Behav Res (2016) 48:1238–1251
DOI 10.3758/s13428-015-0649-1

CRIE: An automated analyzer for Chinese texts

Yao-Ting Sung[1] · **Tao-Hsing Chang**[2] · **Wei-Chun Lin**[1] · **Kuan-Sheng Hsieh**[1] · **Kuo-En Chang**[3]

Published online: 30 September 2015

Abstract Textual analysis has been applied to various fields, such as discourse analysis, corpus studies, text leveling, and automated essay evaluation. Several tools have been developed for analyzing texts written in alphabetic languages such as English and Spanish. However, currently there is no tool available for analyzing Chinese-language texts. This article introduces a tool for the automated analysis of simplified and traditional Chinese texts, called the Chinese Readability Index Explorer (CRIE). Composed of four subsystems and incorporating 82 multilevel linguistic features, CRIE is able to conduct the major tasks of segmentation, syntactic parsing, and feature extraction. Furthermore, the integration of linguistic features with machine learning models enables CRIE to provide leveling and diagnostic information for texts in language arts, texts for learning Chinese as a foreign language, and texts with domain knowledge. The usage and validation of the functions provided by CRIE are also introduced.

Keywords Chinese text analysis · Linguistic feature · CRIE · Readability

✉ Tao-Hsing Chang
changth@kuas.edu.tw

[1] Department of Educational Psychology and Counseling, National Taiwan Normal University, Taipei, Taiwan

[2] Department of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences, 415, Jiangong Road, Kaohsiung City 80778, Taiwan, Republic of China

[3] Graduate Institute of Information and Computer Education, National Taiwan Normal University, Taipei, Taiwan

Textual analysis, which is the analysis of text using algorithmic techniques, is an important and commonly used tool in many areas of language-related research, including discourse analysis, language acquisition, corpus studies, and readability analysis. Textual-analysis tools enable researchers to quickly and efficiently analyze large quantities of data when seeking certain information. For example, constructing or analyzing a text corpus or carrying out a longitudinal analysis of a language-acquisition database requires the use of a textual-analysis system (Bååth, 2010; Marsden, Myles, Rule, & Mitchell, 2003). Automated essay scoring systems must also make use of a large volume of texts in order to carry out accurate assessments of essay grades (Attali & Burstein, 2006; Burstein, 2003; Foltz, Laham, & Landauer, 1999; Rudner & Liang, 2002). Some researchers have investigated how scientific texts address causal relationships by analyzing these texts using linguistic features that are designed to clarify the causal relationships in language expression (Smolkin, McTigue, & Yeh, 2013). All such research relies on text analyzers that can cope with a large volume of text and a large number of features.

Several automated textual-analysis tools have been proposed previously, the best known of which is the Coh-metrix online text analyzer for the English language (McNamara, Louwerse, McCarthy, & Graesser, 2010; McNamara, Graesser, McCarthy, & Cai, 2014). Some researchers (McNamara et al., 2014) believe that analyzing only the superficial linguistic features of a text is insufficient for understanding the complex process of reading comprehension. Instead, an approach that incorporates multiple levels of linguistic features can better represent the reading comprehension process and its different components. Thus, the Coh-metrix applies computational linguistics methods with the aid of a syntax parser and corpora to analyze words, sentences, and large textual structures, such as paragraphs and discourse. The

Coh-metrix currently consists of 108 features, including word information (e.g., lexical frequency and density), syntactical complexity (e.g., noun-phrase density), cohesion (e.g., conjunctions), and semantic relations.

However, the effectiveness and plausibility of textual features vary between languages due to their linguistic peculiarities; thus, textual-analysis tools and analytical features should be developed based on the linguistic structures and properties of a specific language. For example, EsPal (Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013) is an online text analyzer designed to work with Spanish texts. In addition to analyzing texts for common features such as lexical frequency, word type, and parts of speech, it takes into account several dialects of Spanish and analyzes such features as word formation rules, phonological structure, and trigram and bigram relationships between the individual letters of a word.

Despite the Chinese language being used by more than one billion people (Graddol, 2004; Lewis, 2009), tools designed specifically for Chinese texts are still rare. Chinese Coh-metrix and the Chinese Linguistic Inquiry and Word Count (CLIWC) dictionary (Huang, Chung, Hui, Lin, Seih, Lam, Chen, Bond, & Pennebaker, 2012) are two of the few systems for Chinese texts. According to Huang et al. (2012), CLIWC analyzes the thoughts, feelings, and personalities within the texts by counting specific types of words. Among the 80 features used in CLIWC, only 22 of them are categorized as general linguistic features because CLIWC focuses on psychological research. The rest are under the psychological, personalized, or punctuation categories. CRIE, on the other hand, covers a more comprehensive set of features, making a total of 81 linguistic features and one domain-specific feature. Furthermore, CLIWC analyzes texts solely by counting words. CRIE not only uses other methods when analyzing texts, such as ratios and logarithms, but instead of remaining at the lexical level also includes features spanning semantic, syntactic, and cohesion levels.

Chinese Coh-metrix is another system available for Chinese textual analysis. According to the information published on their website, Chinese Coh-metrix was developed based on Coh-metrix, and aims to analyze cohesion and coherence within Chinese texts and therefore incorporate some features designed for Chinese lexical and textual properties, including part-of-speech and frequency, cohesion, word information, connectives, and sentence structures. As far as we know, Chinese Coh-metrix is still under development, and has not yet been made public. The specialized nature of CLIWC and the ongoing development of Chinese Coh-metrix reveal that there is still a dearth of tools that can quantitatively analyze Chinese texts.

The large morphological and syntactic differences between the characteristics of Chinese and alphabetic languages make specific types of textual analyzing tools preferable. In terms of morphological structures, for example, Chinese does not have a clear word boundary as alphabetic writing systems do, so when it comes to identifying individual words, the first step in textual analysis for any language text, an extra analyzing tool for Chinese word segmentation is required. Another notable characteristic that distinguishes Chinese from alphabetic writing systems is its character structure. Chinese uses strokes and components to constitute characters, and words are composed of one or more characters. Each Chinese character is monosyllabic. By contrast, in alphabetic languages, letters are the basic components of words, and words usually are multi-syllabic. Thus in Chinese, word length cannot be evaluated in terms of letters or syllables, both of which are prevalent when analyzing alphabetic writing systems, and consequently such evaluations are doomed to fail if applied to Chinese texts because of the fundamental difference in word/character structures. Similarly, in terms of syntactic structures, Chinese lacks inflectional morphemes, tense, and agreement markers, and relies on strong syntactic constraints on word order with no overt case-marking systems. This makes textual analyzing tools designed for alphabetic writing systems incompatible with Chinese texts. Although some features, such as word counts and average sentence length, are conceptually equivalent across all languages, even these features cannot be directly applied to Chinese texts because such texts must first undergo segmentation and parsing. Further, each language has features unique to it, such as counting for Chinese character strokes.

