A Study on Syntactic Complexity and Text Readability of ASEAN English News

Yusha Zhang
Hunan University of Information Technology
Changsha, China

Nankai Lin✉ and Shengyi Jiang
School of Information Science and Technology
Guangdong University of Foreign Studies
Guangzhou, China
neakail@outlook.com

***Abstract*—English is the most widely used language in the world. With the spread and evolution of language, there are differences in the English text expression and reading difficulty in different regions. Due to the difference in the content and wording, English news in some countries is easier to understand than in others. Using an accurate and effective method to calculate the difficulty of text is not only beneficial for news writers to write easy-to-understand articles, but also for readers to choose articles that they can understand. In this paper, we study the differences in the text readability between most ASEAN countries, England and America. We compare the textual readability and syntactic complexity of English news texts among England, America and eight ASEAN countries (Indonesia, Malaysia, Philippines, Singapore, Brunei, Thailand, Vietnam, Cambodia). This paper selected the authoritative news media of each country as the research object. We used different indicators including Flesch-Kincaid Grade Level (FKG), Flesch Reading Ease Index (FRE), Gunning Fog Index (GF), Automated Readability Index (AR), Coleman-Liau Index (CL) and Linsear Write Index (LW) to measure the textual readability, and then applied L2SCA to analyze the syntactic complexity of news text. According to the analysis results, we used the hierarchical clustering method to classify the English texts of different countries into six different levels. Moreover, we elucidated the reasons for such readability differences in these countries.**

Keywords- Textual Understanding Difficulty; Textual Readability; Syntactic Complexity; Hierarchical Clustering.

I. INTRODUCTION

Language is the carrier of information. English, as a dominant language worldwide, evolves by its nature and adapts to the need of its users, leading to national and regional variation of how English is written and spoken. English news texts mirror a nation's English language level and language habits, reflecting the changes occurred in English. They are easier to understand in some countries than others as a result of different content and wording in respective countries. How to measure the difficulty of English texts in different countries in a scientific and quantitative way has become a research hotspot in linguistics.

Using an accurate and effective method to calculate the difficulty of text would conduce not only to content management for news writers, but to proper materials selection for readers. The study of English text readability is of great significance for reading theory and reading instruction (such as the choice of teaching materials, the choice of reading test materials, the choice of reading psychology research materials, the choice of reading instruction methods, etc.) [1] . Scholars mainly measure the difficulty of English text from the perspectives of text readability and syntactic complexity. Readability, also known as readability or legibility, refers to the degree or nature of text that is easy to read and understand [1] . Syntactic complexity refers to the range and complexity of language forms in language output [2] .

This paper focuses on the quantitative study of news readability and news syntactic complexity in England, America and eight ASEAN countries (Indonesia, Malaysia, Philippines, Singapore, Brunei, Thailand, Vietnam, Cambodia) and compares their differences in the usage of English. We use different indicators to measure the textual readability. They are: Flesch-Kincaid Grade Level (FKG), Flesch Reading Ease Formula (FRE), Gunning Fog Index (GF), Automated Readability Index (AR), Coleman-Liau Index (CL) and Linsear Write Formula (LW). We also use L2SCA to analyze the syntactic complexity.

Based on the analysis results, we use the hierarchical clustering method to classify the English texts of different countries into six different levels.

The remaining part of this paper is organized as follows: Section 2 briefly reviews related studies; Section 3 demonstrates our approach; Section 4 provides information of our data, as well as the results and analysis; Section 5 concludes our work.

