SukatWika: An Analysis Software for Linguistic Properties of Texts

Kathrina Lorraine Lucasan[1], Angelina Aquino[2], Francis Paolo Santelices[3], Dina Ocampo[4]

[1,4]Center for Integrative and Development Studies - Education Research Program
[2,3]Electrical and Electronics Engineering Institute
University of the Philippines, Diliman, Quezon City, Philippines
{kmlucasan, dina.ocampo}@up.edu.ph, {angelina.aquino, francis.santelices}@eee.upd.edu.ph

Abstract

There is a lack of understanding on the qualities of texts that children can read, especially in Philippine languages. Text quality should be informed by an analysis of text difficulty, which can be measured by the linguistic properties of text such as word density, concept load, and phonological weight. The SukatWika analysis tool was developed to automate the extraction of this information for texts written in Filipino, English, Sinugbuanong Binisaya, and Ilokano languages. The results obtained from this software can be used as an aid in the creation of instructional materials which will support reading development among learners.

Keywords: text analysis software, linguistic properties, Philippine languages, literacy assessment

Buod

Mayroong kakulangan sa pag-unawa sa kalidad ng tekstong nababasa ng mga bata, lalo na sa mga wika sa Pilipinas. Dapat nakahango ang kalidad ng teksto sa pagsusuri ng antas nito, na maaaring sukatin mula sa mga katangiang panlinggwistika ng teksto tulad ng haba at dami ng salita, *concept load*, at sa mga uri ng tunog na nakapaloob sa mga salita. Ang SukatWika ay binuo upang mabilisang makuha ang mga impormasyong panlinggwistika ng tekstong nasa wikang Filipino, Ingles, Sinugbuanong Binisaya, at Ilokano. Ang impormasyong maibibigay ng SukatWika ay makatutulong sa paggawa ng mga kagamitang panturo upang malinang ang kakahayan ng mga mag-aaral sa pagbasa.

1. Introduction

Mother Tongue-Based Multilingual Education (MTB-MLE) is an education program for children wherein they learn to read and write first in their mother tongue (MT) and then use their MT as they learn to understand, speak, read, and write in other languages (UNESCO, 2018). MTB-MLE programs may be for maintenance, for transition, or for enrichment (Lin and Man, 2009). In the Philippines, the MTB-MLE program is implemented from Kindergarten to Grade 3. The program is transitional as the goal is to provide learners with the support they need to gradually and effectively move from MT instruction to mostly Filipino and/or English instruction (Department of Education, 2019). Learning literacy in the mother tongue serves as the foundation for learning literacy in other languages (Cummins, 2003).

The Department of Education released several policies for its implementation, with one released in 2009 to institutionalize the program for Kindergarten to Grade 3, and two others in 2012 and 2013 containing the guidelines and the list of official languages, including Filipino, English, Sinugbuanong Binisaya, and Ilokano. Section 5 of Republic Act (RA) 10533 or the Enhanced Basic Education Act of 2013 stipulated the features of the K to 12 curriculum and mandated that it adhere to the principles and framework of MTB-MLE. The most recent policy was released in 2019, which articulated provisions further, including a guide for possible classroom scenarios.

Many studies discuss the advantages of the use of the MT in schools including increased classroom participation, positive affect, and increased self-esteem (UNESCO, 2004), flexibility with learning strategies (Dahm and De Angelis, 2018), faster learning of a second language (Monje et al., 2019), and increased academic achievement (Nguyen, 2017).

Because the MTB-MLE program has only been recently implemented in the Philippines, many aspects for improvement have been observed (Monje et al., 2019). Cordero (2019) for example, includes the availability of assessment tools and resources among topics that need more research.

1.1. Multi-Literacy Assessments for Filipino Learners

Multi-Literacy Assessments for Filipino Learners is a battery of assessment tools which will measure learners' skills in various literacy domains. It is part of a larger research agenda investigating lifespan literacy development of Filipinos. For its first phase, it aims to develop assessment tools for early literacy in Filipino and English and the mother tongues Sinugbuanong Binisaya and Ilokano. The literacy skills to be assessed include oral and written language development, alphabet knowledge, spelling, decoding, and listening and reading comprehension.

Among the initial steps of development was to conduct an inventory of existing assessment tools made by researchers of the university and secure their permission to adopt/adapt their tool. A priority step in the assessment package development process was to find a way to ensure that test items in the tools were drawn from texts and books that typical Kindergarten to Grade 3 elementary students in the Philippines ordinarily encounter.

2. SukatWika Analysis Tool

In line with the goals of this project, the SukatWika (Filipino: lit. "Measure Language") program was developed to

create a corpus of children's language and texts and to facilitate the analysis of linguistic properties in Philippine texts. Specifically, it automates the counting of various metrics at the sub-word, word, and sentence level which are typically used to assess the surface difficulty levels of reading and instructional materials.

