Witchebelles Anata Magcharot kay Mudra na Nagsusuba si Akech: Developing a Rule-based Unidirectional Beki Lingo to Filipino Translator[1]

Nathaniel Oco, Raymart Fajutagana, Christine Mae Lim, Judi Diane Miñon, Julie-Ann Morano, Ryan Christian Tinoco
National University
{nathanoco,raymartfajutagana1,cm_lim23,diane.fetalino}@yahoo.com,
{moranojulieann,ryeryeryeryerye7}@gmail.com

Abstract
Current work is on the development of a word editor plug-in that provides Filipino translations for Beki words or gay lingo words. A known sociolect, Beki speak – as described in literature – is used as a form of code-mixing; shielding speakers from non-speakers. For this work, we focused on Twitter as the domain, which is a representative example of the sociolect's usage. A rule-based engine was used for the translator and the needed language resources – rule file, trigram model, and word list – were developed and detailed in the paper. Tweets from year 2013 were used and various resources were utilized to develop the needed language resources; in particular, the schema provided by the Komisyon sa Wikang Filipino was used. The plug-in was evaluated for recall and results show less than 50 percent recall rate. Although promising, the result can also be interpreted as scarcity of available digital Beki resources and that the sociolect constantly evolves. The work can be extended by providing a toolkit for users to easily add rules and descriptions.

Keywords: machine translation, Beki speak, rule-based, natural language processing, tweets

1. Introduction

The Philippines is a country characterized with 182 living languages and 4 extinct ones[2]. This highlights the growing need for language research – and sociolects are no exception. To be understood by all, a translator can aid non-speakers of the sociolect and applications are not only limited to communication and research. In this paper, we present our work on developing a rule-based translator for Beki speak – a well-known sociolect in the country used primarily by gays (Ponce, 2008). We used LanguageTool, an existing rule-based engine with language support for Tagalog and a language identification module (Naber, 2003; Oco & Borra, 2011), and developed the necessary language resources needed. These resources are the following: the rule file, which is used to translate words; the trigram model, which is used to automatically identify the language of the text input; and the word list, which is used for tagging and

[1] The paper is a continuation of a research work presented in AsiaLex2015 entitled "Lavender Filipino: Computational Models of Twitter Swardspeak".

[2] Data taken from Ethnologue, a catalog of known languages. Website entry for the Philippines: http://www.ethnologue.com/country/PH

look-up. The various language resources can also be used by other researchers in other fields such as linguistics and sociology. The paper is towards contributing to the growing number of literature in gender studies in the country.

The paper is organized as follow: chapter 2 details related works – both in gender studies and in translation; chapter 3 describes the methodology while chapter 4 details the results and in-depth discussion; chapter 5 details the evaluation done; and we conclude our work in chapter 6.

2. Gender Studies and Translators

Most literature on gay studies in the country are focused on the plight of gay individuals (Baytan, 2000; Dones, 2015; Madula, 2014), representation and views (Bautista, 1993), and on discussing the language they use (Casabal, 2008). Madula (2014), in his article, presented "pagrampa" as a theory and detailed qualitative research work on the experiences of a gay individual in a communist group. Dones (2015) focused on the elderly using descriptive methods while Baytan (2000) presented an exploratory study on Chinese Filipinos. As for language, Baytan posits that the use of Beki speak enables the user to "hide things and speak freely in the presence of non-speakers" and to make words "less injurious" (as cited in Casabal, 2008, 75-76). In the international setting, other sociolects such as IsiNgqumo in South Africa (Rudwick & Ntuli, 2008), Polari in the United Kingdom (Baker, 2002), and Gayle in Cape Town (Luyt, 2014) have been studied with emphasis on providing clear linguistic descriptions. In all these sociolects, translators are helpful aids for researchers that are not speakers.

In the country, recent developments in machine translation geared towards data-centric approaches. Research works such as the Philippine component of ASEANMT (Nocon et al., 2014) and the Ilocano-English bidirectional translator (Bautista et al., 2015) utilized statistical phrase-based engines, both garnering a BLEU score of at least 32 units. These were made possible with the availability of existing parallel corpora. In the past, expert-centric systems were developed (Roxas et al., 2008), relying on expert knowledge to develop rules and resources. Each approach has its own advantage and disadvantage depending on the level of text to be translated.

3. Methodology

The methodologies of the research can be summed up using the stages shown in Fig. 1. From a databank of language resources, tweets with Beki words were collected using specific Beki words as filters. Language models were then generated to evaluate the level of Beki speak that the system has to cover. From the collected tweets and language models, the required LanguageTool resources were developed using modern tools and literature as aid. After integrating the language resources, the plug-in was evaluated for recall.

