KOMPARATIBONG ANALISIS NG AKTUWAL NA GAMIT NG WIKA AT MGA PILING PAMANTAYAN SA GRAMATIKA AT ORTOGRAPIYA SA FILIPINO, SEBWANO-BISAYA AT ILOKANO: LAPIT BATAY SA KORPUS[1]

nina Joel P. Ilao, Timothy Israel D. Santos at Rowena Cristina L. Guevara

INTRODUKSIYON

Opisyal nang iniharap ng Kagawaran ng Edukasyon ang inisyatibong Mother Tongue-Based Multilingual Education sa pamamagitan ng sirkular ng kautusang pangkagawaran noong 2009[2], at inilatag na rin ang mga pamamaraan upang ang preparasyon sa pagpapatupad ng programang MTBMLE ay bumilis tungo sa ganap na implementasyon sa Hunyo 2012. Tinukoy ni Diane Dekker (2010) na ang pinakaimportanteng bahagi ng anumang pagsisimula ng Mother Tongue Language Education ay ang pag-unlad ng "sistema sa pagsulat na katanggap-tanggap sa karamihan ng mga nagsasalita ng katutubong wika at sa pamahalaan at paghihikayat sa mga miyembro ng pamayanang pangwika na ipagpatuloy ang pagbasa at pagsulat sa kanilang wika"[3], isang kahingiang muling binanggit ni Ricardo Nolasco (2011) para sa community-based programang MTBMLE sa kaniyang MLE Primer[4].

Kung titingnan ang sirkular ng Kagawaran ng Edukasyon noong 2009 at isasaalang-alang ang mga dokumentong nagpapakita ng matagumpay na pagpapatupad ng MTBMLE, malinaw na mahalagang kahingian sa programang ito ang isang buhay na ortograpiya na malawakang tinatanggap ng nag-aaral na komunidad at naayon sa intelektuwalisasyon ng wika. Ang

pagkakaiba-iba ng wika sa Pilipinas, na may 171 na buhay na wika at mahigit-kumulang sa 500 dayalekto ang isang malaking hamon sa nasabing inisyatiba na ang sistema ng ortograpiya ng bawat lahok na wika ay hindi matiyak. Habang tumatagal ang programang MTBMLE, darating ang pangangailangang kinisin pa ang mga tuntuning gramatikal at ortograpikal na ginagamit sa instruksiyon habang lumalawak ang paggamit nito sa akademya. Pinatutunayan ng mga senaryong ito na kailangan ang isang sistema ng regular na pagmomonitor sa kalagayan ng pag-unlad ng wika sa pamamagitan ng pag-oobserba sa kung paano ito ginagamit ng populasyon. Sa pag-aaral na ito, inihaharap namin ang kasalukuyang kalagayan ng pag-unlad ng wika para sa tatlong pangunahing wika sa Pilipinas, sa pamamagitan ng aming *corpus-based* analysis ng Filipino/Tagalog, Sebwano at Ilokano. Inorganisa ang papel na ito sa sumusunod na mga bahagi: Seksiyon 1, pag-unlad ng ortograpiya ng nabanggit na tatlong pangunahing wika; Seksiyon 2, metodolohiya sa pagtukoy ng grupo ng varyant sa ispeling na nakuha mula sa korpora ng tatlong pangunahing wika; Seksiyon 3, nakalap na mga datos, listahan ng grupo ng varyant sa ispeling na napili mula sa korporang hawak, at ang aming analisis sa resulta ng eksperimento. Nagtatapos ang papel na ito sa aming mga kongklusyon at rekomendasyon para sa mga gawain sa hinaharap.

Sistema ng Ortograpiya ng Tatlong Pangunahing Wika sa Pilipinas Batay sa Populasyon ng Gumagamit

Batay sa sensus na kinalap ng National Statistics Office (NSO) noong 2000, mahigit kumulang 10 sa 171 na buhay na wika sa Pilipinas ang itinuturing na pangunahing wika kung saan ang pangunahing wika ay nangangahulugang may mahigit sa isang milyon ang nagsasalita o gumagamit nito. Ayon din sa datos ng nabanggit na sensus, ang tatlong pangunahing wika, batay sa bilang ng nagsasalita ay Tagalog (~21.5 milyong katutubong nagsasalita o 28.15% ng kabuuang populasyon ng mga Filipino), Sebwano (~20 milyong katutubong nagsasalita o 26.1% ng kabuuang populasyon), at Ilokano (~7.7 milyong katutubong nagsasalita o 10.1% ng kabuuang populasyon). Itinuturing din na rehiyonal na lingua franca ang tatlong wikang ito na, ang Ilokano ay ginagamit ng mga rehiyon sa hilagang bahagi ng Luzon, ang Tagalog sa gitna at timog na bahagi ng Luzon, at ang Sebwano-Bisaya sa isla ng Visayas at ilang bahagi ng

Mindanao. May matatag na tradisyong pasalita at pasulat ang tatlong wika kung kaya angkop na maging wika ng instruksiyon para sa balangkas ng MTBMLE. Sa kabilang banda, itinuturing namang pambansang lingua franca ang Wikang Pambansang Filipino, na naiintindihan at ginagamit ng mahigit sa 96.4% ng mga Filipino batay sa sensus noong taong 2000. Bilang kinikilalang opisyal na wika ng Pilipinas, ginagamit ang Filipino bilang opisyal na midyum ng instruksiyon sa pambansang sistema ng edukasyon simula nang ipatupad ng Kagawaran ng Edukasyon ang *Bilingual Educational Policy* noong 1974. Ipinagpalagay naming ang tuntuning gramatika at ortograpiko ng Filipino ay kapareho ng Tagalog, na pinatutunayan ng sumusunod na analisis at pagtalakay. Inilalarawan ng sumunod na sub-seksiyon ang pag-unlad ng sistemang ortograpiko, at mga halimbawa ng tuntuning ortograpiko batay sa mga napiling materyal na nailathala sa isa sa bawat tatlong wika.

