Malay 26.1 (2014): 53-68

Ang Paggamit ng Trigram Ranking Bilang Panukat sa Pagkakahalintulad at Pagkakapangkat ng mga Wika / *Trigram Ranking: Metric for Language Similarity and Clustering*

Nathaniel Oco, Raquel Sison-Buban, Leif Romeritch Syliongka, Rachel Edita Roxas, Joel Ilao
Pamantasang De La Salle, Pilipinas
nathan.oco@delasalle.ph , rakkisisonb@yahoo.com , leif.syliongka@delasalle.ph , rachel.roxas@delasalle.ph , joel.ilao@delasalle.ph

Ang *trigram* ay tatlong magkakasunod na titik na bahagi ng isang salita. Bilang halimbawa, ang mga trigram na mababuo sa salitang "tatlo" ay ang mga sumusunod: *tat*, *atl*, at *tlo*. Iminumungkahi sa pag-aaral na ito ang paggamit ng *trigram ranking*, isang prosesong gumagamit ng trigram, bilang panukat sa pagkakahalintulad ng mga wika. Sa prosesong ito, [1] kinokolekta ang mga dokumentong gagamitin bilang *training data*; [2] ginagawan ng *trigram profile* gamit ang training data; at [3] kinokompyut ang pagkakahalintulad gamit ang ranggo ng mga *trigram*. Iminumungkahi rin ang paggamit ng *k-means clustering* para pangkatin ang mga wika ayon sa kanilang trigram ranking. Sa pag-aaral na ito, kumolekta ng mga teksto mula sa Internet gamit ang mga awtomatikong pamamaraan: [1] paggamit ng isang xml to text converter para mangolekta ng mga artikulo mula sa English at Tagalog Wikipedia, [2] paggamit ng isang webcrawler para mangolekta ng mga artikulo mula sa mga pahayagan, [3] paggamit ng isang twitter API para kumolekta ng mga tweet, at [4] paggamit ng isang bot para mangolekta ng game chat mula sa Ragnarok, isang online na laro. Kumolekta rin ng mga dokumento mula sa isang parallel na korpus at isang mula naman sa online na korpus. Saklaw sa pag-aaral na ito ang walong wika: Bikol, Cebuano, Hiligaynon, Iloko, Pampanga, Pangasinan, Tagalog, at Waray. Batay sa resulta, galing sa iisang subgrupo ang mga pares ng wika na may magkakalapit na trigram ranking: [1] galing sa iisang subgrupo ang Bikol, Cebuano, Hiligaynon, Tagalog, at Waray at [2] galing naman sa iisang subgrupo ang Iloko at Pangasinan; samantalang [3] nahihiwalay naman ang Pampanga sa isang subgrupo. Maaari ring gamitin ang metrong ito upang sukatin ang pagkakahalintulad ng iba pang wika ng Pilipinas[1].

Mga Susing Salita: pagkakahalintulad ng mga wika, pagpapangkat sa mga wika, trigram ranking, trigram profile, trigram, mga wika sa Pilipinas

A trigram is a 3-letter sequence of a word. As an example, the lists of trigrams that can be generated from the word "tatlo" are the following: tat, atl, and tlo. Presented in this research is trigram ranking, a metric for language similarity. It involves [1] collecting huge amounts of texts as training data, [2] generating trigram profiles from the training data, [3] and computing for language similarity using trigrams. Also presented is the use of k-means clustering to group languages based on their trigram ranking. In this study, the Internet was mined for texts using automatic means: [1] an XML to text converter was used to gather English and

Copyright © 2014 by De La Salle University

Filipino Wikipedia articles; [2] a webcrawler was used to collect online news articles; [3] a twitter API was used to collect tweets; and [4] a bot was used to collect chat logs from Ragnarok, an online game. Documents from a parallel corpus and documents from an online corpus were also collected. The following languages were used as test bed: Bikol, Cebuano, Hiligaynon, Iloko, Pampanga, Pangasinan, Tagalog, and Waray. Based on the results, language pairs with trigram rankings close to each other come from the same subfamily of languages: [1] Bikol, Cebuano, Hiligaynon, Tagalog, and Waray come from one subgroup; [2] Iloko and Pangasinan come from one subgroup; and [3] Pampanga comes from another subgroup. Trigram ranking can be used to measure which Philippine languages are closely-related.

Keywords: *language similarity, language clustering, trigram ranking, trigram profile, trigram, Philippine languages*

PANIMULA

Ang pagkakahalintulad ng mga wika ay tumutukoy sa lalim ng relasyon na nabubuo sa pagitan ng mga ito (Oco, Ilao, Roxas, at Syliongka, "Measuring Language Similarity"). Halimbawa nito ang *0.75* na *lexical similarity* ng Pranses at Espanyol[2]. Ang lexical similarity naman ay isang panukat sa pagkakahalintulad ng mga wika na inilathala ng Ethnologue (Lewis), isang organisasyong sumusubaybay sa mga wika. Mas lumalalim ang relasyon habang papalapit ang valyu sa *1.0*. Upang makuha ang mga valyung kagaya nito, kadalasang ginagamit ng mga dalubhasa ang mga *cognate*. Ang cognate ay mga pares ng salita na may iisang pinagmulan. Halimbawa nito ang Ingles na *night* at ang Alemang *naught* na kapwa tumutukoy sa gabi. Marami na ang naisagawang pag-aaral kaugnay nito. Kabilang na rito ang akda nina McMahon at McMahon; Downey, Hallmark, Cox, Norquest, at Lansing; at nina Do, Roth, Sammons, Tu, at Vydiswaran.

Ipinakita nina McMahon at McMahon ang paggamit ng *phylogenetics analysis* (164) para sukatin ang pagkakahalintulad ng mga wika. Sa prosesong ito, inililista ang mga cognate at iba pang atribyut – hal. kung may *verb aspect* o *reduplication* ang wika – at ginugrupo ang mga wika na may parehas na atribyut. Kung maraming similar na atribyut ang dalawang wika, maaaring sabihin na parehas sila ng pinagmulan o galing sa iisang pamilya. Ito ang madalas na ginagamit upang malaman kung aling wika ang magkakamag-anak. Sa ibang pag-aaral naman (Downey, Hallmark, Cox, Norquest, at Lansing 348; Do, Roth, Sammons, Tu, at Vydiswaran 2), binibilang ang dami ng magkakatulad na titik bilang panukat ng pagkakahalintulad ng mga cognate. Kung maraming magkakatulad na titik, maaaring sabihing galing sa iisang pamilya ang mga wika. Sa lahat ng mga pag-aaral na ito, kinakailangan ang dalubhasang kaalaman at listahan ng cognates. Dagdag pa rito ang pagiging matrabaho at ang oras na gugugulin kung mano-mano ang pagtatala. Sa mga wikang walang nailathalang listahan ng cognates, kailangan ng isang panukat na ginagamitan lamang ng mga makukuhang dokumento at hindi lubos na umaasa sa dalubhasang kaalaman.

Nakasaad sa pag-aaral na ito ang *trigram ranking* (TR), isang prosesong ginawa ng mga may-akda para sukatin ang pagkakahalintulad ng mga wika nang hindi gumagamit ng cognates. Saklaw sa pag-aaral na ito ang walong wika na makikita sa Talahanayan 1: Bikol, Cebuano, Hiligaynon, Iloko, Pampanga, Pangasinan, Tagalog, at Waray. Sa kasalukuyan, wala pang lexical similarity na nailathala para sa mga wikang ito. Patunay lamang sa kakulangan ng didyital na materyal.

