Using Coh-Metrix to Analyse Writing Skills of Students: A Case Study in a Technological Common Core Curriculum Course

Chi-Un Lei, Ka Lok Man and T.O. Ting

***Abstract*—Pedagogy with learning analytics is shown to facilitate the teaching-learning process through analysing student's behaviours. In this paper, we explore the possibility of using a computational linguistic tool Coh-Metrix for analyzing and improving writing skills of students in a technological common core curriculum course. In this study, we mainly focused on the investigation of syntactic simplicity, word concreteness, referential cohesion, and deep cohesion of student's essays. We studied 25 essays from the three-year curriculum students and 26 essays from the four-year curriculum students. Results illustrate the necessity of improving student's writing skills in their university learning, so that they can effectively circulate their ideas to the public in the future.**

***Index Terms*—Information technology in education, general education, learning analytics, educational data mining, computational linguistics, text analysis, writing**

I. Introduction

With recent advances in information technologies, an emerging mode of practices known as the learning analytics (educational data mining) has begun to change the paradigm of higher education [1]–[4]. Learning analytics can be defined as "*the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs*". It has been used to model individual learning trajectories as well as social learning behaviours. Different related studies, such as computational linguistics [5], [6] have been adopted in teaching practices.

Meanwhile, Hong Kong is currently adopting the higher-education transformation from a three-year curriculum to a four-year curriculum. With that in mind, the Department of Electrical and Electronic Engineering at the University of Hong Kong (HKU) has adopted new technologies and practices for teaching facilitation [7]–[10]. In particular, in order to provide key common learning experience for all undergraduate students in HKU and to broaden their horizons beyond their chosen disciplinary fields of study, the department has introduced two one-semester Common Core Curriculum (CCC) courses for students.

One of the major goals of CCC in HKU is to cultivate students to play an active role as responsible individuals in local and global communities. Thus, in order to help students circulate their ideas to the public effectively in the future, CCC is also responsible to help students with their writing process: pre-writing, developing their own research questions, composing thesis statements, and proofreading. However, due to the tight teaching schedule and lack of linguistic consulting experience, it is difficult for us to carefully examine student's writing in a short period of time. Therefore, in the recent semester, we explored the feasibilities of using computational linguistic analysis for developing student's essay writing skills.

In this paper, we give a discussion on applying computational linguistics for analysing and improving student's writing skill. The major contributions of our paper are as follows.

- We used a computational linguistic tool Coh-Metrix [11] to analyse the readability and linguistic features of student's essays. Once these features are identified, we can work with students to help them overcome the obstacles that less cohesive texts might present.
- This study analysed the language varieties and discourse characteristics of writings for differences between three-year curriculum students and four-year curriculum students under the education reformation. Based on the data, we provide a specific training for four-year curriculum students.

Section II describes the course and essays that have been analysed. Section III describes the analysis methodology. Finally, findings and discussions are shown in Section IV.

II. Studied Course and Essays

The studied course, Everyday Computing and the Internet (CCST9003) is a common core (general education) course first offered in 2010. Besides introducing to students a "computational thinking" concept through twelve-weeks teaching, CCST9003 also discusses intensively the societal impacts of computing technologies on our daily life, through surveying of computational methods and analysing usage of computational methods.

In order to learn how to circulate ideas about computational thinking to the public, students have to write a survey essay on a topic related to everyday computing and the internet. The essay should offer knowledge and inspiration to the public as well as engagement with ideas and vice versa..

III. Text Analysis via Coh-Metrix

Computational linguistics study language from a computational perspective. Through knowledge-based or data-driven

Manuscript received December 31, 2013.

C.-U. Lei is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong. Email: culei@eee.hku.hk

K.L. Man is with the Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, China and Baltic Institute of Advanced Technology, Lithuania. Email: ka.man@xjtlu.edu.cn

T.O. Ting is with the Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, China. Email: toting@xjtlu.edu.cn

The research work presented in this paper is partially sponsored by Transcend. It is also partially supported by HKU Strategic Research Theme on "Sciences of Learning".

modeling, linguistic phenomena can be modeled by computational models. These models are often used to explain linguistic behaviour or to provide a working component of a language system. For example, readability has been generally described by the Flesch Reading Ease (FRE) Score

$$\mathrm{FRE} = 206.835 - (1.015 \times \mathrm{ASL}) - (84.6 \times \mathrm{SW}), \quad (1)$$

where ASL is the average sentence length or the number of words divided by the number of sentences; SW is the average number of syllables per word. A higher FRE score indicates the article is easier to read. Generally, an essay has a Flesch Reading Ease score between 6 and 70.

Meanwhile, in this paper, we mainly focused on a few specific linguistic characteristics of texts: syntactic simplicity, word concreteness, referential cohesion, and deep cohesion. These characteristics mainly describe whether the essay is helping the reader mentally connect ideas in the text and easy to comprehend. These characteristics have been discussed in [11], and are outlined in the following subsection. Meanwhile, we used Coh-Metrix to identify these characteristics. In this paper, each of these collected components for a given text has been normalized, according to thousands of sample texts stored in the Coh-Metrix database.