Ipinatutupad kamakailan ang mga pagbabago sa ortograpiya ng Pilipinas na ang layunin ay magkaroon ng iisang ortograpiya[5]. May malawak na arkibo ng mga pangunahing wika sa Pilipinas na nakasulat sa tradisyonal na mga titik[5] at ang modernong sistemang ortograpiko ay may katulad na direksiyon na ring tinatahak sa pagsunod sa pagsulat na Latin. Samantalang ang wika ay kinakatawan ng ortograpiya na siyang may pinakamahalagang ugnayan sa katangiang pasalita at ponetika, ang iba namang pangunahing wika ay may magkaibang anyo ng ortograpiya na may tiyak na mga tuntunin at pamantayan. Inilalarawan ng sumusunod na sub-seksiyon ang pag-unlad ng sistemang ortograpiko, at mga halimbawa ng tuntuning ortograpiko batay sa mga napiling nakalathalang materyal para sa bawat isa sa tatlong wika.

Filipino/Tagalog

Nahubog ang ortograpiyang Filipino sa natural na transpormasyong dulot ng mga pangyayaring sosyo-kultural, at sa mga pambansang reporma sa mga patakarang pangwika. Bago ang panahon ng kolonisasyon o panahon ng mga kaharian, ginagamit sa Pilipinas ang tradisyonal na pagsulat na tinatawag na Baybayin. Ginagamit ang Baybayin sa pagsulat sa Tagalog, Sebwano, at Ilokano kasama ang iba pang mga wika sa Pilipinas ilang siglo bago pa dumating ang kolonyalistang Espanyol hanggang sa pinili na ng mga katutubo na gamitin ang pagsulat na Latin batay sa ortograpiyang Espanyol noong huling bahagi ng ika-16 siglo[5]. Ang Abecedario

na batay sa Espanyol ang pinagmulan ng Abakada, ang alpabetong Filipino na itinakda ng unang Balarilang Wikang Pambansa na ginamit ng mga pampublikong paaralan na siyang ipinaglaban ni Lope K. Santos at inaprobahan sa pamamagitan ng Kautusang Pangkagawaran Blg. 1, s. 1940 ng Pampublikong Instruksiyon.

Ang Abakada ay binubuo ng 20-titik na alpabetong madalas itinuturo gamit ang simpleng tuntunin na kung ano ang bigkas siya ang baybay. Ang ortograpiya ay simple at sapat para sa pambansang lingua franca na batay sa Tagalog ngunit ito rin ay nagdulot ng kabutihan at kakulangan sa sistema ng ispeling. Naging epektibo ang pagtuturo ng wikang Filipino na nagsisimulang lumitaw mula sa pinag-ugatang Tagalog. Dahil ang wikang Filipino ay inakalang payayamanin ng mga lokal at banyagang wika kung kaya't nagkaroon ng ilang pagbabago at dalawang beses binago ang bilang ng titik sa loob ng 30 taon.

Isang memorandum ng DECS noong 1976 (DECS Memorandum Blg. 194, s. 1976) ang nagdagdag ng 11 titik sa tradisyonal na Abakada at matapos ang 10 taon ay binawasan naman at nagresulta sa kasalukuyang ortograpiyang ginagamit na may 28 titik. Ang memorandum na inilabas noong 1976 at 1987 ay sinamahan ng publikasyon ng Tuntunin sa Ortograpiyang Filipino at Alpabeto at Patnubay sa Ispeling ng Wikang Filipino. Ang mga pagbabago ay ginawa bilang tugon sa probisyon ng memorandum noong 1976 at Konstitusyon ng 1987 na paunlarin, pagyamanin at gawing makabago ang pambansang wika. Ang modernong alpabetong Filipino na may 28 titik ay ginawa upang umangkop ang mga salita mula sa ibang uri ng wika sa Pilipinas at sa mga hindi maiiwasang panghihiram ng mga salitang banyaga.

Talaan 1.Pagbabago sa Ortograpiyang Filipino/Tagalog

Pagbabago sa Ortograpiya	Mga Titik na Kasali/Tinanggal
1940 *Abakada*	a b k d e g h i l m n o p r s t u w y
1976 *Tuntunin*	ch f j llñng q rr v x z
1987 *Patnubay*	ch ll rr (mga tinanggal)

Ang pinakahuling rebisyon sa ortograpiya ay ang *Revisyon sa Alfabeto at Patnubay sa Ispeling* noong 2001 ng Komisyon sa Wikang Filipino. Walang pagbabago sa bilang ng mga titik ngunit nagbigay ng pamantayan sa paggamit ng dagdag na mga titik na nagdulot ng kalituhan. Isa sa pinakamahalagang pagbabago sa 2001 Revisyon ay ang pagtanggap sa paggamit ng walong bagong titik sa lahat ng hiram na salita, salungat sa 1976 *Tuntunin* at 1987 *Patnubay* na nagsasaad na limitado lamang ang paggamit sa nasabing mga titik sa mga hiram na terminong siyentipiko/teknikal at mga hiram na salita sa mga wikang rehiyonal. Sinasabi ng mas maluwag na tuntunin na marami sa mga hiram na salita ang maaaring magkaroon ng dalawang anyo, isa ang taglay ang orihinal na ispeling at ang isa naman ang binaybay sa anyo ng Abakada. Tinatanggap ang sumusunod na anyo ng mga salita:

vaso/**b**aso, **f**avorito/**p**a**b**orito, **v**intana/**b**intana

Isa pang malaking kontribusyon ng 2001 Revisyon ay ang pagkakahati ng walong bagong titik sa *phonemic group* at ang *allophone group*. Ang mga titik na f, j, v, at z ay may isa-sa-isang katumbas sa tunog habang ang *allophone group* naman na c, ñ, q, at x ay maaaring maging kinatawan na mahigit sa isang tunog.

Maaaring magbigay-liwanag ang pag-aaral na ito sakaling mayroon pa ring kalituhan dahil sa mga nabanggit na pagbabago sa tuntunin. Maaari nating makita kung ang salita ay nananatili at mas ginagamit.