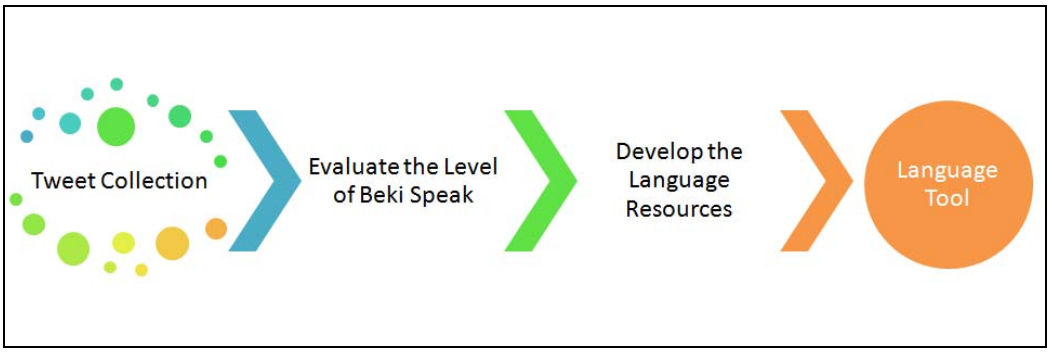

Fig. 1. Stages of the research from tweets rule files in LanguageTool

4. Beki Speak Translator[3]

LanguageTool[4] is a rule-based engine that is freely available online and is primarily used for grammar checking purposes. It supports the following languages: Asturian, Belarusian, Breton, Catalan, Chinese, Danish, Dutch, English , Esperanto, French, Galician, German, Greek, Icelandic, Italian, Japanese, Khmer, Lithuanian, Malayalam, Persian, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish, Tagalog, Tamil, and Ukrainian. For each language support, the following are needed: a rule file, a word list, and a trigram profile. It uses the trigram profile to automatically detect the language of the text input. The word list is used for tagging purposes based on the tags, it uses the rule file to check the grammar and provide suggestions. The succeeding text detail how the Beki speak translator was developed.

4.1. Corpus

Approximately 28 million tweets were taken from a databank of Philippine language resources (Oco et al., 2014). To only get tweets with Beki words, the following were used as filters[5]: jinet, ditey, imbyerna, wetpaks, kayiz, imbey, kemerut, keribels, itey, sinetch, chenelin, churvalu, eklavu, mudrakels, pudra, junakis. Several Beki words were not used as filters as queries result with false positives – tweets that contain Beki words but are not used in a Beki sense (sample tweets are shown in Table 1). The resulting corpus is composed of 4,837 tweets ranging from February 17, 2013 to November 20, 2014.

[3] The resources will be made available at: http://bit.ly/1MpcFoT

[4] LanguageTool is available at: https://languagetool.org/

[5] The initial word list were taken from the YouTube channel of Beki Mon: https://www.youtube.com/channel/UCGJT0kxTg6sV-VdE_7izBYw

Table 1. List of False Positives and Sample Tweets[a]

Beki word	Sample tweets resulting from the query
chaka	• chaka! needed authenticity right up front. so blessed she said yes. pitch to her was that noni would aspire to be like her. #damnnoni • What If Max B & Future made a song together I prolly would just cry & say thank you to the cover like Kanye did to that pic of Chaka Khan .
echos	• Berowka, energy kuno. #echos #berocca #vitaminc (at @XYZ) [pic] — HTTP • Just because its raining. ?????? #NoMorePain #JustLove #Echos ???? regram @XYZ ???? @… HTTP
Chos	• @XYZ it will never be the same wo me chos • Catch my sister (chos) @XYZ in Robinsons Antique later at 11am Thank u sir @XYZ <3
gora	• Gora Galdakaoko gaztiak, kabennnnnnnndio! Mila muxu! :) • När min moster frågar om vi ska göra likadana tatueringar :)))) jag är på!

[a]following research ethics, usernames were removed and replaced with XYZ; URLs were replaced with HTTP.

To evaluate the level of Beki speak that the system has to cover, n-gram models (where *n* ranges from 1 to 3) were generated using SRILM[6]. Modifying the definition provided by Kondrak (2005), the scientific formulation for an n-gram is given in the succeeding text:

> given a string $X = \{x_1...x_k\}$, word sequences $Z = \{z_1...z_n\}$ is an n-gram of X if there exist a strictly incrementing sequence $i_1...i_n$ of indices of X such that for all $j = 1...n$, $X_{i_j} = Z_j$.

