Linguistic Feature Analysis of CEFR Labeling Reliability and Validity in Language Textbooks
(以語言特徵為本的教材分級難度及有效性之驗證)

Hong, Jia-Fei
(洪嘉馡)
National Taiwan Normal University
(國立臺灣師範大學)
jiafeihong@ntnu.edu.tw

Peng, Chun-Yi[1]
(彭駿逸)
Borough of Manhattan Community College, CUNY
(曼哈頓社區學院)
cpeng@bmcc.cuny.edu

Tseng, Hou-Chiang
(曾厚強)
National Taiwan Normal University
(國立臺灣師範大學)
ouartz99@ntnu.edu.tw

Sung, Yao-Ting
(宋曜廷)
National Taiwan Normal University
(國立臺灣師範大學)
sungtc@ntnu.edu.tw

Abstract: Despite the importance of grading language textbooks for teaching and learning, few studies have addressed the issues of reliability, validity, and efficiency of grading texts. This study adopted an automated textbook grading system to examine the grading consistency of five L2 Chinese textbook series labeled with CEFR difficulty levels. Twelve linguistic features were selected to represent the most crucial aspects of text readability: lexicon, semantics, syntax, and cohesion. Both the validity and reliability of grading assignments were tested between and within textbook series. The results suggested that 4 out of the selected 5 textbook series did not assign grading levels accurately reflective of actual text difficulty.

摘要：語言教科書的分級不管對於教學或是學習都是非常重要的一環，但是卻很少討論文本分級的可靠性、有效性的研究。本研究以自動分析教科書等級系統，檢測以 CEFR 作為標示難度的五套華語文教科書的等級一致性。在本研究中，選取了 12 種不同的語言特徵作為最具分級影響力的關鍵指標，分別取自詞彙類、語義類、語法類及篇章凝聚類等四大語言層面。本研究主要探究在不同教科書之間的分級一致性與相同教科書不同等級的分級，其有效性和可靠性。研究結果顯示，在本研究所選定的五套華語文教科書當中，四套華語文教科書沒有依照實際文本難度進行等級分級。

Keywords: Text Readability, reliability, validity, linguistic features, CEFR, language textbooks

[1] Peng, Chun-Yi is the corresponding author.

關鍵詞: 文本可讀性、信度、效度、語言特徵、歐洲共同語文參考標準、語言教科書

1. Introduction

The applications of machine learning have become increasingly important across various disciplines, such as health care (Caruana et al., 2015), education (Chang & Sung, 2019; Hsu et al., 2018; Lin et al., 2019; Lu & Chen, 2019; Lee et al., 2016), and speech recognition (Chen & Hsu, 2019). A crucial application of machine learning in education is the assigning of grade levels to textbooks for adaptive learning (Tseng et al., 2019). With correctly graded materials, educators can better select or even edit existing resources to cater to learners' changing proficiency levels. For learners, the use of appropriately graded materials is also important. It assists them in identifying their proficiency levels, allows them to check their progress, and enhances their learning efficiency. Thus, a standardized text grading system is beneficial for both educators and learners.

L2 Chinese textbooks prove to be a useful example of the necessity for a standardized text grading system. Although many L2 Chinese textbook materials are graded in terms of standards such as the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR; Council of Europe, 2001), the assignment of grade levels is mostly, if not all, carried out by developers based on their own expertise and professional experiences. The variation in expertise and professional experiences runs the risk of inconsistency when standards such as CEFR are applied to language materials grading. That is, the same materials may be assigned by different developers to different difficulty levels within the same set of standards. Such inconsistency creates potential problems when those materials are adopted for teaching and learning. In order to ensure the accuracy and consistency in textbook grading, many highly experienced language educators must be involved in the compilation and grading process. This is often time-consuming and labor-intensive. Furthermore, it can prove to be rather difficult to reach a consensus among educators on a consistent grading scheme.

To this end, this study introduces the use of a standardized textbook grading protocol proposed by Sung et al. (2015b) as a tool for CFL textbook grading. More specifically, we use the CRIE-CFL system, a tool based on Sung et al.'s (2015b) grading protocol, to analyze 5 textbook series that have been graded manually by their developers. Sung et al.'s (2015b) model and the CRIE-CFL system have been shown to be a valid tool in language materials grading. By comparing readings from the tool and the grading levels assigned manually by their developers both within and across those five textbook series, we hope to illustrate the usefulness of such a tool in measuring the accuracy and maintaining consistency of manual gradings to the actual difficulty of language learning materials included in the textbooks.

2. Overview of the CEFR

The CEFR was created by the Council of Europe in 2001 with the aim of providing a unified framework for the teaching, learning, and assessment of all of the languages used within Europe (Fulcher, 2004). The principles of the CEFR framework implies that the 'can-do' statements are unitarily understandable and can be interpreted in only one way which will be the same for everyone in every European country (Vinther, 2013). It provides a set of guidelines for language teaching materials and language evaluation, as well as a point of reference for grading learner levels in order to reduce the barriers of interaction between people speaking different languages within different European countries (Council of Europe, 2001; Little, 2006, 2007). The CEFR has had a profound influence on the design of teaching materials, curriculum planning, and language proficiency testing in several European countries (Hulstijn, 2007). Its role in Europe has evolved from a supportive education tool to a tool used to shape language education policies (Bonnet, 2007; Fulcher, 2007).

The CEFR is a detailed and complex system for evaluating language proficiency levels. It uses "horizontal" and "vertical" dimensions to describe a particular learner's ability to communicate. The horizontal dimension provides a general description of communicative language competency; it consists of several scales that describe various language activities that a learner may encounter, such as context, topic, and purpose (Council of Europe, 2001; Hulstijn, Aldersen, & Schoonen, 2010). The vertical dimension categorizes the language proficiency (i.e. statements of learning objectives) of a learner by using six levels which are organized into three divisions: A1 and A2 (basic users), B1 and B2 (independent users), and C1 and C2 (proficient users). The vertical dimension has various practical applications such as curriculum design and the creation of qualifying examinations (Council of Europe, 2001). The combination of these two dimensions, and their varying definitions, results in communicative language being understood as an amalgamation of the scope of language use (horizontal dimension) and the manifestation of language proficiency (vertical dimension) (Hulstijn et al., 2010). Using both of these dimensions, the CEFR is able to describe and outline the expected reading, listening, speaking, and writing abilities of a learner at each level of proficiency.

The CEFR was officially published in 2001 in both English and French (Little, 2006). In November 2001 a European Union Council resolution recommended using the CEFR as the common system for the recognition of language proficiency. Subsequently, the CEFR became an important system for providing criteria for the validation of foreign language abilities, including Chinese teaching (Figueras, 2012) and second-language teaching in many regions (Hulstijn, Aldersen, & Schoonen, 2010). It also provides reference indicators for second-language learning, assists in the compilation of teaching materials, and supports the assessment of language proficiency (Little, 2006).

Beyond using two distinct dimensions to describe communicative language competency, the CEFR deliberately avoids describing language proficiency in theoretical terms. Instead the CEFR provides general descriptions; this means that its scales for the scope of language use are short, easy to use, and applicable to many different languages (Little, 2007). In addition to providing guidance for the appropriate level of a teaching material or text, the CEFR can also be used to label the difficulty level of language assessments. The ease of use and applicability of the CEFR labeling system to different languages as a common standard has resulted in it being used for defining the difficulty level of language tests developed by various institutions (Alderson, 2007).

3. Feature-based Tools for Grading L2 Teaching Materials

Readability research can be a useful point of departure for L2 text grading. Readability is often understood as *text comprehensibility*, or how well a text can be comprehended by the reader (Klare, 1984). Methods for measuring text readability have long been widely available for alphabetic languages (Dale & Chall, 1948), as are the readability formulas for grading textbooks (Faison, 1951). Traditional readability research assumes that the difficulty level of a text is determined by its semantics and syntax (Collins-Thompson, 2014), and that it is possible to create formulas to predict the difficulty level of a given text based on those two elements. For example, Flesch–Kincaid (1948) readability tests for English, which make use of the number of syllables and words in a sentence to assign grades to English books.

