Kasetsart Journal of Social Sciences

journal homepage: http://kjss.kasetsart.org

Development of automated scoring system for Thai writing ability test of primary education level

Asanee Tongsilp, Kamonwan Tangdhanakanond*, Nhabhat Chaimangkol

Educational Measurement and Evaluation, Department of Educational Research and Psychology, Chulalongkorn University, Bangkok 10330, Thailand

Article Info

Article history:
Received 29 November 2022
Revised 20 September 2023
Accepted 2 October 2023
Available online 30 August 2024

Keywords:
automated scoring system,
automatic essay scoring,
Thai writing ability

Abstract

This research aimed to: (1) develop the automated scoring system for Thai writing ability test of primary education level, and (2) evaluate the efficiency of the system. The study participants were students enrolled in academic year 2020 in schools in Bangkok: (1) 287 sixth-grade students by purposive sampling from schools in Bangkok with the average score of Thai language from O-NET that was high, medium, and low; and (2) 30 sixth-grade students in schools under the Office of the Private Education Commission by purposive sampling because schools participated willingly, having computer labs and internet connection (i.e., to test online system). Research instruments consisted of: an essay test, three evaluation forms, and the online system developed by PHP language and MySQL database. Results were as follows: (1) The automated scoring rubric system for the Thai writing ability test of primary education level was an online system comprised of 3 steps, i.e., data entry answer as text, automated scoring consisting of Thai word segmentation and scoring rubric, and display of output. The testing system found that the score was *M* = 8.52, *SD* = 4.07 and CV = .48, and (2) The efficiency evaluation of the system by using three evaluations forms revealed that rater agreed with the system, which had the highest agreement. The automated scoring rubric was able to predict the test score at .05 level of statistical significant, R^2 = 66.3 to 87.6 percent. M*easure* of agreement of scoring result were with ICC = .88, and RMSE ≤ 3.38.

* Corresponding author.
E-mail address: tkamonwan@hotmail.com (K. Tangdhanakanond).

https://doi.org/10.34044/j.kjss.2024.45.3.05
2452–3151/© 2024 Kasetsart University.

Introduction

An essay test is an instrument used to measure higher-level learning. Efficacy of a subjective test consists of quality of the test, test answers, scoring rubric, rater skills, and scoring monitoring (Kanjanawasee, 2013). Ali and Michael (2010) found that raters had significant influence on scoring. They must be trained, and scoring rubric must have clarity in the meanwhile. In Thailand, an essay test was first included in the Ordinary National Educational Test (O-NET) administered in Thai language subject to 789,951 sixth-grade students in 2016. The problem was the budget of more than 30 million baht is spent on test management, in which partial expenditure is spent on hiring and training 2,800 test proctors across the country; in addition, a long time to score, causing the results to be announced slowly. (Daily new, 2018) In other countries, due to computer technology advancement, problems about budgeting essay scoring are solved through implementing automated essay scoring (AES) to minimize cost, resources, and rater errors (Dikli, 2006).

AES was first created in the United States in 1996 by Page. To date, it has been developed and translated into multiple languages such as languages spoken in India, Korea, and Japan. AES system has been used for national testing and international standardized tests such as TOEFL, GMAT, IELTS, and SAT. AES employs the computer technology called text analysis using natural language processing (NLP), including knowledge based, statistical based, and hybrid approach. There are two analysis methods: lexical and syntactic analyses (Kawtrakul et al., 1999). Research in Thailand shows that there has been the development of AES system using statistical based technique i.e., lexical analysis and NLP approach comparing responses with test answers stored in knowledge base. However, research involving automated scoring which uses Thai writing ability rubrics has not been found. This study research aimed to; (1) develop the Automated Scoring System for Thai writing ability test (ASST system) of primary education level, and (2) evaluate the efficiency of the system. The ASST system is an online program that helps teachers to score Thai writing ability. The benefits of the ASST system are a clear scoring rubric, time saving, and reliability in scoring.

Literature Review

Automated essay scoring was first created by Page, with PEG™ system developed during 1966–1990. The development was then continued by ETS: IEA developed by Peter Foltz and Thomas Landauer in 1997, IntelliMetric® system developed by Vantage in 1998, e-rater® system developed under the collaboration between ETS and Burstein, and meanwhile, the criterion was developed in 1999, BETSY system developed by Rudner in 2002, CRASE® system developed by Mitzel and Lottridge in 2007, and the Hewlett foundation sponsor ASAP (Attali & Burstein, 2006). Afterwards, system development continued to occur in multiple languages: Lahitani et al. (2016) developed a technique to identify terms in Indonesian using TF-IDF; Ke et al. (2016) developed the system in Chinese called CDES; Dascalu et al. (2017) developed the system in Dutch using NLP; and Yamamoto et al. (2018) developed the system in Japanese using machine learning together with rubric, classifying group using SVM via multiple kernels, and predicting a model using a decision tree method. Clearly, during the past 56 years (from 1966–2022), computer technology has been greatly developed, with high-speed internet; text processing for each language, NLP processing via machine learning. These breakthroughs in current technology have inspired this study to develop a Thai written text scoring system further.

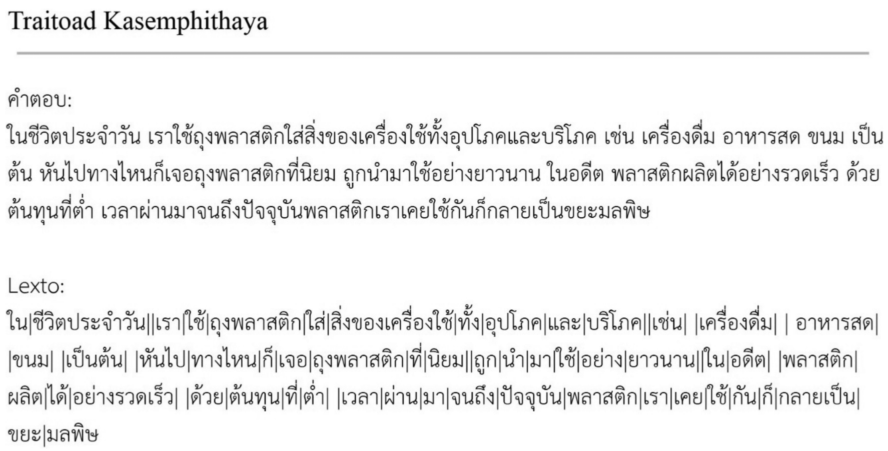
Rubrics for evaluating Thai writing ability was developed from the summary writing of the National Institute of Educational Testing (2018) and synopses writing of the Office of the Basic Education Commission (2018). All rubrics were developed using automated scoring based on (Yamamoto et al., 2018). Rubric comprised paraphrasing, keywords, words for expressing opinion, words for giving examples, words for adding explanations, punctuation, key ideas, pronouns 1 and 2, spelling, sentences in an essay, and a complete sentence. Thai writing ability test via three summary tests consists of the storytelling, the scientific article, and the social and cultural article. This automated scoring system consists of 3 steps: (1) data entry answer as text, (2) automated scoring, and (3) display of output.

