BRIEF COMMUNICATION

Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning

Yao-Ting Sung · Ju-Ling Chen · Ji-Her Cha · Hou-Chiang Tseng · Tao-Hsing Chang · Kuo-En Chang

Published online: 2 April 2014

Abstract Multilevel linguistic features have been proposed for discourse analysis, but there have been few applications of multilevel linguistic features to readability models and also few validations of such models. Most traditional readability formulae are based on generalized linear models (GLMs; e.g., discriminant analysis and multiple regression), but these models have to comply with certain statistical assumptions about data properties and include all of the data in formulae construction without pruning the outliers in advance. The use of such readability formulae tends to produce a low text classification accuracy, while using a support vector machine (SVM) in machine learning can enhance the classification outcome. The present study constructed readability models by integrating multilevel linguistic features with SVM, which is more appropriate for text classification. Taking the Chinese language as an example, this study developed 31 linguistic features as the predicting variables at the word, semantic, syntax, and cohesion levels, with grade levels of texts as the criterion variable. The study compared four types of readability models by integrating unilevel and multilevel linguistic features with GLMs and an SVM. The results indicate that adopting a multilevel approach in readability analysis provides a better representation of the complexities of both texts and the reading comprehension process.

Keywords Readability · Multilevel · Linguistic features · Support vector machine · Validity

Y.-T. Sung · J.-L. Chen · J.-H. Cha · H.-C. Tseng · K.-E. Chang
National Taiwan Normal University, Taipei, Taiwan

T.-H. Chang
National Kaohsiung University of Applied Sciences,
Kaohsiung, Taiwan

Y.-T. Sung (✉)
Department of Educational Psychology and Counseling, National
Taiwan Normal University, 162, Sec. 1, Ho-Ping E. Rd.,
10610 Taipei, Taiwan
e-mail: sungtc@ntnu.edu.tw

Introduction

Readability refers to the degree to which a text can be understood (Dale & Chall, 1949; Klare, 2000). Texts with high readability facilitate comprehension and learning efficiency, and hence, readability has been of long-standing research interest to reading psychologists and educational psychologists (Benjamin, 2012; Klare, 2000). The readability concept has been quantified, with the construction of readability formulae burgeoning after 1950 and having recently been applied in various areas (DuBay, 2007). However, traditional readability research was constrained by technical difficulties encountered in feature selection, formula development, and validation. Furthermore, early approaches took only a few features into account in order to ensure the simplicity and feasibility of the readability formulae (Klare, 1985), with such formulae being criticized for the insufficient and superficial linguistic features not reflecting the complex processes of reading comprehension (Bruce, Rubin, & Starr, 1981; Crossley, Greenfield, & McNamara, 2008; Graesser, McNamara, Louwerse, & Cai, 2004).

Using the Chinese language as an example, the present study considered multilevel characteristics in reading comprehension, and proposed a method for integrating multiple linguistic features. Both unilevel and multilevel linguistic features were used to construct readability models based on both discriminant analysis (DA) and a support vector machine (SVM). The prediction accuracies of the four resulting models—unilevel DA, unilevel SVM, multilevel DA, and multilevel SVM—were also validated.

Multilevel linguistic features that influence reading comprehension

Reading is a cognitive process involving interactions among multiple levels (Graesser, McNamara, & Kulikowich, 2011;

Kintsch, 1998; van den Broek, Risden, Fletcher, & Thurlow, 1996). This study considered the following four levels of linguistic features that are closely related to reading comprehension: word, semantics, syntax, and cohesion levels.

Word level

Character complexity and word length The amount of space occupied by a Chinese character never changes, but the number of strokes can vary markedly. In contrast to adult readers who process characters holistically, child readers tend to do component analysis and exhibit a word-length effect (Samuels, 2006; Chen & Su, 2010; Tan & Peng, 1990).

Word frequency When processing high-frequency words, a reader can perform perceptual integration more rapidly, which allows for easier extraction of lexical semantics from their mental lexicon and results in fewer errors (Gershkoff-Stowe & Hahn, 2007; Liu, Shu, & Li, 2007).

Semantics level

Core meaning Previous studies have shown that readers spend more time contemplating content words (Carpenter & Just, 1983) because of the heavier information load carried by these words. A sentence with a large number of content words may contain more concepts and, therefore, be more difficult to understand.

Pragmatic functions Studies of pragmatics have demonstrated that negative sentences play a crucial role for discourse understanding, since they convey more information than do affirmative sentences (Givón, 1979; Huang, 2000; Jordan, 1998).

Semantic categorization Words with more meanings (i.e., more semantic categories) have more complex semantic structures and are more likely to cause semantic ambiguity. Consequently, the number and complexity of semantic categorizations of the words may influence how readers comprehend sentences (Grainger & Frenck-Mestre, 1998).

Syntax level

Simple and complex sentences Simple sentences are semantically independent syntax units consisting of a subject and a predicate; Chinese writing generally indicates these sentences with a period, exclamation, or question mark. Complex Chinese sentences are formed by two or more simple sentences (i.e., clauses) that are commonly separated by commas.

Phrase The subcomponent of a sentence is the phrase (e.g., noun phrase, verb phrase). A large number of NPs in a sentence indicates the clustering of many concepts (Johansson, 2008). The structure of a sentence is even more complex when the NPs have more or longer modifiers (Ravid & Berman, 2010).

Cohesion level

Cohesion can be described as the syntactic or semantic relationships connecting together a series of sentences to form unified texts that are intelligible and meaningful. It is considered to be crucially and essentially involved in reading comprehension (Graesser et al., 2011).

Conjunctions Conjunctions facilitate the establishment of cohesion relationships (Louwerse & Mitchell, 2003; Sanders, Spooren, & Noordman, 1992). The inclusion of conjunctions in sentences thus facilitates the understanding of the messages and the making of inferences (Singer, Harkness, & Stewart, 1997).

Reference Using pronouns in sentences, either intra- or cross-sententially, may result in multiple interpretations and inferences. Specifically, the excessive use of pronouns in a sentence may cause reference confusion and difficulties in comprehension (van den Broek & Kremer, 2000).

Metaphor Metaphors contain correlations through which known or prior experiences are used to explain new experiences and one object is used to explain another (Lakoff & Johnson, 1980). Ahrens (2006) used sentential contexts to examine time effects when participants respond to visual targets and found that the response latency was longer when one object was used to explain another.

The present study proposed 31 linguistic features (see the Appendix Table 6) for analyzing Chinese texts based on the four levels of linguistic features described above.

Traditional readability research

Traditional readability formulae are usually constructed under the assumption that text difficulty is affected by semantic and syntactic features. Word frequency and sentence length are commonly selected as indicators of semantic difficulty and syntax complexity, respectively. For example, the famous Dale–Chall formula (Dale & Chall, 1948) and the Flesch Reading Ease (Flesch, 1948) were developed on the basis of this assumption.

Doubts have been expressed about whether there is empirical evidence for the validity of predicting readability on the basis of these superficial linguistic features (Bailin & Grafstein, 2001) and whether the statistical assumptions underlying the conventional formulae are valid (Bruce et al., 1981; Schriver, 2000).

Regarding the development of features, there are a variety of methods based on computational modeling proposed to process the deeper structures of texts. For instance, latent semantic analysis (LSA) provides an ability to model the quality of coherence and quantify it by measuring the semantic similarity of one section of text to the next. Therefore, LSA has been widely used in text analysis and to predict the effects of cohesion on comprehension (Foltz, Kintsch, & Landauer, 1998; Louwerse, 2004; Shapiro & McNamara, 2000). Foltz et al. proposed an approach for predicting coherence through reanalyzing sets of texts from two studies that manipulated the coherence of texts and assessed readers' comprehension. Lemaire, Denhiere, Bellissens, and Jhean-Iarose (2006) proposed a computer program based on LSA for knowledge representation that used predication algorithms to simulate the categorization process humans use to comprehend texts. LSA therefore has been used to classify texts containing domain-specific knowledge, such as medicine (Dumais, Furnas, Landauer, Deerwester, & Harshman 1988) and journalism (Cardoso-Cachopo & Oliveira, 2003). Sung et al. (2012) compared the performances of LSA models on texts of language arts and texts of social science. The results suggested that LSA achieved better performance on categorizing texts containing domain-specific knowledge than texts of generic knowledge.