Recently, however, researchers have begun to challenge the way that text difficulty is determined. Collins-Thompson (2014), for instance, points out that only a few shallow linguistic features are actually used in order to estimate text difficulty; these features do not reflect the actual reading process and overly simplify the assessment of text difficulty. As a result, various attempts have been made to approximate the complex process of text understanding (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Louwerse, McCarthy, & Graesser, 2010), such as exploring the relationship between text cohesion indicators and other indicators (Benjamin, 2012), and using computational cohesion and coherence metrics (Crossley & McNamara, 2008; Crossley, Louwerse, McCarthy, & McNamara, 2007; Graesser et al., 2004).

For non-alphabetic languages, Sung et al. (2013; 2015a) developed multi-level Chinese readability models, taking into account features at lexical, syntactic, semantic, and cohesive levels. These models were subsequently extended to determine the difficulty levels of L2 Chinese texts (CRIE-CFL readability model) (Sung et al., 2015b). Sung et al. (2015b) proposed a CRIE-CFL system combining the CEFR grading criteria with the readability assessment methods trained by the support-vector-machine (SVM) technology (Vapnik, 1995). The training data of CRIE-CFL consist of 1,578 texts from 28 CFL textbook series published across 23 countries and regions such as the United Kingdom, Germany, France, Italy, Australia, Mainland China, and Taiwan, etc. where the CEFR standard is often used for learning material grading purposes. The CEFR-graded materials include *Practical Audio-Visual Chinese* (2nd Edition)[2], *Far East Everyday Chinese*[3], and *New Practical Chinese Reader*[4], etc. In order to ascertain the appropriate CEFR level for each text in the training data, expert educators, who had been teaching

[2] National Taiwan Normal University (Eds). 2008. *Practical audio-visual Chinese* (2nd Edition). Taipei: Cheng Chung Bookstore.

[3] Yeh, T. M. (Ed). 2008. *Far East everyday Chinese.* Taipei: Far East Book Company.

[4] Liu, X. (2007). *New practical Chinese reader.* Beijing: Beijing Language and Culture University Press.

CFL for more than 10 years and were familiar with CEFR level grading, read the selected materials and then assigned the corresponding CEFR level. Information on each level is provided in Table 1.

Table 1 Information on Each Level of CRIE-CFL Built-in Texts

CEFR level	No. of texts	No. of characters	No. of characters, mean (*SD*)	No. of words, mean (*SD*)
A1	155	9888	64 (37)	45 (25)
A2	337	48060	143 (64)	101 (45)
B1	470	145006	309 (198)	211 (139)
B2	345	165807	481 (221)	322 (152)
C1	190	122025	642 (358)	425 (253)
C2	81	121900	1505 (978)	1019 (695)
Total	1578	612686	388 (432)	263 (297)

The CRIE-CFL system takes into consideration a variety of text features so that the model is not biased toward a small number of features (McNamara et al., 2002). Sung et al. (2015b) utilized the F-score (Chen & Lin, 2006; Chang & Lin, 2008; Ding, 2009), a commonly used algorithm for selecting relevant features, to determine which features would improve the readability model most significantly. The F-score allows for the predicting power of the model. According to Chen & Lin (2006), the larger the F-score is, the more likely this feature is discriminative. In Sung et al.'s (2015b) study, each text is represented by a series of feature values based on textual complexity. Ideally, texts within the same level should have similar feature values. The algorithm compares those values between and within levels (e.g. the CEFR A1 vs. A2). Features with a high F-score are more useful for assigning grade level.

Eventually, Sung, et al. (2015b) verified the performance of the CRIE-CFL system, which yielded exact-level, adjacent-level, and division accuracies of 75%, 99%, and 90%, respectively. In addition, a trend analysis showed that the values of the 30 indicators that determine the CFL text difficulty level changed significantly with the CEFR levels. This means that the linguistic features data in the current CRIE-CFL corpus have rational validity; moreover, as discussed in Sung et al. (2015b), since the selection of teaching materials for CRIE-CFL is representative of texts from all levels, the quantitative features are valid. The CRIE-CFL itself can, therefore, be considered an anchored teaching material and the data of its various linguistic features can be used as a benchmark for comparison with other teaching materials (Sung, et al., 2015b).

Sung et al. (2016) made use of protocols presented in Sung, et al. (2015a; 2015b) and released a web-based CRIE system[5]. It provides four subsystems: CRIE (Analysis of texts written for native Chinese readers), CRIE-CFL (Analysis of texts written for

[5] c.f. http://www.chinesereadability.net/CRIE/index.aspx?LANG=CHT

learners of Chinese), CRIE-DK (Assesses the knowledge content levels of texts), and WECAn & HanParser (Word segmentation and part-of-speech tagging tools). The CRIE also provided 82 multilevel linguistic features, segmentation, syntactic parsing, and feature extraction. In this study the CRIE-CFL system is applied to examine the grading of five CFL textbook series.

4. Methods

4.1 Instruments

This study utilized the readability analysis system CRIE-CFL developed by Sung et al. (2016) to analyze textbook content. The CRIE-CFL automatically captures the linguistic features of Chinese texts and provides an objective numeric value for each linguistic feature found in the texts. In this study, the CRIE-CFL system is used to obtain quantitative values for each linguistic feature from five L2 Chinese textbook series to examine the consistency of their CEFR grading by their developers.

4.2 Materials for Analysis

In order to maintain consistency in the comparison and interpretation of result data using the CRIE-CFL system, this study selected five CFL textbook series (c.f., Table 2) that have been assigned CEFR proficiency levels. Three of the textbook series were published in the Greater China region because the Chinese-speaking area offers a wide range of CFL materials to select from. The rest two were selected from Europe (i.e., France and Germany) where CEFR was established. The fact that the five textbook series are from different publishers ensures that they are not subject to similar publishing guidelines, which might not represent the actual developments of CFL textbooks in different regions.

As a first attempt to compare grading consistency among different CFL textbooks using the CRIE-CFL system, this study was limited to those where manual CEFR grading by their developers are readily available. It did not include popular textbooks from other regions such as the Integrated Chinese series in North America (Li, Wen, & Xie, 2012), though future research could extend to include textbooks from more regions.

The CEFR scale (A1, A2, B1, B2, C1, and C2) contains six levels; however, Chinese language teaching materials at level C are very rarely seen on the textbook market, and textbook publishers do not tend to give classification to such materials. In addition, texts in *Chinesisch ohne Mühe* (hereafter *Chinesisch*; published in Germany) and *Le chinois par boules de neige* (hereafter *Boules de neige*; published in France) are labeled exclusively with levels B1–B2.

Table 2 Number of Texts at Each Level in the Five Textbooks

Place of publication	Textbook title	Author indications of CEFR levels						Total no. of texts
		A1	A2	B1	B2	C1	C2	
Mainland China	Road to Success[6] 成功之路	167	42	41	24	37	53	364
Taiwan	New Modern Chinese[7] 新時代華語	8	12	18	2	0	0	40
Mainland China	Practical Chinese[8] 实用中文	44	43	49	31	0	0	167
France	Le chinois par boules de neige[9] 雪球	0	16	19		0	0	35
Germany	Chinesisch ohne Mühe[10] 漢語	49		56		0	0	105

4.3 Procedure

4.3.1 Selection of Linguistic Features

The CRIE-CFL has developed 30 linguistic features, which can be divided into four categories: lexicon, semantics, syntax, and cohesion (Sung et al., 2015b). This study selects 12 linguistic features using F-score and Trend Analysis F Value to reflect either the key concepts in the CEFR proficiency level or the unique nature of the Chinese language (cf., Table 3) (Sung et al., 2015b). These 12 features represent the most influential aspects of each of the four categories and are used to determine if the difficulty levels of the five Chinese language textbooks are consistent.

First, lexical category is used to measure the complexity of texts and hence text difficulty. The CEFR scale for overall reading comprehension (Council of Europe, 2001)

[6] Editors of the Road to Success sereis (Ed). (2008-2014). *Road to success (成功之路)*. Beijing Language and Culture University Press. China: Beijing.

[7] NTNU Extension School of Continuing Education (Ed). (2012). *New modern Chinese (新時代華語)*. NTNU Extension School of Continuing Education. Taiwan: Taipei.

