority are relatively small because they are annotated entirely manually. Although our method to semi-automatically annotate our data is fairly straightforward, it nevertheless represents an opportunity to overcome the traditional process of manual annotation, which can be costly in terms of both time and money. This not only benefits sociolinguists who can collect and analyse larger corpora, but also the wider NLP community who can subsequently exploit such corpora for other purposes; e.g. speech recognition, machine translation, or computer assisted language learning.

6. Acknowledgements

We are grateful to the Cambridge Trust, the Philological Society and the Section of Theoretical and Applied Linguistics, University of Cambridge for funding this work. The second author was supported by the Institute for Automated Language Teaching and Assessment (ALTA), University of Cambridge.

7. Bibliographical References

ABS. (2017). *2016 Census: Australian Capital Territory*. Australian Bureau of Statistics.

Adel, H., Vu, N. T., Kirchhoff, K., Telaar, D., and Schultz, T. (2015). Syntactic and semantic features for Code-Switching factored language models. *IEEE Transactions on Audio, Speech and Language Processing*.

Adouane, W., Bernardy, J.-P., and Dobnik, S. (2018). "Improving Neural Network Performance by Injecting Background Knowledge: Detecting Code-switching and Borrowing in Algerian texts". In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 20–28. Association for Computational Linguistics.

Bali, K., Sharma, J., Choudhury, M., and Vyas, Y. (2014). "“I am borrowing ya mixing ?” An Analysis of English-Hindi Code Mixing in Facebook". In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126. Association for Computational Linguistics.

Barman, U., Das, A., Wagner, J., and Foster, J. (2014). "Code Mixing: A Challenge for Language Identification in the Language of Social Media". In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23. Association for Computational Linguistics.

Cacoullos, R. T. and Travis, C. E. (2018). *Bilingualism in the Community: Code-switching and Grammars in Contact*. Cambridge University Press.

Caines, A., Bentz, C., Graham, C., Polzehl, T., and Buttery, P. (2016). Crowdsourcing a Multi-lingual Speech Corpus: Recording, Transcription and Annotation of the CROWDED CORPUS. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Çetinoğlu, Ö., Schulz, S., and Vu, N. T. (2016). Challenges of computational processing of code-switching. In *Proceedings of the Second Workshop on Computational Ap-*

[11] https://github.com/Bak3rLi/CanVEC

proaches to Code Switching, pages 1–11. Association for Computational Linguistics.

Chan, J. Y. C., Ching, P. C., and Lee, T. (2005). Development of a Cantonese-English Code-mixing Speech Corpus. In *INTERSPEECH*.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Deuchar, M., Davies, P., and Donnelly, K. (2018). *Building and using the Siarad corpus: Bilingual conversations in Welsh and English*. John Benjamins.

Dey, A. and Fung, P. (2014). A Hindi-English Code-Switching Corpus Code-Switching in Indian Culture. In *Proceedings of LREC*.

Dorr, B. J., Snover, M., and Madnani, N. (2011). Machine Translation Evaluation and Optimization. In Joseph Olive, et al., editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, chapter 5. Springer, 1st edition.

Hamed, I., Elmahdy, M., and Abdennadher, S. (2017). Building a First Language Model for Code-switch Arabic-English. In *Procedia Computer Science*.

Hofweber, J., Marinis, T., and Treffers-Daller, J. (2016). Effects of dense code-switching on executive control. *Linguistic Approaches to Bilingualism*, 6(5):648–668, October.

Jamatia, A., Gambäck, B., and Das, A. (2015). Part-of-speech tagging for code-mixed English-Hindi twitter and Facebook chat messages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 239–248, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Jurgens, D., Dimitrov, S., and Ruths, D. (2014). Twitter users #codeswitch hashtags! #moltoimportante #wow. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 51–61. Association for Computational Linguistics.

Kidwai, S., Bryant, C., Nguyen, L., and Biberauer, T. (2019). Automatic Language Identification in Code-Switched Hindi-English Social Media Texts. *Cambridge Language Sciences Symposium*.

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press.

Labov, W. (2004). Field methods of the project on linguistic change and variation. In J. Baugh et al., editors, *Language in use: Readings in sociolinguistics*, page 29. Prentice Hall, Englewood Cliffs, NJ.

