Code-Switching in the Digital Domain

A Research Study

HASS: Psychological Approaches to Bilingualism

Agrim Singh / Mendel Oh / Le Vu Huy

Singapore University of Technology and Design

Abstract

This paper aims to explore whether the occurrences of code switching on online platforms can be correlated to the similarity of communications of speech, or away from conventional formal writing. We will study this with focus on code switching in the written domain and the exploration of the digital domain, with particular focus on the Facebook Messenger chat client and conversations occurring on that medium as laid out in the Introduction. This paper will further explain as part of the Methodology how by utilizing Natural Language Processing we were able to detect code switching inter and intra-sentence, with the script available as part of the Appendix, and accordingly present results that have been further discussed in the Discussion section. The conclusions about code-switching in the digital domain have been appropriately bolstered with example conversations.

Introduction

Code switching, or the alternation of usage of words and structures of more than one language by bilinguals, has received a great deal of research and attention from researchers. Despite the large body of literature available on the subject, the bulk of research regarding this linguistic phenomenon has, thus far, been focused primarily on the mixing of spoken language, and, as a result, the majority of theoretical frameworks that address bilingual code switching are based on speech as opposed to writing.

Code switching is thought to be observed in writing with much reduced frequency in comparison with speech, due in part to the greater postulated demands on the language processing capabilities of an individual as well as the formality and contextual demands imposed on the use of written language in various situations. However, it has been noted by various researchers

(Huerta-Macías & Quintero, 1992, Escamilla & Hopewell, 2007) that code switching occurred naturally in written discourse even within developing bilingual children – albeit at a lower frequency than spoken discourse. It was also noted that this was likely because the usage of code switching in written discourse was, in general, less accepted than code switching in spoken discourse.

Early research into the field of bilingualism postulated that language functions were inherently monolingual in nature, and that language systems eventually differentiated into separate structures during bilingual development (Volterra & Taeschner, 1978). This Unitary Language System Hypothesis, as it became known, hypothesized that the occurrence of code switching was primarily due to deficiencies in language system development that led to the random usage of languages in bilinguals due to the inability to differentiate between the two language systems.

While this work was challenged by later studies, and eventually discredited (Genesee, 2003) as it was established that code switching, was neither random, nor due to a deficiency in language development, the foundation laid by early researchers resulted in a pervasive view that linguistic competency was inherently linked to the ability to separate language usage, and that the mixing of language systems was a symptom of linguistic deficiency, particularly in written communication.

In recent times, the rise of discourse over digital media (e-mail, text messaging, etc.) and computer-mediated communication has led to a form of communication described as "written speech" (Ferrara, Brunner & Whittemore, 1991), where it was found that such communications did not conform to the conventional dichotomy between spoken and written language usage,

instead taking various aspects of each form of communication with regards to their contextual usages.

The relative informality, ease of use, and lower regulation and/or stigmatization of discourse in digital forms of communication as compared to formal writing would, in theory, allow for users to engage in, as well as promote greater frequencies of code switching. Such occurrences of code-switching in communications over various forms of computer-mediated discourse have been noted by past studies (Georgakopoulou, A., 1997).

With the rise of "new" digital media allowing for asynchronous real-time mobile chat services (e.g. WhatsApp, Skype, Facebook Messenger) to become prevalent as one of the primary methods of communication over digital devices, the examination of such avenues of communication should be examined in greater detail.

We postulate that, in accordance with the theorized speech-writing continuum that places such communications occurring in and over digital media in-between the established domains of spoken and written discourse (*Yates, S. J. (1996). Oral and written linguistic aspects of computer conferencing. Pragmatics and beyond New Series, 29-46.*), as communications more closely resemble speech in their characteristics, there should be a corresponding rise in the frequency of code-switching occurring between conversing individuals. The brevity of a message as well as the speed at which such messages are exchanged was found to affect its resemblance to speech. The shorter a message, and the faster the rate at which messages were exchanged, the closer their resemblance to spoken interactions (*Yates, S. J. (1996)*).

As such, in this paper, we attempt to determine if the frequency of occurrences of code switching are correlated to the similarity of communications to speech (or, conversely, their departure from conventional formal writing).

Methodology

The focus of this study is on the conversations between bilingual individuals over digital communication channels – primarily instant messaging over social media. For this purpose, a large sample of conversational data between bilinguals over Facebook messaging was collected and analyzed for the purposes of the study.

Message logs from 208 unique conversation threads between 22 Vietnamese-English bilinguals (mean age = 21.6, SD = 2.04) were extracted from the Facebook messaging client for analysis in the study. Messages were dated from 2009 to 2015.

