

# Tagalog-English Code-Switching: Challenges for Automatic Detection\*

Nathaniel Oco  
\*De La Salle University

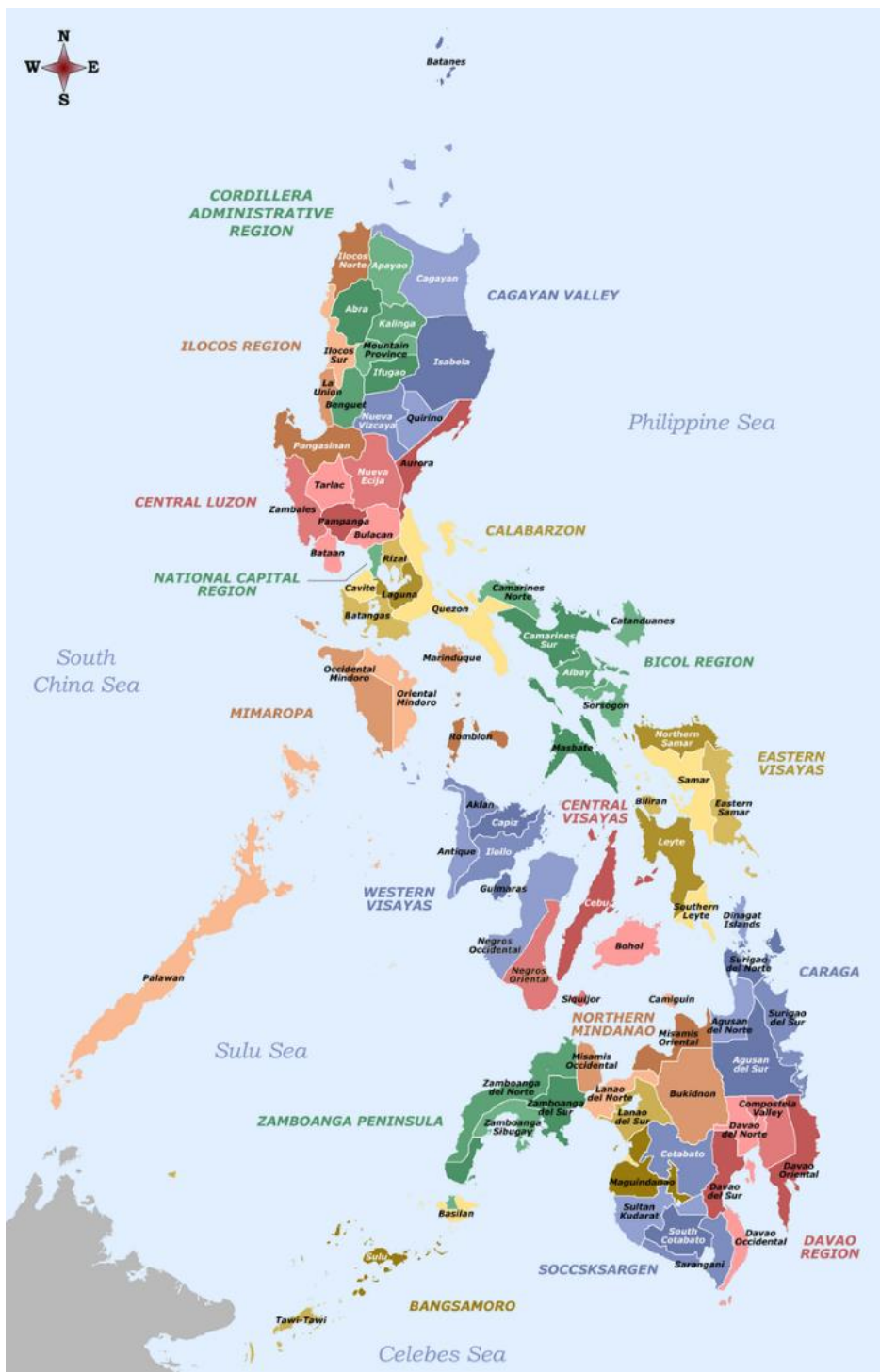
In this poster, I will give a brief overview of the Philippines - a country in Southeast Asia - and discuss Tagalog-English code-switching (TECS). TECS is the use of both Tagalog and English in a discourse. I will also detail existing works to automatically detect TECS and the challenges ahead (e.g., intra-word code-switching and interlingual homographs).

