Research Article

Application of Coh-Metrix 2.0 in Foreign Language Teaching and Research

Shi Xue

Department of Foreign Languages, Luoyang Institute of Science and Technology, Luoyang, China
_____

ABSTRACT

Coh-Metrix2.0 is an online text analysis tool developed by the applied linguists in University of Memphis, which can perform accurate measurement of the readability of the text, vocabulary, syntax, text base and other aspects. It can be used in various aspects of foreign language teaching and research, such as the choice of reading material, reading tasks validity and verification. This paper describes the development motivation and application prospects of the tool, and takes the related project on English reading tasks of Ministry of Education of China as an example to verify the validity of the tool. This paper also introduced the details of the operation and the application of the tool.

Key words: Coh-Metrix2.0; development motivation; operation, application
_____

INTRODUCTION

Over the last decade, the development of technology and corpus linguistics research in computer science and progress made it possible to use computer tools do text analysis. Coh-Metrix series is one of the analysis tools jointly developed by experts from the University Of Memphis. It is a user-friendly, easy to use online tools for analysis of various text features to provide convenient conditions. This tool was developed in 2002 and Coh-Metrix1.0 version launched in 2004 and the latest version 2.0 in 2010. The tool can be used to analyze the English text of 200 to 15,000 words and calculate text difficulty and cohesion in the language, discourse and conceptual level according to user needs on quantify and accurately to reflect the language reading comprehension psychological factors such as decoding, parsing and meaning construction and so on [1]. The analysis results can be stored in various forms, including Text file, Excel file and SPSS files. The tool is in great value for the selection of reading materials, testing efficacy in English reading tasks. This paper describes the development of this tool in motivation, operation, and variable settings. Based on the project " Academic Analysis, Feedback and Guidance System for Primary and Secondary Schools " hosted by the Ministry of Education Basic Education Textbook Development Center, this paper describes the tools in English reading tasks to verify the validity.

DEVELOPMENT MOTIVATION OF COH-METRIX 2.0
Coh-Metrix stands for Automated Cohesion Metric Tool. As the name suggests, this is a computer platform for the convergence of the means to measure the text tool. Based on cohesion and coherence, an important feature of the tool is to quantify the various relationships to predict the convergence of text. It can help readers to interpret the text content of a variety of vocabulary, syntax clues so that readers form a coherent mental representation. Readers utilize the existing cohesive devices and language knowledge and skills to build a variety of coherent relationship in the brain. Therefore, it is the result of a coherent mental representation and processing. It can also say that convergence is the concept of text and coherent is a psychological concept [2]. On the one hand, the development motivation of Coh-Metrix system is based on the existing text readability study of criticism. On the other hand, it comes from the progress of interdisciplinary studies. Scholars' studies found that there are many shortcomings in forty readability formulas developed to measure the text, especially the two most commonly used formulas Flesch Reading Ease Score and Flesch-Kincaid Grade level [3]. Formula are as follows:

Flesch Reading Ease Score ＝206.85－1.105×ASL－84.6×ASW

Wherein, ASL is the average sentence length, which is divided by total number of sentences into the total number of words of text .ASW is the average syllable per word, which is divided by the number of words into the total number of syllables. Readability score will run between 0 and 100. The higher is the score, the easier is the text. Generally it falls between 6 to 70 points.