On the other hand, construction of conventional readability formulae has mostly been based on generalized linear models (GLMs), which presume that the data are independently observed, normally distributed, and large in size, an assumption inconsistent with most real-world data (Feng, Jansche, Huenerfauth, & Elhadad, 2010; Petersen & Ostendorf, 2009). Linear regression is suitable only for continuous dependent variables and is not appropriate for constructing formulae with categorical dependent variables. Therefore, another classification method, DA, has been applied (Lee et al., 2012), due to its ability to deal with categorical dependent variables. However, the accuracy of DA decreases when (1) the data are not normally distributed, (2) the covariance matrices vary between categories, (3) outliers are present, and (4) the data are linearly nonseparable.

Machine leaning approach for readability models

More and more classification tasks rely on machine learning models such as k^{th} nearest neighbor (KNN) (Duda, Hart, & Stork, 2000; Jiang, Pang, Wu, & Kuang, 2012), artificial neural networks (ANNs) (Chae & Nenkova, 2009; Lee et al., 2012), and SVM (Vapnik & Chervonenkis, 1974). The SVM is generally regarded as one of the most effective machine learning methods for classification. It features a hyperplane capable of classifying data and has been widely applied in various fields. The SVM is suitable for text classification for the following reasons: (1) It has the ability to adjust the weights in order to tolerate misclassification, and (2) it allows researchers to select different kernel functions based on the properties of specific sets of data.

SVM employs the structural risk minimization (SRM) principle (Vapnik & Chervonenkis, 1974) to select the subset of data that is most representative of the training data. It then selects a hyperplane most capable of generalization for classification. However, the hyperplanes cannot completely distinguish the named groups when real-world data conform to a linearly nonseparable distribution, which is unfortunately often the case in practice. For example, the four training instances shown in Fig. 1—W_1, W_2, Y_1, and Y_2—cannot be separated by hyperplane L. The SVM adjusts the models by assigning a penalty value to misclassified training data. The penalty value will then be included in the mathematical formula of the SVM to cope with data misclassification. As a result, the problem of linear nonseparablility can be solved effectively by the SVM (Boser, Guyon, & Vapnik, 1992).

The SVM additionally projects the data onto the feature space via a kernel function in order to improve accuracy, as shown in Fig. 2. Four common kernel functions can be used: linear, polynomial, sigmoid, and radial-basis functions (RBFs).

The SVM has recently been employed in readability research for incorporating the relation between textual features and grade levels, both of which were entered into the SVM so that the trained models were able to predict the grade level of new text. In addition, SVM projects the linearly nonseparable data onto the feature space with the kernel function in order to improve accuracy when processing linearly nonseparable data (Feng et al., 2010; Heilman, Thompson, & Eskenazi, 2008).

Both the DA and SVM models are capable of integrating multiple features for training models, but SVM is more appropriate for text classification because (1) DA assumes the data to be normally distributed, whereas SVM does not, and (2) DA constructs discriminant functions and models based on

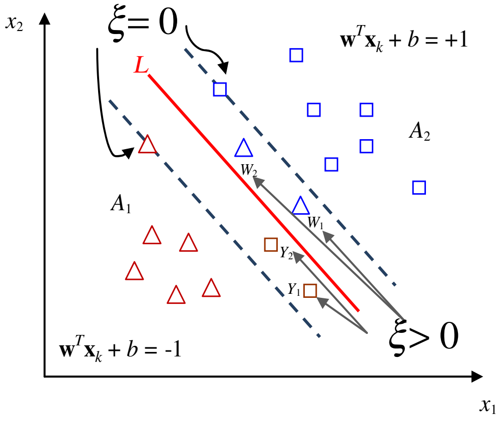
x_2
$\xi = 0$
$\mathbf{w}^T\mathbf{x}_k + b = +1$
L
A_2
W_2
W_1
A_1
Y_2
Y_1
$\xi > 0$
$\mathbf{w}^T\mathbf{x}_k + b = -1$
x_1

Fig. 1 Linearly nonseparable training data

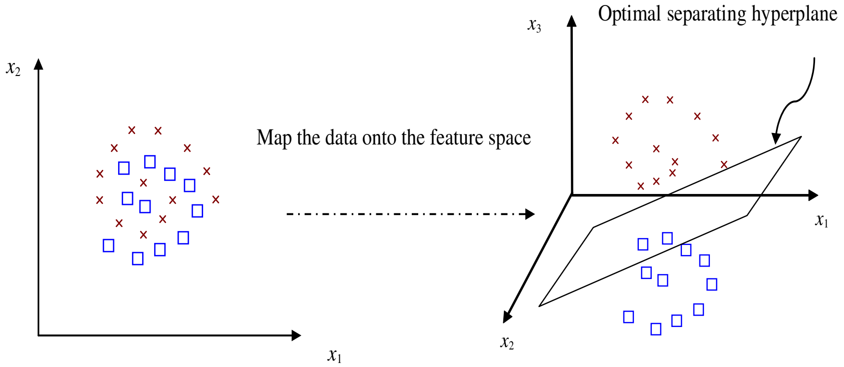
x_2
x_1
Map the data onto the feature space
x_3
Optimal separating hyperplane
x_1
x_2

Fig. 2 A kernel function can project linearly nonseparable data onto the feature space

all of the data and, therefore, may be affected by the outliers, whereas SVM—based on SRM theory—can mitigate the influence of outliers on the models. Hence, SVM can be used to distinguish different categories, and it will outperform DA when the data have similar or identical properties (linearly nonseparable). SVM is therefore better than DA in generalizing text classification (Chen, Zhou, Yin, Marks, & Das, 2007; Vapnik & Chervonenkis, 1974).

Purposes of this study

The purposes of this study were (a) to propose an approach for constructing and validating readability formulae by integrating multilevel linguistic features with the machine learning model (i.e., SVM), and (b) to empirically evaluate the machine learning approach in constructing readability formulae. To achieve these purposes, this study (a) constructed DA and SVM models with unilevel linguistic features, (b) constructed DA and SVM models with multilevel linguistic features, and (c) compared the validities of these models by examining their accuracies in predicting the grade levels of texts.

Method

Figure 3 illustrates the procedure of fivefold cross-validation used for constructing and validating readability models with unilevel and multilevel linguistic features based on DA and the SVM.

We first collected 386 texts (see the next section for details) and then used the Chinese Readability Index Explorer (CRIE) system (see the next section for details) to compute the values of the 31 features for each text. The texts, along with their features, were then divided into five subsets: Four of them were used for training the models, and the fifth was used for testing (i.e., constructing and validating). The process was repeated five times, with each subset used once as the testing data. The accuracy of the model was calculated as the average of the five results.

Tools and materials

CRIE system

Sung et al. (2012; see also Sung et al., 2013) designed the automated Chinese text analyzer, called the CRIE 1.0, which was used in the present study to analyze Chinese texts on the basis of 75 features at the following four levels to cover a wide range of linguistic and discourse representations of a text: words, syntax, semantics, and cohesion. The modules and technologies used in the CRIE 1.0 included lexicons, segmentation, syntax parsers, writing corpora, and other components that are widely used in computational linguistics. The CRIE 1.0 provides linguistic information, readability-level prediction, and writing diagnosis for Chinese native speakers, as well as reading materials for learners of Chinese as a foreign language. In this study, we used the modules of CRIE 1.0 to process texts and produce the values of each feature for the given texts. Then we used the values to construct DA and SVM models and compare their performance. On the basis of the results of this study, we use the empirical evidence about the performance of different readability models as the basis for upgrading the functionalities of CRIE 1.0.