[8] Chinese Time (Ed). (2009). *Practical Chinese (实用中文)*. East China Normal University Press. China: Shanghai.

[9] Bellassen, J., & Liu, J. L. (2011). *Le chinois par boules de neige (Acces raisonne a la lecture du chinois)（雪球）*. Scérén Cndp-crdp. France: Chasseneuil-du-Poitou.

Bellassen, J., & Liu, J. L. (2012). *Le chinois par boules de neige (Niveau elementaire)（雪球）*. Scérén Cndp-crdp . France: Chasseneuil-du-Poitou.

[10] Kantor P. (2004). *Assimil Pack Chinesisch Ohne Mühe（漢語）*. ASSiMiL GmbH. *Volume 1*. Germany: Köln.

Kantor P. (2006). *Assimil Pack Chinesisch Ohne Mühe（漢語）*. ASSiMiL GmbH. *Volume 2*. Germany: Köln.

states that A1- and A2-level learners can understand short texts, whereas C1-level learners can understand detailed, long texts, and C2-level learners can understand a wide range of long texts. Numerical counts of characters and words are used to measure text length as shown in Table 3.

Table 3 Linguistic Features Selected in this Study

	Feature	Definition
Lexical Category	characters	total number of characters
	high-level words	total number of words listed by the 8,000 Chinese Words[11] as being in the vantage or effective operational proficiency levels
	two-character words	number of two-character words
Semantic Category	content words	number of content words
	sentences with complex semantic categories	number of sentences with a number of semantic categories
	complex semantic categories	number of semantic categories from sentences with complex semantic categories
Syntactic Category	average sentence length	average number of words in a sentence
	simple sentence ratio	the number of simple sentences divided by the total number of sentences
	sentences with a complex structure	the number of sentences containing conjunctions and subordinators
Cohesive Category	conjunctions	number of conjunctions
	positive conjunctions	number of conjunctions with positive meanings
	negative conjunctions	number of conjunctions with negative meanings

As seen in Table 3, in addition to the number of characters in a text, the count of two-character words is also applied as a measure of text length. The main component of Chinese is two-character words (Duanmu, 1999; He & Li, 1987). In order to distinguish a

[11] The 8,000 Chinese Words can be found at https://www.sc-top.org.tw/chinese/download.php.

text's difficulty, this study used the number of *characters* (e.g., *shū* 書 book) and *two-character words* (e.g., *zhī jì* 之際 at the time of) in a text as an indicator of length. Moreover, the CEFR scale for overall reading comprehension states that A2-level learners can understand the *highest frequency vocabulary, high-frequency everyday or job-related language.* B2-level learners can understand *low-frequency idioms*. Therefore, a learner who is at a higher CEFR level can understand harder words, at low frequencies. Accordingly, this study incorporated the number of *high-level words* (e.g. *bǎo cún* 保存 preserve) as an indicator of word difficulty in identifying text difficulty. Note that in this study, all the features in Table 3 are calculated independently. Therefore, some words would be counted more than once. For example, *bǎi tuō* (擺脫, to break away from) was counted both as a high-level word and as a two-character word.

The second measure of text difficulty adopted in this study is semantics. To account for semantics, this study selected three semantic features to examine text complexity: 1) the number of *content words* (e.g., *lán qiú* 籃球 basketball), 2) the number of *sentences with complex semantic categories*, and 3) the number of *complex semantic categories*. Content words are words with independent lexical meanings. More content words within a text represent more concepts in that text and thus higher complexity. According to Hong et al. (2016), semantic categories is defined as the number of meanings in a single word. Words with multiple meanings are more likely to cause semantic ambiguity (e.g., *chī bīng qí lín* 吃冰淇淋, which means either 'eat ice cream' or 'look at an eye candy' when used in Taiwan) at the sentence level. In addition, words with larger numbers of semantic categories usually generate more significant lexical semantic variations (e.g., *dǎ diàn huà* 打電話 call someone/ hit the phone). It has been reported that a higher number of semantic categories also increases sentence difficulty (Cheng, 2005), and therefore was included in this study. Furthermore, polysemous words have more lexical meanings which contribute to lexical ambiguities and increase complexity. More semantic categories also imply more complex lexical meanings.

The third category of measuring text difficulty is syntax. Two crucial components of text complexity are sentence length and sentence structure. For example, simple sentences are semantically independent syntactic units that consist of a subject and a predicate. Complex sentences are formed by combining two or more simple sentences (Hong, Sung, Tseng, Chang, & Chen, 2016). Since the meaning of a complex sentence is broader and more intricate, lower-proficiency learners cannot understand texts with a high number of complex sentences. When lower-proficiency learners read texts with a high number of sentences with complex structures, they experience more difficulties.

The last category of features used in this study to measure text difficulty is cohesion. The three cohesion related indicators that are used to examine text complexity in this study are *conjunction* (e.g., *yīn wèi…suǒ yǐ* 因為…所以 because), *positive conjunction* (e.g., *ér qiě* 而且 and), and *negative conjunction* (e.g., *fǒu zé* 否則 otherwise). Conjunctions are employed within a sentence to indicate that subsequent meanings are systematically connected to preceding meanings (Halliday & Hasan, 1976). Therefore, conjunctions facilitate the establishment of cohesive relationships within texts

(Louwerse & Mitchell, 2003). When texts are longer and more complex, more conjunctions are needed to aid a learner's comprehension.

The aforementioned 12 linguistic features categorized by lexical, semantic, syntactic, and cohesive were selected to calculate text complexity in this study.

4.3.2 Quantitative Feature Analysis of Chinese Textbooks

The CRIE-CFL system was used to determine the 12 linguistic features and then to examine whether appropriate CEFR levels were assigned to each of the selected textbooks within this study. A one-way ANOVA was conducted to identify differences in the linguistic features between different levels within each textbook series, including the CRIE-CFL (i.e. accuracy). The CEFR level served as the independent variable and the value of a linguistic feature was identified as the dependent variable. A significant ANOVA result suggests that the value of the linguistic features of at least one level is significantly higher or lower than that of the other levels; this implies that the linguistic features of different levels of teaching materials are not identical. Alternatively, an insignificant ANOVA result suggests that the values of linguistic features of different levels of teaching material are statistically equivalent; this implies that the linguistic features of different levels of teaching material are the same.

When ANOVA results were significant, a trend analysis was conducted to identify if any special trends were present in the linguistic features of each level or if the changes were simply random. The presence of a significant linear trend would indicate that the linguistic features of different levels do change with CEFR levels, and vice versa. If text difficulty changes with the CEFR level, the value of eleven of the twelve linguistic features (except for *simple sentence ratio*) should be lower in lower-level texts (e.g., A1) than in higher-level texts (e.g., B1).

The second stage of the analysis involved investigating whether the authors of the five selected textbooks assigned CEFR levels consistently; this is indicated by their use of linguistic features within textbooks labeled with the same CEFR level. Since level C is absent from the textbooks used in this study, only the textbooks labeled with A and B levels were compared. Another one-way ANOVA analysis was conducted to test whether there were differences in the linguistic features between the six textbook series (i.e. consistency). In this analysis, the CEFR level was the independent variable while the value of a linguistic feature was the dependent variable. A significant ANOVA result would indicate that the value of the linguistic features of at least one teaching material was significantly higher or lower than that of at least one of the others. That is, the linguistic features of the six teaching materials were not identical. On the other hand, an insignificant ANOVA result would suggest that the linguistic features used across the six textbook series were similar. Across series, textbooks labeled with the same CEFR level were expected to yield similar values in their linguistic features.

5. Results

5.1 Comparing Levels of Text Difficulty within the Same Textbook Series

The mean values of the 12 linguistic features within each level of the five Chinese language textbooks are listed in Table 4. The results show that for all five textbooks there are significant differences and significant linear trends in the following six categories: *characters*, *two-character words*, *sentences with a complex structure*, *content words*, *sentences with complex semantic categories*, and *complex semantic categories*. This means that the values of these six linguistic features either increase or decrease as the CEFR level increases. The values of the linguistic features of four out of the five textbooks (*Road to Success*, *New Modern Chinese*, *Practical Chinese*, and *Chinesisch*) increase with the CEFR level, whereas those of *Boules de neige* decrease; for example, there are fewer *characters* in the B1-level and B2-level texts than in the A2-level text.