## The Philippines

Country in Southeast Asia  
Population: 103M  
English, Filipino, Tagalog

Geography  
7,641 islands  
17 regions  
3 island groups

185 Languages  
175 indigenous  
11 dying  
8 non-indigenous  
2 extinct



Source: <https://www.ethnologue.com/country/PH>  
Image from: <https://en.wikipedia.org/wiki/Philippines>

## Tagalog-English Code-switching

Code-switching  
Use of at least two languages in a discourse  
Filipinos are multilinguals  
Code-switching is common  
Causes: Proficiency-driven and deficiency-driven

Tagalog-English Code-switching (TECS)  
Intra-sentential: When I got there, *wala sila* (When I got there, they are not present)  
Nonce-borrowing: *Bisitahan mo 'yung forum* ([you] visit the forum)  
Borrowing: *Diniseybol*, according to management (It was disabled, according to the management)  
Intra-word: Nagla-library

## Approaches: Automatic detection

Tagger Dictionary + Rule-based TECS point detection  
Source: Oco, N., & Roxas, R. E. (2012, November). Pattern matching refinements to dictionary-based code-switching point detection. In Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation (pp. 229-236).

```
<rule id="ENGLISH-TAGALOG" name="Code Switch to Tagalog">
<pattern case_sensitive="no" mark_from="1">
<token postag="ENG.*" postag_regexp="yes"/>
<token postag="TAG.*" postag_regexp="yes"/>
</pattern>
<message>English to Tagalog CSP</message>
<example type="incorrect">I want to be <marker>sundalo</marker>.</example>
<example type="correct">They are soldiers.</example>
</rule>
```

Bigram count  
Source: Oco, N., Wong, J., Ilao, J., & Roxas, R. (2013). Detecting Code-Switches using Word Bigram Frequency Count. In 9th National Natural Language Processing Research Symposium, Quezon City, Philippines.

Create a bigram English model from a large corpus  
Check the bigrams of the input if they are present in the model  
If it is not in the model, it can denote code-switching

of the 242128 in the 156567 and the 57832 on the 50486 for the 40465 by the 40242 at the 37159 as a 35330	from the 33721 is a 32654 as the 28515 of a 26219 it is 24846 in a 24000 is the 22261
--	---

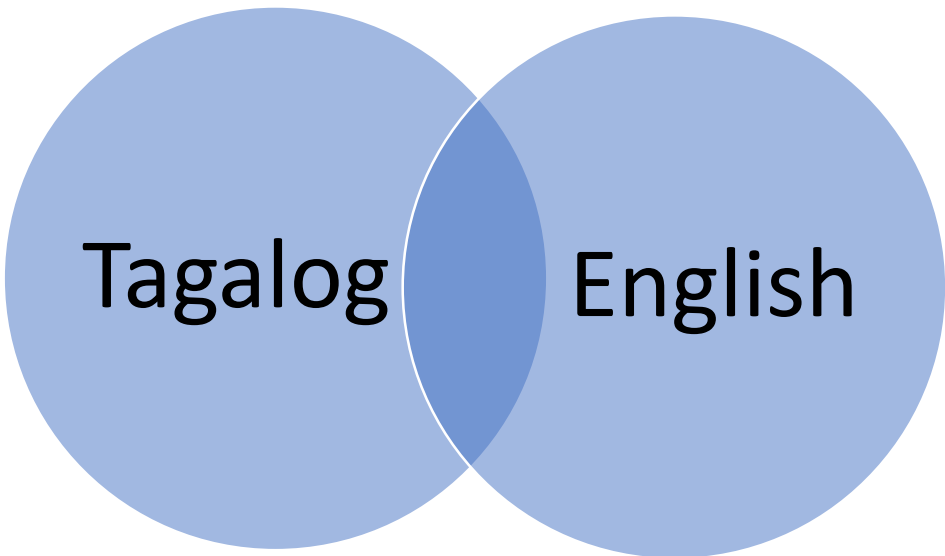
```
{ "While","everyone" }
{ "everyone","can" }
{ "can","access" }
{ "access","maybe" }
{ "maybe","ang" }
{ "ang","magiging" }
{ "magiging","lider" }
{ "lider","lang" }
{ "lang","niyan" }
{ "niyan","ay" }
{ "ay","si" }
{ "si","Nathaniel" }
```

## Challenges: Interlingual homographs, Borrowing, Intra-word Code-switching

Borrowing  
English lexicon; Tagalog orthography  
Intra-word code-switching  
English lexicon; Tagalog affix

Token	Root Word	Aspect
naghahang	hang	Incompleted
i-login	login	Neutral
magredyister	register	Neutral
inedit	edit	Completed
diniseybol	disable	Completed
mali-link	link	Contemplative
ino-note	note	Contemplative
lina-log	log	Incompleted
mag-report	report	Neutral

Interlingual homographs  
Words present in both the English lexicon and the Tagalog lexicon



Token	English POS	Tagalog POS
akin	Adjective	Pronoun
along	Preposition	Noun
at	Preposition	Conjunction
ate	Verb	Noun
away	Adverb	Verb
dating	Verb	Adjective
gusto	Noun	Verb
halos	Noun	Adjective
hanging	Verb	Noun
ho	Interjection	Polite Marker
kilos	Noun	Verb
may	Auxiliary Verb	Existential Marker
naming	Verb	Pronoun
noon	Noun	Pronoun
paring	Verb	Noun
piling	Verb	Noun
raw	Adjective	Enclitic
ring	Noun	Enclitic
sawing	Verb	Adjective
tinging	Verb	Adjective