This article introduces the Chinese Readability Index Explorer (CRIE), which is an innovative Chinese textual analyzer that integrates preprocessing tools for segmentation and parsing, abundant multilevel linguistic features, and text-leveling tools. CRIE has three main distinct characteristics: First, based on multilevel linguistic features, CRIE was designed specifically for the Chinese language and provides users with deeper levels of textual information as well as services such as word segmentation and part-of-speech tagging. Second, to meet the specific needs of Chinese-language teachers and learners of Chinese as a foreign language (CFL), CRIE not only analyzes Chinese texts for native speakers but can also make use of features that characterize foreign-language reading materials for CFL, and calculates the values of the linguistic features that fit such texts. Crossley, Greenfield, and McNamara (2008) point out that features that influence how native speakers and L2 learners comprehend texts may be completely different, and therefore it is worth noting who the text is meant to be read by. Thus, we distinguished two sets of features: general linguistic features, and features meant exclusively for L2 learners. Third, in addition to basic functions of textual analysis, CRIE provides users with advanced functions for presenting information about levels of text difficulty or complexity.

While CRIE supports both traditional and simplified Chinese, with the large majority of functions being identical for both, this study uses traditional Chinese to express and illustrate these functions. In the rare cases where there are

differences for simplified and traditional Chinese, we specifically mention it.

Architecture of the Chinese Readability Index Explorer (CRIE) System

Figure 1 shows the main system architecture of CRIE, which mostly involves preprocessing, textual-features analysis, and applications. The corpora, preprocessing, and textual-features analysis associated with CRIE are described in detail below.

Corpora used by the CRIE system

CRIE includes several data sets. The preprocessing stage employs a collection of 11.68 million Chinese words and 61,087 grammar trees derived from the following four data sets: Sinica Balanced Corpus 4.0 (Huang, Chen, Chen, Wei, & Chang, 1997), Sinica Treebank 3.1 (Chen, Tsai, Chen, & Huang, 1999), CKIP Chinese Electronic Dictionary (CKIP, 1993), and Gigaword (Huang, 2009). We developed dictionaries and lexical information (e.g., frequency and part-of-speech) for traditional Chinese based on the Sinica Balanced Corpus 4.0 and CKIP. We extracted two simplified Chinese corpora, Xinhua News Agency and Lianhe Zaobao from Gigaword, to develop dictionaries and lexical information for simplified Chinese text analysis. Treebank, which contains traditional Chinese data, was used as the source of syntactic rules for the parser.

In preparation for the textual-features analysis stage, we constructed a text corpus of 4,332 texts that was divided into two parts: The first part was the 2,754 texts selected from textbooks of three publishers in Taiwan and covers three domains (language arts, social sciences, and natural sciences) and nine grades (school grades 1–9), and the second part was the 1,578 texts selected from teaching materials for CFL learners.

Preprocessing

In conventional text analyzers, preprocessing involves tagging the parts of speech of words and parsing each sentence into a parsing tree, which are denoted part-of-speech tagging and sentence parsing, respectively. For Chinese texts, preprocessing needs to segment sentences into words because there are no spaces between words; this process is called word segmentation.

Word segmentation and part-of-speech tagging are not straightforward, and this has prompted numerous studies of natural language processing (Manning & Klein, 2002; Petrov, Barrett, Thibaux, & Klein, 2006; Tsai & Chen, 2004). In addition, some systems for sentence parsing have been designed, such as the CKIP word segmentator and parser (Tsai & Chen, 2004), the Stanford parser (Manning & Klein, 2002), and the Berkeley parser (Petrov et al., 2006). However, since these systems do not provide the complete functions needed for Chinese textual analysis, we designed the Word Extractor for Chinese Analysis (WECAn), which is a word segmentation and part-of-speech tagging tool for Chinese text analysis (Chang, Sung, & Lee, 2012), and HanParser, which is a Chinese grammar parser (Chang, Liu, Chen, Sung, & Su, 2013a).

Two previous studies evaluated the performances of WECAn and HanParser. Chang et al. (2012) employed WECAn to segment sentences from the Sinica Balanced Corpus 4.0 into words, with the experimental results indicating that the accuracy rates of word segmentation and part-of-speech tagging using WECAn were 0.93 and 0.92, respectively. Under the same condition, the accuracy of a previous

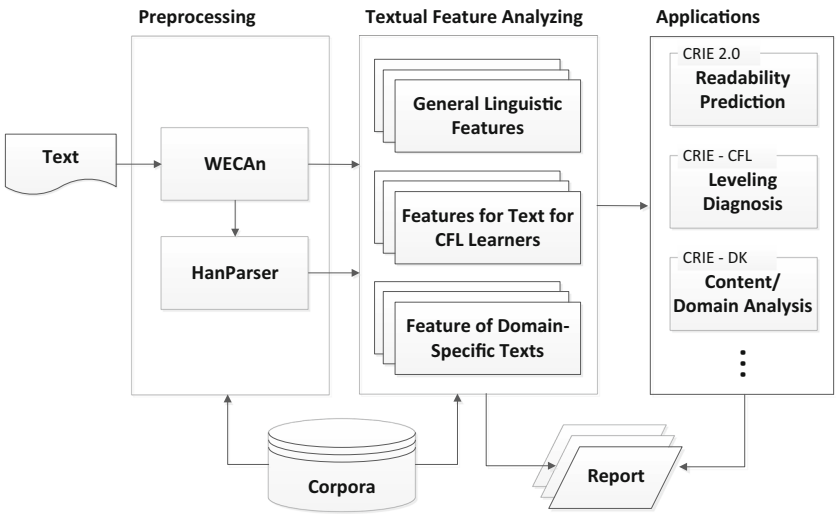

Fig. 1 CRIE framework and applications. WECAn = Word Extractor for Chinese Analysis; CRIE = Chinese Readability Index Explorer; CFL = Chinese as a foreign language; DK = Domain Knowledge

Fig. 2 The Chinese Readability Index Explorer (CRIE) system-login screenshot

approach based on the Conditional Random Field model (Lafferty, McCallum, & Pereira, 2001) was 0.90. In addition, Chang, Liu, Chen, Sung, & Su (2013a) estimated that the accuracy rate of using HanParser to parse sentences from Sinica Treebank 3.0 was 0.86. The following sections describe how these two tools solve a range of problems.

Word segmentation WECAn performs word segmentation in three stages: matching, modification, and unknown-word extraction. The matching stage mainly uses maximum matching algorithms (Xue, 2003) to segment sentences into words. For example, the sentence "今晚馬戲團的表演取消" ("Tonight's circus performance has been canceled") would be divided up into "今晚" ("tonight"), "馬戲團" ("circus"), "的" ("'s"), "表演" ("performance"), and "取消" ("canceled").

