Natural Language Engineering (2019), **25**, pp. 331–361
doi:10.1017/S1351324919000093

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Integrating LSA-based hierarchical conceptual space and machine learning methods for leveling the readability of domain-specific texts

Hou-Chiang Tseng[1,2,3], Berlin Chen[1], Tao-Hsing Chang[4] and Yao-Ting Sung[5,6*]

[1]Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan, [2]Research Center for Psychological and Educational Testing, National Taiwan Normal University, Taipei, Taiwan, [3]Chinese Language and Technology Center, National Taiwan Normal University, Taipei, Taiwan, [4]Department of Computer Science and Information Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan, [5]Department of Educational Psychology and Counseling, National Taiwan Normal University, Taipei, Taiwan and [6]Institute for Research Excellence in Learning Sciences, National Taiwan Normal University, Taipei, Taiwan
*Corresponding author. Email: sungtc@ntnu.edu.tw

(Received 16 May 2017; revised 3 February 2019; accepted 6 February 2019)

Abstract

Text readability assessment is a challenging interdisciplinary endeavor with rich practical implications. It has long drawn the attention of researchers internationally, and the readability models since developed have been widely applied to various fields. Previous readability models have only made use of linguistic features employed for general text analysis and have not been sufficiently accurate when used to gauge domain-specific texts. In view of this, this study proposes a latent-semantic-analysis (LSA)-constructed hierarchical conceptual space that can be used to train a readability model to accurately assess domain-specific texts. Compared with a baseline reference using a traditional model, the new model improves by 13.88% to achieve 68.98% of accuracy when leveling social science texts, and by 24.61% to achieve 73.96% of accuracy when assessing natural science texts. We then combine the readability features developed for the current study with general linguistic features, and the accuracy of leveling social science texts improves by an even higher degree of 31.58% to achieve 86.68%, and that of natural science texts by 26.56% to achieve 75.91%. These results indicate that the readability features developed in this study can be used both to train a readability model for leveling domain-specific texts and also in combination with the more common linguistic features to enhance the efficacy of the model. Future research can expand the generalizability of the model by assessing texts from different fields and grade levels using the proposed method, thus enhancing the practical applications of this new method.

Keywords: Domain-specific text; Machine learning; Readability; Support vector machine; Text mining

1. Introduction

The earliest study of leveraging quantifiable features to analyze text readability can be traced back to Sherman (1893) (see Bailin and Grafstein 2016 for a review). Readability is still a highly active research topic. From the early investigations of literary texts, readability assessment techniques have been widely applied, expanding to other genres and domains including insurance policies, medical awareness pamphlets, jury instructions (DuBay 2004), biology textbooks (Belden and Lee 1961), health education messages (Freimuth 1979), business communication textbooks (Razek and Cone 1981), economics textbooks (Gallagher and Thompson 1981; McConnell 1982), newspapers (Johns and Wheat 1984), and adult learning materials (Taylor and Wahlstrom 1999).

Among the various aspects of readability assessment techniques, the development of new linguistic features to be integrated into readability models is an issue of both theoretical and practical importance (De Clercq and Hoste 2016; Graesser *et al.* 2004; McNamara, Louwerse, and Graesser 2002; McNamara *et al.* 2010; Sung *et al.* 2015a; 2015b; Sung *et al.* 2016a; Tanaka-Ishii, Tezuka, and Terada 2010). Notably, there remains a problem that common general linguistic features (i.e., a set of lexical, syntactic, semantic, and cohesion-related features as described in Appendix) are not capable of reflecting the difficulty levels of the knowledge contained in domain-specific texts. As pointed out by Redish (2000), when a field-specific term appearing in a domain-specific text is also a commonly used word, then the term's difficulty level within that text cannot be accurately assessed by general linguistic features. For example, when appearing in a general text, *shock* is a common and easy word that means "strong, and usually unpleasant, emotion," but in a medical text it refers to "a life-threatening condition that occurs when the body is not getting enough blood flow, which obstructs microcirculation and results in the lack of blood and oxygen in vital organs" (Cecconi *et al.* 2014: 1796). Using the common general linguistic features to measure the word *shock* in the two senses (i.e., in a general vs. a domain-specific text) would lead to the same predicted difficulty levels because both "shocks" have the same part-of-speech, the same word length, and, according to most word lists, *shock* is rated with only a single difficulty score. In other words, apart from the fact that the two "shocks" reside in different domains of knowledge, they are superficially identical. To take another example, in an empirical study analyzing Medical Subject Headings (MeSH) in the US medical database, Yan, Song, and Li (2006) found that general linguistic features such as the number of syllables and word length of a medical term were not related to the term's level of difficulty. From these examples, we can argue that because common general linguistic features are derived from the surface characteristics of text, they are not able to characterize the meaning embodied in the knowledge in particular fields, let alone to represent the relations between different knowledge, or to distinguish their difficulty levels. Researchers need to develop new readability features that are capable of rating the knowledge-oriented difficulty of words. Likewise, Collins-Thompson (2014) argues that the ability to capture the dependencies between concepts is a requisite of knowledge-based readability models for deeper content understanding.

In view of the inability of general linguistic features employed by traditional readability formulas to measure the knowledge of domain-specific texts, how to reasonably and effectively do so becomes a topic worthy of further research. Taking Chinese texts in the natural and social sciences as examples, our study has two main purposes: The first is to design a method to represent the knowledge contained in domain-specific Chinese texts that can identify the knowledge features of different grade levels in the subjects of social science and natural science, and to use these features as important references in defining the readability of texts in these subjects. The second is to compare, in terms of the effectiveness of assessing the readability of domain-specific texts, readability models using knowledge features with models using general linguistic features (e.g., word count, phrase length, and sentence length). Specifically, our study aims to answer the following two questions: (1) How well does a general-linguistic-feature-based readability model perform in predicting readability/difficulty levels of domain-specific texts? (2) Does a knowledge-feature-oriented readability model based on the hierarchical conceptual space extracted by latent semantic analysis (LSA) outperform a general-linguistic-feature-based readability model? In order to compare the different approaches for assessing domain-specific text readability, three sub-studies applying different methodologies were conducted to further validate our findings.

In Study 1, we tested whether general linguistic features are suitable for assessing the readability of domain-specific texts. Readability models that incorporated the general linguistic features were trained through machine learning. Study 1 also provided a baseline for the model validation in Studies 2 and 3. In Study 2, we proposed a hierarchical conceptual space to generate a hierarchy of difficult word lists that correspond to school-grade levels. We then used the word lists to calculate the difficulty distribution of conceptual terms in domain-specific texts. The grade level of a text was estimated based on the difficulty level where most terms are distributed. This is to show that in

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

describing a knowledge, a text employs a great number of domain-relevant terms, whose difficulty levels hence reflect the readability of the text.

Study 3 was an extension of Study 2 and used the difficulty level distribution of conceptual terms in a text as a feature of the readability model, which used a support-vector-machine (SVM)-based classifier. In addition, in order to compare the performance of different readability models, Study 3 experimented with the following three readability models: the model combining the grade-level vectors and general linguistic features, the TF model (term frequency; Salton and Buckley 1988), and the Word2Vec model (Mikolov *et al.* 2013). By comparing the results of the experiments, we contrasted the advantages and disadvantages of using general linguistic features, bag-of-words-based features, and hierarchical conceptual space to determine the readability of domain-specific texts.

2. Literature review

2.1 The development of readability studies

Text readability/difficulty assessment is an important application in text mining field that has been more formally defined as the sum of all elements in textual material that affect a reader's understanding, reading speed, and level of interest in the material (Dale and Chall 1949). The effective assessment of the difficulty of texts can contribute to teaching and learning: readers will understand and learn effectively when they select texts at appropriate readability levels that are suited to their reading ability (Klare 1963, 2000). Texts that are too advanced for one's reading level will significantly increase the reader's cognitive load, leading to feelings of frustration (DuBay 2004). Conversely, a text that is too simple for the reader can lead to a lack of motivation and sense of achievement. In other words, readability assessment is helpful for matching reading materials with reading abilities, through which more successful and joyful reading experiences may be created.

Traditional readability formulas are based on research findings that factors such as semantic, syntactic, and lexical complexities influence the difficulty level of a text (Graesser, McNamara, and Kulikowich 2011; Rubenstein and Aborn 1958). Readability formulas thus employ these linguistic features as key variables to predict text readability. For example, the Flesch Reading Ease Formula (Flesch 1948) used the number of syllables in a word as an indicator of lexical complexity and sentence length as an indicator of syntactic complexity, measuring text difficulty based on average number of syllables per word and average sentence length (Collins-Thompson 2014). The more syllables an average text word has and the longer an average sentence is, the more difficult a text becomes. Chall and Dale (1995) added percentage of difficult words as another variable of text difficulty—the greater the number of difficult words a text contains, the more difficult it is.

As Graesser, Singer, and Trabasso (1994) pointed out, traditional formulas using general linguistic features fail to reflect the actual reading process. This is because the common general linguistic features, which only involve semantic, syntactic, and lexical properties of text, neglect some other essential text features, one of them being the coherence of a text. Collins-Thompson (2014) also suggested that traditional readability formulas focus only on the shallow information of a text, overlooking important deeper content features. This has led to skepticism over the results when such formulas are used to predict text comprehensibility. For example, using word frequency (how often a word occurs in a representative corpus) as the basis of judging whether a word is difficult would not be objective enough, since word usage evolves with time, and new words will always emerge. For this reason, the Thorndike List, which Thorndike and Lorge (1944) developed based on the word frequencies of their time, is not ideal for assessing the difficulty of modern words and renders readability formulas using this feature even less accurate in their predictions. In addition, word frequency is not necessarily an effective way of assessing word difficulty. The word *toothbrush*, for example, is simple enough and commonly understood, but it appears infrequently in written texts, whether it is in newspapers, magazines, or textbooks, and

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

would likely be categorized as a difficult word, which clearly is not true. Using word length as a basis for calculating readability has also often been challenged for the reason that longer words are not necessarily more difficult (Yan *et al.* 2006). Meanwhile, basing syntactic complexity on sentence length and using that as a variable in measuring readability have been criticized as being too intuitive and lacking in sophistication (Bailin and Grafstein 2001).

Scholars have conducted research using various linguistic features. Graesser *et al.* (2004), for example, developed the Coh-Metrix, an online analyzer of text features with categories such as word information, syntax structure, latent semantics, and cohesion. The effectiveness of the Coh-Metrix readability features for text classification is supported by empirical evidence in the field of psychology (Graesser *et al.* 2004; McNamara *et al.* 2002; McNamara *et al.* 2010). Kanebrant, Mühlenbock, Kokkinakis, Jönsson, Liberg, Geijerstam, Folkeryd, and Falkenjack (2015) presented T-MASTER, a tool for assessing students' reading skill on a variety of dimensions. They used the SVIT model for analyzing textual complexity in T-MASTER. The model was constructed using four categories of variables: surface complexity, vocabulary difficulty, syntactic, and idea density.