In simpler terms, an n-gram is an n-word slice of a sentence. The resulting language models are shown in Table 2. Based on the results, "itey" is the most frequently used Beki word followed by "ditey" and "keribels". It can also be noted that the use of Beki speak is limited only to nonce borrowing and not on the full sentence level (as seen in the 3-grams). This is important as this supports the decision to use a rule-based approach instead of a statistical-approach as translation is not conducted on the phrase level.

[6] SRILM stands for Stanford Research Institute Language Modeling toolkit. It is available at: http://www.speech.sri.com/projects/srilm/

Table 2. Top 20 1-gran, 2-grams, 3-grams

Rank	1-gram		2-gram		3-gram	
	n-gram	Freq.	n-gram	Freq.	n-gram	Freq.
1	na	1,835	sinetch itey	379	sinetch itey na	78
2	sa	1,114	ditey sa	223	keribels mo yan	73
3	ko	1,039	ang jinet	220	ang jinet jackson	54
4	itey	979	keribels lang	164	keribels lang yan	32
5	ditey	856	itey na	136	ditey sa bahay	22
6	keribels	813	na itey	106	keribels na yan	22
7	ang	778	keribels na	104	pero keribels lang	20
8	sinetch	760	keribels mo	101	punta ka ditey	20
9	lang	669	jinet jackson	95	ditey sa baler	20
10	ng	665	ko na	93	sinech itey na	20
11	hahaha	634	mo yan	90	ditey sa twitter	19
12	haha	575	na lang	76	ang jinet sa	18
13	ako	571	keribels yan	73	ang jinet jinet	18
14	mo	531	sinetch itech	70	ditey sa balur	17
15	yan	415	ako sa	63	ko ditey sa	15
16	ka	399	ka na	61	mudra at pudra	15
17	imbyerna	396	di ko	60	sinetch itey ang	13
18	naman	369	na yan	59	miss ko na	11
19	pa	396	si pudra	58	happy mudrakels day	10
20	jinet	371	pero keribels	56	ditey na lang	10

As for the distribution of the language model, it follows Zipf's law, as seen in Figure 2. The x-axis refers to the rank while the y-axis refers to the frequency. The result is comparable to that of the Filipino language, indicating that for the training data used, certain Filipino words were replaced with Beki speak supporting the notion that it is a form of code-mixing. A simple word pair can be used for this level of translation.

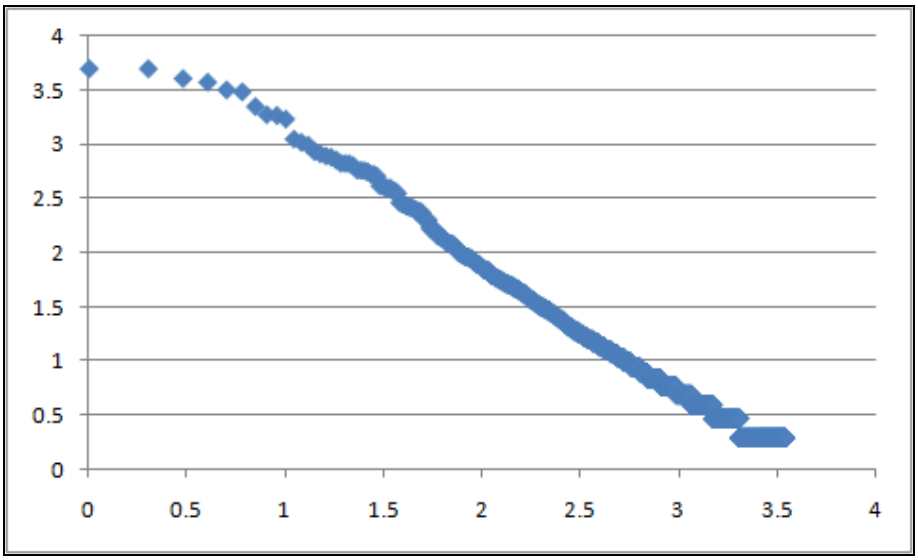

Figure 2. Log-log scatter plot of the 1-grams

4.2. Trigram profile and word list

From the tweets, trigram profiles were then generated as these are needed to perform language identification – the process of automatically classifying which language a text input is in. A trigram profile is composed of character trigrams or 3-character slices of a word. As an example, the list of trigrams that can be generated from the word "beki" are the following: {_be,bek,eki,ki_}. Table 3 shows the top 10 trigrams for Filipino (Oco et al., 2015) and top 10 trigrams for Beki speak. It can be noted that both have the same trigram "ng_" which can be attributed to the determiners "ng" and "ang". However, for Beki speak, the succeeding trigrams reflect Internet language and the sociolect as can be seen with trigrams "aha" and "hah" for the onomatopoeia of laughter (e.g., ahahaha, hahaha) and the trigrams "ey_" and "ite" for the words "itey" and "ditey".