Lavie, A. and Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*.

Li, Y. and Fung, P. (2014). Language modeling with functional head constraint for code switching speech recognition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 907–916, Doha, Qatar, October. Association for Computational Linguistics.

Lo, C.-k. and Wu, D. (2013). MEANT at WMT 2013: A Tunable, Accurate yet Inexpensive Semantic Frame Based MT Evaluation Metric. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 422–428. Association for Computational Linguistics.

Lyu, D.-C., Tan, T.-P., Chng, E., and Li, H. (2015). Mandarin–English code-switching speech corpus in South-East Asia: SEAME. In *Language Resources and Evaluation*, volume 49, pages 1986–1989, 01.

Maharjan, S., Blair, E., Bethard, S., and Solorio, T. (2015). "Developing Language-tagged Corpora for Code-switching Tweets". In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 72–84. Association for Computational Linguistics.

Mave, D., Maharjan, S., and Solorio, T. (2018). "Language Identification and Analysis of Code-Switched Social Media Text". In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 51–61. Association for Computational Linguistics.

McCarthy, M. and Carter, R. (2015). Spoken Grammar: Where Are We and Where Are We Going? *Applied Linguistics*, 38(1):1–20, 01.

Modipa, T. I., Davel, M. H., and de Wet, F. (2013). Implications of Sepedi/English code switching for ASR systems. In *Proceedings of the 24th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*.

Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., and Solorio, T. (2016). "Overview for the Second Shared Task on Language Identification in Code-Switched Data ". In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49. Association for Computational Linguistics.

Muysken, P. (2000). *Bilingual speech : a typology of code-mixing / Pieter Muysken*. Cambridge University Press, Cambridge.

Myers-Scotton, C. (1997). *Duelling languages: grammatical structure in codeswitching*. Oxford: Clarendon.

Nguyen, D. and Doğruöz, A. S. (2013). Word Level Language Identification in Online Multilingual Communication. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Nguyen, L. (2018). Borrowing or Code-switching? Traces of community norms in Vietnamese-English speech. *Australian Journal of Linguistics*, 38(4):1–24, 10.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Petrov, S., Das, D., and McDonald, R. (2012). "A Universal Part-of-Speech Tagset". In *Proceedings of the*

Eighth International Conference on Language Resources and Evaluation (LREC-2012). European Language Resources Association (ELRA).

Plank, B., Søgaard, A., and Goldberg, Y. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418. Association for Computational Linguistics.

Poplack, S. (1980). “Sometimes I’ll start a sentence in Spanish y termino en español”: toward a typology of code-switching. *Linguistics*, 18.

Poplack, S. (1993). Variation theory and language contact. In D. Preston, editor, *American dialect research: An anthology celebrating the 100th anniversary of the American Dialect Society*, page 252. Benjamins, Amsterdam.

Riehl, C. (2005). Code-switching in bilinguals: impacts of mental processes and language awareness. In *ISB4: Proceedings of the 4th International Symposium on Bilingualism*. Cascadilla Press.

Shen, H. P., Wu, C. H., Yang, Y. T., and Hsu, C. S. (2011). CECOS: A Chinese-English code-switching speech database. In *2011 International Conference on Speech Database and Assessments, Oriental COCOSDA 2011 - Proceedings*.

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category - ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. Max Planck Institute for Psycholinguistics, The Language Archive.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of association for machine translation in the Americas*.

Solorio, T. and Liu, Y. (2008). Part-of-Speech Tagging for English-Spanish Code-Switched Text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060. Association for Computational Linguistics.

Soto, V. and Hirschberg, J. (2018). "Joint Part-of-Speech and Language ID Tagging for Code-Switched Data". In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–10. Association for Computational Linguistics.

Tuc, H.-D. (2003). *Vietnamese-English Bilingualism: Patterns of Code-Switching*. Taylor and Francis.

Yilmaz, E., Dijkstra, J., Van De Velde, H., Kampstra, F., Algra, J., Van Den Heuvel, H., and Van Leeuwen, D. (2017). Longitudinal speaker clustering and verification corpus with code-switching frisian-Dutch speech. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*.