2.1. Functionality

As of Version 1.0, SukatWika is capable of analyzing texts written in four Philippine languages: Filipino, English, Sinugbuanong Binisaya, and Ilokano. Given a text document in these languages, the program then provides the lengths and frequencies of lexical and grammatical units, as well as an interface for searching lexical units within the text.

Paragraph length counter. This displays the total number of paragraphs in the text, and enumerates the frequency of paragraph lengths, answering the question "How many paragraphs contain n sentences?"

Sentence length counter. This displays the total number of sentences in the text, and enumerates the frequency of sentence lengths, answering the question, "How many sentences contain n words?"

Phoneme counter. This displays the total number of phonemes in the text, tabulates the frequency of appearance of individual phonemes, and enumerates the frequency of word lengths by phoneme, answering the question "How many words contain n phonemes?" It also gives a list of unique words in the text, ordered by the number of phonemes in each word.

Word frequency counter. This displays the total number of words in the text, tabulates the frequency of appearance of unique words in the text, and sorts the words alphabetically and by frequency.

Word length counter. This displays the total number of syllables in the text, and enumerates the frequency of word lengths, answering the question "How many words contain n syllables?" It also gives a list of unique words in the text, ordered by the number of syllables in each word.

Word searcher. This allows the user to input a string of characters, and gives a list of words containing the string. It also provides options to filter the words displayed by the number of syllables contained in the word, as well as the position of the string in the word (i.e. start, middle, or end of the word).

2.2. Orthography

For the Philippine langauges supported by the program, the rules for syllabication and phonemic transcription of individual words were based on official orthographies for Filipino (Almario (Ed.), 2014), Sinugbuanong Binisaya (Akademiyang Bisaya, 2011), and Ilokano (Komisyon sa Wikang Filipino, 2012). For English, phonemic transcriptions were extracted from the CMU Pronouncing Dictionary (Carnegie Mellon University, 2014), while syllable counts were performed by simply counting the number of vowels in the phonemic transcription, since each syllable in an English word is known to contain only one sonant or vowel sound (Malone, 1957).

Tokenization rules were identical for all four languages: words were tokenized based on whitespace, while sentences and paragraphs were tokenized based on end-of-sentence and newline characters. The parsing rules for each type of analysis were then encoded as Python functions and used in the succeeding scripts to produce the necessary outputs.

2.3. User Interface

In order for the functions to be user-friendly and to be able to visualize the results, a graphical user interface was created based on PyQt5. PyQt5 is a comprehensive set of Python bindings for Qt5. Qt is set of cross-platform C++ libraries that implement high-level APIs (Application Programming Interfaces) for accessing many aspects of modern desktop and mobile systems. Shown in Figure 1 is the starting window wherein no text has been analyzed yet.

Figure 1: Starting window of the SukatWika user interface

Each of the tabs that can be seen in the lower part of the user interface displays each of the features mentioned in the previous section. For example, the test results of the word length counter can be seen in Figure 2. Also seen below the tabs is the status bar which displays the path to the current analyzed file, the language selected, and instructions to export the results. The full SukatWika analysis can be exported to a comma-separated values (.csv) file when the 'Export as .csv' button is clicked. The .csv file can then be opened in a text editor or spreadsheet application.
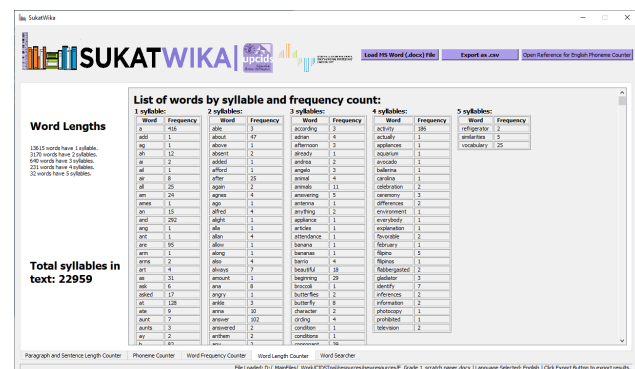

Figure 2: Tab of the word length counter showing the results of the analysis of SukatWika

An added feature that serves as an aid to users is the Phoneme Reference Guide for the English text analysis

based from the CMU Pronouncing Dictionary. This can be accessed by clicking the 'Open Reference for English Phoneme Counter' button found in the upper right portion of the user interface.

Version 1.0 of SukatWika can only support Microsoft Word (.docx) files as input and is compatible with the Windows Operating System. All dependencies are already included in the distribution of SukatWika v1.0.