May tatlong uri ng varyant na maaaring lumitaw sa hiram na salita. Maaaring magkaroon ng varyant dahil sa *transliteration* (arkiyoloji), maaari namang mula sa hindi binagong hiram na anyo (archaeology), at isa naman ay dahil sa mga dahilang tradisyonal o preperensya (arkeolohiya).

Sebwano-Bisaya

Pareho ng ortograpiyang Filipino ang pinagdaanan ng Sebwano-Bisaya. Ang ortograpiyang Sebwano ay nakasulat din sa sinaunang *Baybayin*, ang Abecedario ng mga Espanyol,

Abakada, at kalaunan, ay ang nabagong alpabetong may 28 titik. Habang ang mga titik at letra ay magkapareho, may pagkakaiba naman sa tuntuning ortograpiko ng Sebwano dahil sa naiibang katangian ng wika. Wala pang estandardisadong ortograpiya para sa Sebwano, ngunit ginamit namin ang Tanangkingsing's Functional Reference Grammar for Cebuano [6] at John Wolf's Cebuano-English Dictionary [7].

Ilokano

Pareho ng Sebwano at ng ibang wika sa Pilipinas, ginamit din ng Ilokano ang palapantigan ng Baybayin ilang siglo bago sinunod ang panulat na Latin. Ginamit ang Ilokano na bersiyon ng palapantigan upang maipakita ang *syllable codas*, na nagpapahintulot sa paggamit ng mga pantig na katinig-patinig-katinig (KPK). Inilagay ang bantas na *vowel-killer* "+" bilang *superscript* sa karaniwang *katinig-patinig (KP)* kinatawan ng pantig upang hindi bigkasin ang tunog ng patinig. Sa huli, ang Espanyol na Abecedario ay naging popular, kung kaya ang mga nakatatandang henerasyon ay mas pinili pa rin ang ortograpiyang Espanyol sa kabila ng pagbabago sa ortograpiya ng pambansang wika na nag-uutos sa pagtuturo ng sistemang Abakada at wikang Filipino sa paaralan[8].

Isa sa may pinakamalawak na koleksiyon ng mga materyal sa Ilokano pagkatapos ng panahon ng Espanyol ang magasing *Bannawag*. Nasa sirkulasyon na ito nang mahigit 75 taon, at nagsisilbing lugar kung saan nagsisimula ang mga manunulat sa Ilokano. Hindi man laging konsistent ang sistema ng pagsulat sa *Bannawag*, ito ay isa sa pinakalohikal at ginagamit ng mga manunulat[9]. Ang pag-unlad ng sistema ng *Bannawag* ay hindi dahil sa sadyang pagsisikap upang makabuo ng sistema ng ortograpiya ngunit dala ito ng likas o natural na pagpapalakas na nakaugalian na ng magasin. Dahil dito, maaaring naapektuhan lamang ang sistema ng mga karaniwang manunulat at mga editor. Dahil walang makapangyarihang grupong nakabuo ng estandardisadong ortograpiyang Ilokano, sinangguni namin ang diksiyonaryo/phrasebook[8] at umiiral na ortograpiya batay sa sistema ng *Bannawag*[9].

Talaan 2. Halimbawa ng Ortograpiyang Ilokano

Piniling Anyo ng Trigraph	Halimbawa
ti + V	***tia****n* (stomach)
ts + V	***tsu****per*(driver)
di + V	***dia****ya*(offer)
dy + V	***dyi****p*(jeep), *dyok*(joke)
si + V	***sie****mpre*(of course), *sien*(100)
ni + V	*a****nia*** (what)

Ang mga tunog-ponemiko para sa mga katutubong katinig ay kinakatawan ng isa-sa-isa na tumbasan maliban sa iilang espesyal na kasong tulad ng *dialectal variation* na may <s> at <h> na makikita sa varyant na *haan* para sa salitang *saan* (hindi).

Mayroong sistemang apat na patinig sa Ilokano (a, e, i, o/u). Ang pagpapalit ng 'o' at 'u' sa Ilokano ay hindi nagdudulot ng pagbabago ng kahulugan ng salita.

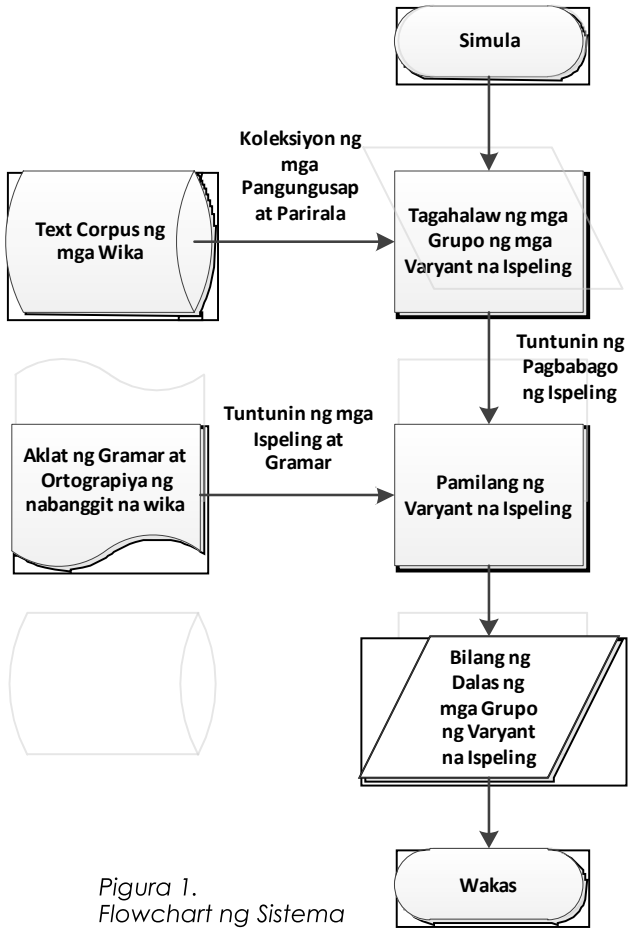
Mga Potensyal na Pagkalito sa mga Wika sa Pilipinas

Sa isang pag-aaral sa disenyo ng ortograpiya ng mga di-nakasulat na mga wika sa Pilipinas, inisa-isa ang listahan ng mga potensyal na mga problema sa ortograpiya ng Pilipinas[10]. Nakalista sa ibaba ang mga posibleng dahilan ng pagkalito batay sa disenyo ng ortograpiya at ilan sa mga madalas na problemang nabanggit sa mga kaugnay na materyal.