II. RELATED WORK

The readability assessment deals with estimating the level of difficulty in reading texts. More and more scholars have carried out text readability analysis on a wide variety of texts. Du Bay and William H introduced the research on readability and the readability formulas [3] . Commonly used text readability evaluation criteria are Flesch Reading Ease Formula (FRE), Automated Readability Index(AR), Gunning Fog Index (GF), Flesch-Kincaid Grade Level (FKG), Coleman-Liau Index (CL) and Linear Write Formula (LW). Rudolph Flesch put forward the simplification of Flesch Reading Ease Formula [4] . Automated Readability Index (ARI) was devised by Smith et al [5] . AR computed the average word length and average sentence length. Appropriate weightings of these factors result in an index reflecting the readability of the passage. Gunning Fog Index (GF) was created by Gunning in 1952 [6] . Kincaid et al recalculated Automated Readability Index, Gunning Fog Index and Flesch Reading Ease Formula for naval purpose and created the Flesch-Kincaid Grade Level [7] . Meri Coleman and Liau proposed a new calculation method Coleman-Liau Index [8] . There is no need to consider syllables since letter count is a better predictor of readability than syllable count. Linsear Write Formula is not actually presented as an index of readability. O'hayre explains that this formula is more about "write ability", that is, it serves the writer, and not the reader. The formula aims at helping writers to use simple, one-syllable words [9] . Betul Karakus et al research the readability analysis of Turkish elementary school textbooks [10] . Scoti A. Crossley et al assessed text readability using cognitively based indices [11] . Solnyshkina et al demonstrated the correlations between the narrativity, abstractness and word concreteness of the texts and Flesch-Kincaid Grade Level [12] . Another study was done by Brenda Lynn Hoketo to see if readability levels printed on recreational reading books were as accurate as when the Fry formula and the Flesch-Kincaid Grade Level were applied to them [13] .

In addition to text readability, text complexity is an important indicator of measuring the difficulty of text. Lu described a computational system L2SCA for automatic analysis of syntactic complexity in second language writing using fourteen different measures that have been explored or proposed in studies of second language development [14] . Moreover, Xiaofei Lu has conducted more in-depth studies in syntactic complexity analysis [15][16][17][18][19][20] . Chen et al. used the academic texts of computer and library/information science as the research object, and compared the readability and complexity of academic texts from the aspects of disciplines, text structure, writer's position and j impact factors for different journals [21] . Wu used a variety of readability evaluation formulas and syntactic complexity analysis tools L2SCA to observe the syntactic complexity and text readability of Chinese journalists in different disciplines [22] .

III. Approach

A. Textual readability analysis

We use different indicators to analyze the readability of English in ten countries from different perspectives.

FRE (Flesch Reading Ease Formula) — FRE is a simple approach to assess the grade-level of the reader. This evaluation method mainly measures the readability of the text from the average sentence length and the average syllable number of words in the text. Flesch believes that the longer the sentence is, the harder it is to read and the more complex the pronunciation of the word is, the harder it is to understand. The formula is:

$$FRE = 206.835 - (1.015 * ASL) - (84.6 * ASW)$$

ASL refers to the average sentence length which is obtained by the number of words divided by the number of sentences, ASW refers to the average number of syllables per word which is obtained by the number of syllables divided by the number of words. FRE means the score of readability. The Flesch Reading Ease formula will output a number from 0 to 100 and a higher score indicates easier reading. The score of 90-100 means "very easy", 80-89 is "easy", 70-79 is "fairly easy", 60-69 is "standard", 50-59 means "fairly difficult", 30-49 means "difficult" and 0-29 is "very confusing". The above scales are assigned to grade level 1-7.

FKG (Flesch-Kincaid Grade Level) — FKG evaluation method is an improvement of Kincaid's FRE method. This method describes the relationship between ASL and ASW on text score more reasonably. This method was originally developed for the purpose of the U.S. Navy and is most suitable for the field of education. FKG outputs a U.S. school grade level which indicates the average student in that grade level can read the text. The formula is:

$$FKG = (0.39 * ASL) + (11.8 * ASW) - 15.59$$

ASL refers to the average sentence length which is obtained by the number of words divided by the number of sentences, ASW refers to the average number of syllables per word which is obtained by the number of syllables divided by the number of words. A score of 9.3 means that a ninth grader would be able to read the document. Because FKG believes that the lowest reading level is third grade level, its output range is between 3-13.