Talahanayan 1. Mga wikang saklaw ng pag-aaral

Wika ayon sa ISO 639-2	Language Code ayon sa ISO 639-2	Dami ng gumagamit ayon sa pinakahuling tala ng Ethnologue
Cebuano	ceb	4,580,000
Hiligaynon	hil	15,810,000
Iloko	ilo	5,770,000
Pampanga	pam	7,016,400
Pangasinan	pag	1,905,430
Tagalog	tgl	1,162,140
Waray	war	24,216,200

Trigram

Ang trigram ay tatlong magkakasunod na titik na bahagi ng isang salita. Bilang halimbawa, ang mga trigram na mabubuo sa salitang "tatlo" ay ang mga sumusunod: *tat*, *atl*, *tlo*. Para mapagkilanlan ang trigram na makikita sa unahan, sa gitna, at sa hulihan ng isang salita, idinadagdag ang isang *underscore* sa unahan at hulihan nito. Sa ganitong paraan, nagiging "_tatlo_" ang salitang "tatlo" at ang mga trigram na mabubuo sa salitang ito ay ang mga sumusunod: *_ta*, *tat*, *atl*, *tlo*, *lo_*. *Trigram profile* naman ang isang listahan na naglalaman ng mga trigram na makikita sa isang teksto. Nakalagay rin dito kung gaano karaming beses lumabas sa teksto ang isang trigram. Kadalasan, kinakatawan ng isang trigram profile ang isang wika.

Hindi ito ang unang pagkakataon na gagamitin ang trigram para pag-aralan ang mga wika sa Pilipinas. Gumamit sina Dimalen at Roxas (147) ng webcrawler upang awtomatikang mangolekta sa internet ng mga dokumento at gumamit sila ng mga trigram profile upang awtomatikong maklasipika ang mga dokumento kung nasusulat ba sa Bikol, Cebuano, o Tagalog. Ginamit naman nina Oco at Roxas (233) ang trigram profile ng Tagalog at Ingles upang awtomatikong malaman ang mga *code-switch* sa isang dokumento. Patunay lamang ito ng dumadaming bilang ng mga pag-aaral na *interdisciplinary*.

Trigram Ranking

Sa prosesong trigram ranking, [1] kinokolekta ang mga dokumentong gagamitin bilang *training data*; [2] ginagawan ng *trigram profile* gamit ang training data; at [3] kinokompyut ang pagkakahalintulad gamit ang ranggo ng mga *trigram*.

Pagkolekta ng Korpus

Training data ang terminong tumutukoy sa mga dokumentong ginagamit upang makagawa ng trigram profile ng isang wika. Inililista ang lahat ng trigram sa training data at binibilang kung gaano ito karami.

Sa pag-aaral na ito, unang kinolekta ang mga teksto mula sa Internet gamit ang mga awtomatikong pamamaraan. Kinolekta ang mga artikulo mula sa English at Tagalog Wikipedia, mga artikulo mula sa mga pahayagan, mga tweet, at mga game chat. Nakasaad sa Talahanayan 2 ang laki ng bawat korpus.

Ang mga artikulo sa Wikipedia naman ay malayang magagamit. Ang buong English Wikipedia ay makukuha sa website na ito: http://dumps.wikimedia.org/enwiki/. Ganoon din ang buong Tagalog Wikipedia na makukuha naman sa website na ito: http://dumps.wikimedia.org/tlwiki/. Gumamit ng isang xml to text converter, (na maaaring ma-download sa website na ito: http://www.evanjones.ca/software/

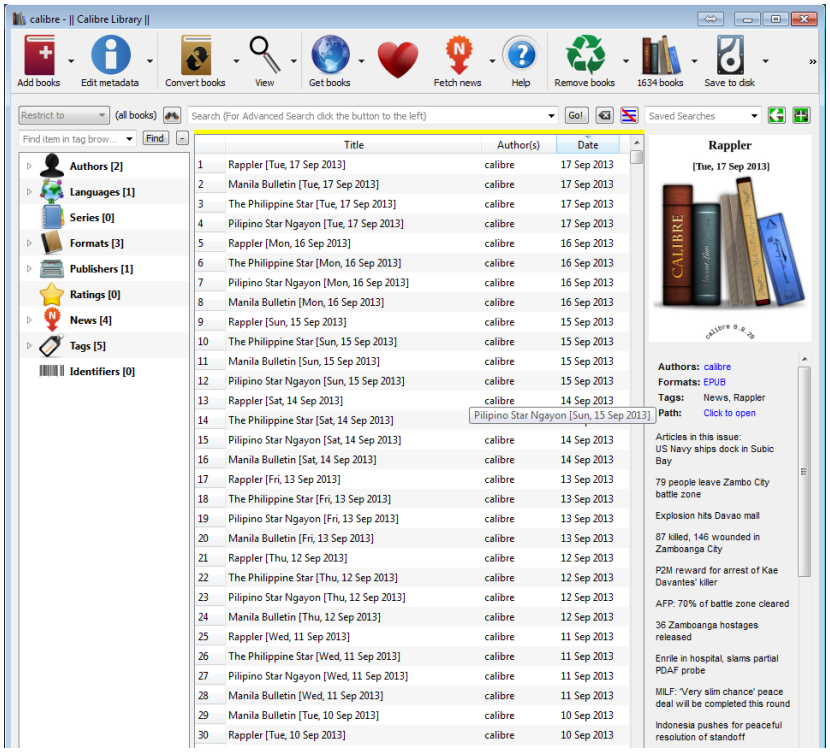
Talahanayan 2. Laki ng bawat korpus

Korpus	Laki ng Korpus	
English Wikipedia	Dami ng salita	10,112,853
Tagalog Wikipedia	Dami ng salita	2,991,474
Mga pahayagan	Dami ng artikulo	1,797
Mga tweet	Dami ng tweet	13,898,375
Mga game chat log	Dami ng linya	842,378

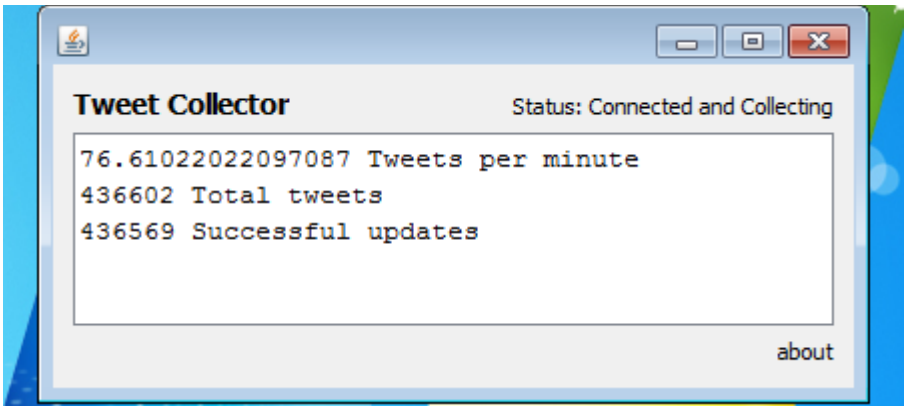
wikipedia2text.html,) para mangolekta ng mga artikulo mula sa English at Tagalog Wikipedia.

Makikita rin sa Internet ang mga artikulo mula sa pahayagan. Ginamit ang Calibre, isang webcrawler o computer program na may kakayahang kumuha sa Internet ng teksto, para mangolekta ng mga artikulo mula sa mga pahayagan. Maaaring namang ma-download ang Calibre sa website na ito: http://calibre-ebook.com/. Makikita sa Larawan 1 ang isang screenshot ng Calibre. Sa pag-aaral na ito, kinolekta ang mga artikulo mula sa *Manila Bulletin, Pilipino Star Ngayon, Rappler,* at sa *The Philippine Star.*

Ang Twitter ay isang uri ng Social Networking Site kung saan maaaring magbasa at magpadala ng mga mensaheng may habang 160 na titik na kung tawagin ay *tweet*. Ang mga tweet ay maaaring publiko o pribado, depende sa taong nagpadala nito. Para makuha ang mga publikong tweet, gumawa ng isang computer program para kumolekta rito. Makikita sa Larawan 2 ang ginawang Tweet Collector, isang computer program na gumagamit ng Twitter4J. Ito ay may kakayahang kumuha sa Internet ng mahigit 70 tweet kada minuto.