- Paggamit ng 'o' at 'u'
- Simbolisasyon ng mga *offglides* (hal. sya, bwaya)
- Simbolo ng *glottal*
- Representasyon ng hugpungan ng "n" at "g" kung hindi nila binubuo ang "ng"
- Pagkamadaling basahin ng mga inuulit na salita (paggamit ng gitling)

- Paggigitling
- Paggigitling at impit
- Pagkakaltas
- Pagdodoble ng katinig
- Pagkakabit ng panghalip
- Salitang tambalan
- Morpoponemiks – mga baryasyon dahil sa mga panlapi
- Asimilasyon - d vs r
- Mga patinig na tinatanggal kapag naglalapi ng mga salita
- Paggamit ng walong bagong titik (ch f j ll ñ ng q rr v x z)
- Mga Hiram na Salita *(phonetic transliteration vs. foreign form)*
- *Interference/Interaction from other local variety*

Corpus-based na Komparatibong analisis ng aktuwal na gamit ng wika at tuntunin sa gramatika para sa filipino/tagalog, sebwano-bisaya at ilokano



Pigura 1.
Flowchart ng Sistema

Ipinapakita sa Pigura 1 ang *flow chart* para sa isang sistema na gumagamit ng corpus-based na komparatibong analisis ng aktuwal na gamit at tuntunin sa gramatika ng Filipino/Tagalog, Sebwano-Bisaya, at Ilokano. Ang koleksiyon ng mga salitang nakasulat sa tatlong pangunahing wikang nabanggit ay kinuha mula sa World Wide Web (WWW) gamit ang ginawang software para sa ganoong layunin[11]. Mula sa *web-mined text corpora* ay nakabuo ng mga listahan ng mga posibleng varyant ng ispeling ng bawat pangunahing wika na kasama sa pag-aaral, batay sa pinakamaraming maaaring umangkop na tuntunin sa pag-edit sa pagbabago ng anyo ng isang salita upang maging lahok na salita sa isang partikular na korpus ng wika.

Pagkatapos ay binuo ang mga Tuntunin ng Transpormasyon sa Ispeling na umaangkop sa uri ng nagkokompetensiyang mga pares ng varyant sa pamamagitan ng manwal na inspeksiyon ng nabuong *Spelling Variant list.* Ang mga *Spelling Transformation Rules* ay tiningnan sa piniling mga reference grammar at ortograpiya ng Filipino/Tagalog, Sebwano, at Ilokano upang matukoy kung gaano ang sinaklaw ng ispeling ay makikita sa mga istandard na tuntunin. Binilang ang dalas ng bawat *Spelling Transformation Rules* na angkop sa mga varyant ng ispeling upang matukoy kung alin ang mas pinipili ng gumagamit ng wika. Inilalarawan sa sumusunod na sub-seksiyon ang mga detalye ng implementasyon ng mga modyul sa Pigura 2.

Koleksiyon ng *Text Corpora* sa Nakatakdang Wika

Gumamit kami ng *corpus mining software* upang magsaliksik at mag-download ng mga dokumento mula sa WWW na nakasulat sa tinukoy na wika[11]. Kinakailangan ang isang *seed document* na nakasulat sa tinukoy na wika para makapag-search nang online ng mga dokumentong nakasulat sa ganoong wika din. Ang mga nakolektang Filipino/Tagalog na text corpus gamit ang corpus mining software ay mas pinalawak pa ng mga Filipinong teksto mula sa korporang mga balita ng *Abante* at *Abante-tonite* (www.abante.com.ph)at *Abante-tonite* (www.abante-tonite.com) na ginagamit sa proyektong *Bantay-Wika*[12]. Ang Sebwano-Bisaya text corpus naman ay nabuo mula sa mga na-download na balita mula sa website ng *Sun Star* (www.sunstar.com.ph). Para naman sa text corpus ng Ilokano, inayos ang corpus mining software upang malayang makapaghanap ng mga dokumentong nakasulat sa Ilokano sa WWW; napansin namin na karaminhan sa mga na-download na dokumento ay mula sa website ng *Tawid News Magasin* (www.tawidnewsmag.com).

Pagputol ng Lexicon, Pagkuha ng Ispeling sa Varyant na Grupo at Pagbuo ng mga Tuntunin ng Transpormasyon ng Ispeling

Ang mga *lexicon* ng tatlong wikang ginamit sa pag-aaral na ito ay binuo sa pamamagitan ng kinalap na text corpora. Dahil hindi napapangasiwaan at hindi limitado ang WWW kung saan ang text corpora ay nakuha, ang mga lexicon ay naglalaman ng mga numero (hal.

petsa, bilang, numero ng telepono), mga maling ispeling at mga termino na hindi istandard (hal. usernames, mga terminong *Jejenese*), ang mga lahok na ito ay kadalasang mula sa mga seksiyon ng mga puna ng mga website ng mga balita, blogs at web fora. Napagpasiyahan naming huwag sundin ang simpleng pagtatanggal ng termino na hindi umaabot sa *threshold* ng pinakamababang bilang ng paglitaw sa korpus sa pagsasala ng mga hindi kailangang lahok dahil ang Sebwano-Bisaya at Ilokano text corpora ay masyadong maliit; ang pagputol ng mga lahok base sa bilang ng paglabas ay siya ring magsasala ng mga balidong lexical na lahok na hindi madalas maganap. Alam na antimano kung ano ang anyo ng isang tipikal na salita sa alinmang wika sa Pilipinas, kung pinili naming linisin ang mga lexicon sa pamamagitan ng pananatili ng mga lexical na lahok na mayroon lamang mga letra at simbolo ng gitling (-).