This study yielded the following additional observations. All of the textbooks except *Boules de neige* show significant positive linear trends between *high-level words* and level, in that the number of *high-level words* increases as the CEFR level increases. *Practical Chinese* and *Chinesisch* show significant positive linear trends between *average sentence length* and level, in that the *average sentence length* increases as the CEFR level increases. Three textbooks (*New Modern Chinese*, *Practical Chinese*, and *Boules de neige*) show significant negative linear trends between *simple sentence ratio* and level, with the *simple sentence ratio* decreasing as the CEFR level increases. All of the textbooks except *Boules de neige* and *Chinesisch* show significant positive linear trends between *conjunctions* and level, in that the number of *conjunctions* increases as the CEFR level increases. All of the textbooks except *Chinesisch* show significant positive linear trends between *positive conjunctions* and level, in that the number of *positive conjunctions* increases as the CEFR level increases. Finally, two textbooks (*Road to Success* and *Practical Chinese*) show significant positive linear trends between *negative conjunctions* and level, with the number of *negative conjunctions* increasing as the CEFR level increases.

According to the results of this study, the 12 linguistic features (12/12) of *Practical Chinese* change in accordance with the change of the CEFR level. For *Road to Success* and *New Modern Chinese*, the results are similar. Both of the texts have 10 linguistic features (10/12) which change according to the CEFR level, but the remaining two linguistic features do not. The results for *Boules de neige* and *Chinesisch* are similar. Both of these texts have eight linguistic features (8/12) which change based on the CEFR level; however, the remaining four linguistic features of each textbook do not follow this pattern. This indicates that at least four of the five textbook authors did not demonstrate their ability to adopt materials with linguistic features that would appropriately reflect the corresponding difficulty levels of texts within the same textbook series.

Table 4 The Mean Values of the 12 Linguistic Features of the Six Different Levels for CRIE-CFL and the Five Textbooks

	CRIE-CFL						*F*	*Road to Success*						*F*	*New Modern Chinese*				*F*	*Practical Chinese*				*F*	*Boules de neige*		*F*	*Chinesisch*		*F*
Linguistic feature	A1	A2	B1	B2	C1	C2	(η^2)	A1	A2	B1	B2	C1	C2	(η^2)	A1	A2	B1	B2	(η^2)	A1	A2	B1	B2	(η^2)	A2	B1/B2	(η^2)	A1/A2	B1/B2	(η^2)
Characters	64	143	309	481	642	1505	353 (.53)	453	836	1231	1612	1128	1776	72 (.50)	67	162	252	316	84 (.88)	110	190	259	627	66 (.55)	404	236	204 (.86)	90	127	32 (.24)
High-level words	2	9	32	62	102	255	552 (.64)	61	127	195	261	165	295	85 (.54)	5	11	25	39	34 (.74)	4	13	27	82	95 (.64)	38	33	n.s. (.09)	6	9	15 (.12)
Two-character words	12	32	76	127	175	403	412 (.57)	124	225	333	425	301	473	75 (.51)	12	36	66	78	71 (.86)	25	45	70	169	66 (.55)	89	76	9 (.22)	17	26	31 (.23)
Average sentence length	6.07	7.85	8.75	9.19	9.88	10.44	130 (.29)	10.29	9.66	10.15	10.21	9.92	10.32	n.s. (.00)	6.55	8.06	8.13	7.27	9 (.44)	7.38	8.47	8.88	9.39	15 (.21)	8.73	8.82	n.s. (.00)	3.96	4.49	9 (.08)
Simple sentence ratio	.97	.85	.62	.45	.32	.37	310 (.50)	.39	.42	.43	.39	.42	.38	n.s. (.01)	.98	.89	.78	.56	18 (.60)	.86	.71	.53	.56	21 (.28)	.74	.53	16 (.33)	1.00	.99	n.s. (.03)
Sentences with a complex structure	1.37	5.33	12.70	20.59	26.49	63.31	309 (.50)	18.41	34.81	52.83	68.96	50.22	75.32	65 (.48)	1.88	7.42	10.39	12.00	28 (.70)	3.36	7.93	11.24	27.55	61 (.53)	16.13	8.32	93 (.74)	1.98	3.38	12 (.10)
Content words	39	85	175	262	346	816	316 (.50)	245	457	672	872	616	959	65 (.48)	46	104	148	183	73 (.86)	67	110	144	351	61 (.53)	239	116	470 (.93)	57	78	27 (.21)
Sentences with complex semantic categories	5.45	8.99	14.96	20.79	23.62	48.96	152 (.33)	16.66	30.81	43.46	52.54	41.14	58.08	44 (.38)	5.00	9.58	14.06	22.00	30 (.72)	7.16	10.16	11.43	25.32	35 (.39)	19.13	8.63	45 (.58)	13.96	16.98	9 (.08)
Complex semantic categories	2.18	3.26	4.98	6.78	7.54	14.82	107 (.25)	4.96	9.33	13.44	15.85	12.81	17.64	40 (.36)	1.90	3.16	4.78	8.13	18 (.60)	2.69	3.51	3.71	7.98	26 (.32)	6.52	2.66	26 (.44)	6.42	7.91	7 (.07)
Conjunctions	0.35	1.85	5.45	10.27	14.25	32.47	346 (.52)	10.14	16.93	27.02	33.58	24.11	37.58	57 (.44)	0.63	2.08	6.11	3.50	17 (.59)	0.89	2.84	6.73	12.94	54 (.50)	4.81	6.21	n.s. (.05)	0.90	1.32	n.s. (.03)
Positive conjunctions	0.25	1.24	3.84	7.10	9.16	20.33	319 (.50)	6.42	10.86	17.98	20.92	15.22	24.19	58 (.45)	0.00	1.25	3.39	2.00	16 (.57)	0.66	2.00	4.94	8.29	39 (.42)	2.63	4.58	6 (.16)	0.51	0.84	n.s. (.03)
Negative conjunctions	0.13	0.56	1.70	3.00	4.29	10.80	197 (.39)	3.28	5.62	8.07	11.54	7.92	12.45	33 (.31)	0.63	0.92	1.83	0.50	n.s. (.20)	0.18	0.86	1.59	4.45	34 (.38)	1.88	1.53	n.s. (.02)	0.49	0.50	n.s. (.00)

Note. *F* = *F* value in the ANOVA test; CRIE-CFL = Chinese Readability Index Explorer for Chinese as a Foreign Language; n.s. = not significant.

More detailed information about the standard deviation and F value in trend analysis can be found on http://140.122.96.190/20171107/table4.pdf.

5.2 Comparing Linguistic Features of Texts with the Same CEFR-labels among Textbooks

Table 5 lists the mean values of the 12 linguistic features of the texts labeled as CEFR A-level and B-level for each of the five textbooks and the CRIE-CFL database. The one-way ANOVA results indicate that the 12 linguistic features of the six teaching materials at level A are significantly different (as indicated by the F values in Table 5). The 12 linguistic features of the six teaching materials at level B are also significantly different (see Appendix 2 for the results of post-hoc comparisons). The results show that textbooks labeled with the same CEFR levels yield different values in terms of their linguistic features, such as *high-level words* and *average sentence length*, and should actually be assigned different difficulty levels. A detailed analysis is presented Table 5.

Many discrepancies can be seen when looking at the lexical feature *characters*. Among all A-level teaching materials, *Road to Success* has the highest number of *characters* (mean = 530 characters) while *Chinesisch* has the lowest number (mean = 90 characters), with a difference of 440 characters. Meanwhile, the average number of *characters* is significantly higher in *Road to Success* and *Boules de neige* than in the CRIE-CFL, and significantly lower in *Chinesisch* than in the CRIE-CFL. Among all B-level teaching materials, *Road to Success* has the highest number of *characters* (mean = 1372 characters) while *Chinesisch* has the lowest number (mean = 127 characters), with a difference of 1245 characters. In addition, the average number of *characters* is significantly higher in *Road to Success* than in the CRIE-CFL and significantly lower in *New Modern Chinese*, *Boules de neige*, and *Chinesisch* than in the CRIE-CFL.