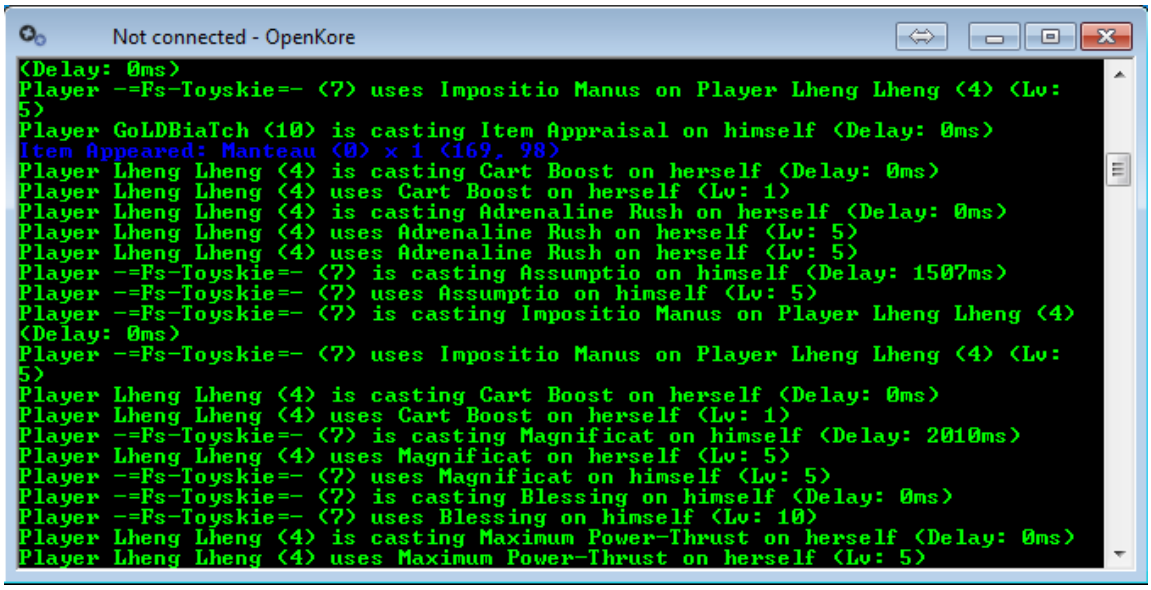


Larawan 1. Halimbawang screenshot ng Calibre



Larawan 2. Screenshot ng Tweet Collector

Maliban sa Wikipedia, pahayagan, at Twitter, nakita rin ang online gaming bilang mapagkukunan ng teksto. Ang Ragnarok, isang uri ng massive multiplayer online role playing game o MMORPG, ay isang sikat na laro sa Pilipinas. Gumamit ng OpenKore, isang uri ng bot, para kumolekta ng game chat o usapan ng mga manlalaro. Makikita sa Larawan 3 ang screenshot ng OpenKore. Ang mga bot ay computer program na may kakayahang maglaro na parang tao.



Larawan 3. Halimbawang screenshot ng OpenKore

Kumolekta rin ng mga dokumento mula sa mga magagamit na korpus: ang parallel na korpus ng ASEAN MT-Phil o ang bahaging nasa-Filipino ng ASEAN MT (Oco, Ilao, Roxas, Sison-Buban, Bonus 12) at ang PALITO (Dita, Roxas, at Inventado 648).

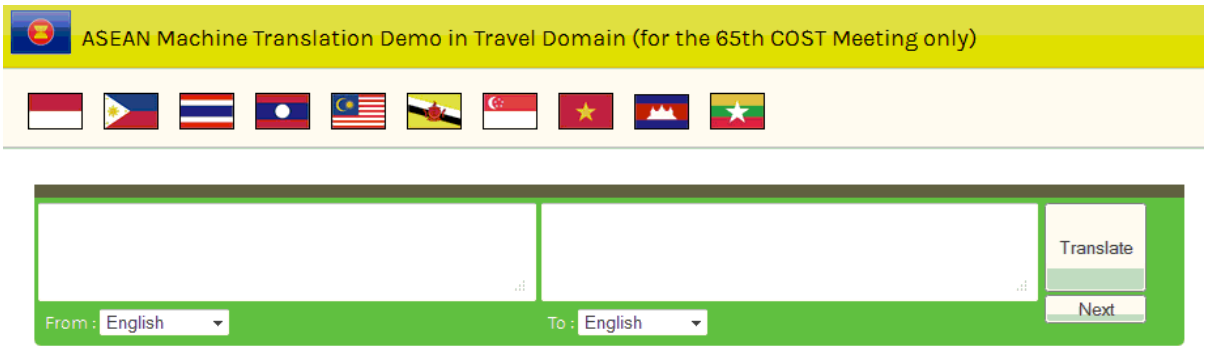
Ang ASEAN MT o ang "Network-based ASEAN Language Translation Public Service Project" ay isang proyektong pinangungunahan ng Thailand National Electronics and Computer Technology Center o NECTEC na naglalayong gumawa ng isang machine translator para sa mga wika sa ASEAN. Ang machine translator

naman ay isang computer program na may kakayahang awtomatikong magsalin mula sa isang wika papunta sa isa pa. Makikita sa Larawan 4 ang isang tumatakbong halimbawa ng ASEAN MT. Makikita rin ito sa website na ito: http://www.aseanmt.org/demo/. Para sa ASEAN MT-Phil, mano-manong isinalin ang mga dokumentong nakuha sa English Wikipedia at ilang dokumentong may kinalaman sa turismo. Makikita sa Talahanayan 3 ang dami ng salita at mapapansin na mas marami ang mga naisalin sa Filipino.

Talahanayan 3. Dami ng salita sa parallel na korpus

Wika	Dami ng Salita
Inggles	3 milyon
Filipino	3.2 milyon

Ang PALITO ay isang website na lagayan ng mga dokumentong hinango sa panitikan at Bibliya. Proyekto ito ng Pamantasang De La Salle-Maynila at puwede itong ma-access sa website na ito: http://ccs.dlsu.edu.ph:8086/Palito/. Makikita sa Larawan 5 ang isang screenshot ng PALITO. Naglalaman din ito ng mga video ng Filipino Sign Language o FSL.



Larawan 4. Screenshot ng ASEAN MT

Larawan 5. Screenshot ng PALITO

Dahil karamihan sa mga tekstong nakuha gamit ang mga awtomatikong pamamaraan ay nakasaad sa Tagalog o Ingles, pati na rin ang Filipinong bahagi ng ASEAN MT, napagdesisyonang gamitin ang mga dokumento sa PALITO na hinango sa Bibliya (Dita, Roxas, at Inventado 648) bilang training data. Makikita sa Talahanayan 4 ang dami ng salita ng training data na ginamit. Hinango ang pangalan ng mga wika at ang language code sa ISO 639-2[3], isang batayan sa pagbibigay-ngalan sa mga wika at grupo ng wika.

Talahanayan 4. Dami ng salita ng training data na ginamit

Wika ayon sa ISO 639-2	Language Code ayon sa ISO 639-2	Dami ng salita
Bikol	bik	120,110
Cebuano	ceb	172,823
Hiligaynon	hil	107,502
Iloko	ilo	88,241
Pampanga	pam	152,733
Pangasinan	pag	124,208
Tagalog	tgl	85,310
Waray	war	121,350

Trigram Profile ng Training Data

May mga computer program na awtomatikong kayang gumawa ng trigram profile. Sa pag-aaral na ito, ginamit ang Apache Nutch[4]. Ginamit din ito nina Oco at Roxas (233) sa kanilang pag-aaral kung saan awtomatikong kinolekta ang mga artikulo sa Wikipedia at ginamit bilang training data.

