Component Model in Discourse Analysis

Haiying Li
University of Memphis
202 Psychology Building
University of Memphis
Memphis, TN, 38152
Tel. 1-901-678-2364
hli5@memphis.edu

Arthur C. Graesser
University of Memphis
202 Psychology Building
University of Memphis
Memphis, TN, 38152
Tel. 1-901-678-4857
graesser@memphis.edu

Zhiqiang Cai
University of Memphis
202 Psychology Building
University of Memphis
Memphis, TN, 38152
Tel. 1-901-678-2364
zcai@memphis.edu

ABSTRACT
Automated text analysis tools such as Coh-Metrix and Linguistic Inquiry and Word Count (LIWC) provides overwhelming indices for text analysis, so fewer underlying dimensions are required. This paper developed an underlying component model for text analysis. The component model was developed from large English and Chinese corpora in terms of results from Coh-Metrix, and English and Chinese Linguistic Inquiry and Word Count (LIWC).

Keywords
Component model, Coh-Metrix, LIWC, principal component analysis

1. INTRODUCTION
With the development of the computational linguistics, automated text analysis tools like Coh-Metrix and Linguistic Inquiry and Word Count (LIWC) have been developed to analyze enormous amounts of data efficiently.

Coh-Metrix provides 53 language and discourse measures at multilevels related to conceptual knowledge, cohesion, lexical difficulty, syntactic complexity, and simple incidence scores (http://cohmetrix.memphis.edu) [1]. Meanwhile, a principle components analysis performed on 37,520 texts of TASA corpus extracts five factors (Coh-Metrix-Text Easability Assessor, TEA, http://tea.cohmetrix.com), including Narrativity (word familiarity and oral language), Referential cohesion (content word overlap), Deep cohesion (causal, intentional, and temporal connectives), Syntactic simplicity (familiar syntactic structures), and Word concreteness (concrete words) [1].

Even though the Coh-Metrix provides the normed five dimensions, no articles describe the details of this model. This paper not only gives a thorough description of this model, but also uses this method to build up the normed dimensions with the text analysis tools of English and Chinese LIWC.

LIWC is a text analysis software program with a text processing module and an internal default dictionary [2]. LIWC classifies words into 64 linguistic and psychological categories. The 2007 English LIWC dictionary contains 4,500 words and word stems.

The Chinese LIWC dictionary was developed by National Taiwan University of Science and Technology based on the LIWC 2007 English dictionary, but some word categories unique to the Chinese language were added to the Chinese LIWC dictionary [3]. The Chinese LIWC dictionary included 6,800 words across 71 categories. The Memphis group converted the traditional Chinese characters in LIWC dictionary to the simplified Chinese characters, which was used in our study.

With the overwhelming features for text analysis, researchers prefer fewer underlying dimensions. The most prevalent method to reduce the dimensionality is the principal component analysis (PCA) in text analysis [4, 5]. However, PCA assumes the ratio of cases to variables, so the corpus with smaller amount of cases is inappropriate to perform PCA [6]. Therefore, the standardized and normed component scores from the large reference corpus are needed.

This paper aims to develop a component model of text analysis with the automated tools of Coh-Metrix and LIWC; thus, the component scores of any coming data set computed with this model will be standardized and comparable.

2. METHOD
Two reference corpora were used in this study. The English corpus used TASA (Touchstone Applied Science Associates, Inc.), randomly-collected excerpts of 37,520 samples, 10,829,757 words with nine genres, including language arts, science, and social studies/history, business, health, home economics and industrial arts.

The Chinese reference corpus was collected according to similar genres in TASA such as classic fiction, modern fiction, history, science. Texts in the Chinese corpus included complete 4,679 documents with 25,184,754 words rather than segmented.

Six factors extracted from LIWC in these two independent corpora showed significantly high correlation on dimensions of cognitive complexity, narrativity, emotions and embodiment [7]. Therefore, these two corpora are able to reflect some common linguistic and psychological features.

The procedure of the component model is described below. First, TASA was analyzed by Coh-Metrix, English LIWC; Chinese corpus was analyzed by Chinese LIWC. Thus, three data sets were generated. Second, PCA was performed to reduce a range of indices from Coh-Metrix (53) and LIWC (English 64; Chinese 71) to fewer potential constructs. The fixed number of dimensions

was decided by the eigenvalue greater than 2. Finally, the mean, standard deviation and coefficient for each category in each dimension were extracted to develop component model.

3. RESULTS AND DISCUSSION

The factorability of the items for the appropriateness of the performance of PCA used such criteria as the ratio of cases to variables, correlations, Kaiser-Meyer-Olkin measure, and Bartlett's sphericity.