Text materials

This study used text passages from elementary-school-level Chinese language arts textbooks published by three publishers in Taiwan: Hanlin, Nanyi, and Kangshuan. For each publisher, there were two textbooks for each of the six grade levels, giving a total of 36 textbooks. Each textbook comprised 14–18 units, giving a total of 386 units. The main text, with the exception of poetry, pictures, and illustrations, was extracted to produce 386 texts. Table 1 lists the number of texts and words for each grade. The textbooks were compiled using the following procedure: (1) The publishers invited a group of experts—including subject-matter experts from elementary schools, professors of Chinese language arts, and educational psychologists—to formulate the syllabus guideline on the basis of the national curriculum and competence indices for

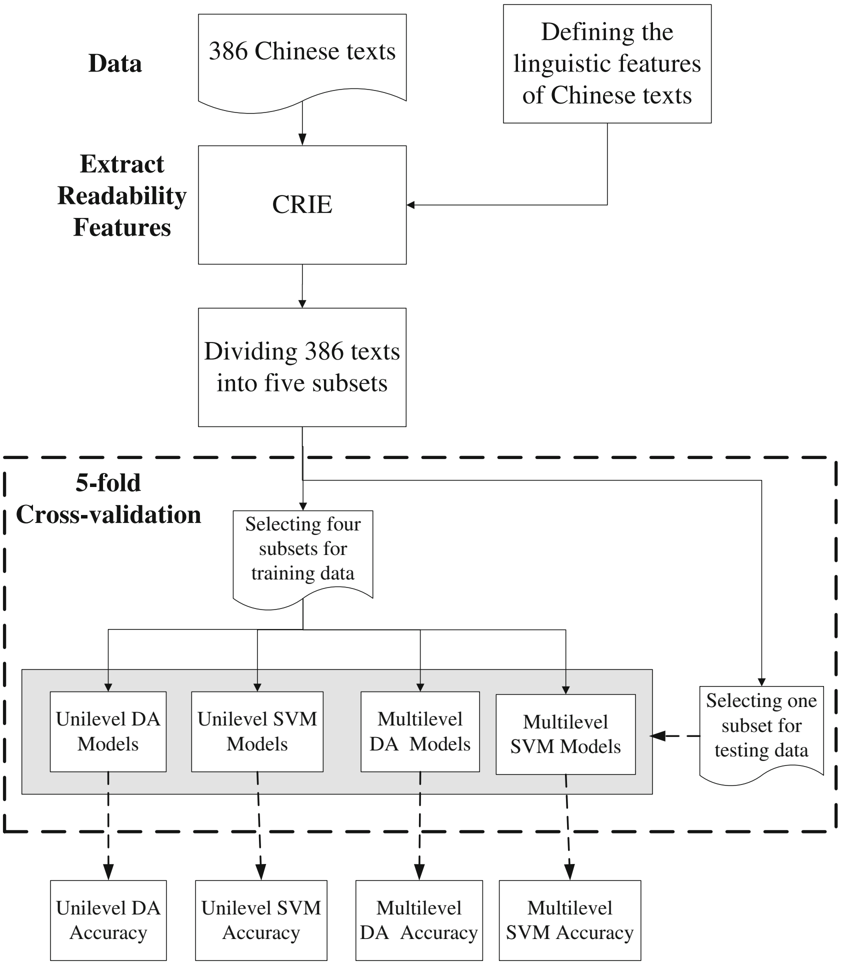

Fig. 3 Procedure for constructing and validating the models

the development of the reading abilities of school children; (2) the publishers invited writers—including elementary-school teachers and professors of language arts—to write the content based on the syllabus guideline; and (c) the publishers revised and finalized the textbooks on the basis o the on-site teaching experiences of teachers. The text difficulty is affected by the linguistic features that are present, and it is common to investigate readability using textbook materials. For example, Beck, McKeown, and Kucan (2002) and Graesser et al. (2011) showed that the number of difficult words in textbooks and the syntax complexity of textbooks increase with the grade level. The content of textbooks used at different grade levels should, therefore, reflect the semantic and syntactic features for those grade levels.

Table 1 Total numbers of texts, with mean number and standard deviation for words, from Chinese textbooks published by three publishers and used in the six grade levels

Publisher	Grade												
	1		2		3		4		5		6		
	N	Mean (SD)	*N*	Mean (SD)	*N*	Mean (SD)	*N*	Mean (SD)	*N*	Mean (SD)	*N*	Mean (SD)	Total
H	18	82.56 (51.22)	24	212.38 (70.12)	23	393.87 (71.43)	24	473.96 (106.51)	22	700.91 (157.91)	22	804.82 (134.19)	133
K	14	69.07 (43.71)	22	222.23 (50.31)	23	367.74 (90.82)	24	506.54 (118.50)	24	683.58 (115.92)	22	782.73 (123.55)	129
N	15	56.40 (29.75)	25	196.24 (51.90)	16	389.81 (40.94)	23	453.61 (86.93)	23	705.26 (183.43)	22	640.23 (191.76)	124
Total	47	70.19 (44.45)	71	209.75 (59.24)	62	383.13 (74.35)	71	478.38 (107.25)	69	696.33 (154.69)	66	742.59 (169.32)	386

Note. Publishers: H, Hanlin; N, Nanyi; K, Kangshuan

Computing multilevel linguistic features

Word segmentation and tagging

Word segmentation is a fundamental issue for computerized Chinese text analysis. Appropriate part-of-speech tagging can be achieved only with correct word segmentation. The present study used WeCan (Chang, Sung, & Lee, 2012) for Chinese word segmentation. When trained by forward maximum matching and Bayes algorithms, the WeCan tagger has an accuracy of at least 92 %.

Feature analysis

Following word segmentation and tagging, the values of the linguistic features of the texts were computed using the CRIE. The 31 analyzed features comprised 9, 5, 6, and 11 features at the word, semantics, syntax, and cohesion levels, respectively (see the Appendix Table 6). These linguistic features were validated by performing trend analysis using grade levels of the texts as the predicting variable and textual features as the dependent variable. The analyses revealed a linear trend between each of the features and the grade levels, with these features either increasing or decreasing as the grade level increases, which suggests that those features are appropriate indicators for the text difficulty (Sung et al., 2013).

Constructing and validating DA and SVM models using fivefold cross-validation

In order to validate the proposed framework, we adopted a fivefold cross-validation to examine accuracy of the readability models.

We first stratified the random sample by dividing the 386 texts into five subsets (A, B, C, D, and E). We allocated an equal proportion of grade levels to each subset, as indicated in Table 2. However, because 386 does not divide evenly by 5, in order to keep the proportions identical, the remainder was placed in subset E, making it have a total of 86 texts instead of 75.

Table 2 Number of texts in the five equal-sized subsets (A–E) according to grade level

Grade	A	B	C	D	E	Total
1	9	9	9	9	11	47
2	14	14	14	14	15	71
3	12	12	12	12	14	62
4	14	14	14	14	15	71
5	13	13	13	13	17	69
6	13	13	13	13	14	66
Total	75	75	75	75	86	386

Next, four of the five subsets in each fold were selected as training data and were used to construct the readability models, with the remaining single subset taken as the testing data to validate the models (see Fig. 4). Considering Fold 1 as an example, subsets A, B, C, and D were taken as training data and were used to construct the Fold 1 readability model, which was validated by subset E (i.e., the testing data set). Then, in Fold 2, subsets A, B, C, and E were, in turn, trained to construct the Fold 2 readability model, and the model was validated by subset D. An analogous construction and validation process was also applied to Folds 3–5. The advantage of this procedure is that each data subset was used for validation exactly once.

To construct the readability models, we entered the training texts into the CRIE to compute the value for each linguistic feature, with the features as the predicting variables and the grade levels as the criterion variable. The constructed models were then validated by entering the testing data into the CRIE for the values of features and then entering the values of features into the readability models for predicting the grade level of each text. Finally, the predicted grade level was compared with the actual grade level in order to determine the prediction (classification) accuracy, which was defined as the number of correctly predicted texts divided by the total number of texts. The procedure was repeated for fivefold cross-validation, and the average of the results was designated as the prediction accuracy of the model.

Constructing unilevel and multilevel DA and SVM models for Chinese text readability

Constructing and validating unilevel DA models

We constructed four unilevel DA models with the linguistic features from each level (Appendix Table 6) as the predicting variables and the grade level as the criterion. The unilevel DA model was constructed by first determining the probability of occurrence of a text. Each text was assumed to have an equal probability of appearing in one of the six grade levels. Each grade level can be represented as a discriminant function, which is a linear combination of each discriminant variable.