Kasama ang letrang 'ñ', may 27 na magkakaibang titik ang alpabetong Filipino. Ang letrang 'ng' naman ay binubuo ng dalawang titik – ang titik 'n' at ang titik 'g'. Dahil dito, halos 21,200 na magkakaibang kombinasyon ng trigram ang pwedeng magawa (Oco, Ilao, Roxas, at Syliongka, "Measuring Language Similarity"); mula *_a_* at *_aa* hanggang *zzy* at *zzz*. Para sa pag-aaral na ito, tanging ang unang 100 na trigram na may pinakamataas na bilang lamang ang ginamit. Ang iba pang mga trigram ay may mababababang bilang at hindi masasabing naglalarawan sa isang wika (Oco, Syliongka, Roxas, at Ilao, "Dice's Coefficient"). Makikita sa Talahanayan 5 ang unang 20 trigram na makikita sa trigram profile ng bawat wika. Nakaranggo ito base sa rami ng bilang o *frequency count*; nangangahulugan na nasa unang ranggo ang may pinakamaraming bilang.

Talahanayan 5. Unang 20 trigram na makikita sa trigram profile ng walong wika

Ranggo	bik	ceb	hil	ilo	pam	pag	tgl	war
1	an_	_sa	ng_	ti_	ng_	an_	ng_	an_
2	_na	ng_	ang	iti	ing	ay_	ang	nga
3	_sa	sa_	_sa	_a_	ang	_na	_sa	_ng
4	_ka	ang	ga_	_ti	an_	_sa	sa_	ga_
5	ng_	an_	_ka	_ke	_ka	_ka	_na	_sa
6	_an	ga_	nga	_it	at_	_ta	_ka	_an
7	_ni	_an	sa_	na_	_at	_si	_an	on_
8	ang	_ka	_ng	_na	_ki	_a_	at_	_ka
9	_ma	nga	_an	_da	kin	_ma	_ng	_na
10	na_	_na	_ma	an_	_ma	na_	an_	_ha
11	sa_	on_	_na	_ma	_in	tan	_ma	_ma
12	_si	_ma	an_	ana	_na	_to	_si	sa_
13	nin	ug_	on_	et_	_di	en_	_pa	_pa

Talahanayan 5

14	in_	_pa	san	nga	_ke	ed_	_at	gan
15	iya	_ug	_si	agi	ala	_ed	ala	ya_
16	nag	_ng	_pa	_ka	_a_	ray	_ni	ay_
17	_pa	ala	ag_	dag	ung	ara	ga_	mga
18	ga_	_si	nag	en_	kan	ang	ya_	_mg
19	ya_	_ni	la_	_ng	din	_pa	ina	ha_
20	kan	_gi	ya_	_pa	ana	to_	na_	Iya

Mapapansin na *an_* ang nangungunang trigram sa Bikol, Pangasinan, at Waray; *_sa* sa Cebuano; *ng_* sa Hiligaynon, Pampanga, at Tagalog, at *ti_* sa Iloko. May kinalaman na rin ito sa mga salitang madalas na ginagamit sa mga wikang ito. Makikita sa Talahanayan 6 at Talahanayan 7 ang sampung madalas na ginagamit na salita. Awtomatikong binilang ito gamit ang Stanford Research Institute Language Modeling Toolkit (SRILM)[5]. Isang computer program ang SRILM na may kakayahang bilangin ang dami ng isang salita sa isang teksto. Para sa pag-aaral na ito, parehas na mga dokumento ang ginamit bilang training data. Batay sa resulta ng SRILM, mapapansin ang mga sumusunod:

- sa Bikol, ang mga salitang *an* at *kan* ay parehas na may *an_* na trigram;
- sa Cebuano, ang nangungunang salita na *sa* ay may *_sa* na trigram;
- sa Hiligaynon, ang mga salitang *ang* at *sang* ay may *ng_* na trigram;
- sa Iloko, ang nangungunang salita na *ti* at ang salitang *iti* ay may *ti_* na trigram;
- sa Pampanga, ang mga salitang *king, ing, ding, ning,* at *nang* ay may *ng_* na trigram;
- sa Pangasinan, ang salitang *tan* ay may *an_* na trigram;
- sa Tagalog, ang mga salitang *ang, ng,* at *kanyang* ay may *ng_* na trigram; at
- sa Waray, ang mga salitang *an, han, ngan,* at *san* ay may *an_* na trigram.

Patunay lamang ito na kinakatawan ng isang trigram profile ang isang wika.

Talahanayan 6. Listahan ng sampung madalas na ginagamit na salita sa bik, ceb, hil, at ilo

Ranggo	bik		ceb		Hil		ilo	
	Salita	**Dami**	**Salita**	**Dami**	**Salita**	**Dami**	**Salita**	**Dami**
1	na	15,280	sa	30,720	nga	17,668	ti	20,684
2	sa	14,847	ang	13,993	sa	17,280	a	20,093
3	an	14,478	nga	10,863	ang	14,616	iti	12,496
4	kan	11,477	ug	9,723	sang	12,904	nga	7,707
5	mga	6,982	mga	7,110	mga	7,046	ni	6,826
6	nin	5,852	ni	4,103	kag	6,884	ket	5,051
7	asin	4,081	si	3,583	iya	5,411	ken	5,007
8	si	2,854	ka	2,755	ni	3,800	dagiti	3,961
9	dai	2,479	iyang	2,686	si	3,647	no	2,365
10	ni	2,464	siya	2,550	ka	3,487	met	2,098

Talahanayan 7. Listahan ng sampung madalas na ginagamit na salita sa pam, pag, tgl, at war

Ranggo	pam		pag		tgl		war	
	Salita	Dami	Salita	Dami	Salita	Dami	Salita	Dami
1	at	10,476	na	8,255	sa	18,620	nga	19,177
2	king	9,510	ed	8,198	ang	17,891	an	12,376
3	ing	8,330	a	7,216	ng	15,228	han	8,488
4	a	8,154	ya	6,750	na	11,119	ha	8,402
5	na	4,761	so	6,268	at	9,747	mga	7,663
6	ding	4,707	tan	6,244	mga	8,535	sa	5,716
7	ning	4,503	to	4,532	ay	3,205	ngan	5,532
8	ya	3,042	say	3,004	ni	2,987	na	3,427
9	nang	2,404	saray	2,999	ito	2,628	san	3,069
10	e	2,055	no	2,839	kanyang	2,302	hin	2,738

Pagkakahalintulad ng mga Wika

Ang paggamit ng trigram para sukatin ang pagkakahalintulad ng mga wika ay hinango sa mga naunang pag-aaral – ang *number of common trigrams* (Oco, Ilao, Roxas, at Syliongka, "Measuring Language Similarity") at ang *Dice's coefficient on trigram profiles* o DCTP (Oco, Syliongka, Roxas, at Ilao, "Dice's Coefficient") – ng ilan sa may-akda. Ipinakita nina Oco, Ilao, Roxas, at Syliongka ("Measuring Language Similarity") na puwedeng gamitin ang mga trigram profile para sukatin kung may pagkakahalintulad ang mga wika. Iminungkahi naman sa isa pang pag-aaral (Oco, Syliongka, Roxas, at Ilao, "Dice's Coefficient") na gamitin ang trigram sa pagsukat ng pagkakahalintulad ng mga wika. Sa DCTP (makikita sa ekwasyon bilang (1)), binibilang ang magkakaparehas na trigram sa dalawang trigram profile at ginagawang porsiyento ito. Dahil pantay-pantay ang bigat ng lahat ng trigram, hindi nabibigyang diin ang mga trigram na may matataas na ranggo.

$$DCTP = \frac{2(X \cap Y)}{X + Y} \qquad (1)$$

Makikita sa ekwasyon bilang (2) ang pormula para makuha ang trigram ranking. Ang *k* ay tumutukoy sa dami ng trigram na ginamit sa bawat trigram profile. Ang *Trigram*(X_i) naman ay tumutukoy sa trigram na nasa ranggong *i* sa wikang *X*. Halimbawa, ang *Trigram*(bik_2) ay tumutukoy sa trigram na *_na*. Ito ang trigram na nasa ranggong *2* sa wikang *bik* (sumangguni sa Talahanayan 5). Ang *Ranggo*(*Trigram*(X_i), *Y*) naman ay tumutukoy sa ranggo ng *Trigram*(X_i) sa wikang *Y*. Halimbawa, ang *Ranggo*(*Trigram*(bik_2),*war*) ay may valyu na *9*. Tumutukoy ito sa ranggo ng trigram na *_na* sa wikang *war* (sumangguni sa Talahanayan 5).