Thai Word Segmentation

Thai Word Segmentation is a process where texts are sensibly segmented. Thai language has complicated sentence structure, where words are written consecutively without space; several clauses are connected, which creates a very long sentence; there is no symbol marking the end of a sentence, and the beginning of a sentence is not always a noun. There are three approaches used to segment Thai words: rule-based, dictionary-based, and corpus-based categorized into probabilistic word segmentation and feature-based word segmentation using the N-gram model (N value ranges from 2-gram, 3-gram, …, n-gram) (Urathumkul & Runapongsa, 2006). Moreover, the algorithms used are categorized into longest matching, maximal matching, probabilistic model, and feature-based approach (Thai Encyclopedia Project Committee for Youth, 2017). Currently, the software used for Thai word segmentation includes word analysis for Thai-SWATH, PyThaiNLP, thainlplib, LexToPlus, and TLex. The current study uses LexToPlus, a dictionary-based software with longest matching technique. It is a software with high accuracy of Thai word segmentation; users can add terms as needed in the meanwhile. It is developed by the National Electronics and Computer Technology Center (NECTEC), the service for Thai word segmentation is provided through API linked with the corpus (National Electronics and Computer Technology Center, 2019). Thai word segmentation is illustrated in Figure 1.

Rubric as Evaluation – Scoring Criteria

Evaluation criteria for Thai writing ability should use analytic scoring rubrics, an approach providing information in detail such as the main idea, supporting ideas, content, spelling, and sentence structure to improve writing ability. Evaluation criteria for Thai writing abilityused as a guideline developed for an automated scoring rubric by a computer are seen in Table 1.



Traitoad Kasemphithaya

คำตอบ:
ในชีวิตประจำวัน เราใช้ถุงพลาสติกใส่สิ่งของเครื่องใช้ทั้งอุปโภคและบริโภค เช่น เครื่องดื่ม อาหารสด ขนม เป็น
ต้น หันไปทางไหนก็เจอถุงพลาสติกที่นิยม ถูกนำมาใช้อย่างยาวนาน ในอดีต พลาสติกผลิตได้อย่างรวดเร็ว ด้วย
ต้นทุนที่ต่ำ เวลาผ่านมาจนถึงปัจจุบันพลาสติกเราเคยใช้กันก็กลายเป็นขยะมลพิษ

Lexto:
ใน|ชีวิตประจำวัน||เรา|ใช้|ถุงพลาสติก|ใส่|สิ่งของเครื่องใช้|ทั้ง|อุปโภค|และ|บริโภค||เช่น| |เครื่องดื่ม| | อาหารสด|
|ขนม| |เป็นต้น| |หันไป|ทางไหน|ก็|เจอ|ถุงพลาสติก|ที่|นิยม||ถูก|นำ|มา|ใช้|อย่าง|ยาวนาน||ใน|อดีต| |พลาสติก|
ผลิต|ได้|อย่างรวดเร็ว| |ด้วย|ต้นทุน|ที่|ต่ำ| |เวลา|ผ่าน|มา|จนถึง|ปัจจุบัน|พลาสติก|เรา|เคย|ใช้|กัน|ก็|กลายเป็น|
ขยะ|มลพิษ

Figure 1 Thai word segmentation of LexToPlus

Table 1 The comparison of rubrics for writing skill

Rubric	Contents	Scale
Writing Skills Pasiphol (2016)	title, content, prioritization, guidelines	4 = Good, 3 = Moderate, 2 = Fair, 1 = Improved
Writing a summary Stansfield (1986)	main idea, supporting idea, not repeating, using your own language, linguistic accuracy	4 = complete, 3 = almost complete, 2 = incomplete, 1 = very few, 0 = none
Essays Yamamoto et al., (2018)	content, structure, evidence, style, skill	A+, A, B, C, D (2 points for each level)
Summary Writing National Institute of Educational Testing (2018)	content, language	3 = all points, 2 = no example/ no further explanation, 1 = quotation marks/suffixes, 0 = out of order
Synopses Writing Office of the Basic Education Commission (2018)	background, content, language, spelling, orderliness	5 = all complete, 4 = complete 4 points, 3 = complete 3 points, 2 = complete 2 points, 1 = complete 1 point
Summary Writing Tongsilp (2022)	content, structure, language	1 = no mistakes 0 = wrong word