The maximum matching method requires the use of a large dictionary. WECAn uses a Chinese dictionary in which lexical information is derived from four of the corpora mentioned above: Sinica Balanced Corpus 4.0, Sinica Treebank 3.1, and the CKIP Chinese Electronic Dictionary. Taking into account the different words in simplified and traditional Chinese, WECAn also uses the Xinhua News Agency and Lianhe Zaobao corpora of Gigaword to increase vocabulary and linguistic information (e.g., part-of-speech and word frequency). WECAn also contains a collection of sayings and proverbs, originally sourced from the Dictionary of Chinese Idioms (http://dict.idioms.moe.edu.tw/cydic/index.htm), which are often used when writing domain-specific texts (e.g., student's essays), and so the use of this collection increases the efficiency and precision of WECAn-based analyses.

The second stage of preprocessing is modification, which involves correcting errors produced in the matching stage. These errors can be classified into two categories. The first is ambiguous segmentation. For example, the sentence "到處有空地" ("There is unused land everywhere") may be segmented into one of two alternatives: (1) "到處" ("everywhere"), "有空" ("have time"), and "地" ("land"), or (2) "到處" ("everywhere"), "有" ("has"), and "空地" ("unused land"). WECAn would select the second alternative at the modification stage. The second type of error is caused by word reduplication, a phenomenon that is frequent in Chinese but extremely rare in Western languages, because reduplication is a common rhetorical strategy for emphasizing mood. For example, in order to emphasize the idea of happiness in the sentence "高興" ("glad") 地 ("-ly") "去" ("go") "購物" ("shopping"), we could instead write "高高興興" ("glad") "地" ("-ly") "去" ("go") "購物" ("shopping"). The reduplicated word "高高興興" ("glad") cannot be identified as a word at the matching stage because it does not appear in the dictionary. WECAn uses a large database of character-reduplication patterns to detect and correct for the many kinds of reduplicated words that occur in Chinese at the second stage. The reduplicated word "高高興興", for example, would be cut into the three following pieces: "高", "高興", and "興". WECAn then uses its large set of rules to detect and correct these three pieces. In this case, the fragments are consistent with the pattern "A-AB-B," of which A and B are both single-character words, and AB is the dual-character word of A combined with B. According to this rule, WECAn will take "高", "高興", and "興" and combine them back into the reduplicated form "高高興興".

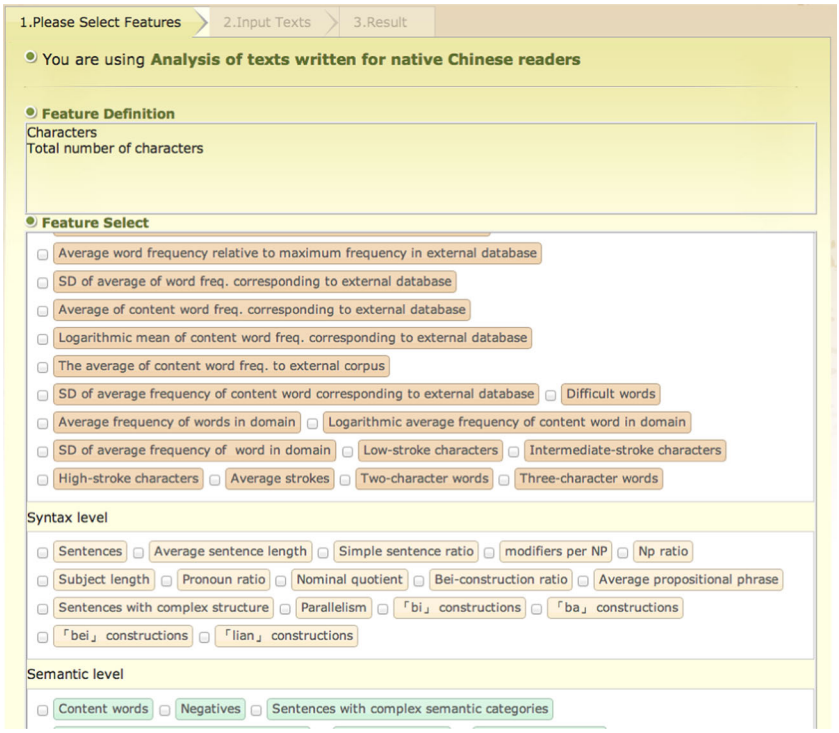

Fig. 3 A screenshot of the feature-selection utility for the Chinese Readability Index Explorer

In addition, sentences containing unknown words—which do not appear in dictionaries—cannot be segmented correctly by maximum matching methods. Moreover, such words are often important proper nouns, and so unknown-word extraction is a crucial step for Chinese textual analysis. WECAn applies the strict phrase likelihood ratio (SPLR) algorithm (Chang et al., 2012) to extract unknown words from the text. The SPLR method is more effective than other algorithms at avoiding treating nonwords as unknown words: the recall and precision rates for recognizing proper names translated from foreign languages into Chinese in SPLR reached 0.84 and 0.90, respectively (Chang et al., 2012).

Part-of-speech tagging Many textual features are related to the parts of speech of words. The part of speech of a word would be tagged by consulting a dictionary if the word is only referred to as one part of speech (e.g., "necklace" is only ever a noun). However, many Chinese words can be tagged with different parts of speech. WECAn utilizes bigram models (Bassiou & Kotropoulos, 2011) to identify the parts of speech

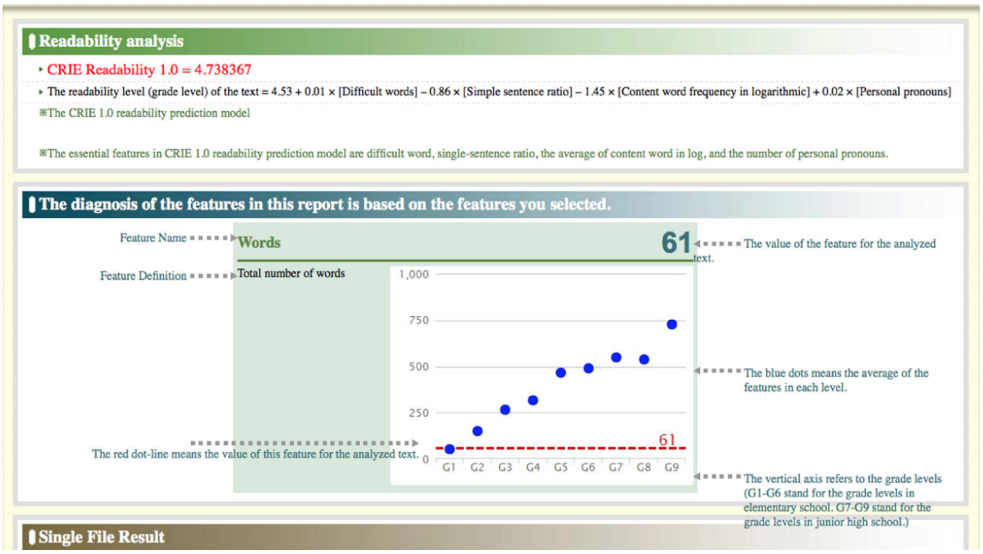

Fig. 4 A screenshot of analysis-results for the Chinese Readability Index Explorer (CRIE)