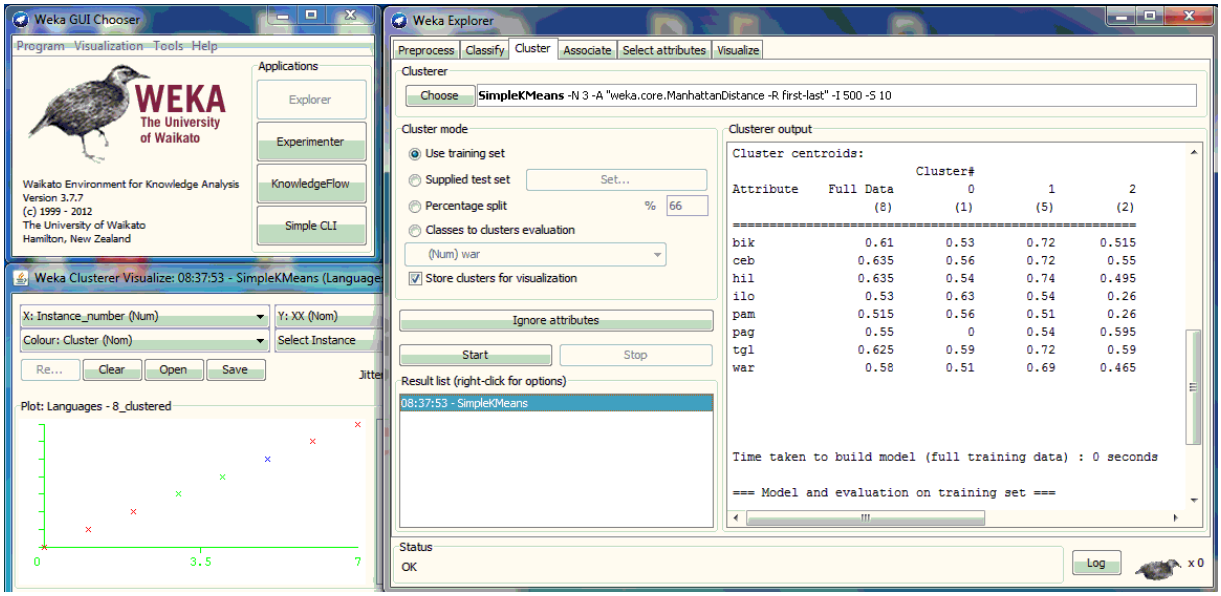
$$TR = \frac{\sum_{i=1}^{k}(k+1-Ranggo(Trigram(X_i),Y)) + \sum_{i=1}^{k}(k+1-Ranggo(Trigram(Y_i),X))}{2*\sum_{j=1}^{k} j} \qquad (2)$$

Sa trigram ranking, binibigyang diin ang mga trigram na may matataas na ranggo. Dito, isinasaalang-alang ang ranggo ng isang trigram. Nangangahulugan na kung mas mataas ang ranggo, mas mataas din ang bigat ng isang trigram. Ang mga sumusunod na proseso ang paraan sa pagkuha ng trigram ranking sa pagitan ng dalawang wika (ang *X* at ang *Y*). Sa bawat trigram ng wikang *X*, kinukuha ang ranggo nito sa wikang *Y*. Kung wala sa wikang *Y* ang naturang trigram, ginagawang *0* ang ranggo nito. Kinukuha rin ang inverted na valyu ng mga ranggo at ang suma total ng mga ito. Hinahati ang nakuha sa 5,050 na siyang suma total naman ng lahat ng ranggo (mula 1 hanggang 100). Pagkatapos, kinukuha

ang average ng mga valyu para sa mga pares ng wika. Ito ang trigram ranking.

Paggrupo sa mga Wika

Matapos makuha ang trigram ranking ng bawat wika, ginagamitan ito ng k-means clustering (Borra, Cheng, Roxas, at Ona 199) para pangkatin ang mga ito batay sa kanilang trigram ranking. Sa prosesong ito, pinapangkat ang mga wika na may magkakalapit na trigram ranking. May mga computer program na kayang awtomatikong magawa nito. Sa pag-aaral ding ito, ginamit ang Weka para kalkulahin ang pagkakapangkat ng mga wika gamit ang k-means clustering na may Manhattan distance. Ang Weka ay isang machine learning software na ginawa ng University of Waikato sa New Zealand. May kakayahan itong magmina ng impormasyon sa malalaking datos. Makikita sa Larawan 6 ang screenshot ng Weka. Maaaring ma-download ang Weka sa website na ito: http://www.cs.waikato.ac.nz/ml/weka/.



Larawan 6. Screenshot ng Weka

Resulta at Diskusyon

Gamit ang training data, kinuha ang trigram ranking ng mga wika at ginamitan ng k-means clustering para makuha ang pagkakapangkat ng mga ito.

Similaridad ng mga Wika

Makikita sa Talahanayan 8 ang trigram ranking ng mga wika; ang mga wika sa kaliwa ang nagsilbing wikang *X* at ang mga wika sa itaas ang nagsilbing wikang *Y*. Symmetrical ang valyu para sa mga wika, nangangahulugan na parehas ang trigram ranking kahit pagbaligtarin ang wikang *X* at wikang *Y*.

Kung pagbabasehan ang kahulugan ng pagkakahalintulad ng mga wika, lumalalim ang relasyon habang papalapit sa *1.0* ang valyu. Nangangahulugang pinakamalapit ang isang partikular na wika sa wikang may pinakamataas nitong trigram ranking. Makikita sa Talahanayan 9 ang walong wika at kung saang wika sila may pinakamataas na trigram ranking.

Talahanayan 8. Valyu ng trigram ranking

Wika	bik	ceb	hil	ilo	pam	pag	tgl	war
bik	1.00	0.72	0.73	0.52	0.51	0.53	0.72	0.69
ceb	0.72	1.00	0.80	0.55	0.55	0.56	0.76	0.71
hil	0.73	0.80	1.00	0.54	0.45	0.54	0.74	0.75
ilo	0.52	0.55	0.54	1.00	0.52	0.63	0.60	0.51
pam	0.51	0.55	0.45	0.52	1.00	0.56	0.58	0.42
pag	0.53	0.56	0.54	0.63	0.56	1.00	0.59	0.51
tgl	0.72	0.76	0.74	0.60	0.58	0.59	1.00	0.65
war	0.69	0.71	0.75	0.51	0.42	0.51	0.65	1.00

Talahanayan 9. Listahan ng walong wika at ang pinakamataas na trigram ranking nila

Wika	Mataas na trigram ranking	
	Wika	Valyu
bik	hil	0.73
ceb	hil	0.80
hil	ceb	0.80
ilo	pag	0.63
pam	tgl	0.58
pag	ilo	0.63
tgl	ceb	0.76
war	hil	0.75