3. Evaluation of Texts

Learner's materials specific for each language and grade level were gathered and collated into one file. These files were then uploaded into the program for analysis with the analysis results discussed in the succeeding section. Filipino and English versions of the tools were developed first and will serve as the basis for the Sinugbuanong Binisaya and Ilokano versions. Table 1 shows the number of learner's materials analyzed which were downloaded from the DepEd Learning Resources Portal. They are a varied collection of learner's materials made by DepEd and its partners.

	Filipino	English
Kinder	9	No texts available[1]
Grade 1	20	6
Grade 2	17	12
Grade 3	16	12

Table 1: Texts analyzed per grade level

3.1. Text Analysis Results

Syllable counter results for both English and Filipino show the progression of the length of words encountered by learners as they moved from one grade level to another. Table 2 shows the results generated by SukatWika in terms of the number of 1- to 3-syllable words in Kindergarten to Grade 3 texts.

Table 2 shows that learners in all grade levels are mostly exposed to 1-syllable words for both Filipino and English texts. The table also shows that the number and percentage of 2- and 3-syllable words increase substantially as learners go on to the next grade levels, signifying increasing complexity of texts.

This is supported by the data from the phoneme counter presented in Tables 3 and 4. Though the tables only show a specific portion of the results, the increasing frequencies of the phonemes from Kindergarten to Grade 3 can still be clearly observed, confirming the increasing level of text complexity shown by the syllable counter data.

The same trend of increasing text complexity can also be observed in the sentence and paragraph length counter results. Table 5 shows part of the sentence length counter results while Table 6 shows part of the paragraph length counter results.

---

[1]The mother tongue is the mode of teaching and learning in Kindergarten.

3.2. Using Text Analysis Results

3.2.1. Development of the Multi-literacy Assessments for Filipino Learners

Though an increase in text complexity is expected as learners moved from one grade level to another, SukatWika results provide the necessary details to make sound decisions on items to be included in the assessment tools. For example, the results of the phoneme counter, word frequency counter, and word length counter influenced the choice of words to be included in phonics and word reading, and spelling assessments. It enabled the assessment materials to provide a progression of word length and complexity based on the data generated by the SukatWika analysis. In identifying the list of words for the word reading test, the frequency counter results served as the basis for inclusion into the assessment tools. When listing the words for the word reading assessment, the syllable counter results and phoneme counter results validated the words that were included in the test. For example, the phoneme /ng/ was excluded from the Kindergarten assessment tools because SukatWika analyses showed that this phoneme occurred more frequently in Grade 1 materials indicating that learners had more experience with this phoneme at that level (See Table 3). It would not have been judicious to include the phoneme /ng/ in Kindergarten tools because the learners at this level can be assumed to have insufficient exposure to it in printed texts.

3.2.2. Other Uses

Sentence and paragraph length counter results may also be used as criteria for those who may want to create stories for a specific grade level. Story writers would simply need to write within the target grade level's analysis results to ensure that target readers will be able to read the text accurately.

SukatWika may also be utilized to determine an existing text's readability level. Existing stories and other learner's materials may be uploaded into the tool and the results of the analysis could then be compared with the results generated from the DepEd learner's materials to establish the material's reading level.

When planning spelling and reading lessons, teachers could also use the word search capability of SukatWika to generate words with the specific consonant blends, digraphs, or phonograms which they are studying in class. This may be especially helpful for reading remediation classes.

4. Conclusion

SukatWika enabled the development of assessment tools based on text and word properties that learners encountered in school. Since the tools are drawn from materials which learners use in school, it will yield accurate assessment results based on the exposure of learners to the printed materials.

Aside from those listed in this paper, SukatWika will have many other possible uses for teaching and assessment. It will be useful for many contexts and will hopefully open more opportunity to support reading development of all Filipino learners.

	Filipino			English		
	1-syllable words	2-syllable words	3-syllable words	1-syllable words	2-syllable words	3-syllable words
Kinder	2,999	1,582	871	No texts available[1]		
Grade 1	11,230	6,250	4,118	13,487	3,041	612
Grade 2	23,503	13,969	9,384	14,689	3,627	776
Grade 3	29,482	23,128	13,813	64,549	20,274	5,494

Table 2: Number of 1- to 3-syllable words in Kindergarten to Grade 3 Filipino and English texts generated by SukatWika

	Filipino							
	Kindergarten		Grade 1		Grade 2		Grade 3	
	Phoneme	Frequency	Phoneme	Frequency	Phoneme	Frequency	Phoneme	Frequency
Consonant phonemes	n	2,432	n	7,907	n	23,276	n	32,618
	t	1,494	ng	6,635	ng	16,045	ng	21,887
	l	1,334	s	5,402	t	13,181	t	17,790
	s	1,288	t	4,741	s	12,455	s	17,709
	k	1,208	l	4,304	m	11,092	l	15,891
Vowel phonemes	a	7,480	a	25,115	a	69,417	a	89,617
	i	1,942	i	8,225	i	21,404	i	32,624
	u	954	o	3,811	o	11,295	o	15,877
	o	868	u	3,280	u	10,342	u	12,449
	e	323	e	1,380	e	2,805	e	4,863

Table 3: Phoneme counter results for Filipino texts

Potential improvements to the software include support for additional Philippine languages, text normalization for special characters such as numbers and mathematical symbols, language identification for bilingual texts, and compatibility with other input file formats and operating systems.