In contrast to Western languages, readability assessment of Chinese text requires readability models of its own that are distinct from those based on alphabetic writing systems. The complexity of Chinese characters cannot be measured by letter or syllable counts, because each Chinese character is comprised of one syllable and most Chinese words are formed by two characters (Hong *et al.* 2016). The uniqueness of the language has led to the inquiry of which readability features are suitable for Chinese text. For example, Sung *et al.* (2016a) developed Chinese Readability Index Explorer (CRIE) as a system for analyzing Chinese text readability, consisting of four categories of indices: word, syntax, semantics, and cohesion. Using multilevel readability features, the model of the CRIE results in improved effectiveness for the assessment of Chinese text readability (Sung *et al.* 2013; Sung *et al.* 2015a, 2015b). Chen, Chen, and Cheng (2013) used the lexical chain technique to examine the suitability of using lexical cohesion as a readability feature for Chinese text. They divided texts in social and life science textbooks into three reading levels: textbooks written for the first and second graders, the third and fourth graders, and the fifth and sixth graders were regarded as the low, middle, and high levels, respectively. The model was highly effective in a coarse-grained manner: the accuracy to distinguish between the low and non-low levels reached as high as 96%, and that between the middle and non-middle levels achieved 85%. Tseng *et al.* (2016) used Word2Vec to obtain a Chinese semantic space and characterized the overall semantic features through vector transformations, thus creating a useful readability feature. When used to assess textbooks in Chinese language arts, social studies, and natural science from the 1st through the 12th grades, the readability model reached an accuracy of 75.99%. However, a large number of dimensions (800 dimensions) were required to build their readability features. As a result, it was very time-consuming to train the model on large data. Moreover, because the testing data consisted of both domain-specific and literary texts, it was not clear whether the accuracy rate was mostly attributed by the prediction of the non-domain-specific texts. Most importantly, the model was uninterpretable. Aside from the leveling result, one has no clue as to what values or benefits these readability features may contribute in the identification of domain knowledge difficulty.

In addition to the concurrent development of readability features, the rise of natural language processing techniques and machine learning algorithms now allow researchers to refine their model algorithms to measure readability of text with a more flexible scope (Feng *et al.* 2010; François and Miltsakaki 2012; Petersen and Ostendorf 2009; Sung *et al.* 2015a; Vajjala and Meurers 2012). Besides the aforementioned studies on document readability, some scholars have started to investigate readability at the sentence level. For example, Vajjala and Meurers (2014) built upon their pioneering research (Vajjala and Meurers 2012) and used 151 features to train readability models on both document and sentence levels. Experimental results indicate that the document-level readability model achieved a Pearson correlation of 0.92, but the sentence-level readability model only achieved an accuracy of 66%. Furthermore, Ambati, Reddy, and Steedman (2016) used an incremental Combinatory Categorial Grammar (CCG) parser to calculate sentence complexity

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

and predict the relative sentence readability. The incremental CCG model outperformed the non-incremental Phrase Structure Trees (PST) model in extracting syntactic features. Apart from the above, Howcroft and Demberg (2017) used the ModelBlocks parser and the program icy-parses to extract 22 features. Four measures (idea density, surprisal, integration cost, and embedding depth) were used to calculate sentence complexity scores and served as the base for training different readability models, in order to figure out which feature set can constitute a readability model that is both effective and efficient. Analysis results indicate that compared with integration cost and idea density, surprisal and embedding depth make the readability model more efficient. The interested reader can refer to Collins-Thompson (2014) for a comprehensive overview of relevant studies that followed this line of research.

2.2 Studies of domain-specific text readability

The content of domain-specific texts is generated through the reproduction of relevant knowledge that humans have developed through documenting recurring and complex problems in life (Hirschfeld and Gelman 1994). For example, having experienced recurring weather conditions such as typhoons, floods, and snowstorms, humans documented the phenomena through sound, image, and text to form specific ideas about them, and attempted to find solutions for them, leading to the development of the domain of meteorology (Hirschfeld and Gelman 1994). Domain-specific texts focus on illuminating the "concepts" of the relevant knowledge. More importantly, it is noticeable that these domain-specific concepts are formed through the convergence of relevant terms. For instance, when explaining the domain-specific concept of "how plants produce nutrients," one inevitably mentions words that underpin the subject such as *photosynthesis*, *chloroplast*, *enzyme*, *glucose*, and *carbon dioxide*. This method of elucidating domain-specific knowledge is different from the structure of descriptive or narrative writing used in general language text.

The more domain-specific terms there are in a domain-specific text, the more domain-specific knowledge the reader will need to have in order to efficiently generate meaningful understanding of the text (Bédard and Chi 1992). Chi, Glaser, and Farr (1988) noted that the difference between an expert and a novice lies in the former's ability to master the domain-specific knowledge and to derive a significant amount of meaningful information from the text. Etringer, Hillerbrand, and Claiborn (1995) have also pointed out that experts possess a broad and deep knowledge base, making it easier for them than for novices to retrieve information from their long-term memory to link fragmentary information together and integrate them into meaningful information. This means that experts possess domain-specific concepts that they can use effectively within the knowledge structure of a domain-specific text; conversely, when faced with a domain-specific text, a novice lacks the corresponding domain-specific concepts for them to retrieve the information contained in the text and is thus unable to generate meaningful understanding. Therefore, the readability of a domain-specific text should be judged by the amount of meaningful information that can be retrieved by most people with similar reading abilities. In other words, the knowledge of a domain-specific text should be within the grasp of most readers at a certain developmental phase (e.g., a certain school grade).

Automatically retrieving, defining, and measuring the impact of various factors on the difficulty of domain-specific texts requires sophisticated methods. Currently in the literature, the four methods or tools most often utilized for calculating the readability of a text, which are reviewed below, are general linguistic features, ontology, word lists, and LSA.

2.2.1 General-linguistic-feature-based assessment of domain-specific texts

Early readability research often employed general linguistic features to build linear regression models to assess the difficulty of domain-specific texts. For example, Razek and Cone (1981)

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

and Gallagher and Thompson (1981) used the Flesch Reading Ease Test to analyze economics textbooks and business communication textbooks, respectively. Miltsakaki and Troutt (2007) proposed the Read-X system to perform readability assessment on different types of website text, using three readability formulas. Kanungo and Orr (2009) focused on predicting the difficulty of web page summaries. However, since they employed general linguistic features, it remains uncertain, and thus further verification is needed to determine whether these models are capable of representing the knowledge structure of domain-specific texts. To offset the shortcomings of general-purpose readability assessment tools, Dell'Orletta, Venturi and Montemagni (2012) proposed a ranking method based on the notion of distance for automatically building genre-specific training corpora. Currently, the most popular readability assessment tool for Mandarin is the CRIE system by Sung *et al.* (2016a). However, the CRIE only employed general linguistic features and was not designed to handle domain-specific texts.

2.2.2 Ontology-based assessment of domain-specific texts

Ontology uses a hierarchical tree structure to represent the relationship between domain-specific concepts (Gruber 1993). Using this tree, it is possible to assess the difficulty of a domain-specific concept by calculating the distance from that concept to the root of the tree; the higher up the concept is on the tree depth, the more difficult it likely is. Yan *et al.* (2006) used the hierarchy of medical symbols in the American MeSH database to determine the complexity of a medical concept in MeSH by calculating the concept's tree depth, that is, the distance of that concept to the root of the tree structure. The scope of a document is regarded as the coverage of the domain concepts in the document. The more terms of the document are identified as domain concepts, the less readable the document tends to be. The study compared how well the readers' level of understanding was predicted by using document scope and other traditional formulas, and the result indicated that document scope is a good predictor of the readability of medical texts. Zhao and Kan (2010) used ontology to construct the domain-specific concept hierarchy in the Math World Encyclopedia, and then calculated the difficulty score of each concept. In a similar vein, Project 2061 (AAAS 2007), a long-term initiative of the American Association for the Advancement of Science (AAAS) that aims to help students become literate in science, mathematics, and technology, serves as a useful ontology tool. Many of the topic maps provided by the AAAS in the Atlas are built from K-12 learning goals. By studying these maps carefully, teachers and other educators can get a better sense of the content and nature of grade-level benchmarks as specific learning goals (AAAS 2007).

The ontology method takes into account the hierarchies of domain-specific concepts, but not every domain has a readily available domain-specific conceptual hierarchy. In such a case, experts are required to produce a conceptual hierarchy in order to apply the ontology approach. Otherwise, it is impossible to assess the conceptual difficulty. Thus, using the ontology method to assess text readability can involve significant costs.

2.2.3 Word-list-based assessment of domain-specific texts

Any collection of words can be viewed as a word list. It is a common practice to define word difficulty by referring to a word list. The assumption underlying the practice of using easy or difficult word lists to assess text readability is that the greater number of difficult words a text contains, the harder it is for the concepts embedded in the text to be understood. One widely known measure of this type is the Revised Dale–Chall formula (Chall and Dale 1995). The formula used the Dale word list, which collected the 3000 words that were familiar to at least 80% of American fourth graders at the time. A word is labeled "unfamiliar" if it does not occur in the list. Accordingly, the more words a text contains that are not present on the word list, the lower its readability is. A similar method was used by Fry (1990), who employed Dale and O'Rourke's Living Word

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

Vocabulary of 43,000 word types to assess the readability of short articles. An example of utilizing word lists to assess domain-specific text readability was Miltsakaki and Troutt (2008), for which measurement of web text readability was based on word difficulty that was rated by word frequency. In order to develop monolingual and bilingual word lists for language learning, Kilgarriff *et al.* (2014) participated in the KELLY project, which created bilingual learning word cards using the 9000 most frequent words in text corpora of 9 languages and 72 language pairs. The above shows that using corpus-derived word frequency to define word difficulty is a common practice. However, there exists many counter-examples of this intuitively correct assumption. For example, some commonly used household items appear infrequently in specific corpora (such as newspapers). Moreover, frequency-based word lists simply categorize words as difficult or simple, the precise definition of which is highly subjective. In an attempt to remedy the aforementioned flaw of frequency-based word lists, Borst *et al.* (2008:73) computed both the "complexity of the category to which a word belongs" and the "word frequency." They then multiplied the two scores to represent word difficulty, which was used as a basis for estimating the complexity scores of sentences, and the readability score of a document corresponds to the average score of its sentences. In a similar vein, Ding (2007) proposed a model to automatically construct knowledge hierarchies that could be helpful to classify the words according to their hierarchical relationships with other words in a knowledge domain. For the current study, we also aim to apply similar ideas to the readability assessment of domain-specific texts to make the assessing process more objective.

2.2.4 LSA-based assessment of domain-specific texts

LSA is a technology for network knowledge representation (Landauer and Dumais 1997; Landauer, Foltz, and Laham 1998). This technique applies singular value decomposition (SVD) to the term-document matrix to obtain a latent semantic space, after which the term, phrase, sentence, or even an entire document can be folded-in to the latent semantic space and be represented as a semantic vector.