Nagbigay si Levenshtein (1966) ng tatlong pamamaraan ng pagsasaayos na maaaring gamitin sa paghahambing ng isang karakter sa isa pang karakter: (1) *pagpapalit,* (2) *paningit* at (3) *pagkakaltas*[13]. Mangyaring sumangguni sa Talaan 3 para sa mga halimbawa ng naedit na operasyon ni Levenshtein na nakita mula sa mga ispeling varyant sa Filipino/Tagalog, Sebwano-Bisaya at Ilokano. Ipinapalagay ang dalawang salita na may *edit distance* na N kung ang N *distinct Levenshtein edit operation* ay dapat na gamitin nang sunod-sunod para matransporma ang isang salita tungo sa ibang salita. Halimbawa, ang salitang tambalan na *mag-asawa* at *nagaasawa* ay mayroong edit distance na 3 dahil ang pagbabago na makikita sa Pigura 2 ay mayroong 3 magkakasunod na hakbang (pansinin na ang letra na naapektuhan ng mga pagbabago ay naka-highlight):

mag-asawa → **n**ag**-**asawa → nagasawa → nag**a**asawa

Pigura 2. Halimbawa ng pagbabago ng salita gamit ang na-edit na operasyon ni Levenshtein

Mula sa nalinis na mga lexicon, kumuha kami ng varyant ng ispeling sa pamamagitan ng paghahanap ng salita na iba sa dalawang naedit na operasyon. Ang nabuong *Spelling Variant cases* ay tiningnang mabuti at Spelling Transformation Rules na nagpapakita ng mga kategoryang kinabibilangan ng mga Spelling Variant ay nabuo.

Talaan 3. Naedit na Operasyon ni Levenshtein at kanilang mga halimbawa ng tatlong wika sa pag-aaral na ito: Filipino/Tagalog (tgl), Cebuano-Visayan (ceb), at Ilokano (ilk). Ang mga halimbawa na naka-bold ay ang mga naapektuhang karakter.

Naedit na Operasyon	Halimbawa
pagpapalit	an**u**-ano → an**o**-ano (tgl) nagmani**h**o →nagman**e**ho (ceb) katalta**l**onan →katalta**l**unan(ilk)
pagkakaltas	**h**inantay →inantay (tgl) makaag**u**wanta →makaagwanta (ceb) pam**m**utbuteng →pamutbuteng(ilk)
paningit	kolehiala→kolehi**y**ala (tgl) nakakuhag→nakakuha**a**g (ceb) nagdadakkel → nag-dadakkel (ilk)

Deskripsiyon ng Text Corpora

Iniharap sa Talaan 4 ang nakalap na text corpora ng iba't ibang wika para sa pag-aaral na ito. Mas malaki ang text corpus ng Filipino/Tagalog kaysa Sebwano-Bisaya at Ilokano text corpora dahil nasimulan na ng mas maaga ang pangangalap ng text corpus ng Filipino/Tagalog sa pamamagitan ng ***Bantay-Wika*** na proyekto sa UP-Digital Signal Processing laboratory, na nakatuon sa pag-obserba sa mga *trend* ng pag-unlad ng wika tungkol sa Pambansang Wika ng Pilipinas. Mas malaki naman ang Ilokano text corpus kaysa sa Sebwano-Bisaya text corpus, dahil nilimitahan namin ang pagda-download ng mga dokumentong Sebwano-Bisaya sa website ng mga balita ng *Sun Star* upang matiyak ang uri/kalidad ng text corpus ng Sebwano-Bisaya. Tandaan na ang corpus ng Ilokano ay binuo na mayroong corpus mining software na naka-disenyo para sa malayang pananaliksik sa WWW.

Makikita sa Talaan 4, na sa pagputol ng lexicon ay napapaliit ang mga lexicon ng mga wikang Filipino/Tagalog, Ilokano, at Sebwano sa 20%, 9.15% at 4.27%. Ipinapakita nito ang

kalidad ng text corpus na binuo mula sa mga dokumento na na-download mula sa WWW ay bumababa tulad din ng pagbaba ng laki nito.

Mga Tuntunin ng Transpormasyon sa Varyant ng Ispeling

Dito namin iniharap ang iba't ibang Tuntunin ng Transpormasyon ng mga Ispeling na naobserbahan sa pamamagitan ng corpus-based analysis ng iba't ibang text corpora. May nakitang 42 Tuntunin ng Transpormasyon mula sa kagyat na pagsisiyasat sa listahan ng mga varyant ng ispeling na gumamit ng metodolohiya na inilarawan sa seksiyon 2.2. Sumangguni sa Apendiks I para sa kompletong listahan ng mga Tuntunin ng Transpormasyon ng Ispeling na naobserbahan mula sa tatlong pangunahing wika na sakop ng pag-aaral na ito.

Mga Varyant ng Ispeling na Nakapaloob sa Aklat ng mga Tuntunin

Sa seksiyong ito, iniharap naming muli ang mga tuntunin at panuntunan para sa ilang mga varyant ng ispeling na nakapaloob na sa aklat ng mga tuntunin at ginagamit na ortograpiya. Ang pagkakaroon ng mga ganitong varyant ng ispeling ay malinaw na nagpapatunay na may ilang bahagi ng populasyon na nag-aalinlangan pa rin sa usaping ortograpiya. Isang malaking tulong sa mag-aaral kung ang mga ito ay binibigyan ng halaga at itinuturo ng mga instraktor/ propesor nang may pag-iingat. Ang mga varyant ng ispeling na nakalista dito ang madalas na naobserbahan na varyant sa tatlong wika at ang iba naman ay makikita lamang sa iisang wika.

Mga Patinig

Napatunayan na mula sa pag-aaral ng tatlong wika ng sinaunang ortograpiya ng Pilipinas (*Baybayin*) na mayroon lamang itong tatlong patinig (a, e/i, o/u). Ang adaptasyon sa pagsulat na Latin at pagsasama ng mga hiram na salita ay siyang nagparami ng bilang ng patinig tungo sa lima. Ang dalawang madalas na kaso ng varyant na ispeling ay ang tuntunin 1 (<i>vs<e>) at 2 (<o>vs<u>).