GF (Gunning Fog Index) — GF is similar to the Flesch scale in that it compares syllables and sentence lengths. FRE and FKG use the average number of bytes to express the difficulty of words in the text, while GF directly calculates the proportion of difficult words in the text, and defines "Foggy" words. "Foggy" words are words that contain 3 or more syllables. A Fog score of 5 is readable, 10 is hard, 15 is difficult, and 20 is very difficult. The formula is:

$$GF = 0.4 * (ASL + PHW)$$

ASL refers to the average sentence length which is obtained by the number of words divided by the number of sentences and PHW is the percentage of hard words in the context. The ideal score for readability with the Fog index is 7 or 8. Anything above 12 is too hard for most people to read. We divide the score into 14 levels. 14th level represents the hardest.

AR (Automated Readability Index) — The Automated Readability Index (AR) is a readability test designed to assess the understandability of a text. It measures the readability of a text from the ratio of the number of characters to the number of words and the ratio of the number of words to the number of sentences. AR outputs a number which approximates the required grade level. For example, if the AR outputs the number 3, it means students in 3rd grade (ages 8-9 yrs. old) should be able to comprehend the text. The formula is:

$$AR = 4.71 * \frac{characters}{words} + 0.5 * \frac{words}{sentences} - 21.43$$

Here $characters$ is the number of letters and numbers, and $words$ is the number of words in the text while the $sentences$ means the number of sentences.

CL (Coleman-Liau Index) — CL relies on characters and sentence length instead of syllables per word. It measures the readability of a text in terms of the average number of characters and sentences per hundred words. This formula will output a grade. For example, 10.6 means your text is appropriate for a 10-11th grade high school student. The formula is:

$$CLI = 0.0588 * L - 0.296 * S - 15.8$$

L is the average number of letters per 100 words. S is the average number of sentences per 100 words. CL is similar to the Automated Readability Index, but unlike most of the other grade-level predictors that rely on syllables per word.

LW (Linsear Write Formula) — Linsear Write Formula is a readability formula for English text, originally developed for the United States Air Force to help them calculate the readability of their technical manuals. Linsear Write Formula is specifically designed to calculate the grade level in the US of a text sample based on sentence length and the number of three-plus syllable words. The LW score computing method is show in Table 1.

IV. Data And Result

A. Data

We selected the news of the authoritative English newspapers of various countries as experimental data. We regarded the official media of official institutions, governments or enterprises and the large media with great influence as the authoritative English newspapers, such as the New York Times, which is the most influential newspaper in the United States, and the Antara News, the official newspaper of Indonesia. The experimental data in this paper is shown in Table 4.

B. Textual readability analysis result

We used six evaluation methods to analyze the text readability of ten countries. The results are shown in Table 5-10. We also present the cumulative results of each evaluation indicator in the form of line charts, as shown in Figures 1-6.

The evaluation results of FRE, FKG, GF and LW are similar. The English news texts are less readable in Cambodia, Indonesia and Singapore than the other seven countries. We regarded the English of America and England as the standard English and we could see that the English news texts of Malay and Thailand are more easy-to-understand than the standard English. Among the AR evaluation results, countries with lowest English readability are Cambodia, Indonesia and Singapore, followed by Vietnam, Philippines and Myanmar, and the other four countries have higher English readability. The CL evaluation indicators show that English texts in Vietnam, Cambodia, Indonesia, and Singapore are more difficult, while British English and American English texts are more readable.

C. Syntactic complexity analysis result

According to the results of the L2SCA syntactic analysis tool, most of the indices in the news texts of Singapore and Cambodia are higher than others. The syntactic complexity of the news texts of these two countries is high, and some indices of Indonesia (C/T, CT/T, CN/T) and Vietnam (MLC, CP/T, CP/C, CN/C) have higher values. Indonesian news texts have higher results when calculating syntactic complexity in units of T while the Vietnamese news text has higher results when calculating the syntactic complexity in units of C. The syntactic complexity of English texts in other countries is low.