The more similar two vectors are, the more similar the two documents are in terms of semantics (Furnas *et al.* 1988). Currently, this computational model is widely used in the field of information retrieval to augment traditional technology (e.g., Boolean search technology), which calculates the similarity between sentences or documents based only on whether the terms they contain coincide. LSA has also been applied to text similarity comparison through the computation of their latent semantic properties. Chang, Sung, and Lee (2013), for example, used LSA to obtain terms at different levels of difficulty from domain-specific textbooks. Kireyev and Landauer (2011) employed LSA to develop a domain-specific concept they called Word Maturity to estimate the grade level of a word, and in turn used these grades to gauge the literacy of students.

In addition to detecting semantic similarities, LSA has been further used to analyze text readability. When classifying documents written for non-native readers of French, François and Miltsakaki (2012) used LSA to compute lexical cohesion as a semantic feature for their readability model. Truran *et al.* (2010) employed the LSA technique to investigate the readability of clinical documents. Graesser *et al.* (2004) developed the Coh-Metrix 3.0, providing eight LSA indices as measures of semantic overlap between sentences or between paragraphs. However, these studies treated the LSA technique as one of the many readability features incorporated into their readability models. They did not explore why LSA was useful for analyzing domain-specific text readability. It is therefore unclear how such a model might differ from models using general linguistic features. In view of this, the current study aims to investigate the independent predictive value of both the LSA technique and the general linguistic features as well as their combined predictive power. Although the methods we devised for the current study are based on the hierarchical conceptual spaces put forward by Chang *et al.* (2013), our work is distinguished from Chang *et al.* (2013) in the following ways. The study by Chang *et al.* (2013) only sketches the concepts of domain-specific knowledge in school textbooks that are supposed to be comprehensible

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

for students in the corresponding grades. More specifically, the sketched domain-specific concepts are isolated from each other. The current study extends the work of Chang *et al.* (2013) by regarding the difficulty levels of the domain-specific concepts in a text as forming a relational pattern. The readability features thus developed are more interpretable and more capable of classifying the readability of Chinese domain-specific texts. Additionally, this paper combines the LSA-derived readability features with the common general linguistic features to investigate their impact on readability models.

3. Proposed method: Using hierarchical conceptual space to calculate grade-level vectors for the readability assessment of Chinese domain-specific texts

Instead of viewing a text as having a fixed (i.e., a single and unchangeable) difficulty level, Chi *et al.* (1988) define readability as a relative notion measured by the degree to which a document is understood by the individual reader. For experts versus novices, a domain-specific document may provide either a massive amount of meaningful content or only trivial information. In other words, the difficulty level of a domain-specific text is determined jointly by the conceptual load of a text and the reading ability of an individual. For example, normally the content of fifth-grade textbooks is easy for ninth graders but difficult for third graders.

Given that each person has a different background, in order to develop a readability model that is widely applicable across age and social groups, we established a common ground of reading abilities according to school education. Nolen-Hoeksema *et al.* (2009) pointed out two approaches for domain-specific concept acquisition: Obtaining the prototype of a concept through everyday experiences and learning the core of a concept through school instructions. At each grade level, students are taught the knowledge of specific domains through subject courses, forming domain-specific concepts and internalizing these concepts into their knowledge system. Accordingly, to assess the right time to start learning a concept, we can use the concepts that students learn from the textbooks as a basis to estimate the time at which concepts of domain-specific knowledge can (or should) be comprehended by most people.

This method is reasonable because most textbooks are written following curriculum standards, and most students are able to understand a certain domain-specific concept after going through a certain stage of learning. Therefore, the domain-specific concepts that they acquire are generally representative of the typical difficulty level of knowledge learned by the average student at a specific learning phase. The word frequency guide of Zeno *et al.* (1995) and the Manulex, a grade-level lexical database from French elementary-school readers developed by Lété, Sprenger-Charolles, and Colé (2004), also used school-grade levels to define the difficulty level of terms. Lété *et al.* (2004) believe that the most important variable in understanding language development and the reading process is the vocabulary that children learn, and that using grade level to quantify children's reading material is valuable to psycholinguists trying to evaluate children's language acquisition. Therefore, in assessing domain-specific text readability, we may use the difficulty level of a word, which can be determined based on the grade level at which it becomes a main theme or topic.

Figure 1 illustrates how instruction of domain-specific concepts develops in the educational process. A fifth-grade natural science text on the topic of oxygen and carbon dioxide, toward the end the article, describes that one of the ways oxygen and carbon dioxide are used is in the photosynthesis process of plants. However, it is not until in a seventh-grade text on the topic of how plants produce nutrients that the whole process of photosynthesis is truly explained. Thus, to which grade-level difficulty should the term *photosynthesis* be assigned? How to automatically and objectively assign difficulty levels for terms thus becomes a problem worth investigating.

Many topics can have various shades of domain-specific concept, which is why the concepts are taught in phases, through textbooks for different grades. At higher grade levels, students are

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

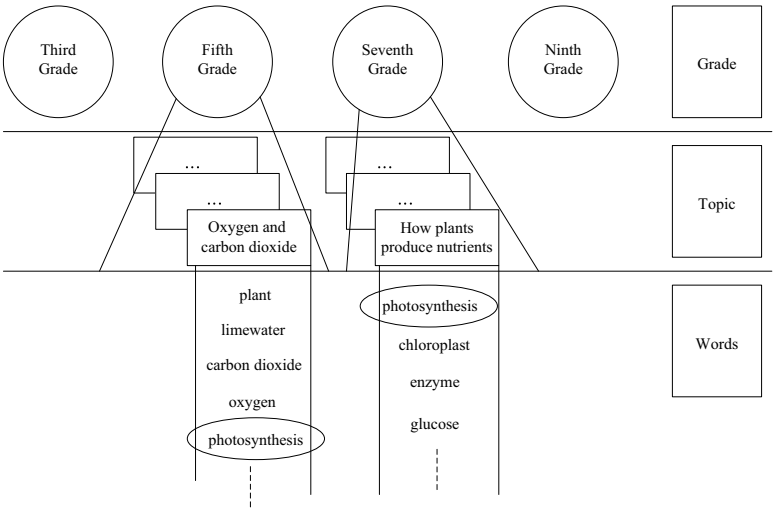

Figure 1. Words taught at different grade levels for different domain-specific concepts.

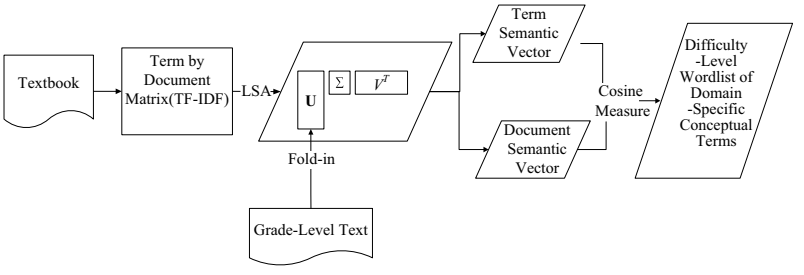

Figure 2. Flowchart for generating the difficulty level of domain-specific conceptual terms.

exposed to a greater number of (and often more sophisticated) terms related to a topic, enabling them to deepen their knowledge. Continuing to use Figure 1 as an example, *photosynthesis* is introduced in a fifth-grade text on the side of *oxygen* and *carbon dioxide*, providing prior knowledge in preparation for the teaching of the seventh-grade lesson on how plants produce nutrients, facilitating the learning process. This arrangement is a typical type of spiral curriculum, in which there is an iterative revisiting of terms, topics, or themes throughout the course. Learners' knowledge may be broadened and deepened through this iteration, in which complex concepts are built up from simple ones (Harden 1999). Taking this example, Chang *et al.* (2013) took 100 terms from elementary school natural science textbooks in Taiwan and used LSA to calculate the semantic relatedness of each term to natural science domain texts from the third to sixth grades. Figure 2 presents the five steps Chang *et al.* (2013) proposed to test which grade level a domain-specific concept is characterized by.

Step 1. Show the relationship between "term" (i.e., word) and "document" (i.e., text) of social science and natural science textbooks, each with a two-dimensional matrix. In the "term-document matrix," the value of each entry represents the frequency at which a term occurs in a document. This matrix is called the "term-document matrix." This step will produce the term-document matrix for social science and natural science texts, respectively.

Step 2. Process the two term-document matrices with the Term Frequency–Inverse Document Frequency (TF–IDF) method. TF–IDF method helps to give greater weight to specifically those

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

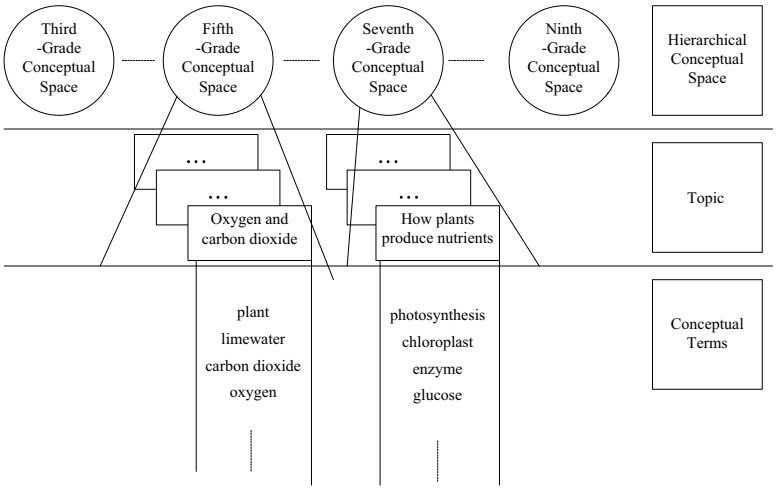

Figure 3. Hierarchical conceptual space with assigned difficulty levels.

words which occur in only some of the documents while down-weighting those which are common to all of the documents (Sparck Jones 1972; Salton and Buckley 1988).

Step 3. Use the SVD (Golub and Reinsch 1970) of the LSA to reduce the dimensions of the term-document matrix, retrieve the latent semantics contained in the terms, and extract semantic clusters to be presented anew. SVD is performed on the term-document matrix ($\boldsymbol{W}$) in order to project all the term vectors and document vectors onto a single latent semantic space with significantly reduced dimensionality $\boldsymbol{L}$. That is, $\boldsymbol{W} \approx \widetilde{\boldsymbol{W}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\boldsymbol{T}}$. In the equation, $\widetilde{\boldsymbol{W}}$ is the rank-$\boldsymbol{L}$ approximation to $\boldsymbol{W}$; $\boldsymbol{U}$ is the left singular matrix; $\boldsymbol{\Sigma}$ is the $\boldsymbol{L}\times\boldsymbol{L}$ diagonal matrix composed of the $\boldsymbol{L}$ singular values; $\boldsymbol{V}$ is the right singular matrix; and $\boldsymbol{T}$ denotes matrix transposition. In the process of reducing dimensions, different meanings and different senses of a word are all lumped together, abstracting information to capture latent semantics of the terms within the document. In this way, words describing related concepts will be close to each other in the latent semantic space even if they never co-occur in the same document, and the documents describing related concepts will be close to each other in the latent semantic space even if they do not contain the same set of words (Lee and Chen 2005).