Conceptual Framework

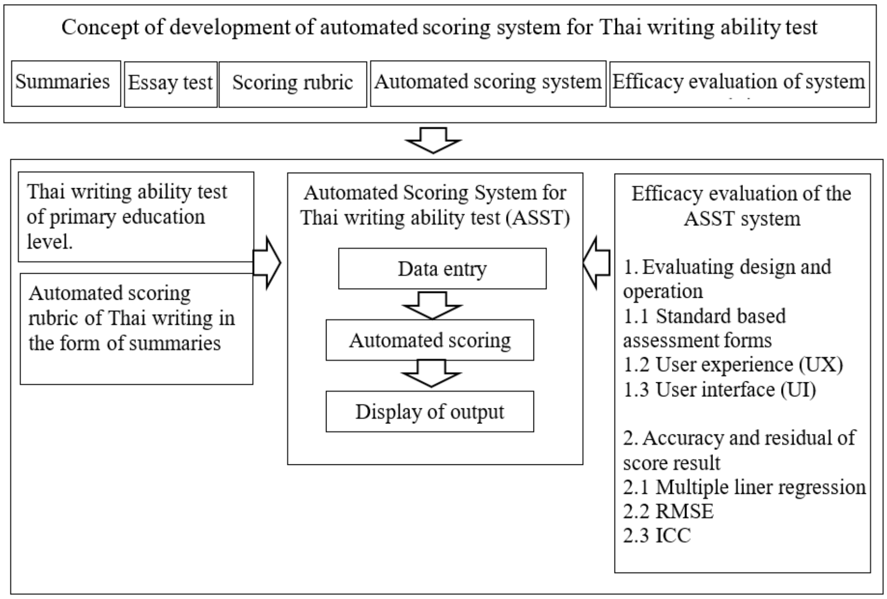

Concept of development of automated scoring system for Thai writing ability test
Summaries
Essay test
Scoring rubric
Automated scoring system
Efficacy evaluation of system
Thai writing ability test of primary education level.
Automated scoring rubric of Thai writing in the form of summaries
Automated Scoring System for Thai writing ability test (ASST)
Data entry
Automated scoring
Display of output
Efficacy evaluation of the ASST system
1. Evaluating design and operation
1.1 Standard based assessment forms
1.2 User experience (UX)
1.3 User interface (UI)
2. Accuracy and residual of score result
2.1 Multiple liner regression
2.2 RMSE
2.3 ICC

Figure 2 Conceptual Framework

Methodology

The current study applied research and development methodology consisting of two steps: (1) developing Automated Scoring System for Thai Writing Ability Test (ASST) for primary school and (2) evaluating efficiency of the ASST system.

Participants

The study participants: (1) 287 sixth-grade students enrolled in academic year 2020 were in schools under the Ministry of Higher Education, Science, Research and Innovation (MHESI), the Department of Education Bangkok (DEB), and the Office of the Basic Education Commission (OBEC) by purposive sampling because of the average score of Thai language from O-NET that was high, medium, and low for representatives; and (2) 30 sixth-grade students in the academic year 2020 at schools under the Office of the Private Education Commission (OPEC) by purposive sampling because schools participated willingly, having computer labs and internet connection (i.e., to test online system).

Process of Automated Scoring

Automated scoring process: (1) *Data entry*, test takers entered their responses into the system; (2) *Automated scoring* was divided into two steps: (2.1) Thai word segmentation, the system transferred test takers' responses to NECTEC corpus via API to segment words using LexToPlus, a dictionary-based software with longest matching technique. After Thai word segmentation was completed, data were transferred back to ASST system for scoring; (2.2) Criterion-referenced scoring, the combination of three methods of word matching was used, including word matching via LexToPlus software, word matching via character count, and word matching via PHP software. Responses were compared with test answers and scored according to all criteria; and (3) *Display of output*, after ASST had checked all activities, the system displayed test result, both for a single item and summative scores according to the evaluation criteria (Angkaseraneekul & Rasakulchai, 2012; Jaihuek & Jaisingh, 2018; Lohraksa, 2007; Premkusonchai, 2006) as shown in Table 2.

Table 2 Formula of automated scoring rubric

No.	Item	Description	formula	Criterion	No.	Item	Description	formula	Criterion
1	*Paraphrasing*	to repeat written using different words from the original	$=\frac{\text{Similar}}{\text{total}}\times 100$	Similarities $\geq 75\% = 0$ $50\%–75\% = 1$ $24\%–49\% = 2$ $< 24\% = 3$	6	Punctuation	Punctuation such as . , - : ;	Punctuation count	word 0 point None 1 point
2	Key words	who, what, where, when, how, results	Word count	None 0 point word 1 point	7	key ideas	Summary key ideas form article	Word count	word 0 point None 1 point
					8	Pronouns 1,2	Use words such as I, we, us, me, me, etc.	Word count	word 0 point None 1 point
3	Words for expressing opinion	No *spoken language*	Word count	word 0 point None 1 point	9	Spelling	Spelling mistake	Word count	word 0 point None 1 point
4	Words for giving example	More examples	Word count	word 0 point None 1 point	10	Sentences in an essay	Writing is an essay	$=\frac{\text{Answer}}{\text{line count}}\times 100$	Sentences 0%–60% = 0 61%–100% = 1
5	Words for adding explanation	Show your own description	Word count	word 0 point None 1 point	11	A complete sentence	Writing is complete sentence	Sentence count	None 0 point Sen. 1 point

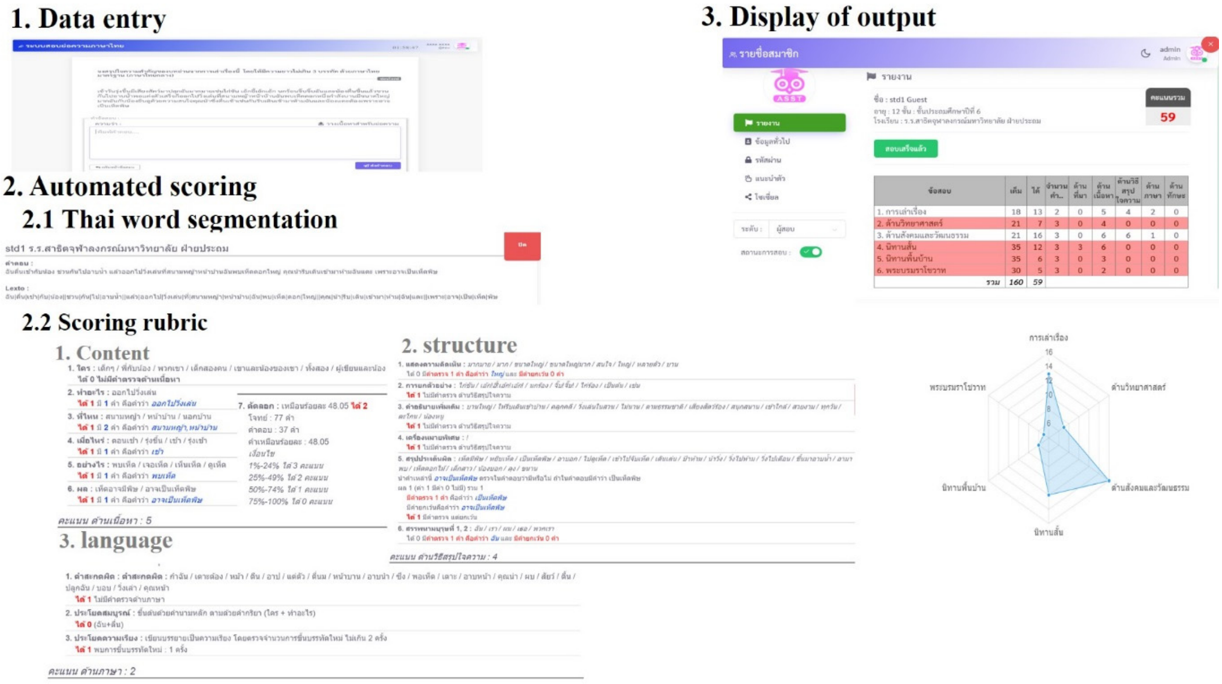
Source: Tongsilp et al. (2022).