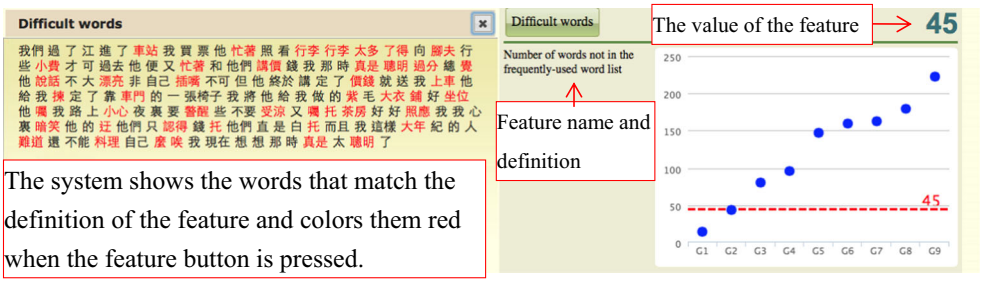

Fig. 5 A screenshot of a diagnostic-report of the Chinese Readability Index Explorer

of words. The model evaluates all possible part-of-speech combinations of a word pair. The probabilities of the occurrence of the combinations are computed and then the two words of the pair are marked with the two parts of speech of the combination that yields the highest probability. WECAn also uses a method based on morphological and contextual rules (Chang et al., 2012) to tag the parts of speech of unknown words.

Sentence parsing For textual analysis, we developed a parsing tool called HanParser that comprises two main modules: a grammar-rule generator and a grammar-tree constructor. Based on the framework of probabilistic context-free grammar (PCFG; Johnson, 1998), the generator extracts PCFG grammatical rules from Sinica Treebank 3.1 (Chen et al., 1999). Moreover, because HanParser also employs the Cocke-Younger-Kasami (CYK) algorithm (Ney, 1991) to increase the efficiency of parsing sentences, the generator revises some rules and ensures that all grammatical rules can be used in the CYK approach.

The grammar constructor of HanParser adopts a bottom-up approach for building the grammar tree of a sentence. This approach is based on the CYK algorithm and dynamic programming approach (Ney, 1991). Previous studies (Manning & Klein, 2002; Petrov et al., 2006) have indicated that these algorithms can construct the grammar tree of a sentence both efficiently and precisely. For instance, HanParser with CYK can achieve the same accuracy as the results without CYK while spending 85 % less time on calculations.

Textual-features analysis

One of the main distinguishing merits of the CRIE system is that its multilevel linguistic features were designed based on theories from reading psychology and linguistics. The linguistic features can be divided into two types: (1) general linguistic features, suitable for analyzing Chinese reading materials for native speakers, and (2) features of CFL texts, aimed at analyzing textual features of reading materials for CFL learners. These linguistic features are listed in Table 1. In addition to linguistic features, we developed a domain-knowledge conceptual vocabulary list in which the words are selected by latent semantic analysis (LSA; Landauer & Dumais, 1997) and represent domain-specific knowledge for readability predictions. The main characteristics of and the idea behind each feature are explained below.

Fig. 6 The results of a Chinese Readability Index Explorer – Domain Knowledge (CRIE-DK) assessment of webpage readability

General linguistic features

A moderately wide range of factors can influence textual structure. In order to create a systematic framework that is supported by theories from reading psychology and linguistics, we analyzed factors that might have affected the reading comprehension process defined by a framework of the following four levels of 70 linguistic features: words, syntax, semantics, and discourse cohesion.

Word level: Word level can be divided into five subcategories: character complexity, word length, word frequency, word count, and lexical richness. The scoring methods for the traditional Chinese and simplified Chinese features are usually the same. Although some conceptual vocabulary differ (e.g., the word "printer" is 表機 in traditional and 印机 in simplified), most words are identical.

Character complexity and word length: Characters and words are closely related, and so we discuss them together. Chinese characters are composed of strokes and components, and the number of strokes used to form single characters varies greatly. Thus, research into Chinese word length should consider both the stroke count and the character count of each word as analytical units (Just & Carpenter, 1987; Su & Samuels, 2010). For this reason we developed two types of features: one is word-count-based (e.g., two-character word count) and the other is word-complexity-based (e.g., intermediate-stroke-characters count); both types of features are unique to the analysis of the Chinese writing system. However, because simplified and traditional characters are morphologically distinct in nature, this intrinsic structural difference causes the results of the features related to counting strokes to vary between the two versions of Chinese. These features are Low-stroke-count characters, Intermediate-stroke-count characters, High-stroke-count characters, and Average strokes.

Word frequency: Many studies have indicated that word frequency is related to word response time, with participants responding faster to words that appear more often (Forster & Chambers, 1973; Whaley, 1978). Some researchers believe that frequency effects are more appropriately presented using logarithmic or exponential functions (Balota & Chumbley, 1984; McCusker, 1977). Fry, Kress, and Fountoukidis (1993) indicated that the most frequent 100 Chinese words can already account for 50 % of most texts' vocabulary, and the most frequent 300 can account for 65 %. Based on this principle, Chall and Dale (1995) used word frequency order to develop a feature for their readability formula (i.e., the Chall-Dale Formula). They then created a list for the 3,000 most common words, and any word not included in these 3,000 words was designated as a difficult word. How many of these difficult words a text contains was then developed into a feature for predicting readability. We considered this method while developing our own feature. After creating a word frequency list for the balanced corpus used in our study (CKIP, 1998), we used the first 3,000 words to create a "commonly used words" list. Words that don't appear in this list were designated as difficult words, which in turn were used to develop our difficult words feature.

Word count: The third subcategory is the word count. Longer texts generally impose a larger cognitive burden on beginner readers, which is why we consider word counts for specific lexical categories, such as verbs and adjectives.

Lexical richness: Lexical richness refers to the degree of variation present among words used in a text. Researchers can determine the lexical richness of a text by counting the usage rate of a specific vocabulary. We quantified the degree of lexical repetition by combining individual word counts and word-type counts to calculate type–token ratios.

Semantic level The semantic level mainly involves the analysis of three aspects of word meaning in texts: core meaning, pragmatic function, and semantic category. It has been shown that readers spend less time processing function words than content words (Carpenter & Just, 1983), and spend more time processing sentences that contain a larger number of content words. Researchers on discourse analysis now believe that negations do not merely serve to express semantic opposition, but also have a pragmatic function. Polysemous words are often the most difficult to deal with but are also the most important linguistic feature and, since they have relatively complex semantic structures, they are more likely to introduce discrepancies. Because words with several meanings have more semantic categories, we used semantic categorization to determine the degree of complexity of polysemes. A text's overall semantic categorization reflects the degree of variation in the text's semantic categories.

Syntactic level The complexity of a sentence was determined mainly based on the diversity of syntactic structures. Complex sentences are usually longer, structurally intense, and impose a higher cognitive burden on the reader, so we developed features that measure structural complexity, including the proportions of simple sentences, noun phrases, prepositional phrases, grammatical subject length, and modifier length.