Step 4. In order to determine whether a certain term is a main concept in a specific grade-level textbook or not, Chang *et al.* (2013) projected each domain-specific term, for example, *atom*, and each grade-level text (all the texts of a grade level were combined to create a single text for a total of 7 texts for the grade levels 3–9) to their respective latent semantic space, generating two vectors: One vector for a specific term and another vector for the grade-level text being compared. For example, we created the term semantic vector for *atom*, and the document semantic vector for each grade-level text. The cosine value of the similarity between the vector of a specific term and the vector of a grade-level text was then calculated. The cosine value fell between 1 and –1, with a higher value meaning more similarity between a term and a text, and a lower value indicating less similarity. This process was repeated for each of the seven grade levels to find which grade *atom* was most similar to. To put it another way, we calculated seven cosine similarities for each domain-specific term, one for each of the grades 3–9. The highest cosine value helped us identify which grade level a domain-specific term belonged to.

Step 5. Based on the degree of similarity between the term vector and document vector, all terms are assigned difficulty of grade level.

After the above-described step, terms were classified to a particular grade-level label. Figure 3 shows part of the result of assigning the domain-specific terms to the levels that most appropriately

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

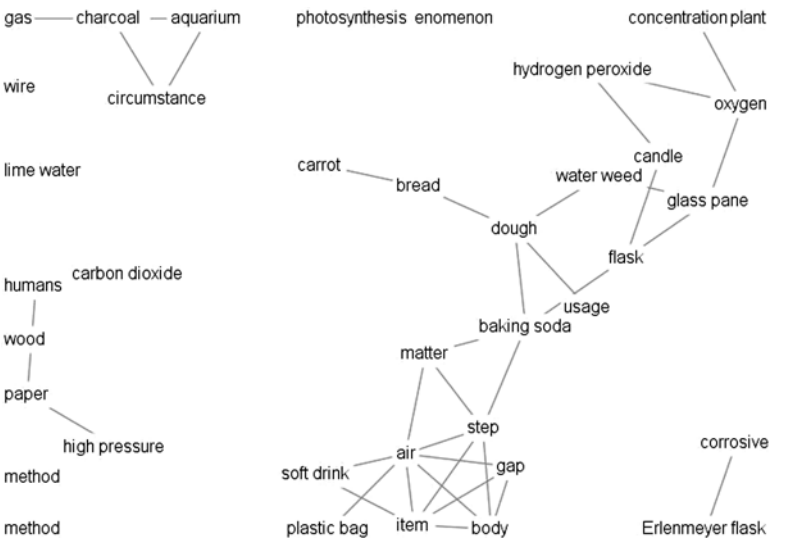

Figure 4. The conceptual space of "oxygen and carbon dioxide" in a fifth-grade natural science text.

characterize their difficulty levels. We are now able to point out that *photosynthesis* is more closely related to the knowledge taught in seventh grade, instead of fifth grade where it could also be found. In other words, *photosynthesis*, *chloroplast*, *enzyme*, and *glucose* are the terms used in describing the knowledge subject of how plants produce nutrients. Learners' knowledge may be treated as a network of concepts (Hunt 2003); however, the content of concepts is delivered by the words/terms used in the texts. This is the reason why we use the "conceptual terms" to represent learners' knowledge of each grade. Once the conceptual terms for all the topics in the textbooks at a certain grade have been gathered, they form the conceptual space unique to this grade, which can be given a corresponding difficulty level. When the conceptual spaces of all the grades are linked together, these spaces together form a hierarchical conceptual space reflecting the spiral curriculum design.

One of the fundamental problems of the word lists developed by past studies is that they do not correspond well with grade levels. As a result, while the past methods may be helpful in creating classification models, they are not able to provide users any sort of justification for why any given article is assigned to a grade level.

Addressing this issue, this article adopts the steps we have discussed above to generate a word list derived from a hierarchical conceptual space. In this way, each conceptual term will be assigned to a grade level of difficulty, and the overall conceptual terms are able to represent the typical knowledge assumed to be learned at each grade.

The relationships between the conceptual terms can be visualized via a Pathfinder network representation (Schvaneveldt, Durso, and Dearholt 1989, 2017). The idea of Pathfinder is to use the cosine matrix of domain-specific concepts (here they are derived from LSA) to determine the distance among concepts. The relationship between the conceptual terms and the conceptual spaces of the topics of "oxygen and carbon dioxide" and "how plants produce nutrient" is represented in Figures 4 and 5, respectively. For the use of Pathfinder, we choose the Threshold Network Type to limit the inter-term relevance of conceptual terms to ensure that only those conceptual terms that are highly relevant to the main idea of the topic are highlighted. If a conceptual term is unrelated to the main idea, it would not appear as connected to other conceptual terms.

Figure 4 shows that the term *photosynthesis* is isolated from all other terms in the fifth-grade text "Oxygen and carbon dioxide," because *photosynthesis* is only referred to as a process that

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

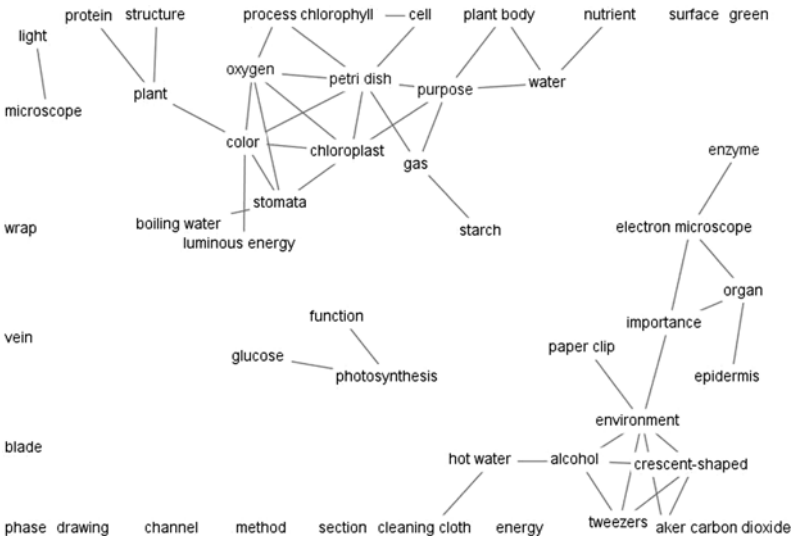

Figure 5. The conceptual space of "how plants produce nutrients" in a seventh-grade natural science text.

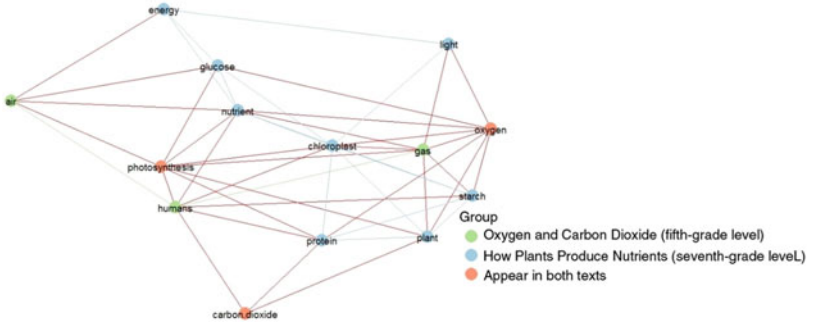

Figure 6. Illustration of a hierarchical conceptual space.

utilizes oxygen and carbon dioxide and is not the focus of the text. By contrast, in Figure 5 because the main topic of the seventh-grade text "How plants produce nutrients" is photosynthesis, we can see that the term *photosynthesis* is related to other terms. These results once again highlight the problem described in Figure 1 and help confirm what is shown in Figure 3 that despite being present in both fifth- and seventh-grade texts, the proper difficulty of the term *photosynthesis* is Grade 7.

In the following, we utilized the GGally module of the R software (Schloerke 2011) to integrate the conceptual spaces of the two texts analyzed above. For the convenience of comparing the relatedness of the two texts, we set the threshold to include only highly relevant conceptual terms. The result is illustrated in Figure 6. Since fifth and seventh grades share the concept of photosynthesis, the two conceptual spaces, following a spiral curriculum design, can be combined to form a hierarchical conceptual space, in which the knowledge expanded both in breadth (conceptual scope: the connection between the conceptual terms of a similar topic) and in depth (conceptual difficulty: the depth of the concepts on the conceptual hierarchy). Through the merging of conceptual spaces, the underlying connection and relations of the conceptual terms will emerge. From Figure 6 we can see that the fifth-grade-level term *humans* is related to the seventh-grade word *glucose* through *photosynthesis*. *Humans* does not even need another word to be connected to

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

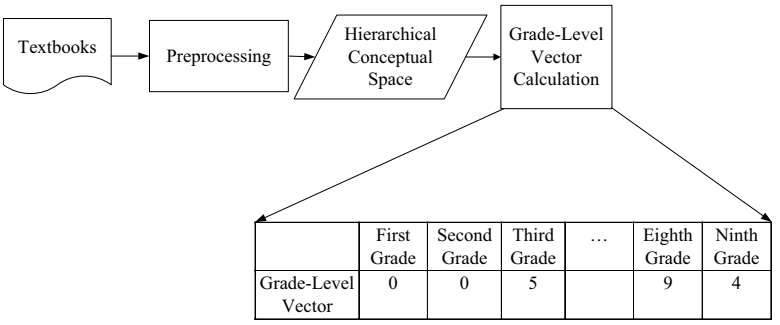

Figure 7. Interpretable grade-level vector diagram.

the seventh-grade word *starch*. Such connections illustrate that the hierarchical conceptual space in which all the levels of knowledge are aggregated not only allows knowledge to be vertically integrated, but can even reveal hidden relationships between domain-specific concepts and help students understand, at a glance, why plants are the foundation of the food chain. Additionally, educators can use this method to observe the relationships between conceptual terms across similar articles, allowing them to prepare teaching materials or add supplementary materials as appropriate.

The example above indicates that the difficulty level of domain-specific texts can be represented by the number and difficulty of conceptual terms which can be transferred to a conceptual space corresponding to grade level. This article therefore hypothesizes that if conceptual space can be effectively matched with the appropriate grade level, then these spaces can be used to estimate the difficulty level of any text in terms of grade level.

After building a hierarchical conceptual space, the text can be matched against it to determine whether the text contains conceptual terms in specific grades. In contrast to the study by Chang *et al.* (2013), which predicts domain-specific readability through a separate mapping of the main concepts of a text with the grade levels, the current study extends their work by treating the distribution of conceptual terms as a relational pattern (i.e., a grade-level vector, which is described below) so that the developed readability features can be more intelligible, providing meaningful interpretation. The relational pattern is formed through the following steps. First, the conceptual terms of each text are tagged for difficulty. The tagged terms for each difficulty level (grade) are then counted, with each term counted only once in each grade, resulting in a difficulty distribution of the text's terms in each grade. The distribution values form a vector, which we call the text's grade-level vector. As shown in Figure 7, people can use grade-level vectors to understand the grade-level distribution of each conceptual term in a text. When leveling a text, this makes it much easier to explain why it is placed at a certain grade. In sum, our study has developed grade-level vectors as a readability feature, and a prediction model incorporating the feature was trained by machine learning. The performance enhancement of the proposed model over the model by Chang *et al.* (2013) was investigated. In addition, this study combined grade-level vectors with common general linguistic features in order to analyze the impact of the former on readability models.