Flesch Kincaid Grade Level ＝0.39 ×ASL ＋11.8×ASW－15.59

Wherein, ASL and ASW have the same meaning as above. The formula is to convert Flesch readability formula, which is in hundred percentage point system, into grade in American schools' K-12 level, so that teachers and parents of students can make judgments on the readability of the text of various materials. For example, a text with 8.2 point score indicates that the text is suitable for grade eight students to read in the United States (average age between 12 to 14 years old).
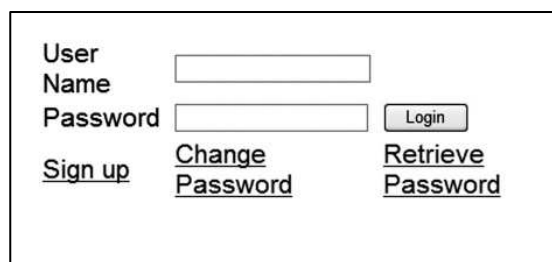
First of all, these scores from readability formulas depend on word length in the text, sentence length and other surface characteristics of the language. Too simple and superficial text may ignore the reader's subjective cognitive potential. Processing and understanding of the text depends not only on text, sentence length but also more on the reader's background knowledge, language skills and other cognitive potential. Alderson believes that the two variables affect the reading are the text and the reader, and artificial separation between the two can lead to distortion of reading research [4]. Some researchers believe that text activation construct and decoding process is an understanding the process of deconstruction from multiple levels of cognition [5]. Discourse psychologist Kintsch distinguished surface code, text base and situational model. He believes that the surface code is code-words and syntax of the text, for which the reader can only hold short-term memory unless these surface codes have a major impact on the code-text content. Text base is retained significance dominant proposition, rather than words and syntax. Text base also include the establishment of local convergence of simple reasoning, and memory can be retained for several hours. Situational model is the content of the text or the microscopic world, such as that a story of scenarios include micro-world characters, scenes and emotions established on the text. Therefore, textbase is built up by readers based on factors such text dominant characteristics, personal background knowledge and reading goals interactive. Text base lasts in memory for several days, months, even years. Accordingly, readers' understanding of the text depends not only on the surface code of the linguistic features of decoding, but also, to a greater extent, depends on the reader library text processing and scene modes. Although the surface characteristics of the text can predict its readability, but it should be seen as an interactive activity between text readability and cognitive potential. Measurement of the surface features of the text does not ensure complete understanding of the text.

Second, this readability formula cannot fully reveal the whole picture of cohesion and coherence of the text. Studies have shown that a better connected text will be easy to understand. It may be considered that scores of readability of such text should be high, but it is not the case. There is sufficient evidence that when comparing better connected sentences with poor connected sentences, the latter has the same or lower than scores than the former in Kincaid readability formulas, but is more difficult to understand. Thus, the average length of the sentence and the number of syllables is not enough to accurately predict the coherence and understanding of the text. In addition, the development of interdisciplinary made the updated text analysis tools possible, including computer linguistics, corpus linguistics, the information research, information retrieval and discourse processing and so on[6]. These studies have made in-depth discussions on cognitive processing and handling text from multiple disciplines, which has been far beyond the surface features of the research, and provided a more accurate prediction for the consistency of the text.

OPERATION OF COH- METRIX2.0
Coh-Metrix is an online analysis tool for academic research and non-commercial study. Its website is http://cohmetrix.memphis.edu/cohmetrixpr/index.html. Users can click on the site; go to Coh-Metrix website of the Department of Psychology at the University of Memphis. Since the latest online version is 2.0, focuses will be on the use of this version.

First, a user registration. User's personal information is sent, and then the site will automatically send the user a password. The user name is usually the name of the registered user. (See Figure 1)
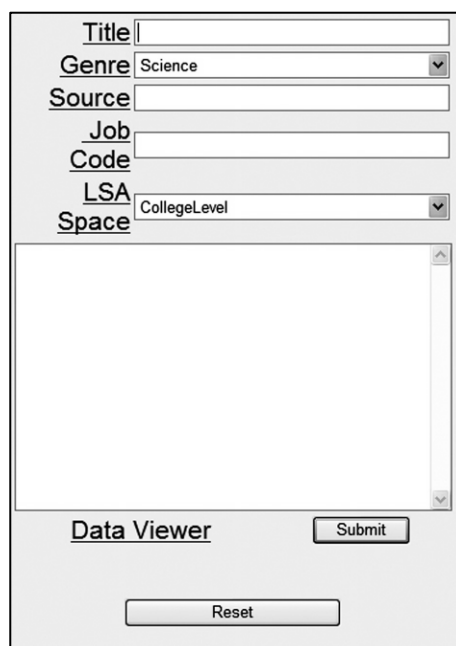
_____

Figure 1: User interface for login

In Figure 1, "Sign up" is used to register. Upon registration, the user name may be virtual, but the mailbox provided must be true, otherwise it is impossible to know your password. User name for login and password can be used continuously, without having to re-register. Then enter the user interface (see Fig. 2).