Cohesion level Cohesion refers to the grammatical and lexical interrelationships between the words in a sentence or text. Such relationships result in a string of sentences together representing a text with a unified meaning. Cohesion is also an important ingredient in the structure of mental models. A reader needs to construct semantic interpretations and mental models in order to obtain a deep understanding of the discourse (Lehnert & Ringle, 1982). This is also why many researchers believe that cohesion affects comprehension (Benjamin, 2012). We focused on three kinds of cohesion—conjunctions, references, and figures of speech—and developed new features for cohesion analysis, including the ratio of similes, and complex conjunctions.

Features of texts for Chinese as a foreign language (CFL) learners

The materials used to teach CFL learners are similar to texts used for native speakers, and hence some of the general linguistic features mentioned above can also be applied to the textual analysis of texts for CFL learners. However, CFL texts exhibit some differences in vocabulary difficulty and sentence patterns, so CRIE includes features developed specifically for the analysis of CFL texts. We defined 50 features that have been recognized as important for the level of difficulty/complexity of CFL texts through literature review. Then, based on a corpus composed of 1,578 CFL texts for CFL that had been leveled by experts according to the proficiency levels in the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001), these 50 features were examined via trend analysis to see if they were sensitive to changes in the text difficulty level (Hsieh, Lin, Dyson, Liu, & Sung, 2014). Thirty-eight features were selected for being the most sensitive to the difficulty and complexity of CFL texts, as listed in Table 1. Below we describe features that are unique to the CRIE-CFL system.

Word level Based on the list of 8,000 Chinese words published by the Steering Committee for the Test of Proficiency-Huayu (SC-TOP) (Chang, 2012), words commonly used in CFL learners' reading materials can be divided into five levels: breakthrough, waystage, threshold, vantage, and effective operational proficiency levels. These five features count the number of vocabularies at the relevant level (e.g., the threshold vocabulary measures the vocabularies that are consistent with the threshold level of SC-TOP). The sixth feature, the average of vocabulary levels, is the average of the vocabulary difficulty ratings of a text.

Difficult texts may contain both difficult vocabulary and words that students have already studied. However, even if difficult vocabulary is relatively rare, such vocabulary can have a disproportionately large impact on the overall difficulty of the text since just one difficult word can make an otherwise simple sentence incomprehensible. A calculation based on averages may therefore underestimate the effect of difficult vocabulary. This is why we also use a weighted difficulty feature to quadratically increase the high-difficulty scores for higher-level vocabulary. The mean square of the vocabulary levels feature takes the sum of squared difficulty values and divides this by the total word count. This approach makes the quantified lexical difficulty more consistent with the degree of text difficulty that is actually experienced by CFL learners.

Syntactic level In addition to the need for vocabulary features designed for CFL learners, some sentence patterns can cause problems for CFL learners because of their particular word order or their implicit meanings. For example, the words "*ba*" ("把") and "*bei*" ("被") are used to construct "verb final" sentence patterns, where the receiver of the action is caused to undergo some effect or enter some state, thus emphasizing the effect of the action on the receiver. CFL learners often find these sorts of sentence patterns difficult to grasp (Cui, 1995; Zhao, 2011). Moreover, Chinese does not express comparisons morphologically; instead, comparisons are generally expressed using any of a variety of prepositions or prepositional phrases. A common expression used to express a difference in quality or magnitude is "*bi*" ("比"). There is also the word "*lian*" ("連"), which suggests an implicit comparison. This type of word is used to indicate an extreme condition and thus creates emphasis, which may sometimes even involve moving the object toward the beginning of the sentence. Sentence patterns that use a specific preposition to express comparison can also pose some problems for CFL learners. We therefore developed syntactic features to detect the number of sentences containing the "*ba*," "*bei*," "*lian*," and "*bi*" constructions.

Features of domain-specific texts

The difference between texts with and without specific knowledge of a domain is that the former aim to transmit a range of specific knowledge. Domain-specific texts employ specific terminology (e.g., proper nouns and lingo) that is not commonly found in everyday language. For example, an explanation of photosynthesis could contain a large number of proper noun words such as "photoautotrophs," "chlorophyll," and "chloroplasts." Such words have domain-specific knowledge, and the words described by them (e.g., photosynthesis) are referred to as "concept words."

According to Chang, Sung, and Lee (2013b), every domain-specific word must have two values: domain specificity and conceptual difficulty. LSA was used in the present study to construct the semantic space of conceptual words of textbooks, which were designed based on curriculum standards, with their content representing the general knowledge levels of students. Chang et al. also proposed a technique for computing these feature values and determining whether or not a conceptual word belongs to a specific domain and its conceptual difficulty. The word count of a domain-specific text will be transformed into vectors that represent the conceptual difficulty of all domain-specific terminology in the text.

Application and validation of the CRIE system

WECAn and HanParser web service

WECAn and HanParser are designed as application programming interface (API) tools for CRIE. The online versions of WECAn and HanParser can be used for Chinese word

segmentation, part-of-speech tagging, and syntax-tree parsing. Users can upload their texts and choose the services they need, and then the processing results will be shown on the website or can be downloaded as an ASCII file.

CRIE 2.0

Textual-features analysis Based on the preprocessing output, CRIE calculates the descriptive statistical information about the 70 linguistic features of word, syntax, semantic, and cohesion in each text or an aggregated result for a batch of texts. Those results may be presented on the webpages or exported to an Excel file.

Text readability analysis CRIE 2.0 was developed to measure the readability of Chinese written texts for native speakers. CRIE 2.0 combines 70 linguistic features with nonlinear mathematical models to create readability prediction models, and then classifies the input text into the corresponding grade level (1–9). The training corpus is the Chinese-language textbooks for Taiwanese primary and secondary schools published by three publishers in Taiwan. CRIE captured the linguistic features of these texts, which were then used to train and test the prediction models. Most conventional readability models or formulae use small numbers of features to create linear formulae, such as stepwise regression, whereas CRIE uses a support vector machine (SVM) as its prediction model. SVMs are common automated classification models that can learn to recognize the relationships between data properties and defined text categories, and can map nonlinear data onto high-dimensionality spaces (Lin & Chen, 2011; Pal & Foody, 2010; Vapnik & Chervonenkis, 1974). In this way, an SVM can integrate many features so as to provide a better classification of nonlinear data.

Previous research (Sung et al., 2013; Sung, Chen, et al., 2015) validated the effectiveness of CRIE 2.0. Using the SVM model along with 24 linguistic features, the readability prediction of language arts textbooks used in grades 1–6 reached an accuracy of 72.92 %. Sung, Chen, et al. (2015) confirmed that the prediction accuracy of the SVM model along with 32 linguistic features outperformed the prediction accuracy of discriminate analysis models in classifying language arts textbooks used in grades 1–6. In a preliminary study, Sung, Lin, and Tseng (2014) further found that CRIE 2.0 achieved similar results for language arts texts used in grades 1–9.