Table 3. Top 10 trigrams for Filipino and top 10 trigram for Beki speak

Rank	Filipino		Beki speak	
	Trigram	Freq.	Trigram	Freq.
1	ng_	706,778	ng_	4,542
2	ang	446,718	aha	4,357
3	_sa	215,713	hah	4,229
4	sa_	206,413	_na	3,607
5	_na	192,546	na_	2,642
6	_ng	181,555	ang	2,623
7	_an	159,343	_ha	2,346
8	_pa	144,798	ey_	2,217
9	ay_	140,278	ite	2,040
10	an_	139,834	ha_	1,942

From the language model, Beki words were manually identified and a table was constructed to serve as word list. The table has three columns – the Beki word, its Filipino translation, and the category. Sample entries are shown in Table 4. For the different categories, the schema provided by the Komisyon sa Wikang Filipino[7] (KWF) was used. For the Filipino equivalent, the YouTube channel of Beki Mon and the resources provided by KWF was used.

Table 4. Sample Entries for the Word List

Beki word	Filipino equivalent	Category
mudra	nanay	Pagpapalit
nota	ari ng lalaki	Panghihiram
thunderbots	matanda	Neolohismo o Paglikha

[7] From the book entitled "Mga Salitang homosekswal : isang pagsusuri", released by KWF in 2004.

4.2. Rule File

The rule file is an XML file composed of the following attributes: *pattern* to be matched, *suggestion* to be provided when the pattern matches the input, and *examples* (a sample is shown in Fig. 1). In terms of translation, pattern refers to the source text while suggestion refers to the target text. In this case, Beki speak and Filipino, respectively. Based on the word list, the different attributes were automatically generated.

```
    <pattern case_sensitive="no" mark_from="0">
        <token>amber</token>
    </pattern>
    <message><suggestion>beer
    </suggestion></message>
    <short>malayang pagdaragdag</short>
    <example correction="beer" type="amber"
    ></example>
</rule>
```

Figure 3. Sample rule showing the Beki speak "amber"

Aside from token to token translations, rules following certain conventions were also developed. Wdiff[8] in Cygwin[9] was utilized for this purpose. It creates a rule file detailing the changes from one word to another. Samples are shown in Table 5.

Table 5. Sample Filipino to Beki speak conventions

Filipino	Beki speak	Change	Wdiff
bago	jogo	Removed "ba" and change to "jo"	[-b a-]{+j o+}
party	jarty	Removed "p" and change to "j"	[-p-]{+j+}
taba	joba	Removed "ta" and change to "jo"	[-t a-]{+j o+}
tae	shoe	Removed "ta" and change to "sho"	[-t a-]{+s h o+}
salsal	balbal	Removed "s" and change to "b"	[-s-]{+b+}

Once the rule file had been added, an OpenOffice[10] plug-in was generated using Ant[11]. A working screenshot of the plug-in is shown in Figure 4. LanguageTool underlines Beki speak and provides the category and the Filipino translation on a right-click mouse event.

[8] More information about Wdiff are available at: http://www.gnu.org/software/wdiff/

[9] Cygwin can be downloaded at: https://www.cygwin.com/

[10] OpenOffice is a word editor that is freely available. The entire suite can be downloaded at: https://www.openoffice.org/

[11] Apache Ant is available at: http://ant.apache.org/

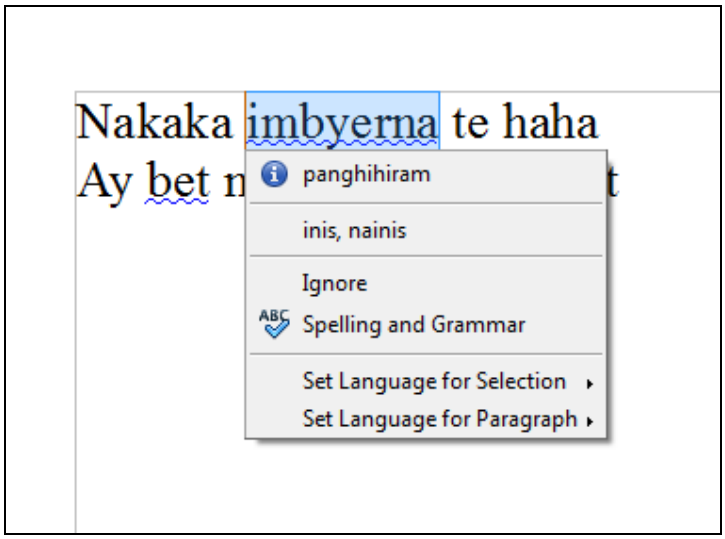

Figure 4. Sample translation for the Beki speak "imbyerna"

5. Evaluation

To evaluate the plugin, 100 sentences – distinct from the Tweet corpus – were used. Out of the 100 sentences, the system was able to detect Beki words in 43 sentences. Analyses of the results reveal that a number of Beki words are not covered by the rule file (e.g., siney, girlalu, girlaloo, watashi, shupatid). These terms are not included in the resources used. There is a need to constantly update the list available as Beki speak constantly evolves and variations exist.