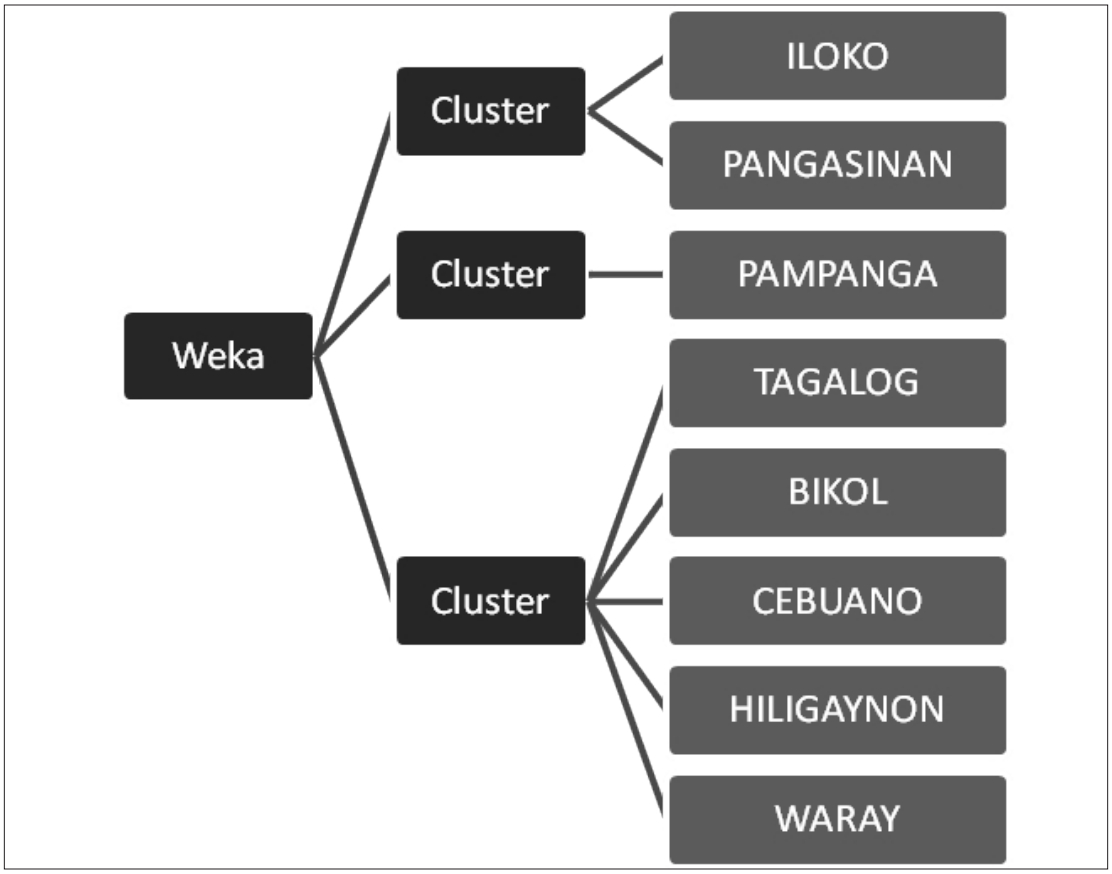
Batay sa dalawang talahanayan, mapapansin ang mga sumusunod:

- pinakamataas ang wikang *bik* sa *hil* na may trigram ranking na *0.73* at hindi nalalayo ang *ceb* (*0.72*), *tgl* (*0.72*), at *war* (*0.69*);
- pinakamataas ang wikang *ceb* sa *hil* na may trigram ranking na *0.80* at hindi nalalayo ang *bik* (*0.72*), *tgl* (*0.76*), at *war* (*0.71*);
- pinakamataas ang wikang *hil* sa *ceb* na may trigram ranking na *0.80* at hindi nalalayo ang *bik* (*0.73*), *tgl* (*0.74*), at *war* (*0.75*);
- pinakamataas ang wikang *ilo* sa *pag* na may trigram ranking na *0.63* at hindi nalalayo ang *tgl* (*0.60*);
- pinakamataas ang wikang *pam* sa *tgl* na may trigram ranking na *0.58* at hindi nalalayo ang *ceb* (*0.55*) at *pag* (*0.56*);
- pinakamataas ang wikang *pag* sa *ilo* na may trigram ranking na *0.63* at hindi nalalayo ang *tgl* (*0.59*);
- pinakamataas ang wikang *tgl* sa *ceb* na may trigram ranking na *0.76* at indi nalalayo ang *bik* (*0.72*), *hil* (*0.74*), at *war* (*0.65*); at
- pinakamataas ang wikang *war* sa *hil* na may trigram ranking na *0.75* at hindi nalalayo ang *bik* (*0.69*), *ceb* (*0.75*), at *tgl* (*0.65*).

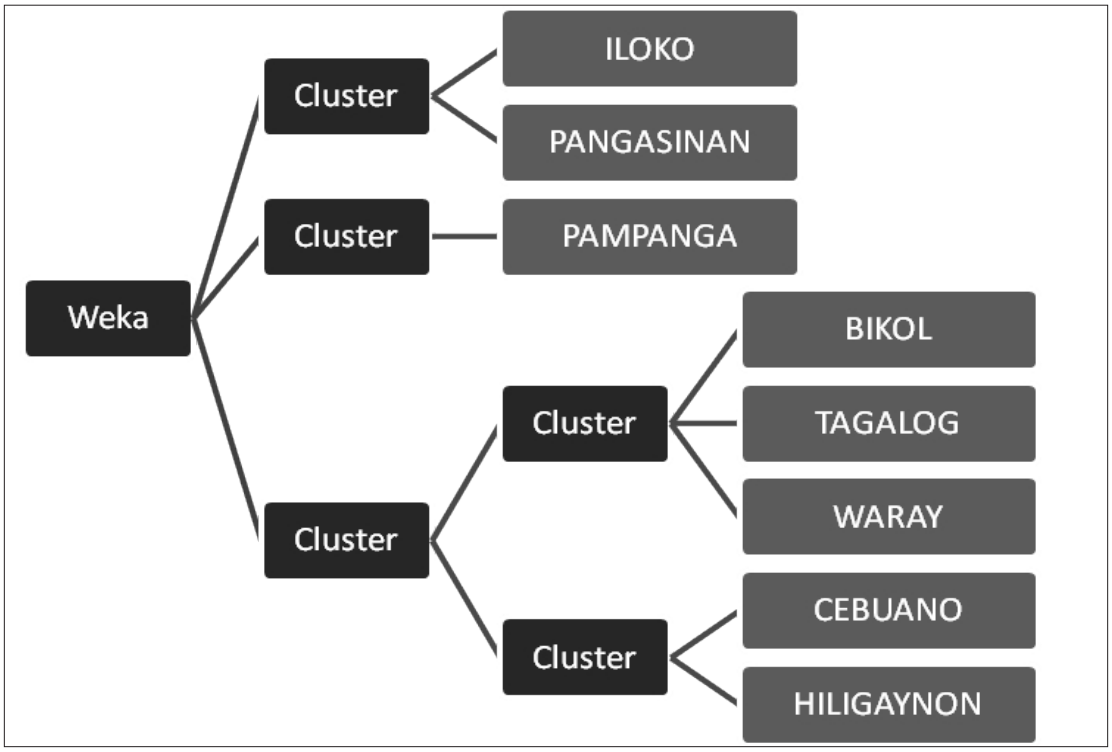
Isa rin sa mga napansin ang pagkakahalintulad ng *hil* sa maraming wika. Maaaring sabihin na maraming trigram ang *hil* na makikita rin sa ibang wika.

Grupo ng mga Wika

Matapos makuha ang trigram ranking ng mga pares ng wika, ginamit naman ito para makuha ang mga "cluster" o pagkakapangkat. Makikita sa Larawan 7 ang nakuhang resulta gamit ang Weka: nasa iisang grupo ang Iloko at Pangasinan samantalang nag-iisa naman ang Pampanga. Magkakasama naman ang Bikol, Cebuano, Hiligaynon, Tagalog, at Waray. Upang lalong makuha ang pagkakapangkat sa limang wikang ito, muling ginamit ang Weka sa limang wika. Makikita sa Larawan 8 ang resulta. Kapansin-pansin na nasa iisang cluster ang Bikol, Tagalog, at Waray samantalang magkasama naman ang Cebuano at Hiligaynon.