4. Study 1: Constructing a readability model for domain-specific texts with general linguistic features as a baseline study

Study 1 examined the effectiveness of a general-linguistic-feature-based model in predicting the readability of domain-specific texts. The performance of this model also served as a comparison baseline for the model performances in Studies 2 and 3.

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

Table 1. Text length distributions among third- to ninth-grade social science and natural science textbooks

Grade	3	4	5	6	7	8	9
Social Science (no. of articles)	80	74	85	81	389	407	325
Mean Character Count (*SD*)	337.8 (*106.4*)	368.8 (*105.5*)	599 (*253.2*)	624.4 (*222.2*)	380.5 (*191.7*)	406.4 (*214.8*)	467.1 (*220.8*)
Mean Word Count (*SD*)	195.3 (*62.4*)	209 (*59.8*)	334.9 (*143.4*)	354 (*126.7*)	215.1 (*107.3*)	230.5 (*121.7*)	259.6 (*123.7*)
Natural Science (no. of articles)	72	67	67	62	172	175	157
Mean Character Count (*SD*)	334.7 (*160*)	443.7 (*257.2*)	693.6 (*400.1*)	681.7 (*345.1*)	1292.9 (*649.1*)	1562.3 (*725.4*)	1642.9 (*736.7*)
Mean Word Count (*SD*)	210.3 (*102.7*)	274.9 (*164.1*)	418.3 (*240.8*)	405.7 (*203.4*)	759.2 (*389.7*)	939.2 (*429.6*)	980.6 (*429.9*)

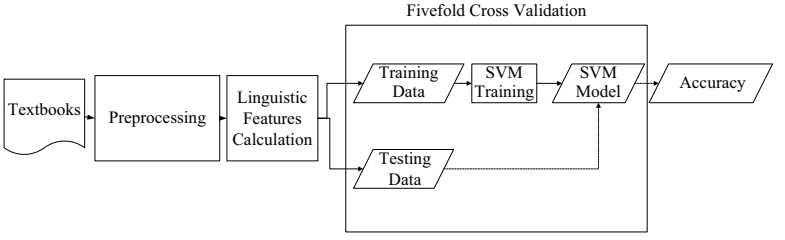

Figure 8. Flowchart of building a readability model with general linguistic features.

4.1 Methods

4.1.1 Materials

The experimental materials for this study were adopted from the third- to ninth-grade textbooks published in 2009 by three major publishers in Taiwan, Nan I (2009), Han Lin (2009), and Kang Hsuan (2009), and included 1441 social science articles and 772 natural science articles. Each article was an independent lesson from one of these textbooks. Each article, in addition to the main content of the text, also contained headings, descriptive text, punctuation marks, and the descriptions of tables/figures. Homework exercises, guided learning questions, and extracurricular content were not considered. In order to increase the accuracy of our articles, we did not use optical character recognition technology to digitize the textbook articles. Instead, we manually inputted the characters content into unformatted plain text files and manually combed through the files for errors. The distribution of text lengths among grade levels can been seen in Table 1. These textbooks were written and edited according to the knowledge/skill levels formulated in the curriculum standards established by the Ministry of Education, Taiwan, and are therefore representative for the average knowledge background of general students from the third to ninth grades. The interested reader can refer to general guidelines for more detailed information about curriculum guidelines of 12-year basic education (Ministry of Education 2014).

4.1.2 Procedure

The experiment procedure of this study is as shown in Figure 8 and is explained in the following subsections.

4.1.2.1 Preprocessing. Segmentation is one of the most basic and important procedures in the preprocessing of texts. Its main function is to ensure the accurate extraction of terms and that they

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

Table 2. Social science texts error matrix of Study 1

		Predicted Grade Level								
Social Science		3	4	5	6	7	8	9	Accuracy	Average Accuracy (Adjacent Accuracy)
	3	52	9	1	6	8	1	3	65.00%	
	4	11	28	3	2	20	3	7	37.84%	
	5	1	3	33	10	24	7	7	38.82%	
Actual Grade Level	6	12	3	16	27	12	3	8	33.33%	55.10% (79.32%)
	7	10	10	12	15	229	65	48	58.87%	
	8	4	2	5	4	59	268	65	65.85%	
	9	4	7	5	10	61	81	157	48.31%	

are assigned the correct part-of-speech according to sentence structure. The WECAn parser was trained on the Sinica Corpus 4.0, and its word segmentation accuracy is 93% (Sung et al. 2016a). This study employed WECAn to perform parsing of Chinese texts to facilitate the subsequent experimental procedure.

4.1.2.2 General linguistic feature calculation. This study used the CRIE, an automatic analysis system for text readability indices (Sung et al. 2016a), to perform numerical computation of the general linguistic features of the social and natural science textbook articles. The 24 general linguistic features employed are the same as those used for the readability model developed by Sung *et al.* (2013), which included four levels of features: the lexical (e.g., number of characters and words), semantic (e.g., number of content words), syntactic (e.g., average of sentence length), and cohesion (e.g., number of pronouns) levels (see Appendix for details). These 24 general linguistic features have been shown to be reasonably accurate (72.92%) when applied to leveling articles in Chinese literature textbooks. Thus, in the current study these 24 general linguistic features were used to train the readability models for social and natural science textbooks to see if these general language features are also suitable for analyzing the readability of domain-specific texts.

4.1.2.3 Training and validating the readability model. Feng *et al.* (2010), Petersen and Ostendorf (2009), and Sung *et al.* (2015a) have all demonstrated that the performance of machine learning based on SVM is superior to that of traditional readability models (i.e., regression models). Therefore, this study used the LIBSVM software (Chang and Lin 2011) to train the readability model and used fivefold cross-validation to verify model effectiveness. To do the fivefold cross-validation, the experimental materials were divided into five subsets: Four of them were used for training the models, and the fifth was used for testing (i.e., validating the model). The process was repeated five times, with each subset used once as the testing data. The accuracy of the model was calculated as the average of the five results.

4.2 Results

The results of the experiment and the classification error matrices are given in Tables 2 and 3.

Tables 2 and 3 show that the prediction accuracy for social science texts is 55.10%, while the accuracy for natural science is even lower, at a mere 49.35%. If we allow plus/minus one level error in the calculation of accuracy (McNamara, Crossley, Roscoe 2013; Sung *et al.* 2015b), the resultant adjacent accuracies for social science and natural science texts are 79.32% and 81.99%, respectively. These results echo past research findings that readability models using general linguistic features do not perform well when used to predict the readability of domain-specific texts (Collins-Thompson 2014).

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

Table 3. Natural science texts error matrix of Study 1

Natural Science		Predicted Grade Level 3	4	5	6	7	8	9	Accuracy	Average Accuracy (Adjacent Accuracy)
	3	51	11	9	1	0	0	0	70.83%	
	4	27	18	12	9	1	0	0	26.87%	
	5	11	13	24	14	0	4	1	35.82%	
Actual Grade Level	6	9	11	12	25	1	3	1	40.32%	49.35% (81.99%)
	7	0	0	4	3	123	25	17	71.51%	
	8	0	0	4	3	41	80	47	45.71%	
	9	0	0	0	3	48	46	60	38.22%	

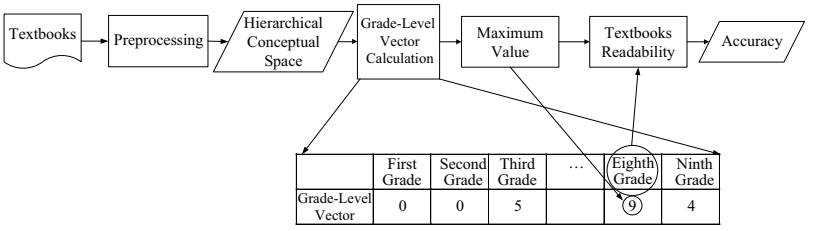

Figure 9. Flowchart of predicting the readability of domain-specific texts using GLVMV.

5. Study 2: Constructing a readability model for domain-specific texts through the hierarchical conceptual space as another baseline study

Study 2 used hierarchical conceptual space to calculate the difficulty distribution of conceptual terms in a domain-specific text, and then matched the text with the grade value that corresponded to the difficulty level where most terms were distributed. This method was then checked for its accuracy in predicting text readability and its performance was built as a baseline for the model performance in Study 3.

5.1 Methods

5.1.1 Materials

The materials were the same as Study 1.

5.1.2 Procedure

This method of extracting grade-level vectors through the hierarchical concept space put forward by Chang *et al.* (2013) shows the difficulty distribution of conceptual terms in a domain-specific text. The model tested in Study 2 was dubbed as "grade-level vector majority vote," henceforth abbreviated as GLVMV. The experimental procedure for this study is shown in Figure 9 and is explained in the following subsections.

5.1.2.1 Preprocessing. The preprocess was the same as Study 1.

5.1.2.2 Generating hierarchical conceptual space for social and natural science texts. We used the method proposed by Chang *et al.* (2013), which was described in Section 3, to obtain the hierarchical conceptual space from training data set for the social and natural science texts.

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

Table 4. Social science texts error matrix of Study 2

		Predicted Grade Level								
Social Science		3	4	5	6	7	8	9	Accuracy	Average Accuracy (Adjacent Accuracy)
	3	69	6	5	0	0	0	0	86.25%	
	4	17	49	5	2	1	0	0	66.22%	
	5	14	24	37	3	2	5	0	43.53%	
Actual Grade Level	6	26	18	1	25	4	5	2	30.86%	46.22% (56.00%)
	7	91	63	100	22	99	12	2	25.45%	
	8	38	49	52	22	19	207	20	50.86%	
	9	36	22	22	52	5	8	180	55.38%	

Table 5. Natural science texts error matrix of Study 2

		Predicted Grade Level								
Natural Science		3	4	5	6	7	8	9	Accuracy	Average Accuracy (Adjacent Accuracy)
	3	55	6	0	0	1	5	5	76.39 %	
	4	11	39	2	1	6	1	7	58.21 %	
	5	9	4	17	5	7	17	8	25.37 %	
Actual Grade Level	6	13	1	3	11	4	18	12	17.74 %	70.98% (83.16%)
	7	0	0	0	2	143	22	5	83.14 %	
	8	1	0	0	1	7	159	7	90.86 %	
	9	0	1	2	6	3	21	124	78.98 %	

5.1.2.3 Calculating and validating grade-level vectors that predict the grade level of domain-specific texts. After building a hierarchical conceptual space, we used it to calculate the grade-level vector of every text (as described in Section 3). The model then identified the grade level that possessed the largest number of conceptual terms in the grade-level vector of a text and assigned the grade level of difficulty to the text accordingly (hence the name "grade-level vector majority vote," GLVMV). In Figure 9, for example, the largest value in the grade-level vector is the nine conceptual terms that belonged to the eighth grade. The amounts of conceptual terms belonging to the other grades were all less than 9 (e.g., only five conceptual terms in the fifth grade). As a result, we assigned the article a difficulty level of eighth grade. This method assumes that the grade level of a domain-specific text can be traced by the concentration of conceptual terms that characterize the knowledge typically acquired at that school age. In the follows, we tested the assumption and assessed the effectiveness of our model by using a fivefold cross-validation. The validation procedure was the same as described in Study 1 (Sung *et al.* 2015a).