Table I. LW Score Computing Flow

(1) Calculate the easy words (defined as two syllables or less) and place a number "1" over each word, even including a, an, the, and other simple words.
(2) Calculate the hard words (defined as three syllables or more) and place a number "3" over each word as pronounced by the dictionary.
(3) Multiply the number of easy words times "1".
(4) Multiply the number of hard words times "3".
(5) Add the two previous numbers together.
(6) Divide that total by the number of sentences.
(7) If the answer of (6) is >20, divide by "2".
(8) If the answer of (6) is <20 or equal to 20, subtract "2" and then divide by "2".

Table II. The Nine Syntactic Complexity Basic Indicators

Index	Meaning
W	Length of text
S	Number of sentence
VP	Number of verb phrases
C	Number of clause
CN	Number of Complex nominals
T	T-unit
DC	Dependent clauses
CT	Complex T-unit
CP	Number of complex nominals

Table III. The Fourteen Syntactic Complexity Evaluation Indicators

Index	Meaning
MLS	Mean length of sentence
MLT	Mean length of T-unit
MLC	Mean length of clause
C/S	Sentence complexity ratio
CT/T	Complex T-unit ratio
C/T	T-unit complexity ratio
DC/C	Dependent clause ratio
DC/T	Dependent clauses per T-unit
T/S	Sentence coordination ratio
CP/C	Coordinate phrases per clause
CP/T	Coordinate phrases per T-unit
VP/T	Verb phrases per T-unit
CN/T	Complex nominals per T-unit
CN/C	Complex nominals per clause

Table IV. The Data That We Used in the Experiments

Country	Abbreviation	Website	Number of news
Indonesia	ID	http://www.antaranews.com/	36001
Malay	MY	http://www.nst.com.my/	76764
Philippines	PH	http://www.malaya.com.ph/	60858
Singapore	SG	http://www.todayonline.com/	56323
Brunei	BN	http://www.brudirect.com/	38313
Thailand	TH	http://www.bangkokpost.com/	104914
Vietnam	VN	http://vov.vn/	59456
Cambodia	KH	https://www.cambodiadaily.com/	38161
English	UK	http://www.theguardian.com/uk	261591
America	USA	http://www.nytimes.com	300094

TABLE V. FRE CALCULATION RESULT (%)

Country	FRE level						
	1	2	3	4	5	6	7
SG	13.78	20.06	17.42	13.56	9.21	10.80	15.17
MY	91.29	7.93	0.40	0.06	0.01	0.02	0.29
PH	82.05	15.39	1.78	0.26	0.07	0.07	0.38
BN	75.01	11.12	0.78	0.12	0.03	0.01	12.95
TH	94.46	4.86	0.30	0.08	0.01	0.01	0.28
VN	80.61	16.39	1.21	0.16	0.04	0.02	1.56
KH	3.30	10.83	17.10	16.95	14.06	17.22	17.22
USA	89.74	9.06	0.84	0.18	0.06	0.04	0.08
ID	8.53	11.49	13.29	12.56	10.47	14.39	29.28
UK	88.07	10.71	0.90	0.15	0.07	0.03	0.07

TABLE VI. FKG CALCULATION RESULT (%)

Country	FKG level										
	3	4	5	6	7	8	9	10	11	12	13
SG	0.73	0.85	2.16	3.61	2.97	5.32	3.96	6.08	6.20	3.82	64.31
MY	13.13	13.01	24.82	23.29	10.42	8.80	2.89	1.97	0.75	0.20	0.70
PH	9.95	9.39	19.82	21.80	12.04	12.46	4.95	4.45	2.20	0.76	2.18
BN	9.94	9.55	18.76	19.18	10.20	10.10	3.73	3.13	1.29	0.44	13.68
TH	18.58	15.88	26.77	20.64	8.02	5.94	1.74	1.19	0.45	0.21	0.58
VN	9.02	9.76	19.25	21.82	11.96	12.19	5.44	4.88	2.25	0.75	2.69
KH	0.10	0.10	0.34	0.76	0.75	1.92	1.61	3.20	3.95	2.84	84.43
USA	20.03	12.77	22.14	19.43	9.24	8.39	3.04	2.46	1.10	0.40	1.01
ID	0.48	0.52	1.46	2.21	1.86	2.94	2.24	3.44	3.69	2.42	78.74
UK	17.51	12.58	21.79	19.58	9.68	9.49	3.63	2.97	1.30	0.45	1.02