Figure2: User interface for online operation

In order to analysis the test, the user can paste the text directly from the source text, or manually type it. When this is done, click the "Submit", and then the background program run the analysis. Generally, after a few seconds, the output will be presented in the right side if the screen in the form of a table. Users can download and store the results of the analysis. Table 1 shows an example of the result of text analysis.

APPLICATION OF COH- METRIX2.0 AND DATA ANALYSIS
The English subject examination in "Academic Analysis, Feedback and Guidance System for Primary and Secondary Schools" is based on curriculum standards in English Grade Four (junior high school students in grade eight). The aim of this project is test students' academic level [7].

Table 1: Output of part of the text analysis results in a longitudinal way

No.	Description	Measure	Text 1	Text 2	Text 3	Text 4	Full description
1	Title	Title	Orlando Low Cohesion	Orlando High Cohesion	Heat Low Cohesion	Heat High Cohesion	Title
2	Genre	Genre	Narrative	Narrative	Science	Science	Genre
3	Source	Source	Coh-Metrix	Coh-Metrix	Coh-Metrix	Coh-Metrix	Source
4	JobCode	JobCode	hclc	hclc	hclc	hclc	JobCode
5	LSASpace	LSASpace	CollegeLevel	CollegeLevel	CollegeLevel	CollegeLevel	LSASpace
6	Date	Date	2006-07-06	2006-07-06	2006-07-06	2006-07-06	Date
7	Causal content	CAUSVP	40	56.777	89.109	88.292	Incidence of causal verbs, links, and particles
8	Causal cohesion	CAUSC	0.727	1.909	0.194	0.808	Ratio of causal particles to causal verbs (cp divided by cv+1)
9	Pos. additive connectives	CONADpi	15.556	20.147	9.901	17.274	Incidence of positive additive connectives

To ensure the validity, scope of inspection includes all the English reading tasks since 2005. The theoretical framework of English tests in reading tasks is a triangulation, which has two basic premises and assumptions.

1. English textbooks for eighth-graders are following the relevant requirements of English curriculum standards of grade four, and the language difficulty is to meet the overall target population. This is a basic premise. Now all the public offering Junior English textbooks are based on the relevant requirements of "English Curriculum Standard" (trial version).

2. If the reading tasks in these English tests is designed in accordance with the relevant requirements of grade four, the text feature of the reading material should has not significantly different correspond with text features used by students. That is to say, the language difficulty of the reading tasks should be in line with the target population. Therefore, in order to prove the above hypothesis and demonstrate the effectiveness of reading task, data and information collection must be done.

Firstly, bring together all reading texts, and then label them according to the genre of the text (such as narrative, expository) and topics (such as school life, culture and customs). Secondly, draw samples from the teaching materials. Topics should be closely related to the student's personal, family and school life, and should include daily life, hobbies, customs and cultural aspects of science topics. To ensure comparability and accurate comparison, reading materials of different genres and authors are selected. The results showed that, 28 texts in the tests (accounted for 66% of the total text) were found with the same genre and topic text in textbooks.

After completing the above steps, we use the Coh-Metrix2.0 tool to measure the above two sets of text variables. Coh-Metrix2.0 has 60 variables, which can be roughly divided into six categories: (1) basic identification information; (2) readability index; (3) the basic vocabulary and text information; (4) syntax index; (5) refers to the semantic index; (6) the profile dimensions.

(1)Basic identification information, which is the information for registration or selection in Figure 2, including the "Title" "Genre" "Source" "Job code" "LSA Space" and so on.