Formula of Automated Scoring by Rubric

The formula of scoring according to the following 11 criteria, 3 components for evaluating writing skills as shown in Figure 3.

1. The content components

"Paraphrasing" The system examined density of responses, comparing with an excerpt: If more than 75 percent of similarities were detected, the system discontinued and the score of 0 was given. Meanwhile,



ชื่อ : std1 Guest
อายุ : 12 ชั้น : ชั้นประถมศึกษาปีที่ 6
โรงเรียน : ร.ร.สาธิตจุฬาลงกรณ์มหาวิทยาลัย ฝ่ายประถม

คะแนนรวม 59

สอบเสร็จแล้ว

ข้อสอบ	เต็ม	ได้	จำนวนคำ..	ด้านที่มา	ด้านเนื้อหา	ด้านวิธีสรุปใจความ	ด้านภาษา	ด้านทักษะ
1. การเล่าเรื่อง	18	13	2	0	5	4	2	0
2. ด้านวิทยาศาสตร์	21	7	3	0	4	0	0	0
3. ด้านสังคมและวัฒนธรรม	21	16	3	0	6	6	1	0
4. นิทานสั้น	35	12	3	3	6	0	0	0
5. นิทานพื้นบ้าน	35	6	3	0	3	0	0	0
6. พระบรมราโชวาท	30	5	3	0	2	0	0	0
รวม	160	59						

Figure 3 Process of automated scoring

when 50–75 percent of similarities were detected, the given score was: (1) 24–49 percent, the given score was; (2) less than 24 percent, the score was; and (3) the system continued the examination.

- "Key words" The system examined density of keywords based on the answers involving Who is doing What, Where, When, and what Outcomes. If less than 50 percent of key words were detected, the system discontinued and the score was given. If more than 50 percent of key words were detected, the system continued the examination.

2. The structure components

"Words for expressing opinion, words for giving example, words for adding explanation, punctuation, key ideas, first and second person pronouns, and spelling" The system compared these words with the answers: If matched word was found, the score was 0; if not, the score was 1.

3. The language components

"Sentences in an essay" was written responses in essay form. The system detected the number of "Enter" presses: If more than two presses were detected, the score was 0; if none or not exceeding two presses, the score was 1.

- "A complete sentence" was sentence writing that began with key nouns (Who) and then followed by verbs (doing What). If there was a complete sentence, the score was 1; if none, the score was 0.

Data collection

This research was approved for human research ethics by the Office of the Research Ethics Review Committee for Research Involving Human Subjects at Chulalongkorn University. Collecting data by using the following instruments: (1) The Thai writing ability test - The number of participants 287 students from Chulalongkorn University Demonstration School, Wichuthit School, and Phibunwet Kindergarten School to examine the quality of an essay test and scoring rubrics; and (2) the ASST system and 3 evaluation forms - The number of participants 30 students from Kasem Pittaya School to examine the quality and efficiency of the ASST system.

Data Analysis

1. The data analysis of the Thai writing ability test and scorning rubrics consisted of: (1.1) The quality of tests and rubrics based on the traditional testing theory, including the content validity, reliability, the difficulty index, discrimination, inter-rater reliability, and intra-rater reliability; (1.2) Automated scoring of the ASST System by descriptive statistics, including the minimum, maximum, skewness, mean, standard deviation, coefficient of variation.

2. The data analysis of the ASST system consisted of: (2.1) Evaluating design and operation using the evaluation forms (standard-based assessment, user experience, and user interface) by the descriptive statistics, including the mean and standard deviation; (2.2) Accuracy and residual of score, including the multiple linear regression (Hair et al., 2010), root mean square error, intraclass correlation coefficient (Fisher, 1954)

Results

The Thai Writing Ability Test and Scorning Rubrics

Quality of the tests and scoring rubrics

According to the results of the tests: (1) the content validity indicated that the tests had an Index of Item-Objective Congruence (IOC) between 0.67 to 1.00; (2) the reliability using Cronbach's alpha coefficient method revealed that both raters had consistent scores and the tests were within acceptable limits. (a = 0.786 and 0.812) (George & Mallery, 2003); (3) the difficulty level of the items was found to be medium, with both raters showing consistent scores for all items (p = .40 to .57); (4) the discrimination results revealed that the raters scores were consistent in items 1 and 3, yielding a very good level of discrimination (B-Index = 0.60 to 0.66), and (5) Pearson correlation revealed a high level of agreement between raters, as evidenced by inter-rater reliability (.92) and intra-rater reliability (.97) (Puangrat, 1997) as shown in Table 3.

The automated scoring of ASST system

A study of the ASST system which included 30 participants, 3 questions and a full score of 60 resulted in an average of 8.52, a standard deviation of 4.07 and a coefficient of variation of 0.48. The data were distributed in a normal curve. (SK = -0.61 and KU = 1.43) (SPSS cited in Rueangtrakul, 2001) as shown in Table 4.