Involved participants were informed and consent for the usage of their conversation logs was obtained. Names were replaced with pseudonyms for anonymity before analysis.

The extracted logs were separated into individual conversations based on time. Sequential chains of messages that were separated by more than 30 minutes were considered to be separate conversations, and treated as such for the purpose of analysis.

After separation into individual conversations, the contents of the resultant conversations were examined, and specific exchanges selected as representative examples for certain patterns of code switching.

Natural Language Processing (NLP) language identification scripts were employed in order to speed up the quantification and classification of conversations by language type, as well as to

evaluate the frequency of inter and intra-sentential code switching that occurred within conversation samples.

The library used in the language identification script was `langdetect`, which is typically used to detect the language in which a section of text is written. The algorithm used calculates language probabilities from features of spelling (Naive Bayes classifier with character n-gram) instead of matching the text against the dictionary of each language. This algorithm was run over each sentence of the conversation, and the resultant probability of the sentence being in English was used as an indicator of the proportion of L2 code switching occurring in an intra-sentential form (henceforth referred to as an "L2 index"). Sentences that were determined to have a nil probability of being in English were considered to be fully L1 for the purposes of data analysis.

Individual conversations were evaluated on a number of variables:

- average L2 index per sentence (L2 index sum divided by number of sentences),
- normalized average L2 index per sentence (L2 index sum divided by number of code-switched sentences),
- Proportion of code-switched L1 sentences against 'full' L1 sentences,
- average sentence length (number of words in individual sentences),
- duration of conversation (time elapsed between first and last messages),
- average message interval (average time elapsed between messages)

Results

Few results and correlations were checked for from the data that was obtained. It was interesting to note that hypotheses that were previously seen in papers such as Georgakopoulou (1997) were put into scrutiny since the results obtained proved otherwise.

1. Proportion of English sentences to Total sentences vs Average Length of sentence

Correlations

		EnglishvsTotal	avg length
EnglishvsTotal	Pearson Correlation	1	-.021
	Sig. (2-tailed)		.399
	N	1691	1691
avg length	Pearson Correlation	-.021	1
	Sig. (2-tailed)	.399	
	N	1691	1691

This was aimed at checking whether number of code-switched sentences would vary with relation to how long one's sentences are in a conversation, more so to suggest comfort with a particular language given that previous studies have attributed code-switching to weaker L2 command. We saw appalling significance (~40%) levels and a near-zero correlation, giving us no relation between the two variables.

2. Proportion of English sentences to Total sentences vs Average time between sentences

Correlations

		EnglishvsTotal	average gap (seconds)
EnglishvsTotal	Pearson Correlation	1	.130**
	Sig. (2-tailed)		.000
	N	1691	1691
average gap(seconds)	Pearson Correlation	.130**	1
	Sig. (2-tailed)	.000	
	N	1691	1691

**. Correlation is significant at the 0.01 level (2-tailed).

This relationship was observed to assess whether the propensity for a code-switch in a conversation would be evident if related to the rate at which messages were being exchanged. While the correlation statistics are significant, the Pearson Correlation coefficient at 0.130 suggests either a non-linear relationship or no correlation whatsoever.

3. Normalized Average Code-Switching score vs Average time between sentences

Correlations

		normalized avg	average gap (seconds)
normalized avg	Pearson Correlation	1	.030
	Sig. (2-tailed)		.217
	N	1691	1691
average gap(seconds)	Pearson Correlation	.030	1
	Sig. (2-tailed)	.217	
	N	1691	1691

Our relation between the Normalized Average Score (L2 index sum divided by number of code-switched sentences) and the time between sentences would have shed light on how likely users are to switch languages in a conversation but a high significance level and a near-zero correlation suggested that there might be no relation here either.

4. Normalized Average Code-Switching score vs Average length of sentence

Correlations

		normalized avg	avg length
normalized avg	Pearson Correlation	1	-.031
	Sig. (2-tailed)		.204
	N	1691	1691
avg length	Pearson Correlation	-.031	1
	Sig. (2-tailed)	.204	
	N	1691	1691

We wanted to see how likely the users were to code-switch in between a sentence and we thus compared the Normalized score to the average length of a conversation sentence but, again, to no avail. The significance levels remained high and correlation was near-zero.

Discussion

The modes of language input used in computers and other digital devices often restrict the use of multiple languages in conversations over digital media, due to the relative inconvenience of switching between language inputs. This primarily affects languages with different writing systems or different alphabets, which require such different input systems, and is much less significant for languages that have significant alphabetical overlap, which would not require such a corresponding change. While such a factor would theoretically curtail or discourage code switching, such patterns were not observed in a qualitative examination of the extracted conversational logs. An example follows below.