5.2 Results

The results of fivefold cross-validation of the conceptual space model for predicting the difficulty levels of 1441 social science articles and 772 natural science articles are presented in Tables 4 and 5.

Tables 4 and 5 show that the prediction accuracy was 46.22% for social science texts and 70.98% for natural science texts. We used McNemar's test (McNemar 1947) to test the statistical significance of the accuracy differences revealed in Study 1 and Study 2. For both social science and

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

Table 6. Grade-level vectors and corresponding conceptual terms in a natural science text

Text Title	A Rich World of Life						
Concept Difficulty	Third Grade	Fourth Grade	Fifth Grade	Sixth Grade	Seventh Grade	Eighth Grade	Ninth Grade
Grade-Level Vector	6	5	7	13	12	2	3
Conceptual Term	Air	Energy	Sun	Ocean	Organism	Ultrasound	Earth
	Plant	Salt	Planet	Oxygen	Photosynthesis	Echo	Region
	Ingredient	Light	Pitcher plant	Microorganism	Respiration		Sea level
	Gas	Insect	Animal	Carbon dioxide	Camouflage		
	Soil	Leaf	Tree branch	Cave	Desert		
	Seed		Forest	Forest	Living creature		
			Prey	Environment	Penguin		
				Atmosphere	Bacteria		
				Raw material	Tree frog		
				Seal	Bat		
				Hot spring	Cactus		
				Land	Humans		
				Kandelia obovate (mangrove plant)			

natural science texts, the prediction accuracy rate was significantly different: $X^2_{1.1} = 23.443, p < .001,$ and $X^2_{1.1} = 82.751, p < .001,$ *respectively*.

These results provide empirical support for the following. First, using the LSA technique to construct a hierarchical conceptual space where difficulty levels of conceptual terms are indexed provides an effective framework of representing the difficulty level of domain-specific texts. Second, compared to the general-linguistic-feature-based model, the accuracy rate of predicting natural science text levels based on this model increased substantially. However, there is still room for improvement when it comes to social science texts.

In our error analysis of a large quantity of misclassified texts, we discovered that the grade level estimated by the maximum value of a grade-level vector of a text is not always a good judge of its readability. Table 6 is an example of a seventh-grade natural science textbook article, but the numbers of conceptual terms that fall into the difficulty levels of grades 6 and 7 are close, which may mean the text should be categorized into the sixth grade instead of the seventh. From this we can see that when domain-specific texts are describing pieces of knowledge, they will use a large number of conceptual terms. If this feature can be utilized alongside a superior classification strategy then it will further bolster the overall readability model's performance. We will discuss this topic further in Study 3.

6. Study 3: Constructing and validating readability models using either hierarchical conceptual space or bag-of-word-based features for domain-specific texts

Although Studies 1 and 2 provided empirical evidence for the better performance in predicting natural science text by using hierarchical conceptual space than by using general linguistic features, there was a critical methodological difference between the two studies. In Study 1, a machine

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

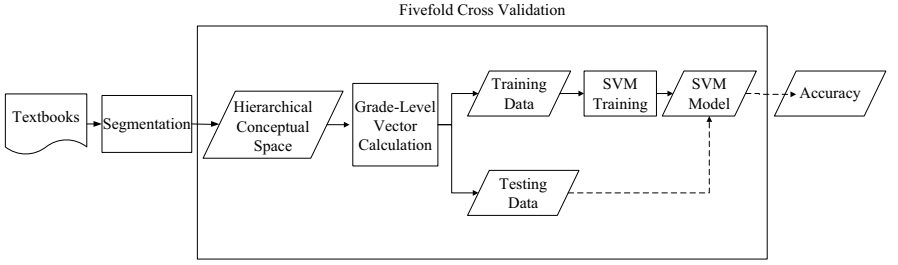

Figure 10. Flowchart of building a readability model using hierarchical conceptual space trained on SVM.

learning approach for training the readability model was used, while in Study 2 a maximum value of grade-level vector approach was employed. Echoing the machine learning method of Study 1 (i.e., using the SVM classifier), in Study 3 we augmented the conceptual-space-based model in Study 2 with the SVM. This approach allows for the influence of different classification strategies (e.g., using the general linguistic features, the LSA-based hierarchical conceptual space, and the machine learning algorithms) on classification accuracy to be observed (Sung *et al.* 2015a).

To expand model capabilities, in Study 3 we experimented optimization of the readability models with more NLP techniques. First, we combined grade-level vectors with general linguistic features and then used SVM to train the readability model. It is of particular interest to determine whether the combination of different features enhances the efficacy of the readability model in tackling domain-specific texts. We also used two bag-of-words-based features to train an SVM model separately. The first bag-of-words feature, TF (term frequency), is often seen in natural language processing. The other, Word2Vec, was recently released by Google and has quickly gained in popularity. We trained using Word2Vec's Continuous Bag-of-Words (CBOW) and Skip-gram approaches on a separate set of training data to derive word vectors. We added up the vectors of words in an article and also averaged the summed vector to represent the vector of the article (Le and Mikolov 2014). Then, the Word2Vec-based readability models were also trained by SVM.

6.1 Methods

6.1.1 Materials

The materials were the same as in Study 1.

6.1.2 Procedure

The experimental procedure for Study 3 is shown in Figure 10 and is explained in the following subsections.

6.1.2.1 Preprocess. The preprocess was the same as Study 1.

6.1.2.2 Generating hierarchical conceptual space for social and natural science texts. The generation of hierarchical conceptual space was the same as Study 2.

6.1.2.3 Training and validating the readability model. The procedure of extracting the grade-level vectors of a text was the same as was done for Study 2. After obtaining the grade-level vectors, this study used the LIBSVM (Chang and Lin 2011) to train the model. The procedure of fivefold cross-validation was the same as in Study 1 (Sung *et al.* 2015a).

6.2 Results

The results of fivefold cross-validation of the expanded readability models for predicting the difficulty level of 1441 social science articles and 772 natural science articles are presented in Tables 7 and 8.

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

Table 7. Social science texts error matrix of Study 3

Social Science / Models(DIM)	Actual Grade Level	Predicted Grade Level 3	4	5	6	7	8	9	Accuracy	Average Accuracy (Adjacent Accuracy)
	3	55	5	1	1	15	1	2	68.75%	
	4	8	32	3	7	21	2	1	43.24%	
	5	1	10	33	0	28	10	3	38.82%	
Grade-Level Vector Model: Grade-level vector + SVM(7)	6	3	3	3	13	20	11	28	16.05%	68.98% (83.62%)
	7	9	6	17	10	285	45	17	73.26%	
	8	0	0	6	2	50	320	29	78.62%	
	9	1	0	1	22	17	28	256	78.77%	
	3	18	2	42	9	5	3	1	22.50%	
	4	1	15	44	3	5	3	3	20.27%	
	5	0	0	76	3	5	1	0	89.41%	
Term frequency + SVM(19960)	6	0	0	32	41	2	4	2	50.62%	66.76% (78.07%)
	7	0	0	133	10	187	49	10	48.07%	
	8	0	0	24	2	2	377	2	92.63%	
	9	0	0	55	5	1	16	248	76.31%	
	3	58	10	0	3	8	1	0	72.50%	
	4	7	55	2	1	9	0	0	74.32%	
	5	0	4	36	6	30	6	3	42.35%	
Word2Vec(CBOW) + SVM(7)	6	3	4	7	31	16	6	14	38.27%	70.85% (86.68%)
	7	4	3	11	8	298	47	18	76.61%	
	8	0	0	2	3	57	319	26	78.38%	
	9	1	1	3	16	42	38	224	68.92%	
	3	45	12	0	3	17	0	3	56.25%	
	4	5	52	3	1	10	0	3	70.27%	
	5	0	9	36	3	28	8	1	42.35%	
Word2Vec(Skip-gram) + SVM(7)	6	4	7	3	29	17	6	15	35.80%	71.48% (85.57%)
	7	8	10	18	9	300	34	10	77.12%	
	8	0	1	4	5	40	328	29	80.59%	
	9	0	3	2	9	32	39	240	73.85%	
	3	71	1	0	0	7	0	1	88.75%	
	4	4	58	1	1	10	0	0	78.38%	
	5	0	3	63	0	16	3	0	74.12%	

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

Table 7. (Continued)

Social Science / Models(DIM)	Actual Grade Level	Predicted Grade Level 3	4	5	6	7	8	9	Accuracy	Average Accuracy (Adjacent Accuracy)
Combined Model: Grade-level vector + General linguistic features + SVM(31)	6	0	1	0	64	8	4	4	79.01%	86.68% (93.41%)
	7	8	2	2	3	344	23	7	88.43%	
	8	0	0	1	2	30	362	12	88.94%	
	9	1	0	0	8	17	12	287	88.31%	

Table 8. Natural science texts error matrix of Study 3

Natural Science / Models(DIM)	Actual Grade Level	Predicted Grade Level 3	4	5	6	7	8	9	Accuracy	Average Accuracy (Adjacent Accuracy)
	3	54	10	5	3	0	0	0	75.00%	
	4	13	44	4	4	0	0	2	65.67%	
	5	9	6	27	11	3	7	4	40.30%	
Grade-Level Vector Model: Grade-level vector + SVM(7)	6	11	4	12	23	3	5	4	37.10%	73.96% (87.95%)
	7	0	0	0	2	152	12	6	88.37%	
	8	0	1	3	2	15	146	8	83.43%	
	9	0	3	4	7	6	12	125	79.62%	
	3	28	4	5	6	6	0	23	38.89%	
	4	4	15	9	2	5	2	30	22.39%	
	5	0	0	21	8	8	1	29	31.34%	
Term frequency + SVM(21701)	6	0	0	11	14	6	0	31	22.58%	56.22% (72.54%)
	7	0	0	0	0	106	5	61	61.63%	
	8	0	0	0	0	6	97	72	55.43%	
	9	0	0	0	0	3	1	153	97.45%	
	3	54	6	8	4	0	0	0	75.00%	
	4	40	14	7	6	0	0	0	20.90%	
	5	15	6	23	18	1	3	1	34.33%	
Word2Vec(CBOW) + SVM(7)	6	16	7	9	26	3	0	1	41.94%	71.37% (89.51%)
	7	1	1	2	1	154	12	1	89.53%	
	8	0	0	2	1	19	141	12	80.57%	
	9	0	0	1	0	10	7	139	88.54%	