Heminasyon

Ang heminasyon o ang pagkakaroon ng magkasunod na katinig ay makikita sa Ilokano. Katulad ng Filipino/Tagalog at Sebwano-Bisaya, mayroong mga salitang inuulit na nasa pangmaramihang anyo. Ngunit, kadalasan ang mga terminong pangkamag-anak ay maaaring gawing salitang nasa anyong pinarami sa pamamagitan ng pagsusunod ng unang katinig sa ikalawang pantig.

- *a**n**ak – a**nn**ak*
- ba**b**ai – ba**bb**ai

Mga Varyant ng Ispeling na Hindi Sakop ng Aklat ng mga Tuntunin

Ang ilang mga naobserbahang varyant ay tila hindi nakapaloob sa mga ginamit na materyal. Susubukan naming suriin ang mga ito at magmumungkahi na magkaroong muli ng masusing pag-aaral upang makita ang mga dahilan ng pagkakaroon nito, at ang mga panuntunang maaaring ipatungkol dito.

False Positives

Dahil sa katangiang aglutinatibong mga wika sa Pilipinas, maraming salita ang tinatanggap pa rin kahit may *small edit distances*. Itinatala ng *automatic spelling variant extractor lists* ang lahat ng mga token na may character edit distance ng 2, at ang ilang kandidatong varyant ng ispeling na itinala ng extractor ay maaari pa ring pares. Halimbawa, ang varyant ng ispeling <g> vs. <n> vs. <m> para sa tanda ng mga Sebwano *gipasalig, nipasalig,* at *mipasalig* ay hindi nakita sa aklat ng mga tuntunin ng ortograpiya dahil iba ang anyo nito sa *salig (tiwala),* kaya morpolohikal ang baryasyon sa halip na ortograpikal. Ang mga varyant ng ispeling ay inalis mula sa pinal na listahan kapag napagtibay na ito ay false-hits (hal. mga nonspelling variant) mula sa pagsangguni sa mga aklat ng gramatika at taong nagsasalita ng katutubong wika.

Varyant ng mga Tuntuning Gramatika (morphosyntax analysis)

Isa pang resulta ng aglutinatibong katangian ng wika ay ang pagsasama ng maraming morpema.

- Mao + ra + ug ->m**ao**rag (ao/o)->m**o**rag (o/a) ->m**o**rag/m**u**rag
- Pagsasama ng unlaping i-, pag-uulit ng pinag, panlaping ma-, kapag

URI NG MGA TUNTUNIN PANG-GRAMATIKA AT ORTOGRAPIKO AYON SA BARYASYON NG PAGGAMIT

Talaan 5. Tuntunin ng mga Transpormasyon sa Filipino/Tagalog ayon sa bilang ng pagkakaiba ng naobserbahang salitang tambalan

Tuntunin ng Transpormasyon	Kabuuang bilang ng naobserbahang salitang tambalan
gitling vs. walang gitling	10342
<o> vs. <u>	4210
Mayroon o walang -<h>-	3693
<i> vs. <e>	3555
<y> vs. <i>	2457

Talaan 6. Tuntunin ng mga Transpormasyon sa Sebwano-Bisaya ayon sa bilang ng naobserbahang salitang tambalan

Tuntunin ng Transpormasyon (Tuntunin Blg.)	Kabuuang bilang ng naobserbahang salitang tambalan
<o> vs. <u>	299
Mayroon o walang -<g>-	233

Gitling vs.	215
<ng>vs. <n>	215
<i> vs. <e>	130

Talaan 7. Tuntunin ng mga Transpormasyon sa Ilokano ayon sa bilang ng naobserbahang salitang tambalan

Tuntunin ng Transpormasyon	Kabuuang bilang ng naobserbahang salitang tambalan
<o> vs. <u>	934
Mayroon o walang gitlapi -<i>-	880
Mayroon o walang -<g>- (No.	667
gitling vs. walang gitling	635
<i> vs. <e>	592

Ipinapakita sa mga Talaan 5, 6, at 7 ang limang pinakamataas na mga tuntunin ng transpormasyon ayon sa bilang ng naobserbahang *distinct case pairs* sa mga wikang Filipino/ Tagalog, Sebwano-Bisaya at Ilokano. Makikita sa mga talaan base sa bilang ng varyant ng ispeling ng salitang tambalang inobserbahan na ang may hindi tiyak na tuntunin sa tatlong wika ay ang paggamit ng simbolong gitling, paggamit ng <o> vs, <u>, at paggamit ng <i>vs. <e>. Sumangguni sa Apendiks I para sa lahat ng mga tuntunin ng pagbabago sa ispeling ng mga kaso ng salitang tambalan na nasa pag-aaral na ito.

Upang mabigyan ng kantidad ang mga lebel ng baryasyon sa gamit ng wika sang-ayon sa kung gaano nagkakalayo ang aktuwal na preperensya sa gamit ng wika, inihaharap namin ang katalogo ng mga baryasyon (V.I.) *metric identified with a Transformation rule*, na ipinaliliwanag sa pamamagitan ng equation(1).

Upang mailarawan, ang *Tuntunin ng Transpormasyon Blg. 2* (<o> vs. <u>) para sa wikang Ilokano.

Mayroong kabuuang 98,783 bilang ng paggamit sa salitang gumagamit ng karakter na <o>, at 25,106 na kabuuang bilang ng paggamit ng salita na ang karakter na <o> ay napalitan ng karakter na <u>.

Dahil dito, ang kabuuang bilang ng paggamit ng Tuntunin ng Transpormasyon Blg. 2 ay 123,889, na mayroong *absolute difference between the all case pairs computed as 73,667*. Kung kaya't ang antas ng V.I. para sa *Tuntunin ng Transpormasyon Blg. 2* ay 40.53%.

Ang *Variation Inde*x ay may antas na mula 0% hanggang 100%, kung saan ang mataas na antas ay nangangahulugan ng mataas na lebel ng hindi pagsang-ayon sa mas kinikilingang gamit para sa ibang varyant ng ispeling na pinapatunayan ng text copus para sa partikular na wika. Ipinapakita ng Talaan 8, 9, at 10 ang limang pinakamataas na *Tuntunin ng Transpormasyon* na inihanay ayon sa kaugnay na antas ng Variation Index. Tandaan na ang pinakamataas na *Tuntunin ng Transpormasyon* ay magkakaiba sa tatlong pangunahing wika sa Pilipinas.