CRIE-CFL

The CRIE-CFL system has three functions: textual-features analysis for CFL texts, carrying out readability leveling in accordance with the CEFR, and providing diagnostic information for texts. These functions are described below.

Textual-features analysis of CFL texts The CRIE-CFL system can be used to analyze the 38 linguistic features especially designed for CFL texts, and it will produce a report of features calculated based on the corpus of CFL texts. Users who are interested in other features beyond these 38 may also use the CRIE 2.0 system—which employed a corpora of Chinese as a native language texts—to analyze the features of interest to them.

Leveling CFL texts Using the SVM model, CRIE-CFL learned the relationships among the 38 features and the CEFR proficiency levels of the 1,578 CFL texts in the corpus, and constructed a readability model for predicting the readability levels of CFL texts. CRIE-CFL creates an assessment report of a text's predicted CEFR level and a description of this level. The text can be further adjusted by language instructors and researchers using the system's diagnostic function and textual-features data. This system can deliver a more objective analysis than the methods that rely on user experience to determine levels.

Textual-features diagnosis Obtaining detailed information about texts is important for CFL text authors, editors, or teachers when they are writing, editing, or selecting teaching materials, respectively. Such users can use this function to indicate the position of specific levels or parts of speech in a text and then use this information in further applications. For example, a teacher could employ this function to modify certain difficult words or sentence patterns, which should help the teacher adjust the content of teaching materials to fit the specific educational requirements of the CFL learners. CFL learners can also use this function to understand what vocabulary is most suitable for themselves.

Cha, Chen, Chang, and Sung (2013) validated the effectiveness of CRIE-CFL. Their results for the CRIE-CFL prediction of CEFR levels showed an accuracy of 72 %. Hsieh et al. (2014) further improved the accuracy of leveling to 75 % by upgrading the preprocessing subsystems, and reached an average accuracy of 90 % when the six CEFR levels were combined into three broad divisions.

CRIE-DK

The CRIE-DK system assesses the knowledge content levels of texts, such as the readability and conceptual difficulty of a webpage or e-book. Previous researchers (e.g., Friedman & Hoffman-Goetz, 2006; Miltsakaki, 2009) used linguistic features to construct readability models to assess the readability of webpages. However, readability levels cannot be determined accurately solely by considering shallow linguistic features. CRIE-DK uses word conceptual difficulty and our SVM readability model to construct a readability model for predicting webpage readability.

Tseng, Chang, Chen, and Sung (2014) validated the performance of CRIE-DK by demonstrating a readability accuracy

of 80.83 % for natural science texts used in grades 3–9. The capability of CRIE-DK is currently being expanded to include texts from other domains.

How to use the CRIE system

The CRIE website can be accessed at http://www.chinesereadability.net/CRIE. After registration, users can choose to use CRIE 2.0, CRIE-CFL, or CRIE-DK, depending on the type of text they wish to analyze. Alternatively, a user can log into WECAn and HanParser directly to have texts segmented and tagged. Figure 2 shows the login screen.

Figure 3 shows the interface displayed to users for choosing the features they need and reading the definition of those features. Users select the textual features to be analyzed by clicking. On the next page, users can paste a single text directly or upload multiple texts compressed in a zip file. Once a text has been uploaded, the system carries out the analysis using whatever textual features the user has selected.

When the text has passed through the preprocessing and textual-features analyses of the two systems described above, an analysis report of the values and readability reference values for each feature is produced. These values can give users a clear idea of the text's structure and the distribution of its features. In addition, the system described in this study can analyze texts of any length, enables users to select many analytical items, and will save analysis results in an Excel file. The analysis results produced by these simple and rapid functions can be used in research.

Figure 4 shows an example CRIE-CFL analysis report, which displays values for all features of the analyzed text, as well as the mean value of each feature. The red dashed line shows the relative position of feature values for the analyzed text. This system also provides a diagnostic function for features that highlights the lexical items with specific features (e.g., which items are difficult words) within a text. Clicking on buttons within the report will open a window in which the words with specific features within the text are colored red, as shown in Fig. 5.

Figure 6 shows a screenshot of using the CRIE-DK system to analyze the leveling of websites. Users log in and enter a keyword that is then sent to the Microsoft Bing search engine. CREI-DK subsequently starts analyzing webpages returned by Bing's keyword search and uses the obtained concept vector to predict the readability level of each returned webpage.

The CRIE-DK report is divided into four columns. The first (entitled "Website Names") contains titles for all the returned webpages. The second ("Website Summary") contains part of the contents of each webpage. The third ("Readability Level") shows the website readability levels (as school grades, from 1–9) as predicted by the SVM readability model. In order to make the prediction of webpage readability more reliable, the webpage's total number of concept words and the proportion of concept words must both be above certain thresholds. Any webpage that meets these criteria will be assigned a readability level. The fourth column contains links to the returned websites. Clicking on the "Go" link will take the user to that website.

Discussion and conclusions

The importance of textual-features analysis has been recognized in various fields, which has prompted proposals for tools that analyze linguistic features (Duchon et al., 2013; McNamara et al., 2010; McNamara et al., 2014). CRIE is the first large-scale text analyzer specifically developed for the Chinese language, and it can analyze multilevel linguistic features of various types of texts. The current evaluation studies indicate that CRIE is an efficient and valid tool for text analysis. We believe that the functions of CRIE will be helpful for researchers in the fields of psychology and language as well as practitioners who are interested in language teaching and learning.

The capacity and reliability of CRIE will be enhanced in three main directions in the future: the accuracy of segmentation and parsing, the development of new features, and the application (applied studies) based on the extracted values of features. First, the accuracy of segmentation greatly influences the performance of CRIE because it is the first step of textual analysis, and so we will improve the segmenting capability of WECAn in detecting various terms such as personal names and four-character Chinese idioms. Second, we will develop a larger number of diverse linguistic features (e.g., the analysis of zero pronouns, cohesion, and genre) in order to deepen and broaden the textual features that can be detected in Chinese texts. This will also improve the ability of CRIE to carry out textual analyses of text domain and genre (e.g., psychology, history, or action). The third direction will involve developing more readability models for various texts and age stages. This requires additional studies of the integration of general linguistic features and domain-specific conceptual features, and studies determining the importance of different linguistic features in different developmental stages.

Acknowledgments This research is partially supported by the Aim for the Top University Project of the National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan, R.O.C. and the International Research-Intensive Center of Excellence Program of NTNU and Ministry of Science and Technology, Taiwan, R.O.C. under Grant NSC 103-2911-I-003-301. The authors are also grateful for the support of the Ministry of Science and Technology, Taiwan, under Grant NSC 101-2511-S-003-047-MY3.