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

Table 8. (Continued)

Natural Science / Models(DIM)	Actual Grade Level	Predicted Grade Level 3	4	5	6	7	8	9	Accuracy	Average Accuracy (Adjacent Accuracy)
	3	49	13	4	6	0	0	0	68.06%	
	4	32	20	8	7	0	0	0	29.85%	
	5	15	2	32	14	0	3	1	47.76%	
Word2Vec(Skip-gram) + SVM(7)	6	11	2	20	23	1	2	3	37.10%	71.24% (91.06%)
	7	0	0	1	1	151	17	2	87.79%	
	8	0	0	2	0	15	146	12	83.43%	
	9	0	0	1	1	8	18	129	82.17%	
	3	56	8	3	5	0	0	0	77.78%	
	4	10	40	7	10	0	0	0	59.70%	
	5	9	8	32	12	1	4	1	47.76%	
Combined Model: Grade-level vector + General linguistic features + SVM(31)	6	9	5	16	27	1	0	4	43.55%	75.91% (91.19%)
	7	0	0	0	1	152	12	7	88.37%	
	8	0	0	2	0	14	146	13	83.43%	
	9	0	0	0	2	6	16	133	84.71%	

Table 9. McNemar test results for social science text readability models

Social Science	Grade-Level Vector Model (Study3)	Combined Model (Study3)
Linguistic Features (Study1)	$X^2_{(1.1)} = 65.565^*$	$X^2_{(1.1)} = 371.380^*$
GLVMV (Study2)	$X^2_{(1.1)} = 220.928^*$	$X^2_{(1.1)} = 503.305^*$
Term Frequency (Study3)	$X^2_{(1.1)} = 1.800$	$X^2_{(1.1)} = 160.070^*$
Word2Vec (CBOW) (Study3)	$X^2_{(1.1)} = 1.499$	$X^2_{(1.1)} = 121.531^*$
Word2Vec (Skip-gram) (Study3)	$X^2_{(1.1)} = 2.876$	$X^2_{(1.1)} = 118.514^*$

* $p < .001$

Tables 7 and 8 show that when the model using grade-level vectors was augmented with SVM machine learning, its prediction accuracies for social science and natural science texts improved from 46.22% and 70.98% to 68.98% and 73.96%, respectively. The accuracy rates of the combined model (i.e., grade-level vector + general linguistic features + SVM) for social science and natural science texts achieved the even higher accuracy rates of 86.68% and 75.91%, respectively, which was the best performance among all the models. The McNemar test results for each readability model are given in Tables 9 and 10, showing that the combined model's performance was significantly different from all other models with most of the *p*-values being under .001.

Furthermore, after using different classification strategies, we can see that the example article in Table 6 was misjudged as belonging to the sixth grade in Study 2, whereas Study 3 correctly classified it as belonging to the seventh grade. The enhancing effect of the SVM indicates the

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

Table 10. McNemar test results for natural science text readability models

Natural Science	Grade-Level Vector Model (Study3)	Combined Model (Study3)
Linguistic Features(Study1)	$X^2_{(1.1)} = 110.935^{***}$	$X^2_{(1.1)} = 139.184^{***}$
GLVMV (Study2)	$X^2_{(1.1)} = 5.319^{*}$	$X^2_{(1.1)} = 12.223^{***}$
Term Frequency (Study3)	$X^2_{(1.1)} = 69.796^{***}$	$X^2_{(1.1)} = 84.448^{***}$
Word2Vec (CBOW) (Study3)	$X^2_{(1.1)} = 1.805$	$X^2_{(1.1)} = 6.531^{*}$
Word2Vec (Skip-gram) (Study3)	$X^2_{(1.1)} = 2.073$	$X^2_{(1.1)} = 7.040^{**}$

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

importance of taking into consideration the difficulty-level distribution of all the (instead of the individual) conceptual terms in an article, as the hallmark of machine learning is the ability to discern meaningful patterns from seemingly unrelated information. The results of Study 3 also suggest that general linguistic features are employed more differentially for social science texts than for natural science texts, which is reflected by the great improvement of the model when incorporating the general linguistic features.

When the Word2Vec models and the grade-level vector model (Study 3) were trained using the same seven dimensions of SVM, the Word2Vec models (whether Skip-gram or CBOW) performed better than the grade-level vector model (Study 3) for social science texts, but the latter performed slightly better than the former for natural science texts. When compared with the TF model, the grade-level vector model (Study 3) showed superior accuracy for both social and natural science texts. This is especially noteworthy given that the huge dimensions of the TF model make using it to train readability models extremely time-consuming.

Overall, these accuracies demonstrate that grade-level vectors, whether applied to social or natural science texts, are a viable method of training readability models. To further explain, from Figure 7 we can see that the vocabulary difficulty distributions made by grade-level vectors are easy for people to understand and to apply to readability assessments. These distributions can also be provided to editors to help develop editing guidelines for new texts. In contrast, Word2Vec and TF cannot offer such assistance.

In recent years, intelligibility has become a vital concern in machine learning. In certain applications, such as health care (Caruana *et al.* 2015), education (Chang and Sung 2019; Hsu *et al.* 2018; Lin *et al.* 2019; Lu and Chen 2019; Lee, Chang, and Tseng 2016), and speech recognition (Chen and Hsu 2019), the intelligibility of a model may far outweigh its accuracy since it could be a helpful feedback for the users (Chang, Sung, and Hong 2015; Sung *et al.* 2016*b*). The reason is obvious—in order for the end-users of a system to trust and act upon the predictions, they need to understand what they are being told (Ribeiro, Singh, and Guestrin 2016; Samek, Wiegand, and Müller 2017). As for text readability, if the model lacks adequate explanatory ability, it would unfortunately be like a mysterious black box whose operation remains unknown and inexplicable. As an attempt to develop readability models that are not only effective, but also interpretable, our study shows that using grade-level vectors to capture the distribution of conceptual terms difficulty among grades can help the reader to determine whether readability predictions are reasonable and also help researchers to further use or improve readability features.

7. Discussion

As in all traditional readability formulas, the general language features used in Study 1 only account for information gleamed from shallow article structures, which cannot effectively reflect the difficulty inherent in a specific domain of knowledge. For example, in Chinese, *morning glory* and *electromagnetic waves* are both three-character words. However, students encounter *morning*

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

glory in third grade, whereas they do not learn about *electromagnetic waves* until in seventh grade. The difficulty of the two terms is clearly different, yet they are assigned the same difficulty by several general linguistic features. This is the reason why general linguistic features are less discriminating when applied to domain knowledge texts. General linguistic features, then, lead to readability model prediction with low accuracy and can even produce serious overestimation or underestimation of text readability (Begeny and Greene 2014).

Therefore, this study uses LSA to produce hierarchical conceptual space as a feature for readability model. Compared with the prediction based on a set of general linguistic features trained by SVM classifier, estimation of the difficulty levels of social science texts using only grade-level vectors with the maximum value (GLVMV) achieved the 46.22% of accuracy, which was only 8.88% less. The accuracy of assessing the difficulty levels of natural science texts improved by 21.63% to reach 70.98%. Given that the model of Study 1 incorporated machining learning techniques while the model in Study 2 (i.e., GLVMV) was not assisted by classification algorithms, the predictive power of the grade-level vectors was very impressive. These experimental results suggest that hierarchical conceptual space is capable of representing the knowledge structure of domain-specific texts. The difficulty distribution of knowledge-relevant terms can be revealed by the grade-level vectors. Our study also indicates that domain-specific texts involve specialized knowledge or domain-specific concepts such that these texts must employ relevant, specific terms in large numbers when describing or explaining a concept. In sum, the results of Study 2 show these grade-level vectors are quite accurate reflections of the difficulty of domain-specific text.

Comparing the experimental results of Study 2 and Study 3, we found that although both use hierarchical conceptual space as the readability feature, using a machine learning algorithm ultimately led to gains in classification accuracy for the social science texts plus 22.76% to achieve the accuracy of 68.98% and for natural science texts plus 2.98% to reach 73.96%. These results show that although grade-level vectors quite accurately reflect the conceptual difficulty of domain knowledge, matching the vector's maximum value with a text's grade level is not the best classification strategy, because this method overlooks the difficulty levels of other conceptual terms in the text, and the readability of a text should be determined by all of the domain-specific concepts that it contains.

Comparing the experimental results of Study 1 and Study 3, we found that when using the same machine learning algorithm, the accuracy of a readability model that uses grade-level vectors as its feature is higher than that of a model that uses general linguistic features. For social science texts, the former exceeds the latter by 13.88% to achieve the accuracy of 68.98%, and for natural science texts, the difference is an even larger percent of 24.61, reaching the high accuracy of 73.96%. These results show that for assessing domain-specific texts, the method of constructing and applying grade-level vectors proposed by the current study is more suitable than using general linguistic features.

A comparison of the readability models in the three studies indicates that their performances vary greatly. For social science texts, the highest accuracy is 86.68%, which is 40.46% higher than the lowest percent of 46.22%. For natural science texts, the highest accuracy is 75.91%, which is 26.56% higher than the lowest accuracy of 49.35%. These results suggest that the selection of readability features and model training methods has a direct impact on the performance of readability models. In this study, the readability model that combines a grade-level vector and general linguistic features performed even better than those that use the word vectors created by Word2Vec as a feature.

The research by Begeny and Greene (2014), which used a word list of 3000 familiar words for the Dale–Chall formula (Chall and Dale 1995), performed the best out of all the traditional readability models used by being able to achieve an accuracy rate of 41.66%. This was compared to the Flesch-Kincaid (Flesch 1948), FOG (Gunning 1952), Forcast (Sticht 1973), Fry (Fry 1968), Lexile (Stenner 1996), PSK (Powers, Sumner, and Kearl 1958), SMOG (McLaughlin 1969), and Spache (Spache 1953) readability models. Begeny and Greene (2014: 13) believe this is because

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

the "percentage of high frequency words may be a good gauge of text difficulty." Word frequency can be used to measure the difficulty of vocabulary and also highlights the importance that word difficulty has on the overall readability of an article. In contrast to the Dale–Chall formula, which uses word frequency to construct its word list, this article uses hierarchical conceptual space to generate the difficulty of conceptual terms and found, through multiple studies, that the resultant readability model is superior to the Dale–Chall formula and outperforms it by +27.31% to achieve an accuracy of 68.98% (compared with the social science model) and +32.29% to reach 73.96% (compared with the natural science model), depending on domain. This shows that the hierarchical conceptual space proposed by this study is superior to solely using a word list derived from word frequency.

Regarding the comparison of hierarchical conceptual space with other bag-of-words methods, such as TF and Word2Vec, we found that the grade-level vector model (i.e., grade-level vector + SVM) performed slightly better than the other methods (i.e., TF + SVM; Word2Vec + SVM) for natural science texts but not for social science texts. One of the possible reasons may be that the natural science texts have more topics with hierarchical conceptual space as shown in Figure 3. However, the conceptual hierarchies in social science may not as obvious as in natural science; therefore, the hierarchical conceptual space and the bag-of-words methods have similar performance in predicting the readability levels. Further evaluation based on more types of texts will be needed to compare the effectiveness of the two methods for the prediction of social science text readability.