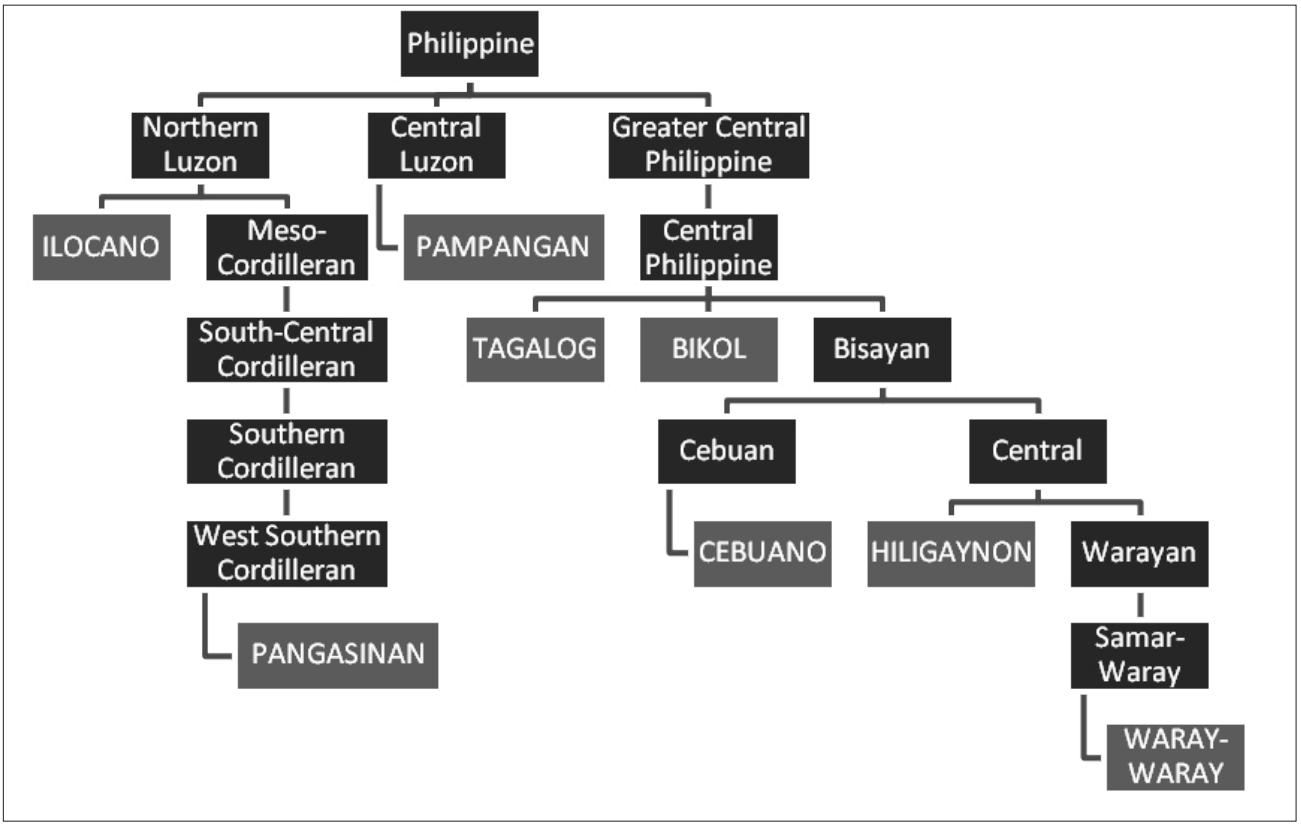
Larawan 7. Mga nakuhang subgrupo gamit ang Weka



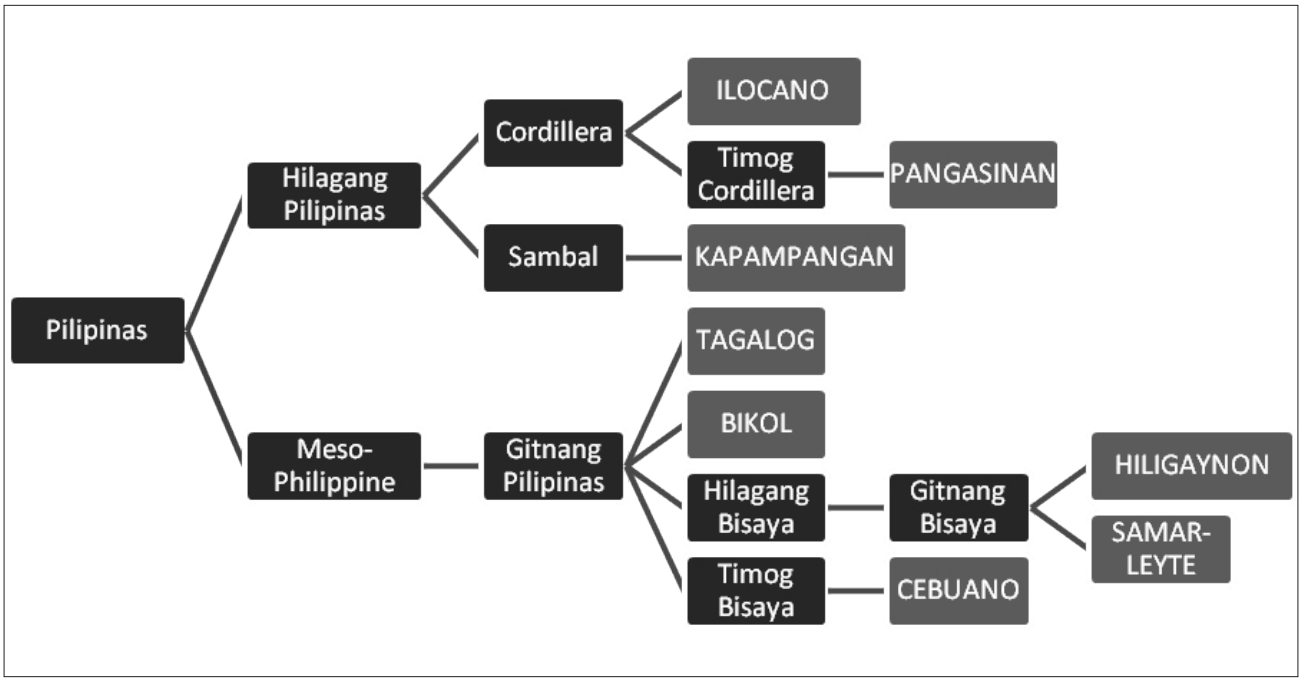
Larawan 8. Mga nakuhang subgrupo sa isang cluster

Inihambing naman ang resulta sa mga *language family tree*. Ayon sa Ethnologue (sumangguni sa Larawan 9), ang *ilo* (*Ilocano*) at *pag* ay galing sa Northern Luzon na pamilya ng wika samantalang ang *bik*, *ceb*, *hil*, *tgl*, at *war* (*Waray-Waray*) naman ay galing sa Central Philippine na pamilya ng wika. Ang *pam* (*Pampangan*) naman, na nakakuha ng mababang trigram ranking sa ibang wika ay galing sa ibang pamilya ng wika – ang Central Luzon. Ayon naman sa balangkas ng subgruping ng mga wika sa Pilipinas (Fortunato 13-17) na ibinatay sa akda ni McFarland (sumangguni sa Larawan 10), ang *ilo* at *pag* ay galing sa subgrupong Cordillera; ang *bik*, *ceb*, *hil*, *tgl*, at *war* (*Samar-Leyte*) naman ay galing sa subgrupong gitnang Pilipinas; at galing sa subgrupong Sambal ang *pam* (*Kapampangan*). Kapwa tumutugma ito sa resultang nakuha gamit ang Weka, na nasa Larawan 7.

Nagkakaiba naman ang resultang nakuha para sa subgrupong kinabibilangan ng *hil* at *war*. Kung susuriin ang dalawang language family tree, mapapansing parehas na nasa gitnang Bisaya ang *hil* at *war* pero lumalabas sa trigram ranking na mas malapit ang *hil* sa *ceb*. Kung susuriin ang mapa ng mga wika ng Pilipinas (sumangguni sa http://en.wikipedia.org/wiki/File:Hiligaynon_language_map.png para sa Hiligaynon at sa http://en.wikipedia.org/wiki/File:Waray-Waray_language_map.png para sa Waray), mapapansin na magkalayo at hindi magkatabi ang lokasyon ng Hiligaynon at Waray samantalang magkakalapit naman ang lokasyon ng Hiligaynon at Cebuano (http://en.wikipedia.org/wiki/File:Distribution-ceb.png), at ng Bikol (http://en.wikipedia.org/wiki/File:Bikol_languages_map.png) at Waray sa isa't isa. Maaaring maisaalang-alang ang ibang uri ng *language contact* sa pagkakaibang ito.