Don
much primitive
KCP Settings
XYsubFilter là clgt
Host
la cai may can lol
xoa het may cai khac di
Don
vẫn chưa được lol
ko thể load sub từ file srt
ko drag and drop srt vào đc
Host
eh
d biet lol

In the conversation, one party code switches into Vietnamese text using the appropriate input method. The second party, however, also code switches into Vietnamese, albeit without switching input methods, opting instead to simply transliterate his messages. This demonstrates

the propensity of individuals to code switch with little issue despite the problems that may arise from the lack of proper textual input.

We also must consider why code-switching occurs for bilinguals. One reason attributed to this would be the inconvenient choice of words. Our research has shown that the shorter length of words in the dominant language might play a role in the choice of words in an online conversation. Alternatively, there might not be an equivalent words existing in the native language, promoting the need for a code switch - a case of convenience.

Samples of conversation:

1/

User:

*Nam may bon no co cai **maze** khac*

*Ma **logic** thi van **reuse** dc =))*

Maze bon no nam nay de vcl

Translation:

They have a different maze this year.

But the logic is reusable.

Their maze this year is very easy.

We see that code switching occurred for the 3 words: “maze”, “logic” and “reuse”.

“Maze”: the Vietnamese equivalent (“Mê cung") is 2 syllables, 6 characters long, and hence more inconvenient to type.

“Logic”: there’s no native Vietnamese equivalent (the equivalent is a transliterated word, “lô-gíc") and thus becomes a natural choice here.

“Reuse": the Vietnamese equivalent (“dùng lại") is longer and hence more inconvenient to type.

There is also a case for domain-specific words which when used naturally promote a code switch. For technical terms in certain domains where a particular language dominates naming conventions, there is little choice but to code switch for two reasons – the lack of an equivalent term in the local language that captures the same meaning and the inconvenience of looking for a word that may do that.

2/

User1:

*Neu ma **connect** dc pptp thi nhanh vc*

hom truoc quen tra cung voi cai pizza

la sao?

dm tai anh tunnel split utorrent

deo biet no co work ko

User2

*nó có nhiều **protocol** để connect to vpn*

*nếu mà anh để **automatic** thì thỉnh thoảng vào đc PPTP*

thì nhanh vc

nếu mà SSTP hay OpenVPN thì cũng chậm bt thôi nhưng mà lúc nào cũng vào đc

Translation:

User1:

It's very fast if connected using PPTP

Should have paid together with the pizza a few days ago.

What?

Since I used tunnel splitting with utorrent

Not sure if it worked

User2:

There are different protocols to connect to VPN

If you set it to automatic then it's occasionally PPTP

Then it's very fast

If SSTP or OpenVPN then it's at normal speed but always available

We see two cases where the Vietnamese equivalent is shorter to type but code switch occurred nonetheless:

"Connect" (Vietnamese equivalent "kết nối"), while the Vietnamese equivalent has equal number of syllables and shorter character-wise, the English version was used. There's a tendency to code switch to English for computing terms due to its complete dominance in the field.

Similarly, the word "Automatic" (Vietnamese equivalent "tự động") can be explained.

"Pizza" (Vietnamese equivalent "pizza") has no alternative in Vietnamese because of the Italian origins of the word itself, and thus is appropriated in this conversation.

Limitations

There are limitations to our findings and research which could be worked on for better results in a subsequent experiment. Firstly, the lack of a common L2 in the team meant that one member was solely responsible for translating collected data and looking for trends while the other members of the team waited. Evidently, a common L2 would have enabled for greater analysis, especially while working with the number of data points that were utilized for this paper.

Secondly, the limitations of data collection. We utilized data collected by Facebook for its chat client and while the data itself can be parsed from the logs they provide, the messages are logged to a minute's accuracy. This meant that our metric for average gap between sentences and average conversation timing was flawed since our data would not be sensitive enough to detect the differences between two messages that were separated by a few seconds or by a minute. This affects our eventual data analysis and perhaps can be improved if chat data is, instead, collected elsewhere.

Lastly, there is the case of understanding the unknowable. There are certain cases of code switching which have not been considered by this paper for the reason that they did not exhibit any particular pattern of switching. Individual reasons for switching are not taken into account in this paper since it would require an extensive psycholinguistic research of those involved in the conversation.