Consider the word-level DA model based on Fold 1 as an example. We first computed the scores of the features from subsets A, B, C, and D in Fold 1 to establish the linear discriminant function for each grade, as shown in discriminant functions 3–8 below, in which D_1,D_2,D_3,D_4,D_5, and D_6 represent the functions for grades 1–6, respectively:

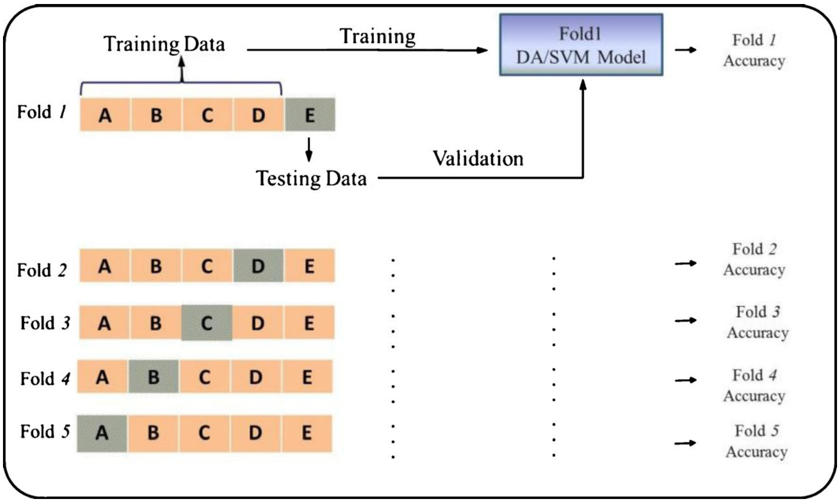

Fig. 4 Procedure for fivefold cross-validation

$$D_1 = -127.448 + 0.365x_1 + 0.512x_2 + 0.005x_3 - 0.668x_4 - 0.492x_5 - 0.093x_6 - 0.446x_7 + 124.144x_8 + 1.730x_9 \quad (3)$$

$$D_2 = -111.264 + 0.341x_1 + 0.472x_2 + 0.007x_3 - 0.569x_4 - 0.460x_5 - 0.047x_6 - 0.387x_7 + 114.871x_8 + 1.778x_9 \quad (4)$$

$$D_3 = -102.396 + 0.349x_1 + 0.467x_2 + 0.002x_3 - 0.581x_4 - 0.458x_5 - 0.031x_6 - 0.337x_7 + 107.813x_8 + 1.800x_9 \quad (5)$$

$$D_4 = -98.401 + 0.377x_1 + 0.490x_2 + 0.004x_3 - 0.592x_4 - 0.491x_5 - 0.014x_6 - 0.331x_7 + 103.167x_8 + 1.926x_9 \quad (6)$$

$$D_5 = -108.434 + 0.426x_1 + 0.548x_2 - 0.037x_3 - 0.700x_4 - 0.521x_5 - 0.035x_6 - 0.307x_7 + 101.972x_8 + 2.228x_9 \quad (7)$$

$$D_6 = -110.859 + 0.447x_1 + 0.578x_2 - 0.031x_3 - 0.748x_4 - 0.561x_5 - 0.023x_6 - 0.301x_7 + 102.773x_8 + 2.206x_9. \quad (8)$$

The Fold 1 word-level DA model was validated by computing the values of the linguistic features from subset E in Fold 1 and then entering the results of each text into discriminant functions 3–8. For example, the six scores for the fourth-grade text N-098-C-042-08 were D_1 = 95.76, D_2 = 101.5, D_3 = 104.5, D_4 = 104.4, D_5 = 102, and D_6 = 101. Since D_3 had the highest score, text N-098-C-042-08 was incorrectly classified as a third-grade reading material.

The identical procedure was applied to the other four folds, and the accuracy of the DA model at the word level was calculated as the average of the five results. We used an identical procedure to establish and validate the other three unilevel DA models.

Constructing and validating unilevel SVM models

The four unilevel SVM models incorporated the same indicators as the unilevel DA models. Consider the word-level SVM model in Fold 1 as an example. We computed the values of the linguistic features for subsets A, B, C, and D and used the LIBSVM library (Chang & Lin, 2011) to construct the SVM models. LIBSVM is integrated software that supports vector classification, regression, distribution estimation, and multiclass classification. We chose the RBF as the kernel function since such a function outperforms other types of kernels (Lin & Lin, 2003) when it is sufficient for model selection (Keerthi & Lin, 2003). In addition, two parameters, gamma and penalty parameters, used in RBF and SVM, were determined using a trial-and-error process (Bazi & Melgani, 2006); for example, the following parameters were used for the unilevel SVM model: gamma = 0.1 and penalty parameters = 1.5.

A decision function (9) was then established for the word-level SVM model based on Fold 1. The Fold 1 word-level unilevel SVM model was validated by computing the values of the linguistic features of testing subset E in (9) and predicted the grade level of each text as follows:

$$f_1(\mathbf{x}) = \operatorname{sgn}\left(\overline{\mathbf{w}}^T\mathbf{x} + \overline{b}\right) = \operatorname{sgn}\left(\sum\nolimits_{k=1}^{N} \overline{\propto}_k y_k \exp\left(-\text{gamma} \times \left\|\mathbf{x}_k^T - \mathbf{x}\right\|^2\right) + \overline{b}\right), \tag{9}$$

Where **x** is a set of N testing patterns, $\mathbf{x}_k^T$ is a set of training patterns with a total number N, and k is the kth training pattern, $\overline{\mathbf{w}}$ is a vector of optimization weights, $\overline{\propto}_k$ is the Lagrange multiplier condition, $\overline{b}$ is the optimization bias, y_k indicates whether the training data belong to a categorization of +1 or –1, and exp() is the exponential function.

The procedure for validating the unilevel SVM models was identical to that of the unilevel DA models, except that decision function (9) was used. For example, decision function (9) correctly predicted the fourth-grade text N-098-C-042-08 to be fourth grade.

The identical procedure was applied to the other four folds, and the accuracy of the SVM model at the word level was calculated as the average of the five results. We used the identical procedure to establish and validate the other three unilevel SVM models.

Constructing and validating multilevel DA models

Multilevel analysis is based on levels of features, instead of individual features. To find the optimal combination of features in readability that enhances the efficacy (validity) and efficiency (latency for computing) of models, we constructed the multilevel models with the *F*-score to identify the linguistic features representative of their corresponding levels.

The *F*-score is a commonly used algorithm for selecting relevant features and enhancing the accuracy of the SVM model (Chang & Lin, 2008; Ding, 2009). A higher *F*-score implies a better discrimination ability for classification. As was mentioned above, the present study was concerned not only with classification issues, but also with the multilevel reading comprehension process. Therefore, in addition to functionality and sensitivity, the ability to reflect the multilevel reading process was also taken into consideration for feature selection. Since there are no clear standards for determining the appropriate *F*-score threshold, we adopted the algorithms suggested in Chen and Lin (2006) to calculate a possible threshold. Hence, we calculated the average validation error with cross-validation and then chose the threshold with the lowest average validation error. This approach needed manual fine-tuning because the selected threshold should cover as many features with higher *F*-scores in each unilevel as possible. On the basis of this approach, we set the threshold value at 0.5 and selected features X_1, X_2, X_3, X_5, X_6, X_7, X_8, and X_9 for the word level, features X_{10}, X_{12}, and X_{13} for the semantics level, features X_{16} and X_{20} for the syntax level, and features X_{25} and X_{26} for the cohesion level. The *F*-score of the 31 features were computed according to formula (A1) and are listed in the Appendix Table 6.