Table 3 Result of item analysis and reliability

Rater 1	Full Score	f(H) = f(L)	f(H)×X	f(L)×X	*p*	B-Index	Difficulty	Discrimination
Item 1	(18)	72	943	117	.51	0.64	Medium	Very Good
Item 2	(21)	72	1108	431	.57	0.45	Medium	Good
Item 3	(21)	72	1114	214	.44	0.60	Medium	Very Good
Cronbach's alpha coefficient (α) = 0.786								
Rater 2	Full Score	f(H) = f(L)	f(H)×X	f(L)×X	*p*	B-Index	Difficulty	Discrimination
Item 1	(18)	72	919	103	.40	0.63	Medium	Very Good
Item 2	(21)	72	1155	239	.46	0.61	Medium	Very Good
Item 3	(21)	72	1211	212	.47	0.66	Medium	Very Good
Cronbach's alpha coefficient (α) = 0.812								
Inter-rater reliability (r_{xy}) = 0.92, Intra-rater reliability (r_{xy}) = 0.97								

Table 4 Descriptive automated scoring

Item	Automated scoring							Weight (100 Point)
	Min	Max	SK	KU	*M*	*SD*	CV	*M*
Item 1 (18 point)	1	14	.23	-1.60	7.47	4.34	0.58	41.48
Item 2 (21 point)	1	15	-.62	-1.11	10.33	4.33	0.42	49.21
Item 3 (21 point)	6	10	-.17	-1.12	7.77	1.19	0.15	36.99
Total (60 point)	1	15	-.61	1.43	8.52	4.07	0.48	42.56

The Data Analysis of the ASST System

Evaluating design and operation

The ASST System evaluated by the 3 evaluation forms results revealed that: (1) Standard based assessment - the system demonstrates a high level of effectiveness across four categories: propriety, utility, feasibility, and accuracy. (*M* = 4.70, 4.50, 4.50, and 3.75, *SD* = 0.11, 0.35, 0.31, and 0.25, respectively); (2) User Experience - system users were overall satisfied at the highest level (*M* = 4.22, *SD* = 0.13); and (3) User Interface - system users were satisfied at high level across four categories: screen, system capabilities, terminology and system information, and learning. (*M* = 3.94, 3.91, 3.83, and 3.81, *SD* = 0.74, 0.93, 0.92, and 1.00, respectively) as shown in Table 5.

Table 5 The evaluation of the design and operation of the ASST system

Evaluation of the ASST system	*M*	*SD*
1. Standard based assessment	4.36	0.15
1.1 Propriety standards	4.70	0.11
1.2 Utility standards	4.50	0.35
1.3 Feasibility standards	4.50	0.31
1.4 Accuracy standards	3.75	0.25
2. User experience: UX	4.22	0.13
3. User interface: UI	4.20	1.21
3.1 Screen	3.94	0.74
3.2 System capabilities	3.91	0.93
3.3 Terminology and system information	3.83	0.92
3.4 Learning	3.81	1.00

Accuracy and Residual of Scoring

The accuracy and residuals of the system were examined using multiple linear regression, root mean square error, and intraclass correlation coefficient, shown as follows.

Multiple Linear Regression

Automatically scored and used as an explanatory variable. On the other hand, the raters score the evaluating 11 criteria in 3 component, 287 participants: (1) Normality of Residuals from Normal p-p plot were distributed as a normal curve; (2) Linearity from Residual Vs Fitted were distributed close to 0, considered valid; (3) Outliers from the Residuals vs Leverage graph showed that the points were slightly outside the cook's distance range (Figure 4) and (4) No multicollinearity problem. In addition, the Variance Inflation Ratio (*VIF*) was less than 5.3 and the Tolerance was greater than 0.19 (Hair et al., 2010). The prediction for Thai writing ability is shown in Table 6.

Table 6 Calculating the weights as the explanatory variable for the automated scoring of 3 items, found that scorning rubrics consist of *paraphrasing;* keyword; words for expressing opinion; words for adding explanation; words for adding explanation; punctuation; key ideas; first and second person pronouns; spelling; sentences in an essay; and a complete sentence have effect on Thai writing ability statistically significant at the 0.01 level, having a multiple correlation coefficient ranging from 66.3 percent to 87.6 percent (R^2 = 0.663 to 0.876), shown as follows:

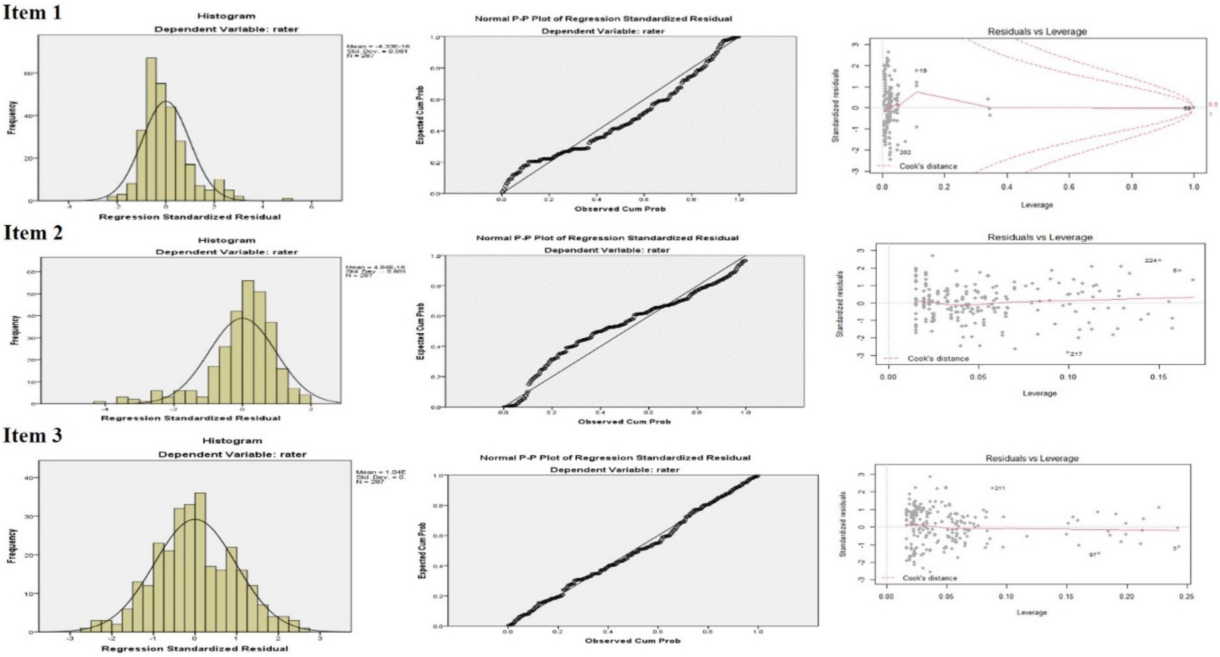


Figure 4 Normality, linearity and outliers

Table 6 Automated scoring rubrics for Thai writing ability

Rubric	b	*SE*	b	*t*-value	*p* value	Tolerance	VIF
Item 1 The story telling (Tel)							
(Intercept)	3.28	0.81		4.05	.00		
key	0.65	0.17	0.20	3.75	.00	0.45	2.24
think	1.53	0.44	0.18	3.51	.00	0.47	2.14
exp	1.43	0.50	0.17	2.86	.01	0.36	2.75
pun	1.20	0.48	0.14	2.52	.01	0.40	2.50
pro	1.22	0.48	0.14	2.54	.01	0.40	2.50
com	1.65	0.57	0.11	2.90	.00	0.79	1.26
R = 0.814, R^2 = 0.663, Adjusted R^2 = 0.649, F-test = 49.165, p = .000							
Item 2 The science article (Sci)							
(Intercept)	3.27	0.84		3.89	.00		
key	0.97	0.14	0.38	7.02	.00	0.31	3.18
exa	0.99	0.55	0.10	1.97	.05	0.31	3.18
exp	1.40	0.63	0.13	2.22	.03	0.25	3.96
spe	0.62	0.32	0.07	1.91	.05	0.69	1.45
sen	3.29	1.12	0.29	2.94	.00	0.21	4.76
R = 0.862, R^2 = 0.743, Adjusted R^2 = 0.733, F-test = 72.327, p = .000							
Item 3 The social and cultural article (Cul)							
(Intercept)	0.39	0.55		0.72	.48		
copy	0.69	0.17	0.09	4.08	.00	0.87	1.15
key	1.17	0.08	0.49	15.05	.00	0.42	2.40
exa	0.82	0.34	0.09	2.43	.02	0.32	3.10
pun	1.12	0.44	0.12	2.52	.01	0.19	5.16
iss	0.85	0.43	0.09	1.96	.05	0.20	4.92
pro	0.75	0.24	0.07	3.15	.00	0.81	1.24
spe	0.99	0.24	0.11	4.14	.00	0.65	1.53
sen	0.84	0.26	0.08	3.21	.00	0.78	1.27
com	0.60	0.24	0.07	2.50	.01	0.66	1.53
R = 0.936, R^2 = 0.876, Adjusted R^2 = 0.871, F-test = 176.948, p = .000							

Note: copy = paraphrasing, key = keyword, think = words for expressing opinion, exa = words for adding explanation, exp = words for adding explanation, pun = punctuation, iss = key ideas, pro = first and second person pronouns, spe=spelling, sen = sentences in an essay, and com = a complete sentence

Item 1 The story telling: 6 criteria, where prediction was statistically significant at the .05 level (F-test = 49.165, p = .00) R^2 = 66.3 percent and Adjusted R^2 = 64.9 percent. As Equation (1).

$$Z_{\widehat{Tel}} = 0.20(\text{key})+0.18(\text{think})+0.17(\text{exp})+ 0.14(\text{pun})+0.14(\text{pro})+0.11(\text{com}) \qquad (1)$$

Item 2 The science article (sci): 5 criteria, where prediction was statistically significant at the .05 level (F-test = 72.327, p = 0.00) R^2 = 74.3 percent and Adjusted R^2 = 73.3 percent. As Equation (2).

$$Z_{\widehat{Sci}} = 0.38(\text{key})+0.10(\text{exa}) +0.13 (\text{exp})+ 0.07(\text{spe})+0.29(\text{sen}) \qquad (2)$$

Item 3 The social and cultural article (cul): 9 criteria, where prediction was statistically significant at the .05 level (F-test = 176.948, p = 0.00) R^2 = 87.6 percent and Adjusted R^2 = 87.1 percent. As Equation (3).

$$Z_{\widehat{Cul}} = 0.09(\text{copy})+0.05(\text{key})+0.09(\text{exa})+ 0.12(\text{pun})+0.09(\text{iss})+0.07 (\text{pro})+ 0.11 (\text{spell})+0.08(\text{sen}) \qquad (3)$$

Root Mean Square Error

The scores of the Thai writing ability for 287 participants, assessed by raters and the ASST system, across the genres of storytelling, science article, and social and cultural article, yielded high correlation scores (r_{xy} = 0.69, 0.84, and 0.83, respectively) and low Root Mean Square Error (RMSE = 3.38, 2.39, and 3.03). (Michailidis, 2019) as demonstrated in Table 7.

Intraclass Correlation Coefficient

The results of the analysis of the consistency values with ICC between rater 1, rater 2, and the ASST system found that all raters gave consistent scores on all items at a good level (ICC = 0.88, 0.79, and 0.80 respectively) (Koo & Li, 2016) as shown in Table 8.

Discussion

The ASST System is an online system built on PHP programming language, using MySQL database. The ASST system goes through the following process of automated scoring: (1) Data entry answers as text;

Table 7 The Pearson *correlation* between r*aters and ASST*

Rubric	Raters		ASST		r_{xy}	RMSE
	M	*SD*	*M*	*SD*		
Item 1 Story telling						
1. Content	4.42	2.00	4.77	1.07	0.40	1.93
2. Structure	2.08	1.78	2.16	2.34	0.72	1.63
3. Language	1.00	0.85	0.67	0.87	0.56	0.88
Total	7.64	4.59	7.60	3.70	0.69	3.38
Item 2 Sscience article						
1. Content	6.87	2.07	6.81	1.59	0.74	1.44
2. Structure	3.27	1.73	3.36	1.69	0.81	1.06
3. Language	1.45	0.83	1.35	0.81	0.66	0.68
Total	11.60	4.38	11.53	3.67	0.84	2.39
Item 3 Social and Cultural article						
1. Content	6.00	2.15	7.42	1.78	0.74	2.03
2. Structure	2.93	2.07	3.32	2.37	0.75	1.63
3. Language	1.45	1.06	0.98	1.00	0.68	0.96
Total	10.38	5.02	11.72	4.56	0.83	3.03

Table 8 measure of agreement of scoring

Item	Rater 1		Rater 2		ASST System		*ICC*	*p* value	agreement
	M	*SD*	*M*	*SD*	*M*	*SD*			
Item 1	8.34	4.81	6.94	4.78	7.60	3.70	0.88	.00**	God
Item 2	11.79	4.18	11.40	5.22	11.53	3.67	0.79	.00**	Good
Item 3	10.06	4.75	10.70	5.89	11.72	4.56	0.80	.00**	Good
Total	10.06	4.97	9.68	5.84	10.28	3.10	0.88	.00**	Good

Note: ** $p < .01$.