The following variances can be seen when analyzing the syntactic feature *average sentence length*. Among A-level teaching materials, *Road to Success* has the longest *average sentence length* (mean = 10.16 words) while *Chinesisch* has the shortest *average sentence length* (mean = 3.96 words), corresponding to a difference of 6.20 words. The *average sentence length* is significantly longer in *Road to Success* than in the CRIE-CFL and significantly shorter in *Chinesisch* than in the CRIE-CFL. Among B-level teaching materials, *Road to Success* has the longest *average sentence length* (mean = 10.17 words) while *Chinesisch* has the shortest *average sentence length* (mean = 4.49 words), corresponding to a difference of 5.68 words. Meanwhile, the *average sentence length* is significantly longer in *Road to Success* than in the CRIE-CFL, and significantly shorter in *New Modern Chinese* and *Chinesisch* than in the CRIE-CFL.

The following observations were made when analyzing the semantic feature *content words*. Among A-level teaching materials, *Road to Success* has the highest number of *content words* (mean = 287 words) while *Chinesisch* has the lowest number (mean = 57 words), corresponding to a difference of 230 words. The number of *content words* is significantly higher in *Road to Success* and *Boules de neige* than in the CRIE-CFL, and significantly lower in *Chinesisch* than in the CRIE-CFL. Among B-level teaching materials, *Road to Success* has the highest number of *content words* (mean = 746 words) while *Chinesisch* has the lowest number (mean = 78 words), corresponding to a difference of 668 words. The number of *content words* is significantly higher in *Road to*

Table 5 The Mean Values of the 12 Linguistic Features in CRIE-CFL and the Five Textbooks at Levels A and B

CEFR	Level A						*F*	Level B						*F*
Linguistic feature	CRIE-CFL	*Road to Success*	*New Modern Chinese*	*Practical Chinese*	*Boules de neige*	*Chinesisch*	(η^2)	CRIE-CFL	*Road to Success*	*New Modern Chinese*	*Practical Chinese*	*Boules de neige*	*Chinesisch*	(η^2)
Characters	118	530	124	149	404	90	350 (.67)	381	1372	258	402	236	127	195 (.48)
High-level words	7	75	9	9	38	6	350 (.67)	45	219	27	48	33	9	256 (.55)
Two-character words	26	144	26	35	89	17	410 (.70)	98	367	67	108	76	26	207 (.50)
Average sentence length	7.29	10.16	7.46	7.92	8.73	3.96	30 (.15)	8.94	10.17	8.05	9.08	8.82	4.49	192 (.48)
Simple sentence ratio	.89	.39	.93	.79	.74	1.00	231 (.57)	.55	.41	.76	.54	.53	.99	46 (.18)
Sentences with a complex structure	4.08	21.70	5.20	5.62	16.13	1.98	316 (.65)	16.04	58.78	10.55	17.56	8.32	3.38	187 (.47)
Content words	71	287	81	88	239	57	295 (.63)	212	746	151	224	116	78	177 (.46)
Sentences with complex semantic categories	7.88	19.50	7.75	8.64	19.13	13.96	97 (.36)	17.42	46.82	14.85	16.81	8.63	16.98	65 (.24)
Complex semantic categories	2.92	5.84	2.66	3.10	6.52	6.42	53 (.24)	5.74	14.33	5.12	5.36	2.66	7.91	48 (.19)
Conjunctions	1.38	11.51	1.50	1.85	4.81	0.90	290 (.63)	7.49	29.45	5.85	9.14	6.21	1.32	166 (.44)
Positive conjunctions	0.93	7.31	0.75	1.32	2.63	0.51	211 (.55)	5.22	19.06	3.25	6.24	4.58	0.84	133 (.39)

Negative conjunctions	0.43	3.75	0.80	0.52	1.88	0.49	148 (.46)	2.25	9.35	1.70	2.70	1.53	0.50	103 (.33)

Note. CRIE-CFL = Chinese Readability Index Explorer for Chinese as a Foreign Language; n.s. = not significant.

More detailed information about the standard deviation can be found at http://140.122.96.190/20171107/table5.pdf.

Success than in the CRIE-CFL, and significantly lower in *New Modern Chinese*, *Boules de neige*, and *Chinesisch* than in the CRIE-CFL.

When examining the cohesion feature c*onjunctions* the following disparities can be seen. Among A-level teaching materials, *Road to Success* has the highest number of *conjunctions* (mean = 11.51 conjunctions) while *Chinesisch* has the lowest (mean = 0.90 conjunctions), corresponding to a difference of 10.61 conjunctions. The number of *conjunctions* is significantly higher in *Road to Success* and *Boules de neige* than in the CRIE-CFL. Among B-level teaching materials, *Road to Success* has the highest number of *conjunctions* (mean = 29.45 conjunctions) while *Chinesisch* has the lowest number (mean = 1.32 conjunctions), corresponding to a difference of 28.13 conjunctions. The number of *conjunctions* is significantly higher in *Road to Success* than in the CRIE-CFL, and significantly lower in *Chinesisch* than in the CRIE-CFL.

6. Discussion

This study used the CRIE-CFL system as a tool to calculate 12 linguistic features in five Chinese textbook series in order to examine their CEFR level grading. It compared textbooks within and between series to examine the accuracy and consistency of their CEFR level grading in relation to the actual text difficulty as measured by the CRIE-CFL system.

6.1 Differences within Series

The linguistic features within the same textbook series did not show a meaningful transition between the levels of difficulty. The data produced in this study indicate that most teaching materials do not match their assigned CEFR levels. The trend analysis shows that of the 1578 texts used for CRIE-CFL training, only *Practical Chinese* demonstrated an increasing or decreasing trend corresponding to the assigned CEFR level. That is, linear trends were not found between linguistic features and CEFR levels in *Road to Success*, *New Modern Chinese*, *Boules de neige*, and *Chinesisch*. The linguistic features in these four teaching materials do not vary in accordance with their assigned difficulty level. Although eight of the linguistic features in *Boules de neige* exhibited positive linear tendencies, six of the linguistic features were inversely correlated with the CEFR levels in the remaining five teaching materials. It should also be noted that there are cases in *Boules de neige* where lower-level texts are explained by words from higher-level texts. For example, in Lesson 2 of the A2-level textbook, the word *buzhibujuedi* (unconsciously; 不知不覺地) is explained by the word *zhuyi* (attention; 注意), which is from Lesson 2 of the B-level textbook.

The mismatch between the vocabulary and CEFR levels calls for standardized leveling criteria to select level-appropriate linguistic features for textbooks. This study provides additional evidence to support the observations by Alderson (2007), Hulstijn (2007), and Hulstijn et al. (2010) that the clarity of CEFR’s definition of each proficiency level should also be improved for educators.

6.2 Differences among Series

Textbooks with the same CEFR level contained significantly different linguistic features. In this study, textbooks that are assigned the same CEFR level are found to have different difficulty levels. For example, the *Road to Success* has the lowest *simple sentence ratio* while *Chinesisch* has the highest *simple sentence ratio*. In other words, with respect to sentence learning, *Road to Success* is more difficult than *Chinesisch*. Furthermore, this study finds that for these two series, their materials at levels A and B yield different values in the linguistic feature analysis. Even though these textbooks are labeled with the same CEFR levels, they actually have different difficulty levels.

As for *Boules de neige*, its A-level materials have relatively more *characters*, *high-level words*, *two-character words*, *sentences with a complex structure*, *content words*, and *conjunctions*. This indicates that these materials are more difficult than their counterparts in other series. On the other hand, *Boules de neige's* B-level materials are comparatively easier as they contain fewer *sentences with a complex structure*, *content words*, *sentences with complex semantic categories*, and *complex semantic categories*. B-level materials in *New Modern Chinese* are also relatively easy as they contain a higher *simple sentence ratio*, a shorter *average sentence length*, and a smaller number of *sentences with a complex structure* and *content words*. Considering these inconsistencies, our linguistic feature analyses suggest that the compilation of teaching materials cannot be based solely on educators' professional experience. A standardized system is required to determine the difficulty level of L2 teaching materials to ensure accuracy within a series and consistency between series.