Paghahambing ng mga Kinikilingang Varyant ng Ispeling na Pare-parehong Ginagamit ng Wikang Filipino/Tagalog, Sebwano-Bisaya at Ilokano

Dalawampu't dalawa (22) sa 42 Tuntunin ng Transpormasyon sa Ispeling ang maaaring maobserbahan sa lahat ng tatlong pangunahing wika sa Pilipinas na pinag-aaralan. Nangangahulugan ito na 52.4% ng mga kategorya ng varyant ng ispeling na natukoy ay ginagamit pareho ng tatlong pangunahing wika. Ipinapaalala lamang ng antas na ito ang pag-aaral ni Paz et al. (2005) na nagsasaad na ang mga wika sa Pilipinas ay mayroong iisang sentrong pinagmulan/*universal nucleus* ng batayang bokabularyo; na may humigit-kumulang 50% ng mga bokabularyo ng wika sa Pilipinas ay magkakapareho sa lahat ng ibang wika sa Pilipinas[14]. May 32 kategorya ng varyant ng ispeling ang Filipino/Tagalog, ang Ilokano ay may 34 kategorya, samantala ang Sebwano-Bisaya ay may 24 na kategorya. Sa tatlong wika, ang Filipino/Tagalog at Ilokano ang may maraming pagkakapareho ng uri ng varyant ng ispeling (26

karaniwang mga salita), samantalang ang mga wikang Ilokano at Sebwano-Bisaya ay may 21 salitang magkakapareho. Sa kabilang banda ang wikang Filipino/Tagalog at Sebwano-Bisaya ay may 22 pagkakapareho ng mga varyant ng ispeling.

Talaan 8. Paghahanay ng mga Tuntunin ng Transpormasyon sa Filipino/Tagalog ayon sa Variation Index (V.I.)

Tuntunin ng Transpormasyon	V.I. antas (%)
<ch>vs. <ts>	98.59
<uw>vs. <w>	98.35
<y> vs. <i>	96.23
<ll> vs. <ly>	82.44
<ng> vs. <n>	81.31

Talaan 9. Paghahanay ng Tuntunin ng Transpormasyon sa Ilokano ayon sa Variation Index (V.I.)

Tuntunin ng Transpormasyon	V.I. antas (%)
Mayroon o walang -<g>-	97.85
<ll> vs. <ly>	97.78
<ng> vs. <n>	94.12
Mayroon o walang -<m>-	93.11
<ng> vs. <g>	92.96

Talaan 10. Paghahanay ng Tuntunin ng Transpormasyon sa Sebwano-Bisaya ayon sa Variation Index (V.I.)

Tuntunin ng Transpormasyon	V.I. antas (%)
<uw> vs. <w>	96.83
<v> vs. <b>	90.59

<aw> vs. <ao>	84.50
<ch> vs. <ts>	83.33
<y> vs. <i>	81.40

Marami sa mga karaniwang Tuntunin ng Transpormasyon ng Ispeling ay bahagi ng sistemang ortograpiya ng lumang Espanyol, na simulang napalitan ng modernong sistema ng ortograpiya sa simula ng ika-20 siglo. Labing-isa (11) sa 13 kategorya ng varyant ng ispeling na maoobserbahan sa lahat ng tatlong wika ay maiuugnay sa sistema ng pagsulat ng lumang Espanyol. Lahat ng wika ay may mga varyant ng ispeling na may kinalaman sa paggamit ng mga patinig (i.e.<o> vs. <u> at <i> vs. <e>), sa paggamit ng simbolong gitling, ganoon din sa mga kasong ukol sa mga kambal-patinig.

Bagama't nagkakatulad ang tatlong wika sa karamihan ng mga nabanggit na salita ng varyant ng ispeling, ilan sa mga ito ay makikita lamang sa isa o dalawang wika.

May uri ng varyant sa ispeling sa wikang Ilokano na nabubuo mula sa geminate na makikita sa ponetikong imbentaryo (tuntunin blg. 10 - 13), isang kategorya ng varyant ng ispeling na hindi makikita sa dalawang pangunahing wikang isinaalang-alang.

Makikita sa ating text corpora, na ang mga kategorya ng varyant ng ispeling na dahilan ng pagsasama ng inulit na salita at paglalapi (tuntunin blg. 29 - 33) ay madalas makikita sa wika ng Filipino/Tagalog, katamtaman naman ang makikita sa Wikang Ilokano, samantalang wala namang makikita sa Wikang Sebwano.

MAAARING MGA PALIWANAG SA PAGKAKAROON NG MGA VARYANT NG ISPELING

Euphony at mga Morpoponemik

Ang tuntunin na kung ano ang bigkas siya ang baybay ay isang problema sa multilingual

na sitwasyon ng wika. Maaari itong magawa sa monolingual na kalagayang makikita sa implementasyon ng Abakada sa simula ng taon ng Filipino, na ang lexicon ay karamihang binubuo ng mga salitang Tagalog. Sa pagdami ng mga hiram na salita na isinama sa lexicon, at sa likas na pag-unlad ng wikang Filipino bilang pambansang lingua franca na ginagamit ng mga mananalita ng iba't ibang katutubong wika, ang tuntunin ay hindi naging kapaki-pakinabang sapagkat may mga baryasyon na nakakahadlang sa ibang mga katutubong wika. Kung ano ang maganda sa pandinig ng isang grupo ng wika ay hindi naman maganda sa isa.

Ang mga salita at mga panlapi ay may ibang anyo kapag inilalagay sa ibang katangiang ponemika. Nangyayari ito upang mas madaling bigkasin ang mga salita o ito ay mas magandang pakinggan. Ang mga pagbabagong ito sa pagbigkas ay matagal nang nangyayari at napunta sa kaugalian ng pagsusulat.