A. *Syntactic Simplicity*

Syntactic simplicity reflects the degree to which the sentences in the text contain fewer words and use simpler, familiar syntactic structures, which makes readers less challenging to understand. Coh-Metrix measures syntactic simplicity through several indices. In particular, texts with fewer clauses and words per sentence, and fewer words before the main verb/clause will give a text a higher score for syntactic simplicity.

B. *Word Concreteness*

Concrete words (e.g. apple, bottle, car and dog) are words that stimulate sensory response in the reader. In other words, we can imaginatively use our senses to experience what the words represent. On the other hand, abstract words (e.g. love, success, freedom and joy) usually refer to the ideas or concepts with no physical referents. A text with relatively high numbers of concrete words is easier to read, thus will have a high word concreteness score. Coh-Metrix can compute the average Word Concreteness through a rating database of 4293 unique words. For example, words “protocol” (264) and “difference” (270) are recorded as less concrete than “ball” (615) in the database.

C. *Referential Cohesion*

A text with high referential cohesion contains words and ideas that overlap across sentences and the entire text, forming explicit threads that connect the text for the reader. In other words, when sentences and paragraphs have similar words or conceptual ideas (i.e. high referential cohesion), the text is connected explicitly by threads, thus it is easier for readers to deduce connections between those ideas as well as to understand the essay easily. On the other hand, low cohesion text is typically more difficult to process because there are fewer connections that tie the ideas together for readers. Referential cohesion can be measured by the overlap between verb, noun, argument, word stem and content word from one sentence to the other.

TABLE I
TYPES OF THESE CONNECTIVES.

Type	Connectives
Time	after, earlier, before, during, while, later
Causal	because, consequently, thus
Additive	both, additionally, furthermore, moreover
Logical	actually, as a result, due to
Adversative	but, yet, however, although, nevertheless

D. *Deep Cohesion*

Deep cohesion measures how well the events, ideas and information of the whole text are tied together. This can be measured by connectives and types of words that connect different parts of a text. For example, adversative connectives are words that connect two phrases or notions that conflict with each other, such as “My favourite subject is operational management however I studied engineering.” or “Tomato is a fruit, yet it is used in savoury.” Different types of these connectives are shown as in Table I.

E. *Connectivity*

Connectives play an important role in the creation of cohesive links between ideas and clauses and provide clues about text organization. Connectivity reflects the degree to which the text contains explicit causal (e.g. because, so), logical (e.g. and, or), adversative (e.g. although, whereas), temporal (e.g. first, until) and additive (e.g. and, moreover) connectives to express relations in the text. This component reflects the number of logical relations in the text that are explicitly conveyed. This score is likely to be related to the readers deeper understanding of the relations in the text.

F. *Temporality*

Texts that contain more cues about temporality and that have more consistent temporality (e.g., tense, aspect) are easier to process and understand. In addition, temporal cohesion helps readers understand the situation of the event in the text.

G. *Length of Sentences and Paragraphs*

The organization of an essay can also be described by the architecture of sentences and paragraphs, the mean number of sentences in paragraphs and the mean number of words in sentences. A higher value indicates that the section may have more complex syntax and may be more difficult to process. For example, a large standard deviation of the mean length of paragraphs indicates that the essay may contain some very short and some very long paragraphs, posing understanding difficulty for most readers.

IV. RESULTS

We studied 25 essays from the three-year curriculum students and 26 essays from the four-year curriculum students. Furthermore, effect size has been calculated to show the strength of the relationship between variables. Results are shown in Table II. These metrics can assess students whether

TABLE II
RESULTS OF ANALYSIS. (SD: STANDARD DERIVATION)

Metric	4-Year (Mean)	4-Year (SD)	3-Year (Mean)	3-Year (SD)	Effect Size
Flesch Reading Ease	50.71	7.06	**50.76**	7.74	-0.01
Number of paragraphs	16.6	8.26	10.46	7.38	0.78
Number of sentences	49.4	16.37	38.38	12.34	0.76
Number of sentences in a paragraph (Mean)	3.78	3.63	4.56	1.88	-0.27
Number of sentences in a paragraph (Standard Deviation)	3.69	6.42	2.36	1.10	0.29
Number of words in a sentence (Mean)	14.88	2.49	17.11	2.58	-0.88
Number of words in a sentence (Standard Deviation)	9.40	2.41	9.43	2.42	-0.01
Syntactic simplicity (Normalized)	**68.65**	13.34	60.94	19.30	0.46
Word concreteness (Normalized)	**25.64**	17.22	25.47	22.46	0.01
Referential cohesion (Normalized)	29.18	20.50	**33.62**	23.91	-0.20
Deep cohesion (Normalized)	66.92	18.54	**72.29**	19.31	-0.28
Verb cohesion (Normalized)	19.53	18.70	**39.57**	25.58	-0.89
Connectivity (Normalized)	**6.42**	9.48	5.93	13.85	0.04
Temporality (Normalized)	**45.28**	22.96	40.62	22.25	0.21
Noun overlap (Adjacent sentences)	0.38	0.13	**0.39**	0.12	-0.11
Argument overlap (Adjacent sentences)	0.46	0.13	**0.48**	0.13	-0.16
Stem overlap (Adjacent sentences)	0.49	0.13	**0.50**	0.09	-0.10
Noun overlap (Adjacent sentences)	**0.30**	0.12	0.29	0.12	0.03
Argument overlap (Adjacent sentences)	0.37	0.11	**0.37**	0.13	-0.04
Content word overlap (Adjacent sentences)	0.09	0.03	**0.09**	0.04	-0.26
Content word overlap (All sentences)	0.06	0.02	**0.07**	0.03	-0.26