Conclusion

We were unable to find any correlation between the frequency of occurrences of code switching and the similarity of communications to speech (or, conversely, their departure from conventional formal writing). Sitting somewhere in between, we posit that code switching on online platforms occurs partly due to convenience or due to the lack of equivalent terms in the dominant language; there needs to be more research for the frequencies of messaging, code switching and speed of message exchange to conclusively prove whether speech is indeed modelled by online conversations. Furthermore, we conclude that code switching on online platforms was independent of input methods and would continue to occur regardless of input alternatives, overridden by the user's convenience.

References

Escamilla, K., & Hopewell, S. (2007, April). The role of code-switching in the written expression of early elementary simultaneous bilinguals. In Annual Meeting of American Educational Research Association.

Genesee, F. (2003). Rethinking bilingual acquisition. Multilingual Matters, 204-228.

Ferrara, K., Brunner, H., & Whittemore, G. (1991). Interactive written discourse as an emergent register. Written communication, 8(1), 8-34.

Georgakopoulou, A. (1997). Self-presentation and interactional alliances in e-mail discourse: the style-and code-switches of Greek messages. International Journal of Applied Linguistics, 7(2), 141-164.

Huerta-Macías, A., & Quintero, E. (1992). Code-switching, bilingualism, and biliteracy: A case study. Bilingual Research Journal, 16(3-4), 69-90.

Volterra, V., & Taeschner, T. (1978). The acquisition and development of language by bilingual children. Journal of child language, 5(02), 311-326.

Danet, B. and Herring, S. C. (2003), Introduction: The Multilingual Internet. Journal of Computer-Mediated Communication, 9: 0. doi: 10.1111/j.1083-6101.2003.tb00354.x

Androutspoulos, J. (2013). Networked multilingualism: Some language practices on Facebook and their implications. International Journal of Bilingualism, 0(0), 1-21.

Lončarić, J. (2014). Croatian-English Code-Switching Patterns of Croatian Facebook Users. University of Zagreb, Faculty of Humanities and Social Sciences, Department of English.

Appendix

1. Script used for Natural Language Processing

The script used to parse through Facebook's chat log to provide something meaningful. "messages.htm" is the file in which Facebook provides all its messages.

```
from __future__ import division
from datetime import datetime, timedelta
import xml.etree.ElementTree as ET
import langdetect
import re
from nltk import word_tokenize
import csv
from multiprocessing import Process, Queue
def reformatFile(filename, newFilename):
      tree = ET.parse(filename)
      root = tree.getroot()
      body = root[1]
      contents = body[1]
      newRoot = ET.Element("root")
      for threadGroup in contents:
            for thread in threadGroup:
                  newRoot.append(reformatThread(thread))
      newTree = ET.ElementTree(newRoot)
      newTree.write(newFilename)
def reformatThread(thread):
      previousTime = datetime.today()
      deltaRef = timedelta(minutes = 30)
      fThread = ET.Element("thread", {'users': thread.text})
      for m in xrange(len(thread)//2):
            metaElement = thread[2*m][0]
            user = metaElement[0].text
            timeString = metaElement[1].text
            time = datetime.strptime(timeString, '%A, %B %d, %Y at %I:%M%p
%Z+%S')
            if previousTime - time > deltaRef:
                  ET.SubElement(fThread, "conv")
            ET.SubElement(fThread[-1], 'msg', {'user': user, 'time' :
timeString}).text = thread[2*m+1].text
            previousTime = time
      return fThread
def scoreConv(conv, users):
      if len(conv) < 5:
            return
      engScores = []
```

```
	sentenceLength = []
	zeroCount = 0
	for msg in conv:
		try:
			score = langdetect.detect_langs(msg.text)
			sentenceLength.append(getSentenceLength(msg.text))

			isEnglish = False
			for lang in score:
				if lang.lang == 'en':
					engScores.append(lang.prob)
					isEnglish = True
					break
			if not isEnglish:
				zeroCount += 1
				engScores.append(0)
		except:
			pass
	totalSeconds = (datetimeFromMsg(conv[0]) - datetimeFromMsg(conv[-
1])).total_seconds()
	avgSeconds = totalSeconds / (len(conv) - 1)
	if (len(engScores) - zeroCount):
		average = sum(engScores)/len(engScores)
		switchedAverage = sum(engScores) / (len(engScores) - zeroCount)
		averageSentenceLength = sum(sentenceLength) / len(sentenceLength)
		return [users, average, switchedAverage, len(engScores),
zeroCount, averageSentenceLength, totalSeconds, avgSeconds]
	else:
		return
def datetimeFromString(timeString):
	return datetime.strptime(timeString, '%A, %B %d, %Y at %I:%M%p %Z+%S')
def datetimeFromMsg(msg):
	timeString = msg.attrib["time"]
	return datetimeFromString(timeString)
def getSentenceLength(sentence):
	return len(re.findall(r'\w+', sentence))
def csvWriter(q, csvFilename):
	with open(csvFilename, 'wb') as csvfile:
		writer = csv.writer(csvfile)
		writer.writerow(['users', 'avg', 'normalized avg', 'sentence
count', 'non-English sentence count', 'avg length', 'duration (seconds)',
'average gap(seconds)'])
		while True:
			result=q.get()
			if result:
				try:
					writer.writerow(result)
				except:
					pass
```

```
                  else:
                        break
def generateCsv(filename, csvFilename):
      queue = Queue()
      p = Process(target=csvWriter, args=(queue, csvFilename, ))
      p.start()
      tree = ET.parse(filename)
      root = tree.getroot();
      threadCount = len(root)
      offset = 0
      for index, thread in enumerate(root[offset:]):
            print index, threadCount - offset
            users = thread.attrib['users'].encode('ascii', 'ignore')
            for conv in thread:
                  result = scoreConv(conv, users)
                  if result: queue.put(result)
      queue.put(None)
if __name__ == "__main__":
      reformatFile("luccan.htm", "luccan.xml")
      generateCsv("luccan.xml", "luccan.csv")
```