MGA KONGKLUSYON AT REKOMENDASYON

Ang mga iniharap na mga varyant ng ispeling sa papel na ito ay simula lamang ng mahabang listahan; ang pagpili ng mga salita ay hindi puspusan at sa halip ay eksploratoryo lamang. Gaya ng nabanggit, ang pag-aaral na ito ay maaaring mapabuti pa sa sumusunod na rekomendasyon:

- Ilan sa mga kategorya ng mga varyant ng ispeling ay maaaring maiuri sa isang tiyak na tuntunin ng wika na nagiging resulta ng mga Tuntunin ng Transpormasyon. Halimbawa, ang paggamit ng gitling (Tuntunin blg. 27) ay mayroong pitong hiwalay na kaukulang tuntunin na makikita sa[15], kung saan makikita ang antas ng mas kinikilingang gamitin. Maaaring pagtuunan ng mga susunod na pag-aaral ang pagtukoy ng mga sub-kategoryong ito para sa mas mahusay at mas masinsinang analisis ng aktuwal na paggamit ng wika na korpora.
- Kailangan pang imbestigahan ang mga *euphonious pairs* at ikonekta sa pangingibabaw ng mga varyant ng ispeling. Mga halimbawa ay ang gemination (hal. *dakkel* vs. *dakel*)

at ang paggamit ng unlapi na <i>- (hal. *ipinasa* vs. *pinasa)*.

- Ang mga datos na natipon ay walang *date stamps,* kung kaya ito ay hindi posibleng maiugnay sa mga varyant ng ispeling sa isang partikular na panahon. Mas mauunawaan kung paano nagbabago sa paglipas ng panahon ang mas higit na ginagamit na varyant ng ispeling sa pamamagitan ng analisis ng pangkasaysayang korpora.
- Ang paraan para sa awtomatik na pag-alis ng mga varyant na ispeling na ginamit sa pag-aaral ay may limitasyon na maaaring alisin sa mga susunod na pag-aaral: (1) ang Levenshtein edit distance of 2 ay maaaring hindi gamitin dahil hindi naman lahat ng mga varyant ay maaaring masakop ng ganitong lapit, at (2) ang mga operasyon ng pag-eedit ay maaaring magresulta ng pagkakaiba.
- Sa ibang bahagi ng baryasyon ng gramatika at ortograpiya na nangangailangan ng karagdagang impormasyon para sa pagpapasiya (hal. bahagi ng pananalita, lemma) ay hindi maaaring masiyasat na gamit ang aming lapit. Maaari pang palawakin ang sakop ng mga salita sa teknolohiyang ginamit na Computational linguistic (hal. POS Taggers at morphological analyzers), tulad ng paggamit ng *nang* vs. *ng* sa Filipino/ Tagalog (kung mayroong POS Tagger) o *byag* vs. *biyag* (gamit ang morphological analyzers sa pagtatapat-tapat ng mga varyant na *agbyag* at *panagbiyag*).

Sa huli, matapos mapabuti ang mga tuntuning gagabay sa wastong paggamit ng wika, kami naman ay magtataguyod ng mga mekanismo para sa pagpapatatag ng wastong edukasyon sa pamamagitan ng mga regular na publikasyon ng aklat ng gramatika at terminolohiya na may istandard na ispeling.

MGA TALA

[1] Salin ng papel na *Comparative analysis of actual language usage and selected grammar and orthographical rules for Filipino, Cebuano-Visayan and Ilokano: a Corpus-based Approach* na binasa sa 2nd Philippine Conference Workshop on Mother Tongue-Based Multilingual Education noong 16-18 Pebrero 2012 sa Lungsod Iloilo.

[2] *Department of Education - Philippines Order No. 74 s. 2009,* Department of Education, 2009.

[3] Dekker, Diane. "Key components of a MTBMLE Program," sa *First MLE Conference*, Cagayan De Oro, 2010.

[4] Nolasco, Ricardo., 21 reasons why Filipino children learn better while using their mother tongue: A Primer on Mother Tongue-based Multilingual Education (MLE) & Other Issues on Language and Learning in the Philippines, Guro Formation Forum, University of the Philippines, 2009.

[5] Comandante Jr., B. "Ancient Baybayin: Early Mother Tongue-Based Education Model," sa *1st MLE Conference, "Reclaiming the Right to Learn in One's Own Language"*, Capitol University, Cagayan de Oro, Feb 18-20, 2010.

[6] Tanangkingsing,M. A functional reference grammar of Cebuano: from a discourse perspective, Volume 1. LAP Lambert Academic Publishing, 2011.

[7] Wolff, J.U. Cebuano-Visayan Dictionary, The Linguistics Society of the Philippines and the Southeast Asia Program. Cornell University, 1972.

[8] Rubino, C.G. Ilocano-English / English/ Ilocano Dictionary and Phrasebook, 2005 ed. New York: Hippocrene Books Inc., 1998.

[9] Benosa, S. E. "An Ilocano Orthography for MTB-MLE," Diliman, Quezon City.

[10] Stone, R. at N. Zamora. "Designing an Alphabet for an Unwritten Language," sa *1st MLE Conference, "Reclaiming the Right to Learn in One's Own Language"*, Capitol University, Cagayan de Oro, Feb 18-20, 2010.

[11] Ilao, Joel at Rowena Guevara. "Mining Filipino-English Corpora from the Web," sa *International Symposium on Multimedia and Communication Technology*, Manila, September 8-10, 2010.

[12] Ilao, Joel, et al. "Bantay-Wika: towards a better understanding of the dynamics of Filipino culture and linguistic change," sa *9th Workshop on Asian Language Resources*, Chiang Mai, Thailand, 2011.

[13] Levenshtein. "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady,* vol. 10, blg. 8, pp. 707-710, 1966.

[14] Paz.Consuelo J. *Ang Wikang Filipino: atin ito,* Quezon: UP - Sentro ng Wikang Filipino, 2005.

[15] Zafra, Galileo S. et al., Gabay sa Ispeling, Quezon City: Sentro ng Wikang Filipino - UP Diliman, 2008.