this system can record answers in two ways: typing and copying text from the question. The copying text are suitable for users who are primary school students because they are not fluent in typing and want to finish the exam in time. But different answer recording formats will affect the accuracy of the scores. Therefore, teacher users of the system should pay attention to this issue before using scores to judge students' Thai writing ability; (2) Automated scoring consisting of (2.1) Thai word segmentation and scoring rubric; Thai word segmentation based on LexToPlus by dictionary-based method. LexToPlus is suitable for this tests because the tests were created according to the basic vocabulary list of the Ministry of Education. Therefore, this program helps to reduce meaningless words (the National Electronics and Computer Technology Center, 2016); it is stated that the highlight of the LexToPlus program is that it is a dictionary-based word segmentation program using the Longest Matching technique that uses a system to divide words for the Thai language with high accuracy. Users can add word lists as needed. To cut the words appropriately for use; and (2.2) The scoring rubric found that all evaluation criteria could be adopted through the ASST system to score Thai written tasks, highlight of scoring process that combines three methods including word matching, word count, and matching key words with word lists given in the rubrics. The scoring rubric for the Thai writing ability test of the primary education level including plagiarism; key words; words for expressing opinion; words for giving example; words for adding explanation; punctuation; key ideas; first and second person pronouns; spelling; essays; and complete sentences consistent with Yamamoto et al. (2018) who developed AES for Japanese writing scoring; and (3) Display of output; the system displays test results by criterion, which provides students detailed feedback on areas of improvement.

The efficiency evaluation of the system consists of: (1) Evaluating design and operation with evaluation forms revealed that all users were highly satisfied with the system of standardized assessments, user experience assessments, and user interface assessments because students can use feedback from the system in order to improve their ability, and teachers follow up with students in Thai writing ability. The system shows results according to criteria for both components and items. In addition, the system clearly specifies the user's Thai writing ability, and it has the agreement before the testing. The system also helps teachers save time in scoring be said that the system meets the appropriateness evaluation standards, focuses on formalizing the agreement, has a clear evaluation report, and gives importance to the right to receive information (Finn et. al., 1997; Madaus & Stufflebeam, 1989; Stufflebeam & Shinkfield, 1990; Pitiyanuwat, 1998; Karnjanawasri, 1994); and (2) Accuracy and residual of scoring via Multiple Linear Regression (MLR), Root Mean Square Error (RMSE), and Intraclass Correlation Coefficient (ICC). In this study, the results of predicting Thai writing ability using the automatic scoring rubric found that all scoring rubrics have effect on Thai writing ability statistically significant at the .01 level, it is possible to predict Thai writing ability from 66.3 percent to 87.6 percent. ($R^2 = 0.663$ to 0.876). The ability to predict scores depends on two rubrics: key words and key ideas, which have more effect on Thai writing ability, shown as follows: the rubric "key words" because a writer must understand the main ideas of the reading assignment and then rewrite them with the deletion of unnecessary context. And, it was the "key ideas" rubric which required test takers to accurately identify key ideas and write their responses in similar order to the given test. For example, a test answer is "Life cannot be disconnected from nature." Test takers will not obtain any score if they write the following summary: "It is not possible to disconnect life from nature.", "Life and nature cannot be disconnected.", or "Cannot disconnect life from nature." This is because the system is unable to detect word sequence. For the analysis of congruence between raters and ASST system, acceptable level of congruence was found. That is consistent with IntelliMertric and IEA systems (Attali & Burstein, 2006). For error found in the system, the content aspect had the highest level of error. That is because this aspect involves detecting key words, consisting of what words, word chunks, and statements. As a result, the system was required to have additional steps to operate word check, which slowed the system down. Moreover, there were synonyms which caused error in scoring some words. For the measure of agreement of the scoring with ICC and the scoring error with RMSE between rater 1, rater 2, and the ASST system, it was found that all raters gave consistent scores on all items at a good level (ICC = 0.88, 0.79, and 0.80) and a low scoring error (RMSE $\leq$ 3.38), consistent with Attali et al. (2010) who developed the Intelli Mertric system. The consistency results were between 0.80–0.84. The error in scoring found that the content aspect had the greatest discrepancy. Because the content aspect is keywords, which include words, groups of words, and text, the system must add step-by-step functionality to check words that are not just single words but groups of words, text, or possibly similar words, such as synonyms, which increases inaccuracy in a score. Therefore, scores from the system will have a high error score.

Conclusion and Recommendation

The ASST system can be used by educational institutions, teachers, and parents as learning media to improve students' writing skills. That is because it is an immediate automated scoring which provides instructional feedback on summarizing skills at a given criterion. Students will learn their weaknesses to eventually improve such areas. The development of an automatic scoring system for Thai writing ability can be processed to analyze words in the system immediately without having to link to API, capturing key words together with the rubric for scoring Thai writing ability. However, if in the future the system can be developed to analyze grammar instead of capturing keywords, it will allow the system to analyze sentences and learn about the principles of sentence structure, meaning, and relationships between words in sentences of the language Furthermore, the system should be developed towards assessment for learning that focuses on developing learners by providing feedback to correct deficiencies in depth.

Conflict of Interest

The authors declare that there is no conflict of interest.

Fundings

The 90th Anniversary of Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund).

Acknowledgments

I would like to express my gratitude to my advisor and the faculty members of the Department of Educational Measurement and Evaluation for their contribution, which greatly benefitted my study. My sincere gratitude goes to Chulalongkorn University for funding "The 90th Anniversary of Chulalongkorn University Fund" (Ratchadaphiseksomphot Endowment Fund). Also, I would like to thank the National Electronics and Computer Technology Center (NECTEC) for providing linkages to database.

References

Ali, R., & Michael, L. (2010, January). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*(1), 18–39.

Angkaseraneekul, S., & Rasakulchai, C. (2012). *Automatic examination of subjective Thai language examinations*. Kasetsart University.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning, and Assessment*, *4*(3), 131. https://ejournals.bc.edu/index.php/jtla/article/view/1650

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *The journal of Technology, Learning and Assessment*, *10*(3), 1–17.