TABLE VII. AR CALCULATION RESULT (%)

Country	AR level													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
SG	0.01	0.02	0.03	0.04	0.06	0.16	0.39	0.71	0.98	1.44	1.97	2.31	2.72	89.16
MY	0.11	0.09	0.15	0.34	1.15	3.08	6.60	10.61	13.09	13.36	12.02	9.21	7.39	22.81
PH	0.11	0.11	0.20	0.42	1.09	2.30	4.60	7.33	9.62	10.98	10.91	9.55	8.26	34.54
BN	0.04	0.04	0.16	0.29	0.96	2.58	4.86	7.58	8.78	9.28	8.32	7.16	6.55	43.39
TH	0.06	0.20	0.44	0.71	1.71	4.08	7.45	10.76	11.45	10.57	8.72	7.59	8.01	28.26
VN	0.04	0.04	0.06	0.21	0.65	1.93	3.57	5.61	6.37	6.37	6.33	6.15	7.12	55.56
KH	0.01	0.01	0.10	0.10	0.10	0.42	0.60	0.10	0.18	0.29	0.57	0.74	1.18	96.78
USA	1.06	0.73	0.69	1.14	2.64	5.22	8.49	11.73	13.68	13.37	11.61	9.02	6.44	14.18
ID	0.04	0.01	0.01	0.02	0.03	0.10	0.18	0.37	0.54	0.77	1.04	1.24	1.40	94.27
UK	1.07	0.63	0.78	1.30	2.15	4.27	7.72	11.08	12.79	12.79	11.38	9.34	7.44	17.26

TABLE VIII. GF CALCULATION RESULT (%)

Country	GF level									
	1	2	3	4	5	6	7	8	9	10
SG	0.03	0.11	0.32	1.12	1.30	3.13	2.84	4.93	20.07	66.13
MY	0.47	2.56	5.65	17.49	16.41	24.84	12.59	11.46	7.97	0.57
PH	0.45	2.21	4.07	12.64	12.68	21.95	13.51	14.77	15.50	2.23
BN	0.50	2.48	4.28	13.79	13.31	22.28	13.24	14.47	14.22	1.44
TH	1.53	4.01	7.45	21.62	18.05	23.79	10.14	8.11	4.87	0.43
VN	0.26	1.78	3.95	12.92	12.72	21.98	13.54	14.63	16.72	1.49
KH	0.01	0.02	0.05	0.13	0.19	0.59	0.61	1.70	10.83	85.87
USA	2.18	5.41	7.42	17.81	14.80	20.98	10.83	10.35	9.06	1.17
ID	0.09	0.10	0.17	0.66	0.91	1.98	1.74	2.97	11.68	79.70
UK	1.76	4.24	6.61	17.47	14.47	20.96	11.16	11.16	10.71	1.21

TABLE IX. LW CALCULATION RESULT (%)

Country	LW level											
	1	2	3	4	5	6	7	8	9	10	11	12
SG	0	0.01	0.04	3.12	41.39	41.94	11.47	1.87	0.17	0.01	0.01	0
MY	0.01	0.02	0.57	43.45	55.67	0.21	0.05	0.02	0.01	0.01	0	0
PH	0	0.01	0.51	31.03	67.18	1.26	0.01	0	0	0	0	0
BN	0	0	0.67	31.53	65.31	2.48	0.01	0	0	0	0	0
TH	0	0.02	1.75	50.08	47.94	0.21	0.01	0	0	0	0	0
VN	0	0.01	0.37	28.34	70.33	0.94	0.02	0.01	0	0	0	0
KH	0	0	0.01	0.64	25.41	56.87	15.73	1.29	0.04	0	0	0
USA	0.01	0.09	2.83	49.64	46.98	0.40	0.06	0.01	0.01	0	0	0
ID	0	0.02	0.06	1.48	23.11	46.26	23.23	5.34	0.49	0.02	0.03	0
UK	0.02	0.09	2.48	47.74	49.23	0.42	0.04	0.01	0.01	0	0.01	0.01