The findings of this study provide empirical evidence indicating the inconsistency in the difficulty levels of different Chinese teaching materials. In order to account for these inconsistencies, scholars have independently developed rubrics for evaluating vocabulary, grammar, and reading sections of language textbooks, such as using metrics based on the vocabulary load, vocabulary difficulty, and word frequency (Rahimpour & Hashemi, 2011; Williams, 1983). However, these rubrics take the form of questionnaires, which are still predicated on a subjective evaluation. As Sung et al. (2015b) pointed out, the manual leveling of learning materials presents three problems: high demands on both time and effort, difficulty in reaching a consensus, and ambiguity in the interpretation of leveling criteria. These problems also appeared in the CFL textbooks that were analyzed in this study.

6.3 Types of Linguistic Features that Affect Text Difficulty

According to the CEFR scale for overall reading comprehension (Council of Europe, 2001), learners at levels A1 and A2 should be able to understand high-frequency words, and B2 learners should possess a larger vocabulary than A1. Based on our linguistic feature analyses with CRIE-CFL, A-level materials in *Road to Success* and *Boules de neige* have a considerably higher number of *characters*, *high-level words*, and *two-character words* than other textbook series. Therefore, beginners may find these textbooks difficult. B-level materials in *Road to Success* suffer from the same problem. Textbooks compilers should adjust the difficulty levels of these textbooks by selecting proficiency-appropriate vocabulary at a given level and then ensure that the vocabulary consistently increases as a learner's proficiency level increases (Rahimpour & Hashemi, 2011).

Regarding syntax, A1 and A2 learners should be able to understand short and simple sentences, while C1 and C2 learners are able to understand long and complex sentences. Our linguistic feature analyses show that the *average sentence length* and the *simple sentence ratio* in *Road to Success* do not change significantly in relation to the CEFR level. Compared to other series, texts at levels A and B in *Road to Success* have a higher *average sentence length* and a higher number of *sentences with a complex structure* but a lower *simple sentence ratio*. Lower-level texts in *Road to Success* tend to be comprised of longer and more complex sentences, which may cause comprehension difficulty for beginners. On the other hand, texts at levels A and B in *Chinesisch* do not show any significant differences in their *simple sentence ratio*, which means that the sentences in the B-level material in *Chinesisch* may be too short and should increase their complexity.

In terms of semantic features, higher-level learners should be able to understand more content words and more semantically complex sentences. For most of the teaching materials, the values for the three semantic features tend to be higher for the higher-level texts than for the lower-level ones; *Boules de neige* is the only exception, in that its three semantic linguistic features move in opposite directions, which suggests that the vocabulary of this textbook needs to be adjusted. *Road to Success* has higher values for the three semantic features than the other textbooks. This indicates that its texts are more difficult and include a larger number of concepts that require more time to process.

The linguistic features adopted by the CRIE-CFL correspond to those in the CEFR reading comprehension grading standards, and also include an additional three linguistic features: *conjunctions*, *positive conjunctions*, and *negative conjunctions.* These three features were added because conjunctions help learners to establish cohesion when reading a text (Louwerse & Mitchell, 2003). Cohesion is an important component of reading comprehension (Benjamin, 2012; Graesser, McNamara, & Kulikowich, 2011; Graesser et al., 2004; McNamara, Louwerse, McCarthy, & Graesser, 2010). The combination of cohesive sentences, consistent text, and cohesive semantics contribute to the creation of texts that are more readable to learners (Gernsbacher, 1990; McNamara & Kintsch, 1996; McNamara, Kintsch, Songer, & Kintsch, 1996).

Our analyses of the five textbooks and the texts in the CRIE-CFL system have shown that the number of *negative conjunctions* in the CRIE-CFL training data set, *Road to Success*, and *Practical Chinese* increases with the CEFR level. These results suggest that higher-level texts contain more transitions, as more cognitive resources are required to process the complex relationships between sentences in the texts. However, no significant differences in the three cohesive features were found between different level textbooks in *Chinesisch.* This suggests that *Chinesisch* did not take into account the effect of conjunctions on reading comprehension.

6.4 Pedagogical Implications

Findings of this study have several pedagogical implications. First, developers of language materials who are looking to incorporate the CEFR scale should carefully consult the statements regarding the horizontal dimension of the CEFR scales in order to obtain an in-depth understanding of various topics, the scope of language use, and language proficiency. These statements define the level of proficiency in listening, speaking, reading, writing, and translation. Additional training led by experienced experts in CEFR and language learning material grading may help users to better understand the criteria in the lexical, syntactic, and semantic aspects of language proficiency. Language-specific proficiency standards should also be developed to make the description of each standard more objective and precise.

Secondly, language educators need to increase their awareness of the influence of linguistic features, such as *characters and words*, *semantics*, *syntax*, and *cohesion*, on text difficulty and reading comprehension. The awareness of linguistic features enhances an educator's ability to select proficiency-appropriate materials for leaners. Such awareness also facilitates the process of selecting CEFR-graded teaching materials, comparing textbooks published by different publishers or in different regions, and complying with the language proficiency standards in the CEFR.

Lastly, analytics tools, such as the CRIE-CFL, can be useful for quantifying linguistic features to ensure that textbook contents are consistent with both the vertical and the horizontal dimension statements of the CEFR scales. The automatic analysis functions of CRIE-CFL can also help educators efficiently develop parameters that reflect a text's level of difficulty and therefore enhance the objective evaluation of textbook levels.

7. Conclusion

Despite the importance of grading language textbooks for teaching and learning, few studies have addressed the issues of consistency, accuracy, and efficiency in the grading of texts. Based on the CEFR framework and the analytic tool CRIE-CFL (Sung et al., 2016), this study examined the accuracy and consistency of text grading within and between textbook series. Based on our linguistic feature analyses, we found that most of the textbooks we examined did not use the linguistic features reflective of their corresponding proficiency levels. The language used in these textbooks does not always increase in difficulty as the level increases. Our analyses also show that even textbooks labeled with the same CEFR level yielded different values in terms of their use of linguistic features, therefore indicating a varying level of difficulty. The results of this study call for a standardized system for educators to use in determining the difficultly level of teaching materials as manual text grading is no longer effective or reliable.

Finally, there are four major Chinese proficiency standards adopted across continents: ACTFL (US), CEFR (EU), HSK (China), and TOCFL(Taiwan). As conversions among these proficiency standards are, in fact, rather straightforward (e.g.

CEFR A2 would be equivalent to ACTFL Intermediate or HSK 4[12]), future research will extend the CRIE-CFL model to other proficiency standards to further validate the findings of this study. Thus, with CRIE-CFL, educators as well as textbook developers will be able to make use of the tool when they select and compile texts suitable for students at various proficiency levels.

Acknowledgments: This work was supported by National Taiwan Normal University's Chinese Language and Technology Center. The center is funded by Taiwan's Ministry of Education (MOE), as part of the Featured Areas Research Center Program, under the Higher Education Sprout Project.

References

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*(4), 659–663. doi: 10.1111/j.1540-4781.2007.00627_4.x

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, *24*(1), 63–88. doi: 10.1007/s10648-011-9181-8

Bonnet, G. (2007). The CEFR and education policies in Europe. *The Modern Language Journal*, *91*(4), 669–672. doi: 10.1111/j.1540-4781.2007.00627_7.x

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721-1730), Sydney, Australia.

Chang, T. H., & Sung, Y. T. (2019). Automated Chinese essay scoring based on multi-level linguistic features. In X. Lu, & B. Chen (Eds.), *Computational and Corpus Approaches to Chinese Language Learning* (pp. 258-274). Singapore: Springer.

Chang, Y. W., & Lin, C. J. (2008, December). Feature ranking using linear SVM. In *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008* (pp. 53–64), Hong Kong, China,.

Chen, B., & Hsu, Y. C. (2019). Mandarin Chinese mispronunciation detection and diagnosis leveraging deep neural network based acoustic modeling and training techniques. In X. Lu, & B. Chen (Eds.), *Computational and Corpus Approaches to Chinese Language Learning* (pp. 219–237). Singapore: Springer.