The procedure for constructing the multilevel DA models was identical to that used for the unilevel DA models. Consider the multilevel DA models of Fold 1 as an example. We first computed the scores of the linguistic features from the texts in subsets A, B, C, and D of Fold 1 and then constructed the discriminant function for each grade level as shown in discriminant functions (10)–(15) below, in which D_1, D_2, D_3, D_4, D_5, and D_6 represent the functions for grades 1–6, respectively:

$$\begin{aligned} D_1 = &-133.102 + 0.226x_1 + 0.327x_2 + 0.239x_3 - 0.128x_5 - 0.030x_6 \\ &- 0.410x_7 + 127.319x_8 + 1.562x_9 - 0.288x_{10} - 0.119x_{12} \\ &+ 0.639x_{13} + 8.861x_{16} - 0.396x_{20} - 0.315x_{25} - 0.977x_{26} \end{aligned} \tag{10}$$

$$\begin{aligned} D_2 = &-115.727 + 0.220x_1 + 0.311x_2 + 0.205x_3 - 0.144x_5 + 0.022x_6 \\ &- 0.361x_7 + 118.869x_8 + 1.650x_9 - 0.240x_{10} - 0.120x_{12} \\ &+ 0.435x_{13} + 1.790x_{16} - 0.420x_{20} - 0.258x_{25} - 0.989x_{26} \end{aligned} \tag{11}$$

$$\begin{aligned} D_3 = &-106.867 + 0.223x_1 + 0.300x_2 + 0.198x_3 - 0.129x_5 + 0.050x_6 \\ &- 0.316x_7 + 112.294x_8 + 1.701x_9 - 0.244x_{10} - 0.109x_{12} \\ &+ 0.297x_{13} - 1.882x_{16} - 0.484x_{20} - 0.248x_{25} - 0.981x_{26} \end{aligned} \tag{12}$$

$$\begin{aligned} D_4 = &-102.917 + 0.248x_1 + 0.320x_2 + 0.209x_3 - 0.157x_5 + 0.075x_6 \\ &- 0.313x_7 + 107.907x_8 + 1.853x_9 - 0.237x_{10} - 0.179x_{12} \\ &+ 0.318x_{13} - 4.542x_{16} - 0.613x_{20} - 0.316x_{25} - 0.845x_{26} \end{aligned} \tag{13}$$

$$\begin{aligned} D_5 = &-112.186 + 0.267x_1 + 0.346x_2 + 0.185x_3 - 0.124x_5 + 0.070x_6 \\ &- 0.292x_7 + 106.219x_8 + 2.213x_9 - 0.286x_{10} - 0.0005x_{12} \\ &- 0.196x_{13} - 4.150x_{16} - 0.624x_{20} - 0.288x_{25} - 0.773x_{26} \end{aligned} \tag{14}$$

$$\begin{aligned} D_6 = &-114.649 + 0.282x_1 + 0.366x_2 + 0.203x_3 - 0.154x_5 + 0.087x_6 \\ &- 0.284x_7 + 107.193x_8 + 2.143x_9 - 0.301x_{10} - 0.014x_{12} \\ &- 0.211x_{13} - 4.103x_{16} - 0.457x_{20} - 0.307x_{25} - 0.871x_{26}. \end{aligned} \tag{15}$$

The Fold 1 multilevel DA models were validated by computing the values of linguistic features in subset E and then entering the results of each text into discriminant functions (10)–(15). For example, the six scores for the sixth-grade text K-098-C-062-05 were D_1 = 82.17, D_2 = 95.18, D_3 = 103.5, D_4 = 109.7, D_5 = 114.1, and D_6 = 113.7. Since D_5 had the highest score, text K-098-C-062-05 was incorrectly predicted to be fifth grade.

Constructing and validating multilevel SVM models

The procedure for constructing the multilevel SVM models was identical to that used for constructing the multilevel DA models. Decision function (16) was then established for the multilevel SVM model based on Fold 1. The parameters of the multilevel SVM model were gamma = 0.1 and penalty parameter = 1.5. The Fold 1 multilevel SVM model was validated by computing the values of the linguistic features of testing subset E in (16) and predicting the grade level of each text. For example, decision function (16) correctly predicted the sixth-grade text K-098-C-062-05 to be sixth grade:

$$f_2(\mathbf{x}) = \operatorname{sgn}\left(\overline{\mathbf{w}}^T\mathbf{x} + \overline{b}\right) = \operatorname{sgn}\left(\sum\nolimits_{k=1}^{N} \overline{\propto}_k y_k \exp\left(-0.1 \times \left\|\mathbf{x}_k^T - \mathbf{x}\right\|^2\right) + \overline{b}\right). \tag{16}$$

Results

Prediction accuracies of the unilevel and multilevel DA and SVM models

Table 3 lists the average prediction accuracies of the fivefold cross-validation based on the unilevel and multilevel DA and SVM models. The prediction accuracies of the unilevel DA models were between 44.73 % and 58.79 %, whereas those of the unilevel SVM models were between 43.97 % and 65.13 %. The prediction accuracies for both types of unilevel model were best at the word level, followed by the semantics, syntax, and cohesion levels. The prediction accuracies of the multilevel DA and SVM models were 57.76 % and 71.75 %, respectively. The error matrix and prediction accuracies of multilevel SVM and DA models are shown in Table 4.

Planned comparisons were performed according to Glass and Hopkins (1984), with one-tailed *t*-tests adopted to investigate the differences in the average accuracies of the DA and SVM models between levels and the differences in the average accuracies at each level between the two types of model. Among the DA models, the average prediction accuracy was significantly higher for the multilevel model than at the syntax, $t(8) = 2.83$, $p < .05$, and cohesion, $t(8) = 4.40$, $p < .01$, unilevels, but the multilevel accuracy did not differ significantly from the accuracies at the word, $t(8) = -0.57$, $p > .05$, and semantics, $t(8) = 0.48$, $p > .05$, unilevels. The average prediction accuracy was significantly higher at the word unilevel than at the syntax, $t(8) = 2.74$, $p < .05$, and cohesion, $t(8) = 4.27$, $p < .05$, unilevels, but it did not differ significantly between the word and semantics unilevels, $t(8) = 0.78$, $p > .05$. The average prediction accuracies were significantly higher at the semantics, $t(8) = 3.24$, $p < .05$, and syntax, $t(8) = 2.04$, $p < .05$, unilevels than at the cohesion unilevel.

Table 3 Prediction accuracies (%) of the unilevel and multilevel DA and SVM models

Level	Fold					Average
	1	2	3	4	5	
Word						
DA	61.33	57.33	53.33	62.67	59.30	58.79
SVM	64.00	65.33	62.67	72.00	61.63	65.13
Semantics						
DA	53.33	58.67	49.33	60.00	61.63	56.59
SVM	54.67	64.00	50.67	58.67	61.63	57.93
Syntax						
DA	58.67	50.67	50.67	46.67	52.33	51.80
SVM	56.00	52.00	54.67	53.33	51.16	53.43
Cohesion						
DA	44.00	34.67	45.33	52.00	47.67	44.73
SVM	45.33	37.33	38.67	52.00	46.51	43.97
Multilevel						
DA	56.00	60.00	56.00	58.67	58.14	57.76
SVM	70.67	77.33	65.33	73.33	72.09	71.75

Among the SVM models, the prediction accuracy was significantly higher for the multilevel model than at the word, $t(8) = 2.14$, $p < .05$, semantics, $t(8) = 4.48$, $p < .01$), syntax, $t(8) = 5.57$, $p < .001$, and cohesion, $t(8) = 8.34$, $p < .001$, unilevels. The prediction accuracy was significantly higher at the word unilevel than at the semantics, $t(8) = 2.39$, $p < .05$, syntax, $t(8) = 6.17$, $p < .001$, and cohesion, $t(8) = 6.76$, $p < .001$, unilevels and significantly higher at the semantics, $t(8) = 3.88$, $p < .01$, and syntax, $t(8) = 3.34$, $p < .05$, unilevels than at the cohesion unilevel.