Appendix

Table 1 Linguistic Features Implemented in CRIE

Feature	Definition
Word	
Character complexity	
Low-stroke-count characters+*	Total number of characters containing 1 to 10 strokes
Intermediate-stroke-count characters+*	Total number of characters containing 11 to 20 strokes
High-stroke-count characters+*	Total number of characters containing more than 20 strokes
Average strokes+*	Average number of character strokes
Word length	
Two-character words+*	Total number of two-character words
Three-character words+*	Total number of three-character words
Lexical count	
Characters+*	Total number of characters
Words+*	Total number of words
Nouns+	Total number of nouns
Verbs+*	Total number of verbs
Adjectives+	Total number of adjectives
Adverbs+*	Total number of adverbs
Ancient Chinese words+	Number of ancient Chinese words used in text
Proportion of ancient Chinese words+	Proportion of ancient Chinese words
Proportion of single-character words+	Proportion of words composed of a single character
Average of vocabulary levels*	Total number of difficulty scores (according to SC-TOP), divided by the total word count
Mean square of vocabulary levels*	Sum of squared difficulty scores (according to SC-TOP) divided by the total word count
Breakthrough vocabulary*	Total number of words found in the list of 8000 Chinese words at the breakthrough level
Waystage vocabulary*	Total number of words found in the list of 8000 Chinese words at the waystage level
Threshold vocabulary*	Total number of words listed in SC-TOP Level 3
Vantage vocabulary*	Total number of words listed in SC-TOP Level 4
Effective operational proficiency vocabulary*	Total number of words listed in SC-TOP Level 5
Frequency	
Average word frequency+	Average word frequency
Average logarithmic word frequency+	Logarithm of the average word frequency
Average frequency skewness+	Measure of the asymmetry of the probability distribution of the average word frequency
SD of average word frequency+	Standard deviation of the average word frequency
Average word frequency according to external database+	Average word frequency according to Academia Sinica database
Average logarithmic word frequency according to external database+	Logarithm of the average word frequency according to Academia Sinica database
Average word frequency relative to maximum frequency in external database+	Average word frequency divided by the maximum frequency according to the Academia Sinica database
SD of average word frequency according to external database+	Standard deviation of the average word frequency according to Academia Sinica database
Difficult words+*	Total number of words listed in Academia Sinica database of 3000 difficult words
Lexical richness	
Type–token ratio+	Degree of lexical diversity
Content-word density+	Density of content words

Table 1 (continued)

Feature	Definition
Semantics	
Semantic complexity	
Content words+*	Total number of content words
Content-word frequency+	Average frequency of content words
Average frequency of domain content words+	Average frequency of content words with domain knowledge
Average logarithmic frequency of content words+*	Logarithm of the average frequency of content words, according to Education Ministry word frequency list
Average logarithmic frequency of domain content words+	Logarithm of the average frequency of content words with domain knowledge
SD of frequency of domain content words+	Standard deviation of the average frequency of content words with domain knowledge
SD of frequency of content words according to external database+	Standard deviation of the average frequency of content words according to Academia Sinica database
Average content-word frequency according to external database+	Average frequency of content words according to Academia Sinica database
Logarithm of content-word frequency according to external database+	Logarithm of the average frequency of content words according to Academia Sinica database
Average content-word frequency relative to maximum frequency in external database+	Average frequency of content words divided by the maximum frequency according to the external database
Negations+	Total number of negation words
Sentences with complex semantic categories+*	Total number of sentences with complex semantic categories
Number of intentional words+	Total number of words denoting intention
Density of proper nouns+	Ratio of proper nouns to total word count
Density of words in natural-science fields+	Ratio of words with specific meanings in natural-science fields relative to total word count
Ratio of content to function words+	Ratio of content words to function words
Complex semantic categories+*	Total semantic category scores from complex sentences
Density of words in social-science fields+	Ratio of words with specific meanings in social-science fields relative to total word count
Syntax	
Sentences+	Total number of sentences
Average sentence length+*	Average number of words per sentence
Simple sentence ratio+*	Proportion of simple sentences
Modifiers per NP+	Total number of adjectives or adverbs before head noun in noun phrases
Subject length+	Average number of characters in grammatical subjects
NP ratio+	Ratio of noun phrases to total number of sentences
Prepositional phrases+*	Average number of prepositional phrases per sentence
Pronoun ratio+	Average number of pronouns per NP
Nominal quotient+	Total number of nouns and prepositions, divided by total number of pronouns, verbs, and adverbs
Intentional construction ratio+	Ratio of constructions conveying intentional meaning to total number of sentences
Parallelism+	Total number of sentences with a parallel construction
Sentences with complex structure+*	Total number of sentences constructed with conjunctions and subordinators
Ba construction (把)*	Total number of expressions containing ba (把) construction
Bei construction (被)*	Total number of expressions containing bei (被) construction
Bi construction (比)*	Total number of expressions containing bei (比) construction
Lian construction (連)*	Total number of expressions containing lian (連) construction
Cohesion	
Reference words	
Pronouns+*	Total number of pronouns
Personal pronouns+*	Total number of personal pronouns

Table 1 (continued)

Feature	Definition
First personal pronouns+*	Total number of first personal pronouns
Third personal pronouns+*	Total number of third personal pronouns
Conjunctions	
Conjunctions+*	Total number of conjunctions
Positive conjunctions+*	Total number of positive conjunctions
Negative conjunctions+*	Total number of negative conjunctions
Causal conjunctions+*	Total number of causal conjunctions
Condition conjunctions+	Total number of conjunctions expressing conditionality
Hypothetical conjunctions+	Total number of conjunctions expressing hypotheticality
Purposive conjunctions+	Total number of conjunctions conveying purpose or intention
Concessive conjunctions+	Total number of conjunctions conveying concession
Conjunction/complex sentence ratio+	Proportion of conjunctions in complex sentences
Figure of speech	
Simile+	Total number of simile words between adjacent sentences

Note: Features marked by plus (+) and asterisk (*) symbols are those used in the CRIE 2.0 and CRIE-CFL systems, respectively

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *Journal of Technology, Learning and Assessment, 4*(3), 1–31.

Bååth, R. (2010). ChildFreq: An online tool to explore word frequencies in child language. *LUCS Minor, 16,* 1–6.

Balota, D. A., & Chumbley, J. J. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 340–357.

Bassiou, N., & Kotropoulos, C. (2011). Long distance bigram models applied to word clustering. *Pattern Recognition, 44*(1), 145–158.

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

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Mahwah: Lawrence Erlbaum Associate, Inc.

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: Academic Press.

Cha, J. H., Chen, J. L., Chang, T. H, & Sung, Y. T. (2013). *The development of an automated analyzer for Chinese text.* Paper presented at the 23rd meeting of Society of Computers in Psychology (SCiP 2013), Toronto, Canada.

Chall, J. S., & Dale, E. (1995). *Readability revisited and the New Dale-Chall Readability Formula*. Cambridge: Brookline Books.

Chang, L. P. (2012). The study of the vocabulary size at the CEFR levels for CFL/CSL learners. *Journal of Chinese Language Teaching, 9*(2), 77–96.