Daily news. (2018, February 3). NSTDA is confident that subjective examination programs are correct and accurate. *Dailynews online*. https://www.dailynews. co.th/education/625229

Dascalu, M., Westera, W., Ruseti, S., Trausan-Matu, S., & Kurvers, H. (2017). ReaderBench learns Dutch: Building a comprehensive automated essay scoring system for Dutch language. In E. André, R. Baker, X. Hu, Ma. M. T. Rodrigo, B. du Boulay (Eds.), *Artificial Intelligence in Education* (pp. 52–63). https://doi.org/10.1007/978-3-319-61425-0_5

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology*, *Learning and Assessment*, *5*(1), 1–36. https://ejournals.bc.edu/index.php/jtla/article/view/1640

Finn, C. E., Stevens, F. I., Stufflebeam, D. L., & Walberg, H. J. (1997). A meta-evaluation. In H. Miller (Ed.), The New York City public schools integrated learning systems project. *International Journal of Educational Research*, *27*, 159–174. https://doi.org/10.1016/S0883-0355(97)90031-8

Fisher, R. A. (1954). *Statistical methods for research workers* (12th ed. rev). Oliver and Boyd.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference*. 11.0 Update (4th ed.). Allyn & Bacon.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed). Pearson Prentice Hall. Retrieved October 22, 2021, from https://www.drnishikantjha.com/papersCollection/Multivariate%20Data%20Analysis.pdf.

Jaihuek, S., & Jaisingh, S. (2018). Development of a program to automatically score subjective exams using word matching methods. *Kasalongkham Research Journal*, *12*(2), 81–93.

Jang, E. S., Kang, S. S., Noh, E. H., Kim, M. H., Sung, K. H., & Seong, T. J. (2014, April). KASS: Korean automatic scoring system for short-answer questions [Paper presentation]. *International Conference on Computer Supported Education* (Vo.2, pp. 226–230). https://doi.org/10.5220/0004864302260230

Karnjanawasri, S. (1994). *Evaluation theory*. Chulalongkorn University.

Karnjanawasri, S. (2013). *Traditional test theory* (7th ed.). Chulalongkorn University.

Kawtrakul, A., Nakasiri, K., Manomaiphibul, W., Tangteing, S., Burapacheep, T., & Burapacheep, T. (1999). *Grammar and style checking automatically for Thai sentences*. Faculty of Engineering, Kasetsart University.

Ke, X., Zeng, Y., & Luo, H. (2016). Autoscoring essays based on complex networks. *Journal of Educational Measurement*, *53*(4), 478–497. https://doi.org/10.1111/jedm.12127

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016, April 26–27). *Cosine similarity to determine similarity measure: Study case in online essay assessment* [Paper presentation]. The 4th International Conference on Cyber and IT Service Management, Bandung, Indonesia. https://doi.org/10.1109/CITSM.2016.7577578

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/0163853980 9545028

Lohraksa, C. (2007). *Automatic scoring of Thai essay writing using latent meaning analysis techniques & artificial neural network techniques.* Thammasat University.

Madaus, G. F., & Stufflebeam, D. L. (1989). Appraising and recording student progress. In G. F. Madaus, D. L. Stufflebeam (Eds.), *Educational evaluation: Classic works of Ralph W. Tyler. Evaluation in Education and Human Services.* Springer, Dordrecht. https://doi.org/10.1007/978-94-009-2679-0_5

Michailidis, M. (2019). *Regression metrics' guide*. https://www.h2o.ai/blog/regression-metrics-guide/Cagatay

National Electronics and Computer Technology Center. (2019, September 10). *NECTEC launches AI for Thai, a Thai AI platform at NECTEC-ACE 2019.* https://www.nectec.or.Th/news/news-pr-news/nectec-ace2019-aiforthai-2.html

National Institute of Educational Testing. (2018) *Summary of the results of the primary national educational test (O-NET) grade 6 academic year 2017* [Unpublished manuscript]. http://www.newonetresult.niets.or.th/AnnouncementWeb/PDF/SummaryONETP6_2560.pdf [in Thai]

Office of the Basic Education Commission. (2018). *Measurement and evaluation guide "The ability to read and write" of primary school students*. Ministry of Education. http://122.154.253.83/chiangrai1/supervisor/57000001/files/20190612141136Q73AMKH.pdf

Pasiphol, C. (2016). *Measurement and evaluation learning outcomes*. Printing House of Chulalongkorn University.

Phandi, P., Chai, K. M. A., & Ng, H. T. (2015, September). *Flexible domain adaptation for automated essay scoring using correlated linear regression* [Paper presentation]. Conference on empirical methods in natural language processing (pp. 431–439). https://aclanthology.org/D15-1049.pdf

Pitiyanuwat, S. (1998). *Methods of educational assessment*. Chulalongkorn.

Premkusonchai, S. (2006). *Evaluating the quality of Thai language transcription by analyzing hidden meanings.* Thammasat University.

Puangrat, T. (1997). *Research methods in behavioral and social sciences.* Srinakharinwirot University.

Ruangtrakul, P. (2001). *Development and analysis of the quality of developmental score measurement methods based on traditional testing theory and test response theory*. Department of Educational Research, Chulalongkorn University.

Stansfield, C. (1986). A history of the test of written English: The developmental year. *Language Testing*, *3*(2), 224–234. https://doi.org/10.1177/026553228600300209

Stufflebeam, D. L., & Shinkfield, A. J. (1990). *Systematic evaluation*. Kluwer–Nijhoff.

Thai Encyclopedia Project Committee for Youth. (2017). *Thai Encyclopedia for youth by King Bhumibol Adulyadej*. Thai Encyclopedia Project for Youth by King's wishes.

Tongsilp, A., Tangdhanakanond, K., & Chaimongkol, C. (2022). Development of the supply type test and automated scoring rubric for Thai summary writing of primary education level. *Srinakharinwirot Research and Development Journal (Humanities and Social Sciences)*, *14*(27), 206–218. https://ejournals.swu.ac.th/index.php/swurd/article/view/14499

Urathumkul, P., & Runapongsa, K. (2006) Improved rule-based and new dictionary for Thai word segmentation. Faculty of Engineering, Khon Kaen University.

Ke, X., Zeng, Y., & Luo, H. (2016). Autoscoring essays based on complex networks. *Journal of Educational Measurement*, *53*(4), 478–497. https://onlinelibrary.wiley.com/doi/abs/10.1111/jedm.12127

Yamamoto, M., Umemura, N., & Kawano, H. (2018). Automated essay scoring system based on rubric. In R. Lee (Eds.), *Applied Computing & Information Technology*. ACIT 2017. Studies in Computational Intelligence (Vol. 727). Springer. https://doi.org/10.1007/978-3-319-64051-8_11