Regarding the differences in the prediction accuracies between SVM and DA models, the results of one-tailed *t*-tests showed that the prediction accuracy was significantly

Table 4 The error matrix and predication accuracies (%) of the models of multilevel SVM and DA

		Predicted Grade Level (SVM/DA)						Accuracy (SVM/DA)
		1	2	3	4	5	6	
Actual Grade Level	1	45/39	2/8	0/0	0/0	0/0	0/0	95.74/82.98
	2	6/11	61/48	4/12	0/0	0/0	0/0	85.92/67.61
	3	0/0	5/12	42/33	15/17	0/0	0/0	67.74/53.23
	4	0/0	3/0	10/15	50/44	8/8	0/4	70.42/61.97
	5	0/0	0/0	0/1	11/10	37/29	21/29	53.62/42.03
	6	0/0	1/1	2/2	3/4	18/29	42/30	63.64/45.45

higher for the SVM models than for DA models at the multilevel, $t(8) = 6.67$, $p < .001$, and at the word unilevel, $t(8) = 2.93$, $p < .01$, but did not differ between the SVM and DA models at the semantics, $t(8) = 0.40$, $p > .05$, syntax, $t(8) = 0.76$, $p > .05$, and cohesion, $t(8) = -0.20$, $p > .05$, unilevels.

Prediction errors of DA and SVM models

Table 5 lists the prediction errors of the DA and SVM models. There were 277 texts correctly classified by the multilevel SVM models, of which 199 were correctly classified by multilevel DA models (71.84 %). There were 223 texts correctly classified by the unilevel DA models, of which 199 were correctly classified by unilevel SVM models (89.24 %). SVM models correctly classified most of the texts that DA models also correctly classified. Meanwhile, SVM models correctly classified 78 of the 163 texts (47.85 %) that were misclassified by DA, while DA models correctly classified 24 of the 109 texts (22.02 %) that were misclassified by the SVM. In other words, SVM models were more capable of classifying the texts that DA had misclassified. A similar pattern was revealed by comparisons of the SVM and DA models at the word level. These results indicate that the accuracy in text classification was higher for readability models constructed using the SVM than for readability models constructed using DA.

Table 5 Comparison of the prediction errors of the DA and SVM models

Level	DA Correct		Total	DA Incorrect		Total
	SVM Correct	SVM Incorrect		SVM Correct	SVM Incorrect	
Multilevel	199	24	223	78	85	163
Word	191	36	227	60	99	159
Semantics	188	31	219	36	131	167
Syntax	178	22	200	28	158	186
Cohesion	137	36	173	33	180	213

Discussion

Multilevel-features-based readability evaluation

Previous readability formulae that have included few linguistic features, such as average sentence length, word counts, and syllable counts, may have advantages of parsimony and intuitiveness. However, such an approach fails to represent text complexity and results in low-accuracy readability predictions. This situation prompted researchers to construct readability models incorporating multiple levels of features (Crossley, Allen, & McNamara, 2011; Heilman et al, 2008).

The present study adopted the approach of integrating multilevel linguistic features to construct readability models. Four levels of linguistic features were included: word, semantics, syntax, and cohesion. We found that multilevel-features-based readability models performed significantly better than their counterparts based on unilevel features, with this advantage being more obvious when the models were constructed using the SVM. Our findings support the claim that multilevel-features-based readability models are more effective for predicting texts at different grade levels.

Our approach is consistent with some of the previous research reported in the literature. For example, Heilman et al. (2008) found that integrating word and syntactic features in statistical language models significantly enhanced the accuracy in predicting text grade levels. François and Fairon (2012), on the other hand, proposed a formula, called AI readability, for French as a foreign language. Their formula was based on SVM and integrated 46 textual features of four levels. The AI readability achieved an average adjacent accuracy rate of 79 % for six levels of L2 texts. Furthermore, Crossley et al. (2011) found that Coh-Metrix L2 reading indices containing features related to word frequency, syntax, and cohesion had higher validity for predicting text complexity (at three levels) for learners of English as a foreign language. The Coh-Metrix L2 reading indices achieved an accuracy rate of 60 %, which is 19 % higher than that of the Flesch–Kincaid formula. Our study extended the levels of features and the text grade levels (to levels) and found that the SVM readability models with multilevel linguistic features outperformed other models in terms of exact prediction accuracy by 6.63 % to 27.78 %. The results of our study provide further evidence supporting the claim that the multilevel approach in readability analysis better represents the complexities of texts and the comprehension process (Graesser & McNamara, 2011; Graesser et al., 2004; McNamara, Louwerse, McCarthy, & Graesser, 2010).

In addition to the finding that multilevel-features-based readability had better performance for classifying texts, there were several noteworthy findings about the relationships between unilevel features and text readability. First, word-level features were the most influential in readability models (Liu et al., 2007), since readability models containing word-level features outperformed other unilevel-features-based models in classifying texts. Given that beginning readers have not yet mastered word decoding, curriculum experts tend to emphasize features at the word level in low-grade-level texts. The results of *F*-scores selection also showed that the discrimination ability among grade levels was better for most of the features at the word level than at the other three linguistic levels. Our finding that word-level features played a more significant role in text readability is consistent with other research of French (François & Fairon, 2012). This finding also concurred with previous research that found that word-level features and related knowledge substantially impact comprehension (Graesser & McNamara, 2011; Perfetti, 2007; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001). This role is also consistent with the viewpoint that word recognition is the first step to successful reading comprehension (Noble, Rowland, & Pine, 2011; Verhoeven & Perfetti, 2008; Yang, Wang, Chen, & Rayner, 2009).

Second, reading comprehension involves word decoding, and therefore, the characteristics of the writing system used in a particular language should be taken into account. We designed a few exclusive features based on our experiments involving Chinese texts. The characters used in the Chinese logographic writing system are distinct from alphabetic systems and can be divided into several components. Some of these components provide semantic information, while others provide phonetic information. The character strokes also vary. Phonomorphologically, each character corresponds to a syllable and a morpheme, rather than to a phoneme. With respect to morphology, Chinese has few inflections or derivations but a large number of compound words, such as two-character words (e.g., 太陽, meaning “the sun”). These characteristics mean that the processes involved in learning to read Chinese differ from those used when learning to read alphabetic languages (e.g., Anderson & Chen, 2012; McBride-Chang et al., 2005). Interestingly, similar to the important role of word length for readability models of alphabetic languages (e.g., English), we found that character complexity and word length also play important roles in the readability of Chinese text. We conclude that although the present study took Chinese as an example, the proposed method and procedure employed could be applied when constructing and validating readability models for other writing systems.

The third important finding from our study is related to the role of cohesion features. The cohesion of text is an important factor affecting the mental representation of readers and can be established through reference and inference. Nonetheless, the prediction accuracies of the DA and SVM models at the cohesion level were not as expected. This was probably due to cohesion within text being established through using both reference and ellipsis. The number of cohesive ties (e.g., pronouns) cannot therefore adequately represent the cohesion relationship in a text. Furthermore, reference in Chinese can still be achieved by using a “zero pronoun”—that is, if there is no obvious structural word in the sentence. Therefore, calculating the occurrence of a certain type of cohesive ties (e.g., conjunctions or pronouns) may underestimate the influence of cohesion. Previous studies have suggested that texts with high cohesion facilitate the coherence of mental representations (van den Broek & Kremer, 2000), and future studies should investigate the influence of cohesion from the perspective of readers.

Constructing readability models using machine learning methods

In addition to the importance of multilevel features themselves, this study found that methods for discriminating and integrating those features also played a critical role when constructing valid readability models. Among the four types of models constructed, it was found that combining multilevel features with SVM models provided the highest accuracy in predicting text grades. Furthermore, multilevel-features-based SVM models outperformed unilevel-features-based SVM models. In contrast, although the accuracies were significantly higher for the multilevel DA models than for the unilevel DA models at the semantics and syntax levels, no significant differences were found between multilevels and the word unilevel. Furthermore, SVM models could accurately predict 90 % of the texts that DA accurately predicted, while DA models could accurately predict only 70 % of the texts that SVM accurately predicted (see Table 4). For the texts that were misclassified by DA, the correct prediction percentage for the SVM model was 50 %, whereas the DA models achieved only 20 % accuracy for texts misclassified by the SVM model.

The superior performance of the SVM over DA can be explained by the influence of outliers being mitigated in SVM model training. In contrast, DA constructs discriminant functions based on all of the data points, and these are potentially misled by the extreme (i.e., less deterministic) values. This characteristic makes the SVM model especially preferable when the training data have linearly nonseparable properties. Both texts N-098-C-042-08 and K-098-C-062-05 (described for the word-level and multilevel DA models, respectively; see the Method section) illustrate how outliers may put the DA model at a disadvantage. In contrast, the SVM models were able to categorize both of these texts correctly. All of the results demonstrate that the SVM is better able to cope with linearly nonseparable data and is less affected by outliers, which give it a better generalization ability than DA.