Chen Y. W., & Lin, C. J. (2006) Combining SVMs with various feature selection strategies. In I. Guyon, M. Nikravesh, S. Gunn, & L. A. Zadeh (Eds), *Feature extraction. Studies in fuzziness and soft computing* (Vol. 207, pp. 315-324). Springer, Berlin, Heidelberg.

Cheng, C. C. (2005). Computing the degree of difficulty in lexical semantics and sentence reading. In *The 6th Chinese lexical semantics workshop* (pp. 261-265), Fujian: Xiamen University. 261–265. [鄭錦全. (2005) 詞彙語意與句子閱讀難易度計量, 261-265.]

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*,

[12] https://www.actfl.org/sites/default/files/reports/Assigning_CEFR_Ratings_To_ACTFL_Assessments.pdf

165(2), 97-135.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, UK: Cambridge University Press.

Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal*, *91*(2), 15–30. doi: 10.1111/j.1540-4781.2007.00507.x

Crossley, S. & McNamara, D. (2008). Assessing L2 reading texts at the intermediate level: An approximate replication of Crossley, Louwerse, McCarthy & McNamara (2007). *Language Teaching*, *41*(3), 409–429. doi: 10.1017/S0261444808005077

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, *27*(2), 37–54.

Ding, S. (2009, November). Feature selection based F-score and ACO algorithm in support vector machine. In C. Zhao, Y. Wu, J. Wang, & Q. Liu. (Eds), *2009 Second International Symposium on Knowledge Acquisition and Modeling* (Vol. 1, pp. 19-23). IEEE Computer Society 10662 Los Vaqueros Cir, Los Alamitos, CA 90720-1314.

Duanmu, S. (1999). Stress and the development of disyllabic words in Chinese. *Diachronica*, *16*(1), 1–35. doi:10.1075/dia.16.1.03dua

Faison, E. W. (1951). Readability of children's textbooks. *Journal of Educational Psychology*, *42*(1), 43. doi: 10.1037/h0060329

Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, *66*(4), 477–485. doi: 10.1093/elt/ccs037

Flesch, R. F. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221-233.

Fulcher, G. (2004). Deluded by artifices? The common European framework and harmonization. *Language Assessment Quarterly: An International Journal*, *1*(4), 253–266. doi:10.1207/s15434311laq0104_4

Fulcher, G. (2007, December). The reification of the Common European Framework of Reference (CEFR) and effect-driven testing. Paper presented at the 14th International Conference of Applied Linguistics, Thessaloniki, Greece.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Earlbaum.

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. doi: 10.3102/0013189X11413260

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. doi:10.3758/BF03195564

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

He, K. & Li, D. (1987). *Three thousand most commonly used words in modern Chinese*. Beijing, China: Beijing Normal University Press. [何克抗, & 李大魁. (1987). *现代汉语三千常用词.* 中国北京: 北京师范大学出版社.]

Hong, J. F., Sung, Y. T., Tseng, H. C., Chang, K. E., & Chen, J. L. (2016). A multilevel analysis of the linguistic features affecting chinese text readability. *Taiwan Journal of Chinese as a Second language*, *13*, 95–126.

Hsu, F. Y., Lee, H. M., Chang, T. H., & Sung, Y. T. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, *54*(6), 969-984.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language Proficiency. *The Modern Language Journal*, *91*(4), 663–667. doi:10.1111/j.1540-4781.2007.00627_5.x

Hulstijn, J. H., Alderson, J. C., & Schoonen, R. (2010). Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them?. In I. Bartning, , M. Maisa, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 11–20). Amsterdam, Netherlands: Eurosla.

Klare, G. R. (1984). Readability. In P. D. Pearson, R. Barr, M. L. Kamil, & P. B. Mosenthal (Eds.), *Handbook of reading research*. New York: Longman

Lee, L. H., Chang, L. P., & Tseng, Y. H. (2016). Developing learner corpus annotation for Chinese grammatical errors. In Y. Lu (Ed), *Proceedings of the 20th international conference on Asian language processing* (pp. 254–257). Tainan, Taiwan.

Li, Y., Wen, X. H., & Xie, T. W. (2014). CLTA 2012 Survey of College-Level Chinese Language Programs in North America. *Journal of the Chinese Language Teachers Association*, *49*(1), 1-49.

Lin, S. Y., Chen, H. C., Chang, T. H., Lee, W. E., & Sung. Y. T. (2019). CLAD: A corpus-derived Chinese lexical association database. *Behavior Research Methods*, *51*(5), 2310-2336. doi:10.3758/s13428-019-01208-2

Little, D. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching*, *39*(3), 167–190. doi: 10.1017/S0261444806003557

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *Modern Language Journal*, *91*(4), 645–655. doi: 10.1111/j.1540-4781.2007.00627_2.x

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*, *35*(3), 199–239. doi: 10.1207/S15326950DP3503_1

Lu, X., & Chen, B. (2019). Computational and corpus approaches to Chinese language learning: an introduction. In X. Lu, , & B. Chen (Eds.), *Computational and corpus approaches to Chinese language learning* (pp. 6–14). Singapore: Springer.

McNamara, D. S. & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*, 247–287. doi:10.1080/01638539609544975

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1–43. doi:10.1207/s1532690xci1401_1

McNamara, D. S., Louwerse, M. M., & Graesser, A. C. (2002). *Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension*. Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010).

Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*(4), 292–330. doi:10.1080/01638530902959943

Rahimpour, M., & Hashemi, R. (2011). Textbook selection and evaluation in EFL context. *World Journal of Education*, *1*(2), 62–68. doi:10.5430/wje.v1n2p62

Sung, Y. T., Chang, T. H., Lin, W. C., Hsieh, K. S., & Chang, K. E. (2016). CRIE: An automated analyzer for Chinese texts. *Behavior Research Methods*, *48*(4), 1238–1251. doi:10.3758/s13428-015-0649-1

Sung, Y. T., Chen, J. L., Cha, J. H., Tseng, H. C., Chang, T. H., & Chang, K. E. (2015a). Constructing and validating readability models: The method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods*, *47*(2), 340–354. doi: 10.3758/s13428-014-0459-x

Sung, Y. T., Lin, W. C., Dyson, S. B., Chang, K. E., & Chen, Y. C. (2015b). Leveling L2 Texts Through Readability: Combining Multilevel Linguistic Features with the CEFR. *Modern Language Journal*, *99*(2), 371–391. doi:10.1111/modl.12213

Sung, Y. T., Chen, J. L., Lee, Y. S., Lee, Y. S., Cha, J. H., Tseng, H. C., Lin, W. C., Chang, T. H., & Chang, K. E. (2013). Investigating Chinese text readability: Linguistic features, modeling, and validation. *Chinese Journal of Psychology*, *55*(1), 75–106. doi: 10.6129/CJP.20120621. [宋曜廷, 陳茹玲, 李宜憲, 查日龢, 曾厚強, 林維駿, 張道行, & 張國恩. (2013). 中文文本可讀性探討：指標選取、模型建立與效度驗證. *中華心理學刊, 55*(1), 75-106.]

Tseng, H. C., Chen, B., Chang, T. H., & Sung, Y. T. (2019). Integrating LSA-based hierarchical conceptual space and machine learning methods for leveling the readability of domain-specific texts. *Natural Language Engineering*, *25*(3), 331-361.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York, NY: Springer Verlag.

Vinther, J. (2013). CEFR - in a critical light. In J. Colpaert, M. Simons, A. Aerts, & M. Oberhofer (Eds.), *Language testing in Europe: Time for a new framework?* (pp. 242-247), University of Antwerp, Belgium.

Williams, D. (1983). Developing criteria for textbook evaluation. *ELT Journal*, *37*(3), 251–255. doi:10.1093/elt/37.3.251

Appendix 1

Textbooks used in this study

Bellassen, J., & Liu, J. L. (2011). *Le chinois par boules de neige (Acces raisonne a la lecture du chinois)* （雪球）. Chasseneuil-du-Poitou, France: Scérén Cndp-crdp.

Bellassen, J., & Liu, J. L. (2012). *Le chinois par boules de neige (Niveau elementaire)* （雪球）. Chasseneuil-du-Poitou, France: Scérén Cndp-crdp.