TABLE X. CL CALCULATION RESULT (%)

Country	CL level												
	1	2	3	4	5	6	7	8	9	10	11	12	13
SG	0	0.01	0.01	0.01	0.01	0.03	0.12	0.32	1.07	2.72	5.50	9.12	81.12
MY	0	0	0.01	0.01	0.02	0.09	0.36	1.00	2.30	4.10	6.68	8.15	77.30
PH	0	0	0	0.01	0.03	0.18	0.51	1.15	2.08	3.62	6.40	9.02	77.00
BN	0	0	0	0	0.01	0.05	0.26	0.84	1.69	3.15	5.71	7.94	80.36
TH	0	0	0	0.01	0.01	0.05	0.27	0.85	1.95	3.72	6.25	9.01	77.88
VN	0	0	0	0	0.01	0.03	0.11	0.28	0.78	1.42	2.48	3.74	91.15
KH	0	0	0	0.01	0.01	0.03	0.18	0.67	2.19	4.59	7.85	10.81	73.66
USA	0	0.01	0.02	0.08	0.29	0.83	2.11	4.90	8.73	13.03	15.79	15.05	39.17
ID	0	0.01	0.01	0.01	0.01	0.01	0.01	0.04	0.09	0.45	1.19	2.95	95.26
UK	0	0.01	0.04	0.11	0.34	0.89	2.10	4.14	7.06	10.72	13.61	13.96	47.04

TABLE XI. SYNTACTIC COMPLEXITY INDEX FOR EACH COUNTRY

Index	Country									
	UK	USA	ID	MY	TH	PH	VN	BN	SG	KH
W	822.22	680.47	332.14	341.35	487.96	422.08	265.98	302.15	567.22	455.34
S	39.81	31.42	13.95	15.25	21.93	17.84	10.57	13.42	15.40	13.81
VP	98.32	84.23	35.54	39.60	56.81	45.96	24.85	34.45	65.27	57.94
C	76.20	63.67	27.10	29.50	42.10	34.15	17.05	24.71	47.79	44.49
T	43.37	36.28	12.87	15.77	23.61	18.83	11.08	14.17	18.55	16.89
DC	27.91	24.25	10.61	10.99	15.47	12.70	4.91	9.31	23.96	20.56
CT	19.82	17.44	6.61	7.69	11.06	8.93	3.93	5.99	11.60	9.77
CP	18.73	15.37	7.28	7.47	11.31	10.52	9.66	6.46	11.95	8.87
CN	98.42	82.51	42.04	40.95	59.59	52.57	34.37	34.48	75.37	58.18
MLS	22.06	23.08	24.15	23.34	23.78	25.22	26.46	24.05	46.58	35.98
MLT	20.22	20.18	27.53	23.78	22.49	24.22	25.59	24.10	39.97	29.54
MLC	11.65	11.50	13.01	12.54	12.72	13.71	17.48	15.54	13.80	10.67
C/S	1.94	2.05	1.93	1.93	1.94	1.91	1.59	1.71	3.76	3.40
VP/T	2.29	2.38	2.86	2.63	2.47	2.48	2.26	2.39	4.40	3.66
C/T	1.76	1.78	2.18	1.94	1.82	1.82	1.52	1.66	3.21	2.79
DC/C	0.36	0.37	0.38	0.36	0.35	0.35	0.26	0.34	0.54	0.46
DC/T	0.66	0.69	0.88	0.73	0.68	0.68	0.43	0.61	1.70	1.33
T/S	1.09	1.15	0.90	1.00	1.07	1.04	1.04	1.01	1.32	1.24
CT/T	0.46	0.49	0.52	0.50	0.47	0.48	0.35	0.37	0.72	0.59
CP/T	0.47	0.46	0.60	0.51	0.52	0.63	0.98	0.51	0.84	0.58
CP/C	0.28	0.27	0.29	0.28	0.31	0.37	0.68	0.32	0.30	0.21
CN/T	2.48	2.51	3.50	2.87	2.79	3.04	3.31	2.65	5.39	3.84
CN/C	1.42	1.42	1.64	1.51	1.57	1.71	2.24	1.65	1.83	1.38