Larawan 9. Pamilya ng mga wika, hango sa Ethnologue[6]



Larawan 10. Pamilya ng mga wika, hango kay Fortunato (13-17)

KONGKLUSYON

Ipinapakita sa pag-aaral na ito ang trigram ranking bilang isang panukat sa pagkakahalintulad ng mga wika. Lumalabas sa mga resulta na ang mga pares ng wika na may matataas na trigram ranking ay magkakalapit at may pagkakahalintulad. Ipinakita rin na maaaring gamitin ang isang clustering algorithm para makuha ang pagkakapangkat ng mga wika gamit ang trigram ranking. Lumalabas na ang mga pares ng wika na may magkakalapit na trigram ranking ay napapangkat sa iisang cluster. Tugma rin sa mga language family tree ang mga nakuhang resulta. Galing sa iisang subgrupo ang *bik*, *ceb*, *hil*, *tgl*, at *war*. Samantala, magkapares naman ang *ilo* at *pag*. Mag-isa sa iisang subgrupo naman ang *pam*. Iminumungkahi na gamitin ang trigram ranking bilang panukat sa pagkakahalintulad ng iba pang wika sa Pilipinas gamit ang mas malaking training data.

Maituturing na malaki ang potensiyal ng pag-aaaral na ito sa iba pang mga larangan tulad ng lingguwistika, sosyo-lingguwistika, at pagplaplanong pangwika. Batay sa pananaliksik na ito, maaaring magamit ang trigram ranking upang lalong mapaigting ang pag-aaral sa iba pang pagkakahalintulad ng iba pang mga wikang bernakular sa Pilipinas. Sa larangan naman ng sosyo-linggwistika, maaaring magamit ang pag-aaral na ito upang masuri ang iba't ibang estilo ng diskursong namamayani sa mundo ng cyberspace. Samantala, malaki naman ang posibilidad na magamit ito upang makita ang pagkakahalintulad sa preferens ng mga praktisyuner at ordinaryong ispiker ng Filipino pagdating sa ispeling upang ma-standardize ang patakaran sa ortograpiyang Filipino.

Pagkilala at Pasasalamat

Ang proyektong ito ay sinuportahan ng University Research Coordination Office ng Pamantasang De La Salle (na may numerong 09 IR U 2TAY12-2TAY13). Nagpapasalamat ang mga may-akda sa mga sumusunod na naging bahagi ng proyekto: G. Jason Wong at G. Wilhelm Paul Martinez.

DULONG TALA

[1] Lumalabas sa resulta ng mga language model na mas mataas ang bilang ng trigram na "pil" kumpara sa trigram na "fil". Nangangahulugan lamang na mas inilalarawan ng Pilipinas, at hindi ng Filipinas, ang wikang Filipino.

[2] Makikita ang lexical similarity ng Pranses at ng iba pang wika sa website na ito: www.ethnologue.com/show_language.asp?code=fra. Para sa iba pang wika, palitan lamang ang "fra" ng language code ng wika.

[3] Makikita ang mga language code sa website na ito: http://www-01.sil.org/iso639-3/codes.asp

[4] Maaaring ma-download ang Apache Nutch sa website na ito: http://nutch.apache.org/. Kinakailangan ng kaalaman sa Java programming language para magamit ito.

[5] Maaaring ma-download ang SRILM sa website na ito: http://www.speech.sri.com/projects/srilm/. Kinakailangan ng kaalaman sa UNIX commands para magamit ito.

[6] Nakatala sa Ethnologue ang pamilya ng mga wika. Makikita ang tala para sa Pilipinas sa website na ito: http://www.ethnologue.com/subgroups/philippine

SANGGUNIAN

Borra, Allan, Charibeth Cheng, Rachel Edita Roxas, at Sherwin Ona. "Information Extraction and Opinion Organization for an e-Legislation Framework for the Philippine Senate." *Conference on Human Language Technology for Development*. 2-5 Mayo 2011, Alexandria, Egypt. Alexandria: HLTD Organizing Committee, 2011. 196-204. Limbag.

Dimalen, Davis, at Rachel Edita Roxas. "Autocor: A Query Based Automatic Acquisition of Corpora of Closely-related Languages." *21st Pacific Asia Conference on Language, Information and Computation*. 1-3 Nobyembre 2007, Seoul, South Korea. Eds. Hee-Rahk Chae, Jae-Woong Choe, Jong Sup Jun, Youngchul Jun, Eun-Jung Yoo. Seoul: Korean Society for Language and Information, 2007. 146-54. Limbag.

Dita, Shirley, Rachel Edita Roxas, at Paul Inventado. "Building Online Corpora of Philippine Languages." *23rd Pacific Asia Conference on Language, Information and Computation*. 3-5 Disyembre 2009, Hong Kong, China. Ed. Olivia Kwong. Hong Kong: City University of Hong Kong Press, 2009. 646-53. Limbag.

Do, Quang, Dan Roth, Mark Sammons, Yuancheng Tu, at V.G.Vinod Vydiswaran. "Robust, Light-weight Approaches to compute Lexical Similarity." *Computer Science Research and Technical Reports, University of Illinois*. 2009. Web. 13 Mayo 2013. <http://cogcomp.cs.illinois.edu/papers/DRSTV09.pdf>.

Downey, Sean, Brian Hallmark, Murray Cox, Peter Norquest, at J. Stephen Lansing. "Computational feature-sensitive reconstruction of language relationships: developing the ALINE distance for comparative historical linguistic reconstruction." *J. Qualitative Linguistics* 15 (2008): pp. 340–369. Limbag.

Fortunato, Teresita. *Mga Pangunahing Etnolinggwistikong Grupo sa Pilipinas*. Manila, Philippines: De La Salle University Press, 1993. Limbag.

Lewis, Paul. *Ethnologue: Languages of the World*. 16th ed. Dallas, Texas: SIL International, 2009. Limbag.

Lloyd, Stuart. "Least Squares Quantization in PCM." *IEEE Trans. Information Theory* 28.2 (2008): pp. 129–137. Limbag

McFarland, Curtis. *A Linguistic Atlas of the Philippines*. Manila, Philippines: Linguistic Society of the Philippines, 1983. Limbag.

McMahaon, April, at Robert McMahon. *Language Classification by the Numbers*. Oxford, UK: Oxford University Press, 2005. Limbag.

Oco, Nathaniel, at Rachel Edita Roxas. "Pattern Matching Refinements to Dictionary-based Code-switching Point Detection." *26th Pacific Asia Conference on Language, Information and Computation*. 7-10 Nobyembre 2012, Bali, Indonesia. Jakarta: Faculty of Computer Science, Universitas Indonesia, 2012. 229-36. Limbag.

Oco, Nathaniel, Joel Ilao, Rachel Edita Roxas, Raquel Sison-Buban, at Don Erick Bonus. "ASEAN MT-Phil: Philippine Component of the ASEAN Machine Translation Project." *9th National Natural Language Processing Research Symposium.* 07 Marso 2013, Lungsod Quezon, Pilipinas. Manila, Pilipinas: De La Salle University, 2013. 11-13. Limbag.

Oco, Nathaniel, Joel Ilao, Rachel Edita Roxas, at Leif Romeritch Syliongka. "Measuring Language Similarity using Trigrams: Limitations of Language Identification." *3rd International Conference on Recent Trends in Information Technology*. 25-27 Hulyo 2013, Chennai, India. Chennai, India: Anna University, 2013. Limbag

Oco, Nathaniel, Leif Romeritch Syliongka, Rachel Edita Roxas, at Joel Ilao. "Dice's Coefficient on Trigram Profiles as Metric for Language

Similarity.” *16th International Oriental COCOSDA Conference.* 25-27 Nobyembre 2013, Gurgaon, India. Gurgaon, India: IEEE Xplore. Web. 17 Enero 2014