(2)Readability indexes. There are two, namely the two formulas: Flesch readability formula and Flesch Kincaid grade level formula. The calculation of the length of sentence and word in the two formulas is based on CELEX corpus database. The database contains 17.9 million words corpus of COBUILD Corpus 1991 edition, of which 100 million were spoken English corpus, and the other for written materials.

(3)The basic vocabulary and text information, which include a total of 14 variables. It includes basic counting, word frequency, degree of physical vocabulary, verbs and nouns Hyponymy.

(4)Syntax index of 22 variables are used to measure the complexity of the text syntax, syntactic categories and syntactic composition and specific constituents. In general, the more complex sentence structure, the more embedded components contains. The high the structural density, the more complexity the cognition is and difficult to understand. There are three ways for measuring syntactic complexity. ①Calculate the average number of qualifiers of noun phrases, including adjectives, adverbs, qualifiers defining the center of the word. ②Calculate the average level component of each sentence. That is to calculate the number verb phrase in a complex sentence, because different verb phrase control different number of words in a speech. ③In a complex structure, the calculated number of words in front of the main verb clause, because that the different number of words will have impact on readers' memory. Syntax index include: ①components of parts of speech, ② pronouns, signs of class and form and ratio of personal pronouns and nouns ③all kinds of conjunctions indicating progressive, time, logic, cause and effect and other cohesion and relations.

(5)Coreference and semantic index with a total of 10 variables. Coreference means that nouns, pronouns, or noun phrases are used to refer to another component. Semantic index is the similarity of a sentence or paragraph to others in semantic or conceptual aspects, which was divided into three cases. ① Anaphora, including the neighbor sentence anaphora, and anaphora with more than five sentences. ②Same referent, including full nouns same referent, stem same referent etc..③ Latent semantic analysis, including adjacent sentences, all sentences and paragraphs semantic analysis.

(6)Text base dimension refers to the contents of the text or the creation of microscopic world, with a total of six variables, which are divided into four categories. ① Causal dimension, which is used more in science and technology text analysis. It is mainly based on the WorldNet database [8]. ② Object dimensions is more used for the story or narrative passages, suitable for a living individual to perform certain actions in order to achieve certain purposes of analysis. ③ Time dimension is used for texts with a variety of table time as cohesive methods. ④ Spatial dimensions are used for texts with a variety of spatial relationships as cohesive methods.

To ensure the reliability and validity of the study, the measurement of the text should include features from words to sentences, various dominant features (such as counting language units) and recessive trait (such as moving, noun hyponymy relationship). At the same time, in order to study the operability, 14 variables of 54 variables were extracted in addition to the basic identifiable, including Hyponymy readability, the average length of words, verbs and nouns and noun phrases defined before syntactic structure of word similarity average, average sentence length, structural similarity adjacent syntactic text all the sentences of the adjacent sentence anaphora, etc., covering variables in 5 categories. Social statistical software SPSS is used to perform the T-test for two texts of 14 independent variables of the samples. The results shown in the following table.

Table 2: Statistical result of T test on independent samples from reading texts in examinations and teaching materials

Variables	Reading texts in exams		Reading texts in teaching materials		Df.	T value	Sig. (2-tailed)
	Mean	SD.	Mean	SD.			
Flesch Reading Ease Score	81.22	11.08	84.04	7.03	53	-0.83	0.41
Flesch Kincaid Grade Level	4.43	1.64	4.45	1.66	53	-0.03	0.98
average syllable per word	1.36	0.14	1.30	0.08	53	1.526	0.14
average sentence length	10.00	2.93	11.68	2.45	53	-1.702	0.10
Average sentences in the text	16.27	6.98	16.47	5.01	53	-0.09	0.93
Hyponymy of nouns	5.18	0.69	16.47	5.01	53	0.218	0.83
Hyponymy of verbs	1.51	0.27	1.53	0.68	53	-0.31	0.76
Occurrence rate of nouns	320.76	32.68	298.64	33.93	53	1.819	0.08
Qualifiers before nouns	0.74	0.21	0.71	0.18	53	-0.527	0.60
Similarity in sentence structure of adjacent sentences	0.75	0.04	0.77	0.04	53	-1.77	0.08
Similarity in sentence structure of all sentences	0.16	0.05	0.14	0.04	53	1.208	0.24
Anaphora of adjacent sentences	0.20	0.26	0.13	0.03	53	1.309	0.31
Anaphora of all sentences	0.43	0.22	0.47	0.15	53	-0.638	0.53
Thematic contact ratio	0.37	0.21	0.45	0.16	53	-0.219	0.23