First, the ratio of cases to variables at least 521:1 was satisfied. Then the majority of correlations among indices were above .50. Secondly, the overall Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy of Coh-Metrix sets was higher than .50 and the Bartlett test of sphericity was statistically significant across three data sets. All indices were included in the analyses. The varimx rotation was used in the analysis.

The initial eigen values greater than 2 indicated the appropriate fixed number of components. One reason why we used eigen values greater than 2 rather than 1 was that too many components were extracted with eigen values greater than 1. In TASA Coh-Metrix, 6 components were extracted explaining 58% of the total variance. In TASA LIWC, 6 components were extracted and explained 40% of the total variance. In Chinese LIWC, 7 components were extracted and explained 53% of the total variance.

The components were labeled based on the linguistic or psychological features of the highly loaded categories in the component. For the English Coh-Metrix data set, the components were labeled from the first to the fifth in order as Narrativity; Referential Cohesion; Syntactic Simplicity, Word Concreteness, and Deep Cohesion. The last component only had 3 variables, so we removed that component from the model. For the English LIWC data set, the components were labeled from the first to the sixth in order as Narrativity; Processes, Procedures, Planning; Social Relations; Negative Emotion; Embodiment; Collection. For the Chinese LIWC data set, the components were labeled in the order from the first to the seventh as Processes, Procedure, Planning; Narrativity; Space and Time; Embodiment; Positive Emotion, Negative Emotion; and Personal Concerns.

The component composite score for the coming data set will be computed through an automated tool developed according to the formula of Component Model. Component Model will be obtained by the following formula,

$$y = \sum_{1}^{n} \left(\frac{x-\mu}{s}\gamma\right)$$

among which y is a component score for a coming corpus (CC); x is the value of each category on a document of CC; μ is the mean of each category from reference corpus (RC) which includes TASA Coh-Metrix, TASA LIWC or Chinese LIWC; s is the standard deviation of each category from RC; γ is coefficient of each category from RC. *1* to *n* means the number of categories in each component. $\sum$ means the sum of all the scores of the categories on each component.

For example, a teacher would like to look at the composite component score of Negative Emotion from the students' writings in English with LIWC. The teacher only has 15 subjects, so this data set is inappropriate to perform PCA. Therefore, the English LIWC Component Model should be used. First, the teacher should analyze the writing with English LIWC to obtain the score of all the indices (64). Then the mean and standard deviation of indices in all the categories, the corresponding coefficients of Negative Emotion component in the Component Model should be obtained from the reference corpus.

For instance, the "verb" score for one subject is 1.5. According to the model, the mean of the "verb" is 1.37, standard deviation 1.29, and the coefficient -0.06. Thus, the value of the "verb" in the component score is [(1.5-1.37)/1.29](-0.06) = 0.01 for this subject. We need compute the value of all the other categories in this way, then sum them, and finally obtain the value of the Negative Emotion composite score for this subject.

Thus, each component composite score from any coming corpus will be computed and standardized based on this component model from these three component models.

4. CONCLUSION

This study developed three component models for text analysis with Coh-Metrix component model, English LIWC component model and the Chinese LIWC component model. The component model can be used to generate the composite component scores when the data set has a small sample size and PCA is inappropriately performed. The results are comparable across different data sets.

The limitation of this study is that we didn't evaluate the model with human judgment. In the future, the evaluation of the model will be carried out.

5. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (BCS 0904909) for the Minerva project: Languages across Culture.

6. REFERENCES

[1] McNamara, D.S., Graesser, A.C., McCarthy, P., and Cai, Z. in press. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press, Cambridge.

[2] Pennebaker, J. W., Booth, R. J., and Francis, M. E. 2007. *LIWC2007: Linguistic Inquiry and Word Count.* Austin, Texas: LIWC.net

[3] Huang, J., Chung, C. K., Hui, N., Lin, Y., Xie, Y., Lam, Q., Cheng, W., Bond, M., and Pennebaker, J. W. 2012. 中文版語文探索與字詞計算字典之建立 [The development of the Chinese Linguistic Inquiry and Word Count dictionary]. *Chinese Journal of Psychology, 54(2),* 185-201.

[4] Biber, D. 1988. *Variation across speech and writing*. Cambridge, England: Cambridge University Press.

[5] Lee, D. Y. W. 2004. *Modeling variation in spoken and written English*. Routledge, London/New York.

[6] Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. 2009. *Multivariate data analysis*. Prentice Hall, New Jersey.

[7] Li, H., Cai, Z., Graesser, A.C., and Duan, Y. 2012. A comparative study on English and Chinese word uses with LIWC. In *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*. (California, US, May 23 – 25, 2012), 238-243.