5. Acknowledgements

6. Bibliographical References

Akademiyang Bisaya. (2011). *Cebuano Phonetics and Orthography*. Author, Cebu.

Almario (Ed.). (2014). *KWF Manwal sa Masinop na Pagsulat*. Komisyon sa Wikang Filipino, Cebu.

Carnegie Mellon University, (2014). *Carnegie Mellon Pronouncing Dictionary, version 0.7b*. Retrieved from http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Cordero, G. (2019). The basic education research agenda of the department of education and system assessment in the k to 12 basic education curriculum. In D. Ocampo et al., editors, *Key Issues in Curriculum, Assessment, and ICT in Basic Education*, pages 11–30, Quezon City. University of the Philippines Center for Integrative and Development Studies.

Cummins, J. (2003). Bilingual education: Basic principles. In A. Housen J. M. Daelewe et al., editors, *Bilingualism: Language and Cognition*, pages 56–66, England. Multilingual Matters.

Dahm, R. and De Angelis, G. (2018). The role of mother tongue literacy in language learning and mathematical learning: Is there a multilingual benefit for both? *International Journal of Multilingualism*, 15:194–213.

Department of Education, (2019). *Policy Guidelines on the K to 12 Basic Education Program*. Retrieved from https://www.deped.gov.ph/wp-content/uploads/2019/08/DO_s2019_021.pdf.

Komisyon sa Wikang Filipino. (2012). *Tarabay iti Ortograpia ti Pagsasao nga Ilokano*. Author, Manila.

Lin, A. and Man, E. (2009). *Bilingual Education: Southeast Asian perspectives*. Hong Kong University Press, Hong Kong.

Malone, K. (1957). Syllabication. *College English*, 18(4):202–207.

Monje, J. D., Orbeta, A., Francisco-Abrigo, K., and Capones, E., (2019). 'Starting where the children are': A process evaluation of the mother tongue-based multilingual education implementation. Retrieved from https://pidswebs.pids.gov.ph/CDN/PUBLICATIONS/pidsdps1906.pdf.

Nguyen, M. (2017). Bilingual education helps to improve the intelligence of children. *World Journal of English Language*, 7:11–17.

	English					
	Grade 1		Grade 2		Grade 3	
	Phoneme	Frequency	Phoneme	Frequency	Phoneme	Frequency
Consonant phonemes	t	3,756	t	4,256	t	22,881
	n	3,568	n	3,979	n	21,540
	d	2,574	d	2,937	r	16,012
	l	2,398	s	2,750	d	14,284
	s	2,310	m	1,856	l	13,380
Vowel phonemes	i (ink)	2,693	i (ink)	3,160	i (ink)	17,016
	e (modest)	1,938	a (apple)	1,828	e (modest)	11,224
	a (apple)	1,626	a (align)	1,585	a (align)	8,747
	a (align)	1,477	e (egg)	1,498	a (apple)	8,693
	e (egg)	1,456	o (button)	1,103	e (egg)	6,898

Table 4: Phoneme counter results for English texts

	Filipino			English		
	Sentences with 2 words	Sentences with 3 words	Sentences with 4 words	Sentences with 2 words	Sentences with 3 words	Sentences with 4 words
Kinder	82	89	155	No texts available[1]		
Grade 1	562	580	303	832	614	515
Grade 2	967	483	565	592	358	491
Grade 3	873	893	763	1,880	2,107	1,512

Table 5: Sentence length counter results

	Filipino			English		
	Paragraphs with 2 sentences	Paragraphs with 3 sentences	Paragraphs with 4 sentences	Paragraphs with 2 sentences	Paragraphs with 3 sentences	Paragraphs with 4 sentences
Kinder	138	22	9	No texts available[1]		
Grade 1	989	144	68	1,698	51	37
Grade 2	2,276	435	125	912	125	76
Grade 3	1,607	586	281	3,026	80	257

Table 6: Paragraph length counter results

UNESCO, (2004). *The importance of mother tongue-based schooling for educational quality.* Retrieved from `https://unesdoc.unesco.org/ark:/48223/pf0000146632`.

UNESCO, (2018). *MTB MLE resource kit: Including the excluded: Promoting multilingual education.* Retrieved from `https://unesdoc.unesco.org/ark:/48223/pf0000246278`.