Chinese Time (Ed). (2009). *Practical Chinese (实用中文)*. Shanghai: East China Normal University Press.

Editors of the Road to Success sereis (Ed). (2008-2014). *Road to Success(成功之路)*. Beijing: Beijing Language and Culture University Press.

Kantor P. (2004). *Assimil Pack Chinesisch Ohne Mühe (漢語) Volume 1.* Köln, Germany: ASSiMiL GmbH.

Kantor P. (2006). *Assimil Pack Chinesisch Ohne Mühe (漢語) Volume 2.* Köln, Germany: ASSiMiL GmbH.

Liu, X. (2007). *New Practical Chinese Reader*（新实用汉语課本）. Beijing : Beijing Language and Culture University Press.

National Taiwan Normal University (Ed). (2008). *Practical Audio-Visual Chinese* (2nd Edition) *(新版視聽華語).* Taipei: Cheng Chung Bookstore.

NTNU Extension School of Continuing Education (Ed). (2012). *New Modern Chinese (新時代華語)*. Taiwan: NTNU Extension School of Continuing Education.

Yeh, T. M. (Ed). 2008. *Far East Everyday Chinese (遠東生活華語).* Taipei: Far East Book Company.

Appendix 2 The Results of the Post-hoc Comparisons Between the Five Textbooks and the CRIE-CFL

The values across the horizontal rows and down the vertical columns are for the linguistic features in level-A and level-B textbooks, respectively. The upper right-hand and lower left-hand corners of the table provide the post-hoc comparisons of level-A and level-B textbooks, respectively.

Characters

		CRIE-CFL	Road Success	to New Chinese	Modern Practical Chinese	Boules neige	de Chinesisch	000
		118	530	124	149	404	90	
CRIE-CFL	381	—	***			***	**	
Road to Success	1372	***	—	***	***	***	***	
New Modern Chinese	258	***	***	—		***		
Practical Chinese	402		***	***	—	***	***	
Boules de neige	236	***	***		***	—	***	
Chinesisch	127	***	***	***	***	***	—	

High-level words

		CRIE-CFL	Road Success	to New Chinese	Modern Practical Chinese	Boules neige	de Chinesisch
		7	75	9	9	38	6
CRIE-CFL	45	—	***			***	
Road to Success	219	***	—	***	***	***	***
New Modern Chinese	27	***	***	—		***	
Practical Chinese	48		***	***	—	***	*
Boules de neige	33	***	***		*	—	***
Chinesisch	9	***	***	***	***	***	—

Two-character words

		CRIE-CFL	Road Success	to New Chinese	Modern Practical Chinese	Boules neige	de Chinesisch	000
		26	144	26	35	89	17	
CRIE-CFL	98	—	***		*	***	***	
Road to Success	367	***	—	***	***	***	***	
New Modern Chinese	67	***	***	—		***		
Practical Chinese	108		***	***	—	***	***	
Boules de neige	76	***	***		*	—	***	
Chinesisch	26	***	***	***	***	***	—	

Average sentence length

		CRIE-CFL	Road Success	to New Chinese	Modern Practical Chinese	Boules neige	de Chinesisch
		7.29	10.16	7.46	7.92	8.73	3.96
CRIE-CFL	8.94	—	***				***
Road to Success	10.17	***	—		***		***
New Modern Chinese	8.05	***	***	—			*
Practical Chinese	9.08		***	***	—		***
Boules de neige	8.82		***	*		—	**
Chinesisch	4.49	***	***	***	***	***	—

Note. CRIE-CFL = Chinese Readability Index Explorer for Chinese as a Foreign Language.

* $p < .05$，** $p < .01$， *** $p < .001$.

(continued)

Simple sentence ratio

		CRIE-CFL	Road Success	to New Chinese	Modern Practical Chinese	Boules neige	de Chinesisch 000
		0.89	0.39	0.93	0.79	0.74	1
CRIE-CFL	0.55	—	***		**	**	***
Road to Success	0.41	***	—	***	***	***	***
New Modern Chinese	0.76	***	***	—	**	**	
Practical Chinese	0.54		**	***	—		***
Boules de neige	0.53			**		—	***
Chinesisch	0.99	***	***	***	***	***	—

Sentences with a complex structure

		CRIE-CFL	Road Success	to New Chinese	Modern Practical Chinese	Boules neige	de Chinesisch
		4.08	21.7	5.2	5.62	16.13	1.98
CRIE-CFL	16.04	—	***			***	***
Road to Success	58.78	***	—	***	***	***	***
New Modern Chinese	10.55	***	***	—		***	**
Practical Chinese	17.56		***	***	—	***	***
Boules de neige	8.32	***	***	.052	***	—	***
Chinesisch	3.38	***	***	***	***	***	—

Content words

		CRIE-CFL	Road Success	to New Chinese	Modern Practical Chinese	Boules neige	de Chinesisch 000
		71	287	81	88	239	57
CRIE-CFL	212	—	***			***	*
Road to Success	746	***	—	***	***	***	***
New Modern Chinese	151	***	***	—		***	
Practical Chinese	224		***	**	—	***	**
Boules de neige	116	***	***	***	***	—	***
Chinesisch	78	***	***	***	***	***	—

Sentences with complex semantic categories

		CRIE-CFL	Road Success	to New Chinese	Modern Practical Chinese	Boules neige	de Chinesisch
		7.88	19.5	7.75	8.64	19.13	13.96
CRIE-CFL	17.42	—	***			***	***
Road to Success	46.82	***	—	***	***		***
New Modern Chinese	14.85		***	—		***	***
Practical Chinese	16.81		***		—	***	***
Boules de neige	8.63	***	***	***	***	—	
Chinesisch	16.98		***			***	—

Note. CRIE-CFL = Chinese Readability Index Explorer for Chinese as a Foreign Language.

* *p* < .05，** *p* < .01， *** *p* < .001.

(continued)

Complex semantic categories

		CRIE-CFL	Road to	New Modern	Practical	Boules de	Chinesisch	000
			Success	Chinese	Chinese	neige		
		2.92	5.84	2.66	3.1	6.52	6.42	
CRIE-CFL	5.74	—	***			**	***	
Road to Success	14.33	***	—	***	***			
New Modern Chinese	5.12	***	***	—		**	***	
Practical Chinese	5.36		***		—	**	***	
Boules de neige	2.66	***	***	***	***	—		
Chinesisch	7.91	***	***	***	***	***	—	

Conjunctions

		CRIE-CFL	Road to	New Modern	Practical	Boules de	Chinesisch
			Success	Chinese	Chinese	neige	
		1.38	11.51	1.5	1.85	4.81	0.9
CRIE-CFL	7.49	—	***			**	
Road to Success	29.45	***	—	***	***	***	***
New Modern Chinese	5.85		***	—		**	
Practical Chinese	9.14		***	*	—	*	**
Boules de neige	6.21		***			—	**
Chinesisch	1.32	***	***	***	***	***	—

Positive conjunctions

		CRIE-CFL	Road to	New Modern	Practical	Boules de	Chinesisch	000
			Success	Chinese	Chinese	neige		
		0.93	7.31	0.75	1.32	2.63	0.51	
CRIE-CFL	5.22	—	***			*	*	
Road to Success	19.06	***	—	***	***	***	***	
New Modern Chinese	3.25	**	***	—		*		
Practical Chinese	6.24		***	***	—		**	
Boules de neige	4.58		***			—	**	
Chinesisch	0.84	***	***	***	***	***	—	

Negative conjunctions

		CRIE-CFL	Road to	New Modern	Practical	Boules de	Chinesisch
			Success	Chinese	Chinese	neige	
		0.43	3.75	0.80	0.52	1.88	0.49
CRIE-CFL	2.25	—	***			*	
Road to Success	9.35	***	—	***	***	**	***
New Modern Chinese	1.70		***	—			
Practical Chinese	2.70		***		—		
Boules de neige	1.53		***			—	.055
Chinesisch	0.50	***	***	**	***	*	—

Note. CRIE-CFL = Chinese Readability Index Explorer for Chinese as a Foreign Language.

* *p* < .05，** *p* < .01， *** *p* < .001.