Our results are consistent with previous comparisons of the validity of machine learning models and traditional readability models by Feng et al. (2010) and Petersen and Ostendorf (2009). Those studies could have been influenced by confounding factors, since the readability indices they compared used different sets of features. In contrast, the SVM and DA models that we compared consisted of the same set of linguistic features. By comparing the performance and validity of models constructed by DA and SVM with unilevel and multilevel features, our study has provided more reliable evidence of the advantages of combining SVM models with multilevel features for readability prediction.

Our study found that SVM-based readability models are more effective for predicting text grades than are DA-based models, but this does not mean that DA-based models are not useful. Similar to other GLMs, such as linear regression, DA-based models can use equations to represent the relationships among text grades and different linguistic features. This makes is easy for readers to identify relationships between the criterion variable (e.g., the grades) and the predicting variables (e.g., the features); moreover, the weight of each feature provides straightforward information about the relative importance of the features in the model. A disadvantage of SVM-based models is that they do not provide the above information, which makes it difficult for readers to interpret the relationships among the included features. In contrast to the vast majority of readability studies using English as the research material, the present study used Chinese as an example. Because different writing systems may share similar processes in reading comprehension (Graesser, Singer, & Trabasso, 1994) and most of the linguistic features used in the present study have similar functions in alphabetic languages, we suggest that the approach of constructing readability models by combining multilevel linguistic features and the SVM model may also be suitable for nonlogographic languages. The validation of this implication is especially worthy of further study. Since features in different languages may not play the same roles in determing the complexity or readability of texts, how to optimize ways of assembling features with SVM to maximize the efficacy of readability models would be of theorectical and practical value.

In addition to the methodological implications above, our approach of using grade levels to be the criteria for leveling texts through readability models also have practical implications. From the perspective of adaptive reading, using the readability levels revealed through the readability assessment, teachers will be able to choose texts appropriate for their students' grade levels. The publishers may have a more objective reference framework for writing/compiling their language arts textbooks and avoiding the possibility of mismatching texts to grade levels higher or lower than their corresponding target learners. Future research could consider integrating more features relevant to readability, such as content and domain-specific knowledge, with LSA and reading developmental stages. Prior research has shown that LSA-based features, such as cohesion and semantic spaces, are capable of differentiating texts containing different domain knowledge. These functions will be especially helpful for enhancing the performance of leveling domain-related texts in different ages or grades. In addition, the importance of particular textual features to reading comprehension is likely to change as readers develop their reading abilities and skills. Therefore, it is essential to establish readability models with ensembles of different features to make them more appropriate for different reading developmental stages. Such a method would also help optimize the functionality of readability models.

Acknowledgement The authors appreciate the funding supports from the Aim for the Top University Project of Ministry of Education, Taiwan and the National Science Council, Taiwan (102-2911-I-003 -301; 102-2511-S-003 -001 -MY3; 101-2511-S-003 -058 -MY3; 101-2511-S-003 -047 -MY3; 99-2631-S-003-003).

Appendix

Table 6 Linguistic features and *F*-scores

Level	Aspect	Feature		Definition	*F*-score
Word	Character complexity	X_1	Few-stroke characters	Total number of characters containing 1 to 10 strokes	3.762
		X_2	Moderate-stroke characters	Total number of characters containing 11 to 20 strokes	3.910
	Word length	X_3	Two-character words	Total number of two-character words	4.135
		X_4	Three-character words	Total number of three-character words	0.471
	Lexical count	X_5	Words	Total number of words	3.575
		X_6	Verbs	Total number of verbs	2.855
		X_7	Adverbs	Total number of adverbs	3.065
	Frequency	X_8	Logarithm of the content-word frequency	Logarithm of the average content-word frequency	0.937
		X_9	Difficult words	Number of words not in the list of frequently used words	1.221
Semantics	Semantics complexity	X_{10}	Content words	Total number of content words	3.343
		X_{11}	Negations	Total number of negation words	0.240
		X_{12}	Sentences with complex semantic categories	Sentences with complex semantic categories	2.131
		X_{13}	Complex semantic categories	Average number of semantic categories of sentences	1.165
		X_{14}	Intentional words	Total number of words denoting intentional actions, events, and animate agents	0.159
Syntax	Syntax complexity	X_{15}	Average sentence length	Average sentence length	0.411
		X_{16}	Simple-sentence ratio	Proportion of simple sentences	0.583
		X_{17}	Modifiers per noun phrase	Number of adjectives, and adverbs before noun phrases	0.132
		X_{18}	Prepositional phrase ratio	Proportion of prepositional phrases to total sentence number	0.047
		X_{19}	Parallelism	Number of parallel construction sentences	0.088
		X_{20}	Sentences with complex structure	Number of sentences with complex structure	2.571
Cohesion	Referential words	X_{21}	Pronouns	Total number of pronouns	0.437
		X_{22}	Personal pronouns	Total number of personal pronouns	0.356
		X_{23}	First-person pronouns	Total number of first-person pronouns	0.166
		X_{24}	Third-person pronouns	Total number of third-person pronouns	0.279
	Conjunctions	X_{25}	Positive conjunctions	Number of conjunctions with positive meanings	1.475
		X_{26}	Negative conjunctions	Number of conjunctions with negative meanings	0.583
		X_{27}	Causal conjunctions	Number of conjunctions with causality	0.318
		X_{28}	Purpose conjunctions	Number of conjunctions with purpose or avoidance meanings	0.094
		X_{29}	Concession conjunctions	Number of conjunctions with the intention of concession	0.339
		X_{30}	Conditional conjunctions	Number of conjunctions with conditional relations	0.090
	Metaphor	X_{31}	Figures of speech, metaphors	Number of metaphorical expressions	0.094

References

Ahrens, K. (2006). The effect of visual target presentation times on lexical ambiguity resolution. *Language and Linguistics, 7*(3), 677–696.

Anderson, R. C., & Chen, X. (2012). Chinese reading development in monolingual and bilingual learners: Introduction to the special issue. *Scientific Studies of Reading, 17*(1), 1–4.

Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language & Communication, 21*(3), 285–301.

Bazi, Y., & Melgani, F. (2006). Toward an optimal SVM classification system for hyperspectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing, 44*(11), 3374–3385.

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York, NY: Guilford Press.

Benjamin, R. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24*(1), 63–88.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In D. Haussler (Ed.), *5th Annual ACM Workshop on Computational learning theory* (pp. 144–152). Pittsburgh, PA: ACM Press.

Bruce, B., Rubin, A., & Starr, K. S. (1981). Why readability formulas fail. *IEEE Transactions on Professional Communication, PC-24,* 50–52.

Cardoso-Cachopo, A., & Oliveira, A. L. (2003). An empirical comparison of text categorization methods. In M. A. Nascimento, E. S. de Moura, & A. L. Oliveira (Eds.), *String processing and information retrieval* (pp. 183–196). Heidelberg: Springer.

Carpenter, P. A., & Just, M. A. (1983). What your eyes do while your mind is reading. In K. Rayner (Ed.), *Eye movements in reading: Perceptual and language processes* (pp. 275–307). New York, NY: Academic Press.

Chae, J., & Nenkova, A. (2009). *Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text.* Paper presented at the Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece.

Chang, Y. W., & Lin, C. J. (2008). Feature ranking using linear svm. Causation and Predicition Challenge in Machine Learning, Volume 2, 47.

Chang, C. C. & Lin, C. J. (2011). LIBSVM [a library for support vector machines]. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/libsv

Chang, T. H., Sung, Y. T., & Lee, Y. T. (2012, November). *A Chinese word segmentation and POS tagging system for readability research*. Paper presented at the 42nd Annual Meeting of the Society for Computers in Psychology, Minneapolis, MN.

Chen, Y. W., & Lin, C. J. (2006). Combining SVMs with various feature selection strategies. *Studies in Fuzziness and Soft Computing, 207,* 315–324.