6. Conclusion and Recommendation

We have presented in this paper the development of a Beki speak to Filipino translator. The needed resources, issues, challenges, and computational characteristics were detailed. Evaluation results also reveal that a number of Beki words are not covered by the rule file and not found in the resources used. Being a sociolect that penetrates deeply in the society, there is a need to constantly update existing lists as it evolves in an exponential rate. This work can be extended by developing a digital toolkit for users – both speakers and non-speakers – to be able to add new rules and descriptions. Also, an in-depth analysis is needed to be able to computational crack conventions and etymology of each word. *And world peace, we thank you*.

References

Baker, P. (2002). *Polari – The lost language of gay men*. London: Routledge.

Bautista, J. (2015). Bi-directional Ilocano-English language translator using customized Moses statistical machine translation system (SMTS), In *Proceedings of the 11th National Natural Language Processing Research Symposium*, 18-25. National University, Manila.

Bautista, M. R. (1993). *Bata, Bata ano ang bakla? Isang pag-aaral ukol sa pananaw ng batang Filipino hinggil sa mga bakla* (Published Diploma thesis). De La Salle University, Manila.

Baytan, R. (2000). Sexuality, ethnicity and language: Exploring Chinese Filipino male homosexual identity. *Culture, Health, and Sexuality* 2(4): 391-404.

Casabal, N. (2008). Gay language: Defying the structural limits of English language in the Philippines. *Kritika Kultura* 11: 74-101.

Dones, D. (2015). *Lola na ang lolah mo: Isang deskriptibong pag-aaral at pagtatanghal ng kabaklaan sa panahon ng katandaan* (Published PhD Dissertation). De La Salle University, Manila.

Kondrak, G. (2005) N-gram similarity and distance. In *Proceedings of the 12th International Conference on String Processing and Information Retrieval (SPIRE 2005)*, 115-126. Buenos Aires, Argentina.

Luyt, K. (2014). *Gay language in Cape Town: A study of Gayle – attitudes, history and usage* (Published MA Dissertation). University of Cape Town, Cape Town.

Madula, R. (2014) Ka Laya Rampadora: Mga tala ng kakaibang pagrampa ng isang bakla. *Malay* 26(1): 104-116.

Naber, D. (2003) *A rule-based style and grammar checker* (Published Diploma thesis). Bielefeld University, Bielefeld.

Nocon, N., Oco, N., Ilao, J., & Roxas, R.E. (2014). Philippine component of the network-based ASEAN language translation public service. In *Proceedings of the 7th IEEE International Conference Humanoid, Nanotechnology, Information Technology Communication and Control, Environment and Management (HNICEM)*. Hotel Centro, Puerto Princesa, Palawan.

Oco, N. & Borra, A. (2011). A grammar checker for Tagalog using LanguageTool. In *Proceedings of the 9th Workshop on Asian Language Resources*, 2-9. Shangri-La, Chiang Mai.

Oco, N., Sison-Buban, R., Syliongka, L.R., Roxas, R. E., & Ilao, J. (2014) Ang paggamit ng trigram ranking bilang panukat sa pagkakahalintulad at pagkakapangkat ng mga wika. *Malay* 26 (2): 53-68.

Oco, N., Syliongka, L.R., Allman, T., & Roxas, R. E. (2015) Building resources for Philippine languages. Paper presented at *MAPLEX 2015*. Yamagata, February 09-10.

Ponce, M. J. (2008). Framing the Filipino diaspora: Gender, sexuality, and the politics of criticism. *Philippine Studies* 56(1): 77-101.

Roxas, R. E., Borra, A., Cheng, C., Lim, N.R., Ong, E.C., & Tan, M.W. (2008). Building language resources for a multi-engine English-Filipino machine translation system. *Language Resources and Evaluation* 42(2): 183-195.

Rudwick, S. & Ntuli, M. (2008) IsiNgqumo – Introducing a gay Black South African linguistic variety. *Southern African Linguistics and Applied Language Studies* 26(4): 445–456.