Statistical results showed that there was no significant difference ($p > 0.05$) between two groups of text in the readability of the text, the average word length, Hyponymy verbs and nouns and noun phrases before defining the

word average, adjacent syntactic structure similarity, all text syntactic structure of sentences similarity adjacent sentence anaphora, etc. It can be concluded that the difficulty of reading text is in line with that of teaching materials, which is suitable to the language level of the testee [9]. In other words, the reading task design in the project is effective and achieves the purpose of academic tests.

CONCLUSION

From the description, operation and application of Coh-Metrix2.0, we can see that the tool is used as a free online tool, which is powerful, user-friendly. It can give an accurate measurement on the dominant features of the text (and the related convergence, as) and the complexity of the recessive trait vocabulary, sentence structure, sentence interpersonal relations (and coherence related to a number of variables such as the relationship between meaning, sentence or paragraph semantic index). Users can obtain accurate quantitative data to provide scientific basis for decision making.

There are some drawbacks of Coh-Metrix 2.0. First of all, reading comprehension is a complex cognitive process, in addition to the variable design tools, readers use strategies, emotional when they are reading (such as motivation, anxiety, etc.). And other factors will affect the understanding of the text. Secondly, the analyzing tool for the type of the text genre is in broad terms. Apart from that it can has accurate analysis of science and technology, research and sociological narrative style, but all others are classified into "other" column. When argumentative and narrative texts are compared, cognitive processing of the former is more complex [10]. Furthermore, the calculation of the result data is more complicated to deal with. Regardless of the number of variables user selected, the tool will show data of all 60 variables, which increased the workload. What's more, these data cannot be used directly. Users can only collect these basic data and utilize some other statistical software in order to make more scientific and accurate decision-making.

Coh-Metrix 2.0 is indeed an easy-to-use analysis tool. It can provide very accurate, comprehensive data of text feature, and promote more in-depth academic study of the text.

REFERENCES

[1] Crossley S A, Greenfield J and McNamara. *TESOL Quarterly*, V. 42, n. 3, pp.475-493, January, **2008**.
[2] Louwerse M M. *Cognitive Linguistics*, V. 12, n.12, pp.291-315, December, **2002**.
[3] Graesser A C, McNamara D S, Louwerse M M. *Behavior Research Methods, Instruments & Computers*, V. 36, n. 6, pp.193-202, February, **2004**.
[4] Marcus M, Santorini B & Marcinkiewicz M. *Computational Linguistics*, V.24, n.19, pp.313-330, October, **2003**.
[5] Lehnert W G. *Discourse Processes*, V.24, n.23, pp.441-470, December, **2007**.
[6] Belew R K. *Information Retrieval*, V.12, n.5, pp.269-278, May, **2002.**
[7] Deerwester S, Dumais S T, Furnas G W. *Journal of the American Society for Information Science*, V.48, n.41,pp.391-407 December,**2006**.
[8] Voorhees E. *Natural Language Engineering*, V. 12, n. 7, pp.361-378, July, **2001**.
[9] Green A, and Weir C, *Language Testing*, V.24, n. 2, pp 191-211, January, **2010**.
[10] Kintsch W, *Comprehension：A paradigm for cognition* [M]. Cambridge：Cambridge University Press，**1998**.