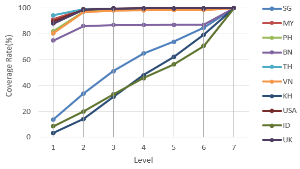

Figure 1. FRE cumulative coverage rate.

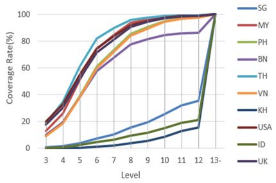

Figure 2. FKG cumulative coverage rate.

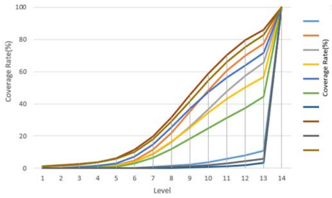

Figure 3. AR cumulative coverage rate.

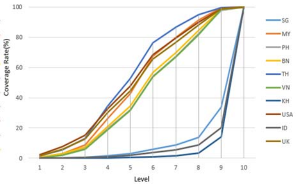

Figure 4. GF cumulative coverage rate.

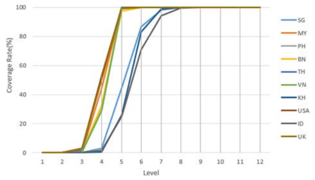

Figure 5. LW cumulative coverage rate.

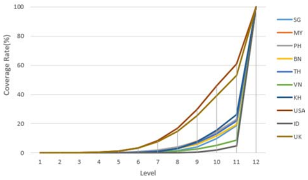

Figure 6. CL cumulative coverage rate.

D. Overall result analysis

There are different indicators in text readability and syntactic complexity. In order to comprehensively consider each indicator, we used all the indicators to cluster the English texts of each country. We utilized the average of each indicator of text readability and the average of the various indicators of syntactic complexity, and applied the hierarchical clustering method to cluster the English texts of each country. We implemented this with the scipy[1] tool. The parameters of the model we used are shown in Table 12. The result is shown in Figure 7.

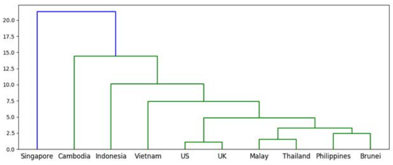

Figure 7. Hierarchical clustering result.

TABLE XII. THE PARAMETERS OF THE METHOD

Parameter	Value
metric	euclidean
optimal_ordering	false
method	weighted

[1] https://github.com/scipy/scipy

TABLE XIII. THE RESULT OF THE CLASSIFICATION

Level	Country
1	Malay Thailand Philippines Brunei
2	America England
3	Vietnam
4	Indonesia
5	Cambodia
6	Singapore

Based on the results of hierarchical clustering, we divided the difficulty of English texts in ten countries into six levels. Level 1 represents the easiest level and level 6 is the most difficult level. The result of the classification is shown as table 13. We can see that, as standard English, the English texts' difficulty in America and England is at the second level. The results show that Vietnam, Indonesia, Cambodia and Singapore's news are more difficult than the news of America and England.

We conducted an in-depth study of the grading results. English is highly popular in Singapore, but its distinctive "Singapore English" is more difficult for non-Singapore people to read. Cambodia pays attention to the promotion of Khmer in the country. The use of English in Cambodia is not as high as Khmer, so it also has a certain reading difficulty. Due to its high openness and the absorption of multiple languages, Indonesia has reduced the standardization of language use, which makes reading in English more difficult.