they can write organized, rich and complex essays that are easy to understand. Thus, these metrics can be used to investigate problems in student's writing, for example,

- A low concreteness score indicates students may not be able to explain abstract ideas clearly.
- A low referential cohesion score means students might have trouble on building sentences on each other.
- A low deep cohesion scores indicates students have difficulties to comprehend how the ideas, events or information of the text as a whole fit together.

Some observations are as follows.

- Four-year curriculum students tend to write more paragraphs with less sentences and words in each paragraph, comparing to three-year curriculum students. However, usually it is not easy to develop a concrete idea in a paragraph with three or four sentences only, due to the lack of supporting detail.
- Essays from four-year curriculum students tend to possess more syntactic simplicity and temporality, comparing to those from three-year curriculum students. Meanwhile, three-year curriculum students work better in developing referential/deep/verb cohesion relationships in their essays. Therefore, the analysis slightly indicate the necessity of improving writing skills of four-year curriculum students, due to lessening of writing training in the new six-year high school curriculum.
- However, comparing to samples in the Coh-Metrix database written by other students all around the world, students in HKU tend to write essays that are difficult to be understood. In particular, a low overall cohesion and a significantly low connectivity indicate their essays tend to be less-organized and less easy to be understood.
- Essays from three-year students tend to have slightly more content overlap in terms of argument and content word. However, comparing to samples in the Coh-Metrix database, students in HKU are still not yet applicable to write essays with enough content overlap. Thus, ideas developed by students may not be effectively circulated to the general public. It indicates the necessity of improving student's writing skills in their learning.

V. CONCLUSION

In this paper, we used Coh-Metrix to analyse the organization of students' essay in a technological Common Core Curriculum course. The evaluation illustrates the necessity of improving student's writing skills in their university learning stage. The adopted analysis methodology can be extended to the determination of several advanced metrics, namely: latent semantic analysis, lexical diversity, syntactic complexity, and syntactic pattern density. Eventually, we would like to develop an automated essay assessment system, such that teaching staffs can focus on designing effective learning activities, but not routine marking process.

REFERENCES

[1] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," *Educause Review*, vol. 46, no. 5, pp. 30–32, 2011.

[2] R. Ferguson, "The state of learning analytics in 2012: A review and future challenges," *Knowledge Media Institute, Technical Report KMI-2012-01*, 2012.

[3] R. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[4] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 40, no. 6, pp. 601–618, 2010.

[5] D. Jurafsky, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. MIT Press, 2000, vol. 2.

[6] R. Ferguson and S. B. Shum, "Learning analytics to identify exploratory dialogue within synchronous text chat," in *Proc. ACM Intl. Conf. on Learning Analytics and Knowledge*, 2011, pp. 99–103.

[7] C.-U. Lei, H.-N. Liang, and K. L. Man, "Advancements in using a machine design project for teaching introductory electrical engineering," in *Teaching, Assessment and Learning for Engineering (TALE), 2013 IEEE International Conference on*, 2013, pp. 556–559.

[8] C.-U. Lei, "Applying the problem-based learning approach in teaching digital integrated circuit design," in *Proc. Enhancing Learning Experiences in Higher Education: Int. Conf.*, Dec. 2010, pp. 1–5.

[9] C.-U. Lei, H. K.-H. So, E. Y. Lam, K. K.-Y. Wong, R. Y.-K. Kwok, and C. K. Chan, "Teaching introductory electrical engineering: a project-based learning experience," in *Proc. IEEE Intl. Conf. on Teaching, Assessment and Learning for Engineering*, 2012, pp. 335–339.

[10] C.-U. Lei, K. L. Man, H.-N. Liang, E. G. Lim, and K. Wan, "Building an intelligent laboratory environment via a cyber-physical system," *International Journal of Distributed Sensor Networks*, vol. 2013, 2013.

[11] A. C. Graesser, D. S. McNamara, and J. M. Kulikowich, "Coh-metrix providing multilevel analyses of text characteristics," *Educational Researcher*, vol. 40, no. 5, pp. 223–234, 2011.