Chen, J. L., & Su, Y. F. (2010). The effect of decoding ability on character complexity effect and word length effect in Taiwanese beginning reader. *Bulletin of Educational Psychology, 41*(3), 579–604.

Chen, S., Zhou, S., Yin, F. F., Marks, L. B., & Das, S. K. (2007). Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis. *International Journal of Medical Physics Research and Practice, 34,* 3808–3814.

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language, 23*(1), 84–101.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42*(3), 475–493.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin, 27,* 37–54.

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English, 26,* 23.

Ding, S. (2009). Feature selection based F-score and ACO algorithm in support vector machine. *Second International Symposium on Knowledge Acquisition and Modeling, 1,* 19–23.

DuBay, W. H. (2007). *Smart language: Reader, readability, and the grading of text*. Costa Mesa, CA: Impact Information.

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification* (2nd ed.). NewYork: Wiley.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve information retrieval. Proceedings of *the SIGCHI Conference on Human Factors in Computing Systems*, 281–285. Washington, D.C.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment. Proceedings of *the 23rd International Conference on Computational Linguistics*, 276-284.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–223.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes, 25,* 285–307.

François, T., & Fairon, C. (2012). An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 466-477). Jeju Island, Korea: Association for Computational Linguistics.

Gershkoff-Stowe, L., & Hahn, E. R. (2007). Fast mapping skills in the developing lexicon. *Journal of Speech, Language, and Hearing Research, 50,* 682–697.

Givón, T. (1979). *On understanding grammar*. New York, NY: Academic Press.

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psycholgoy*. Englewood, NJ: Prentice-Hall.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multi-level discourse comprehension. *Topics in Cognitive Science, 3,* 371–398.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multi-level analyses of text characteristics. *Educational Researcher, 40*(5), 223–234.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193–202.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Science, 17*(8), 684–691.

Grainger, J., & Frenck-Mestre, C. (1998). Masked priming by translation equivalents in proficient bilinguals. *Language and Cognitive Processes, 13*(6), 601–623.

Heilman, M., Thompson, K. C., & Eskenazi, M. (2008). *An analysis of statistical models and features for reading difficulty prediction*. Paper presented at the Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, Columbus, OH.

Huang, S. (2000). The story of heads and tails— on a sequentially sensitive lexicon. *Language and Linguistics, 1*(2), 79–107.

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). A improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications, 39*(1), 1503–1509.

Johansson, V. (2008). *Lexical diversity and lexical density in speech and writing*. Lund, Sweden: Lund University.

Jordan, M. P. (1998). The power of negation in English: Text, context and relevance. *Journal of Pragmatics, 29,* 705–752.

Keerthi, S. S., & Lin, C. J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation, 15*(7), 1667–1689.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Klare, G. R. (1985). Matching reading materials to readers: The role of readability estimates in conjunction with other information about comprehensibility. In T. L. Harris & E. J. Cooper (Eds.), *Reading, Thinking, and concept development*. New York: College Entrance Examination Board.

Klare, G. R. (2000). The measurement of readability: Useful information for communicators. *ACM Journal of Computer Documentation, 24*(3), 107–121.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. London, UK: University of Chicago Press.

Lee, Y. S., Tseng, H. C., Chen, J. L., Peng, C. Y., Chang, T. H., & Sung, Y. T. (2012, July). *Constructing a novel Chinese readability classification model using principal component analysis and genetic programming*. Paper presented at the 12th International Conference on Advanced Learning Technologies (ICALT), Rome, Italy.

Lemaire, B., Denhiere, G., Bellissens, C., & Jhean-Iarose, S. (2006). A computational model for simulating text comprehension. *Behavior Research Methods, 38*(4), 628–637.

Lin, H. T., & Lin, C. J. (2003). *A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods*. Technical report, Department of Computer Science and Information Engineering, National Taiwan University.

Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behavior Research Methods, 39,* 192–198.

Louwerse, M. M. (2004). Semantic variation in idiolect and sociolect: Corpus linguistic evidence from literary texts. *Computers and the Humanities, 38,* 207–221.

Louwerse, M. M., & Mitchell, H. H. (2003). Toward a taxonomy of a set of discourse markers in dialog: A theoretical and computational linguistic account. *Discourse Processes, 23,* 441–470.

McBride-Chang, C., Cho, J-R., Liu, H., Wagner, R. K., Shu, H., Zhou, A., … Muse, A. (2005). Changing models across cultures: Associations of phonological awareness and morphological structure awareness with vocabulary and word recognition in second graders from Beijing, Hong Kong, Korea, and the United States. *Journal of Experimental Child Psychology, 92*, 140-160.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47,* 292–330.

Noble, C. H., Rowland, C. F., & Pine, J. M. (2011). Comprehension of argument structure and semantic roles: Evidence from English–learning children and the forced–choice pointing paradigm. *Cognitive Science, 35,* 963–982.

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11,* 357–383.

Petersen, S. E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech and Language, 23,* 89–106.

Ravid, D., & Berman, R. A. (2010). Developing noun phrase complexity at school age: A text-embedded cross-linguistic analysis. *First Language, 30,* 3–26.

Rayner, K., Foorman, B., Perfetti, C., Pesetsky, D., & Seidenberg, M. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2*(2), 31–74.

Samuels, S. J. (2006). Toward a model of reading fluency. In S. J. Samuels & E. E. Farstrup (Eds.), *What research says about reading instruction* (pp. 24–46). Newark, DE: International Reading Association.

Sanders, T. J. M., Spooren, W. P. M., & Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes, 15,* 1–35.

Schriver, K. (2000). Readability formula in the new millennium: What's the use? *ACM Journal of Computer Documentation, 24*(3), 138–140.

Shapiro, A. M., & McNamara, D. S. (2000). The use of latent semantic analysis as a tool for the quantitative assessment of understanding and knowledge. *Journal of Educational Computing Research, 22,* 1–36.

Singer, M., Harkness, D., & Stewart, S. T. (1997). Constructing inferences in expository text comprehension. *Discourse Processes, 24,* 199–228.

Sung, Y. T., Chen, J. L., Lee, Y. S., Cha, J. H., Tseng, H. C., Lin, W. C., & Chang, K. E. (2013). Investigating chinese text readability: Linguistic features, modeling, validation. *Chinese Journal of Psychology, 55*(1), 75–106.

Sung, Y. T., Chen, J. L., Lee, Y. T., Lee, Y. S., Peng, C. Y., Tseng, H. C., & Chang, T. H. (2012, July). *Constructing and Validating a Readability Modal with LSA : A case study of Chinese and social science textbooks*. Paper presented at the 22th Annual Meeting of Society for Text and Discourse Process, Montreal, Canada.

Tan, L. H., & Peng, D. L. (1990). The effects of semantic context on the feature analyses of single Chinese characters. *Journal of Psychology, 4,* 5–10.

van den Broek, P., & Kremer, K. E. (2000). The mind in action: What it means to comprehend during reading. In B. M. Taylor, M. F. Graves, & P. v. d. Broek (Eds.), *Reading for meaning: Fostering comprehension in the middle grades* (pp. 1–31). Newark, DE: International Reading Association.

van den Broek, P., Risden, K., Fletcher, C. R., & Thurlow, R. (1996). A "landscape" view of reading: Fluctuating patterns of activation and the construction of a stable memory representation. In B. K. Britton & A. C. Graesser (Eds.), *Models of understanding text* (pp. 165–187). Mahwah, NJ: Erlbaum.

Vapnik, V. N., & Chervonenkis, A. Y. (1974). *Teoriya raspoznavaniya obrazov: Statisticheskie problemy obucheniya* [Theory of pattern recognition: Statistical problems of learning]. Moscow, Russia: Nauka.

Verhoeven, L., & Perfetti, C. (2008). Advances in text comprehension: Model, process and development. *Applied Cognitive Psychology, 22,* 293–301.

Yang, J., Wang, S., Chen, H., & Rayner, K. (2009). The time course of semantic and syntactic processing in Chinese sentence comprehension: Evidence from eye movements. *Memory & Cognition, 37*(8), 1164–1176.