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

Chang, T. H., Liu, C. L., Chen, B., Sung, Y. T., & Su, S. Y. (2013a). *A grammar parser for automated essay scoring of Chinese as a second language: Development and implementation*. Paper presented at the 21st annual meeting of the International Association of Chinese Linguistics, Taipei, Taiwan.

Chang, T. H., Sung, Y. T., & Lee, Y. T. (2013b). Evaluating the difficulty of concepts on domain knowledge using latent semantic analysis. In G. Fu, H. Qi, M. Dong, M. Zhang, & Y. Aibaidula (Eds.), *Proceedings of the 2013 International Conference on Asian Language Processing (IALP 2013)* (pp. 193–196). Los Alamitos: IEEE.

Chen, F. Y., Tsai, P. F., Chen, K. J., & Huang, C. R. (1999). Sinica Treebank. *Computational Linguistics and Chinese Language Processing (CLCLP), 4*(2), 87–104.

CKIP (1993). *Chinese Electronic Dictionary.* Technical Report, No. 93-05, Taiwan: Academia Sinica.

CKIP (1998). *Accumulated word frequency in CKIP corpus. Technical Report, No. 98–02*. Taipei, Taiwan: The Association for Computational Linguistics and Chinese Language Processing.

Council of Europe (2001). *Common European Framework of Reference for Language: Learning, teaching and assessment.* Cambridge: Cambridge University Press. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

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

Cui, X. (1995). Some syntactic and semantic puzzles concerning the Ba-construction. *Chinese Teaching in the World, 33*(3), 12–21.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods, 45*(4), 1–13.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2), 939–944.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12,* 627–635.

Friedman, D. B., & Hoffman-Goetz, L. (2006). A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Education & Behavior, 33*(3), 352–373.

Fry, E. B., Kress, J. E., & Fountoukidis, D. L. (1993). *The reading teacher's book of lists* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Graddol, D. (2004). The future of language. *Science, 303,* 1329–1331.

Hsieh, K. S., Lin, W. C., Dyson, S. B., Liu, P. C., & Sung, Y. T. (2014). *Leveling L2 texts through readability: Combining multilevel linguistic features with CEFR.* Paper presented at the 4th Annual Asian Conference on Technology in the Classroom (ACTC 2014), Osaka, Japan.

Huang, C. R. (2009). *Tagged Chinese Gigaword Version 2.0 LDC2009T14*. Philadelphia: Linguistic Data Consortium.

Huang, C. R., Chen, K. J., Chen, F. Y., Wei, W. J., & Chang, L. L. (1997). Design criteria and content of segmentation standard for Chinese information processing. *Applied Linguistics, 1,* 92–100.

Huang, C. L., Chung, C. K., Hui, N., Lin, Y. C., Seih, Y. T., Lam, B. C. P., … Pennebaker, J. W. (2012). The development of the Chinese Linguistic Inquiry and Word Count dictionary. *Chinese Journal of Psychology, 54*(2), 185–201.

Johnson, M. (1998). PCFG models of linguistic tree representations. *Computational Linguistics, 24*(4), 613–632.

Just, M. A., & Carpenter, P. A. (1987). Orthography: Its structure and effects on reading. In M. A. Just & P. A. Carpenter (Eds.), *The psychology of reading and language processing* (pp. 287–325). Newton: Allyn and Bacon.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In C. E. Brodley & A. P. Danyluk (Eds.), *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 282–289). San Francisco: Morgan Kaufmann.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Lehnert, W. G., & Ringle, M. H. (1982). *Strategies for natural language processing*. Hillsdale: Lawrence Erlbaum Associates.

Lewis, M. P. (2009). *Ethnologue: Languages of the world* (sixteenthth ed.). Dallas: SIL International.

Lin, K. P., & Chen, M. S. (2011). On the design and analysis of the privacy-preserving SVM classifier. *IEEE Transactions on Knowledge and Data Engineering, I., 23*(11), 1704–1717.

Manning, C., & Klein, D. (2002). Fast exact inference with a factored model for natural language parsing. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Proceedings of Advances in Neural Information Processing Systems 15 (NIPS 2002)* (pp. 3–10). Vancouver: Neural Information Processing Systems Foundation.

Marsden, E., Myles, F., Rule, S., & Mitchell, R. (2003). Using CHILDES tools for researching second language acquisition. *British Studies in Applied Linguistics, 18,* 98–113.

McCusker, L. M. (1977). *Some determinants of word recognition: Frequency.* Paper presented at the 24th Annual Convention of the Southwestern Psychological Association, Fort Worth, TX.

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.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. New York: Cambridge University Press.

Miltsakaki, E. (2009). Matching readers' preferences and reading skills with appropriate web texts. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations session, Greece, 12*, 49–52. doi:10.3115/1609049.1609054.

Ney, H. (1991). Dynamic programming parsing for context-free grammars in continuous speech recognition. *IEEE Transactions on Signal Processing, 39*(2), 336–340.

Pal, M., & Foody, G. M. (2010). Feature selection for classification of hyperspectral data by SVM. *IEEE Transactions on Geoscience and Remote Sensing, 48*(5), 2297–2307.

Petrov, S., Barrett, L., Thibaux, R., & Klein, D. (2006). Learning accurate, compact and interpretable tree annotation. In *Proceedings of the 21st COLING and the 44th Annual Meeting of the Association for Computational Linguistics* (pp. 433–440). Sydney, Australia: Association for Computational Linguistics.

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment, 1*(2), 1–22.

Smolkin, L. B., McTigue, E. M., & Yeh, Y.-F. (2013). Searching for explanations in science trade books: What can we learn from Coh-metrix. *International Journal of Science Education, 35,* 1367–1384.

Su, Y. F., & Samuels, S. J. (2010). Developmental changes in character-complexity and word-length effects when reading Chinese script. *Reading and Writing: An Interdisciplinary Journal, 23,* 1085–1108.

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., Lin, W. C, & Tseng, H. C. (2014). Analyzing the readability of textbooks at different learning stages. Paper presented at the 2014 International Workshop on Linguistic Features Analysis, Taipei, Taiwan.

Sung, Y. T., Chen, J. L., Cha, J. H., Tseng, H. C., Chang, T. H., & Chang, K. E. (2015). Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods, 47*(2), 340–354.

Tsai, Y. F., & Chen, K. J. (2004). Reliable and Cost-Effective PoS-Tagging. *International Journal of Computational Linguistics and Chinese Language Processing, 9*(1), 83–96.

Tseng, H. C., Chang, T. H., Chen, B. L. & Sung, Y. T. (2014). *Analyzing textbooks by a readability model based on concepts and support vector machine*. Paper presented at the Asian Conference on Language Learning (ACLL 2014), Osaka, Japan.

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

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior, 17*(2), 143–154.

Xue, N. (2003). Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing, 8*(1), 29–48.

Zhao, Y. (2011). Review article: A tree in the wood: A review of research on L2 Chinese acquisition. *Second Language Research, 27*(4), 559–572.