2. Sample of Facebook's conversation log

This log is a Bahasa-English code switched conversation which was eventually not utilized due to our research group not having anyone who knew the language. This short snippet shows the formatting that the log is provided in. The names have been hidden for confidentiality.

1636201724@facebook.comWednesday, February 9, 2011 at 9:46pm UTC+08

me? no la too lazy. Your english name Henry huh? :) u signed?

[redacted]

[redacted]Monday, May 19, 2014 at 8:25am UTC+08

[redacted]Sunday, May 18, 2014 at 11:29pm UTC+08

Kl gitu gw Ga usah nggombal dong :))

[redacted]Sunday, May 18, 2014 at 11:29pm UTC+08

Kok tau diri

[redacted]Sunday, May 18, 2014 at 11:29pm UTC+08

?

[redacted]Sunday, May 18, 2014 at 11:29pm UTC+08

Beratri Lu Ga berasa fotogenik

[redacted]Sunday, May 18, 2014 at 8:13pm UTC+08

[redacted]Sunday, May 18, 2014 at 8:12pm UTC+08

ohh! yaaay! hahah

3. People studied for this research paper and their details

Names have been anonymized for confidentiality.

name	gender	age	grew up in	years stay in L2 speaking country	conversation count	sentences count	avg sentence	TotalEnglishSentence	NormalizedTotalSum	WeightedScore
Kant	m	22	vietnam	6	16	220	13.75	39	23.20303908	0.59494972
Xie	m	23	vietnam	6	55	939	17.0727273	167	110.5053217	0.661708513
Sonic	m	22	vietnam	6	17	314	18.4705882	53	40.42559234	0.762747025
Jay	m	22	vietnam	6	39	741	19	218	159.6876564	0.732512185
Hyu	m	22	vietnam	2	94	2087	22.2021277	273	181.7505395	0.665752892
Guisess	m	22	vietnam	2	42	998	23.7619048	163	111.5774423	0.684524186
Tommy	m	22	vietnam	6	336	8412	25.0357143	4911	3939.95368	0.802271163
Teddie	f	21	vietnam	5	6	156	26	29	20.14281275	0.69457975
KK	m	18	vietnam	0	11	288	26.1818182	19	11.28568517	0.59398343
Don	m	21	vietnam	5	45	1187	26.3777778	302	249.9991833	0.827811865
Anna	m	20	vietnam	4	4	106	26.5	16	12.85709492	0.803568432
Anya	m	21	vietnam	5	283	7987	28.2226148	2359	1872.655856	0.793834615
Jane	f	27	vietnam	12	44	1247	28.3409091	329	231.5699371	0.703859991
Tron	m	21	vietnam	5	509	16057	31.546169	1931	1452.641606	0.752274265
Van	m	22	vietnam	6	24	950	39.5833333	101	67.85683348	0.671849836
Narnia	m	23	vietnam	7	6	268	44.6666667	17	11.42853909	0.672267005
Tom	f	16	vietnam	0	4	181	45.25	9	5.857126294	0.65079181
Que	m	22	vietnam	6	46	2514	54.6521739	1249	167.1257829	0.133807672
Mango	m	22	vietnam	0	83	5145	61.9879518	348	238.2980986	0.684764651
Mini	f	22	vietnam	0	14	1138	81.2857143	54	25.57139598	0.47354437
Quin	f	22	vietnam	4	13	1133	87.1538462	135	103.9162473	0.76974998