V. CONCLUSION

This paper focuses on the quantitative study of the difficulty of English texts in the United Kingdom, the United States and ASEAN countries, and researches the differences in the English news text readability and syntactic complexity between these countries. We used different indicators that including FKG, FRE, GF, AR, CL and LW to measure the textual readability, and then used L2SCA to analyze the syntactic complexity of news text. According to the analysis results, we used the hierarchical clustering method to classify the English texts of different countries into six different levels according to the difficulty of the text. Moreover, we analyzed the reasons for the differences in the difficulty of English texts in these countries. In the future, we will further analyze the text readability from other dimensions.

ACKNOWLEDGEMENT

This work was supported by the 13th Five-Year Plan Project of Educational Science for Hunan Province (No．XJK18CGD044). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] S. Shan, "Overview of legibility research," Journal of PLA University of Foreign Languages, vol. 23, pp. 1-5, 2004.

[2] L. Ortega, "Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing," Appl. Linguist., 2003.

[3] W. Dubay, "The Principles of Readability," CA, vol. 92627949, pp. 631–3309, 2004.

[4] R. Flesch, "New Facts about Readability," Coll. English, vol. 10, pp. 225-226, 1949.

[5] E. A. Smith and R. J. Senter, "Automated readability index.," AMRL-TR. Aerosp. Med. Res. Lab., 1967.

[6] R. Gunning, The Technique of Clear Writing, McGraw-Hil. New York, 1968.

[7] J. P. Kincaid, J. Fishburne, R. Robert P., C. Richard L., and Brad S., "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," 1975.

[8] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring," J. Appl. Psychol., 1975.

[9] J. O'Hayre, "Gobbledygook Has Gotta Go," D.C.: U.S. Dept. of the Interior, Bureau of Land Management, Washington, 1966.

[10] B. Karakus, G. Aydın, and I. R. Hallac, "Distributed Readability Analysis of Turkish Elementary School Textbooks," Int. Conf. Inf. Technol. Comput. Sci., 2015.

[11] S. A. Crossley, J. Greenfield, and D. S. McNamara, "Assessing text readability using cognitively based indices," TESOL Q., 2008.

[12] M. I. Solnyshkina, R. R. Zamaletdinov, L. A. Gorodetskaya, and A. I. Gabitov, "Evaluating Text Complexity and Flesch-Kincaid Grade Level," www.jsser.org J. Soc. Stud. Educ. Res. Sos. Bilgiler Eğitimi Araştırmaları Derg., 2017.

[13] B. Lynn Hoke, "Comparison of Recreational Reading Books Levels Using the Fry Readability Graph and the Flesch-Kincaid Grade Level," 1999.

[14] X. Lu, "Automatic analysis of syntactic complexity in second language writing," Int. J. Corpus Linguist., 2010.

[15] X. Lu, "A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development," TESOL Q., 2011.

[16] H. Ai and X. Lu, "A corpus-based comparison of syntactic complexity in NNS and NS university students' writing," 2014.

[17] W. Yang, X. Lu, and S. C. Weigle, "Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality," J. Second Lang. Writ., 2015.

[18] X. Lu and H. Ai, "Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds," J. Second Lang. Writ., 2015.

[19] X. Lu, Q. Xu, "L2 Syntactic Complexity Analyzer and its applications in L2 writing research," Foreign Language Teaching and Research, vol. 48, pp. 409-420, 2016.

[20] X. Lu, "Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment," Lang. Test., vol. 34, no. 4, pp. 493–511, 2017.

[21] L. Chen, X. Li, C. Zhao, "Analysis of the Readability and Complexity of Academic Texts," Digital Library Forum, vol. 168, pp. 64-68, 2018.

[22] X. Wu, "A Study on Syntactic Complexity and Text Readability of International Journal Articles by Chinese Scholars," Journal of PLA University of Foreign Languages, vol 40, pp.11-19, 2017.