8. Conclusions and implications

How well does a general-linguistic-feature-based readability model predict the difficulty level of domain-specific text? Our research provided empirical support for the idea that employing common general linguistic features is not suitable for use on domain-specific texts, because the accuracy rate of prediction was less than 60%. In contrast, a method that used a hierarchical conceptual space, which represented the domain-specific knowledge learned by students in different learning stages, effectively estimated the readability of domain-specific texts, outperforming the general-linguistic-feature-based model by 13.88% and 24.61% to achieve accuracies of 68.98% and 73.96% for social science and natural science texts, respectively. The model combining grade-level vectors with general linguistic features outperformed a general-linguistic-features-only model by 31.58% and 26.56% to achieve accuracy rates of 86.68% and 75.91% for social science and natural science texts, respectively. This indicates that the readability features presented in this paper are not only suitable for representing domain-specific texts, but can also complement linguistic features that are commonly used for general text analysis. In other words, readability models trained only on common general linguistic features are not suitable for assessing the readability of domain-specific texts. However, when combined with suitable readability features, such as grade-level vectors, the general linguistic features can indeed enhance the performance of readability models. The findings above have the following implications for further research and practices.

Firstly, hierarchical conceptual space, which was extracted from the knowledge base of students in different learning stages, is an appropriate and valid tool for assessing the readability/difficulty levels of domain-specific texts and can be combined with machine learning algorithms to predict the readability of both social and natural science texts with good performance. A hierarchical conceptual space can not only serve as a readability feature, but, through the use of data visualization software, also present a text's conceptual space diagrammatically. This could serve as a tool for instructors or students when working through a text and could potentially supplement any outlines or summaries provided for the text.

Secondly, teachers, book editors, or publishers should consider combining the linguistic-features-based approach with conceptual-space-based approach of readability models when

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

leveling their teaching/learning materials, especially when the targeted materials are domain specific or the targeted learners are content-area readers.

Thirdly, since the readability models proposed by our research are not equally effective for text in different domains and are constructed based only on reading materials for third to ninth graders, future research can address the issue of model generalizability by applying the model to texts in more domains, or expanding the grade levels for higher grades of learners.

Finally, the research can certainly stand improvement in its effort to refine readability assessment models and improve its ability to explain the logic behind the distribution of domain-specific concepts in text. As a case in point, the accuracy of our predicting sixth-grade social science text difficulty lagged behind that of the other grades (see Tables 4 and 7). Upon examination, we found that these sixth-grade texts covered a wide range of topics. For example, one of them introduces the world of ancient civilizations and discusses culture, economics, martial, and religious issues, such as river valley civilizations and Egyptian, Greek, Roman, Mayan, Indian, Islamic, and Chinese cultures, to name a few. The broad subject of this text leads to the result that many conceptual terms were not classified into the sixth grade. For example, the term *democratic politics* is a word that, while mentioned within this text, is fully discussed in 23 eighth- and ninth-grade articles. This resulted in the concept of democratic politics being highly related to the eighth and ninth grades. The readability model ultimately predicted this sixth-grade text as a ninth-grade one, because it used many conceptual terms the model has labeled as ninth-grade concepts. How to more accurately predict the difficulty levels of texts with such broad topics is a subject worthy of future research. Another potential avenue for improvement is to change the grade-level vectors proposed herein to soft assignment (e.g., retain the complete cosine similarities between all conceptual terms and grade levels) to generate the difficulty distribution of conceptual terms. This may retain even more information which, in turn, could help a readability model classify texts that cover a variety of topics.

In the future, we hope to develop even more powerful readability models. Currently, our proposed readability model can accurately rate the difficulty levels of texts within a single domain. For our future research, we aim to design models with which articles from a wide range of domains can be put together and then be rated at the same time. Development of such a general-purpose readability model could benefit from the incorporation of more domain-specific features (e.g., grade-level vector of different domain) and generic readability features (e.g., Word2Vec and GloVe (Pennington, Socher, and Manning 2014)), the advancement in NLP techniques, and a better understanding of reading process (Crossley *et al.* 2017).

Acknowledgements. The collection of empirical data had been supported by the Ministry of Science and Technology (MOST-107-2511-H-003-022-MY3; MOST-108-2622-8-003-002-TS1) and the Higher Education Sprout Project of Ministry of Education, Taiwan.

References

Ambati B.R., Reddy S. and Steedman M. (2016). Assessing relative sentence complexity using an incremental CCG parser. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, California, United States, pp. 1051–1057.

American Association for the Advancement of Science, & National Science Teachers Association (2007). Atlas of Science Literacy: Project 2061. Washington: AAAS.

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

Bailin A. and Grafstein A. (2016). *Readability: Text and Context*. London: Palgrave Macmillan.

Bédard J. and Chi M.T.H. (1992). Expertise. *Current Directions in Psychological Science* **1**(4), 135–139.

Begeny J.C. and Greene D.J. (2014). Can readability formulas be used to successfully gauge difficulty of reading materials? *Psychology in the Schools* **51**(2), 198–215.

Belden B.R. and Lee W.D. (1961). Readability of biology textbooks and the reading ability of biology students. *School Science and Mathematics* **61**(9), 689–693.

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

Borst A., Gaudinat A., Grabar N. and Boyer C. (2008). Lexically-based distinction of readability levels of health documents. *Acta Informatica Medica* **16**(2), 72–75.

Caruana R., Lou Y., Gehrke J., Koch P., Sturm M. and 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*, Sydney, Australia, pp. 1721–1730.

Cecconi M., De Backer D., Antonelli M.l., Beale R., Bakker J., Hofer C., Jaeschke R., Mebazaa A., Pinsky M.R., Teboul J.L., Vincent J.L. and Rhodes A. (2014). Consensus on circulatory shock and hemodynamic monitoring. Task force of the European Society of Intensive Care Medicine. *Intensive Care Medicine* **40**(12), 1795–1815.

Chall J.S. and Dale E. (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge, MA: Brookline Books.

Chang C.C. and Lin C.J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2**(3), 1–27.

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

Chang T.H., Sung Y.T. and Lee Y.T. (2013). Evaluating the difficulty of concepts on domain knowledge using latent semantic analysis. In *Proceedings of International Conference on Asian Language Processing*, Urumqi, China, pp. 193–196.

Chang T.H., Sung Y.T. and Hong J.F. (2015). Automatically detecting syntactic errors in sentences written by learners of Chinese as a foreign language. International Journal of Computational Linguistics and Chinese Language Processing 20(1):49–64.

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

Chen Y.T., Chen Y.H. and Cheng Y.C. (2013). Assessing Chinese readability using term frequency and lexical chain. *International Journal of Computational Linguistics & Chinese Language Processing* **18**(2), 1–18.

Chi M.T.H., Glaser R. and Farr M. (eds) (1988). *The Nature of Expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Collins-Thompson K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics* **165**(2), 97–135.

Crossley S.A., Skalicky S., Dascalu M., McNamara D.S. and Kyle K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes* **54**(5–6), 340–359.

Dale E. and Chall J.S. (1949). The concept of readability. *Elementary English* **26**(1), 19–26.

De Clercq O. and Hoste V. (2016). All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics* **42**(3), 457–490.

Dell'Orletta F., Venturi G. and Montemagni S. (2012). Genre-oriented readability assessment: A case study. In *Proceedings of the Workshop on Speech and Language Processing Tools in Education*, Mumbai, India, pp. 91–98.

Ding L. (2007). A model of hierarchical knowledge representation? Toward knowware for intelligent systems. *Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII)* **11**(10), 1232–1240.

DuBay W.H. (2004). The principles of readability. Available at http://files.eric.ed.gov/fulltext/ED490073.pdf (accessed January 2017).

Etringer B.D., Hillerbrand E. and Claiborn C.D. (1995). The transition from novice to expert counselor. *Counselor Education and Supervision* **35**(1), 4–17.

Feng L., Jansche M., Huenerfauth M. and Elhadad N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, Association for Computational Linguistics, Stroudsburg, PA, pp. 276–284.

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

François T. and Miltsakaki E. (2012). Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations*, Association for Computational Linguistics, Stroudsburg, PA, pp. 49–57.

Freimuth V.S. (1979). Assessing the readability of health education messages. *Public Health Reports* **94**(6), 568–570.

Fry E. (1968). A readability formula that saves time. *Journal of Reading* **11**(7), 513–578.

Fry E. (1990). A readability formula for short passages. *Journal of Reading* **33**(8), 594–597.

Furnas G.W., Deerwester S., Dumais S.T., Landauer T.K., Harshman R.A., Streeter L.A. and Lochbaum K.E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, NY, USA, pp. 465–480.

Gallagher D.J. and Thompson G.R. (1981). A readability analysis of selected introductory economics textbooks. *The Journal of Economic Education* **12**(2), 60–63.

Golub G.H. and Reinsch C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik* **14**(5), 403–420.

Graesser A.C., Singer M. and Trabasso T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review* **101**(3), 371–395.

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

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

Gruber T.R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition* **5**(2), 199–220.

Gunning R. (1952). *The Technique of Clear Writing*. New York, NY: McGraw-Hill.

Harden R.M. (1999). What is a spiral curriculum? *Medical Teacher* **21**(2), 141–143.

Han L. (2009). Available at https://www.hle.com.tw/ (accessed March 2018).

Hirschfeld L.A. and Gelman S.A. (1994). *Mapping the Mind: Domain-Specificity in Cognition and Culture*. New York, NY: Cambridge University Press.

Hong J.F., Sung Y.T., Tseng H.C., Chang K.E. and Chen J.L. (2016). A multilevel analysis of the linguistic features affecting Chinese text readability. *Taiwan Journal of Chinese as a Second Language* **2**(13), 95–126.

Howcroft D.M. and Demberg V. (2017). Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain, pp. 958–968.

Hunt D.P. (2003). The concept of knowledge and how to measure it. *Journal of Intellectual Capital* **4**(1), 100–113.

Hsu F.Y., Lee H.M., Chang T.H. and 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.

Johns J.L. and Wheat T.E. (1984). Newspaper readability: Two crucial factors. *Journal of Reading* **27**(5), 432–434.

Kanebrant E., Mühlenbock K.H., Kokkinakis S.J., Jönsson A., Liberg C., Geijerstam Å., Folkeryd J.W. and Falkenjack J. (2015). T-MASTER – A tool for assessing students' reading abilities. In *Proceedings of the 7th International Conference on Computer Supported Education (CSEDU 2015)*, Lisbon, Portugal, pp. 220–227.

Kang H. (2009). Available at https://www.knsh.com.tw/ (accessed March 2018).

Kanungo T. and Orr D. (2009). Predicting the readability of short web summaries. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, ACM, New York, NY, USA, pp. 202–211.

Kilgarriff A., Charalabopoulou F., Gavrilidou M., Johannessen J.B., Khalil S., Kokkinakis S.J., Lew R., Sharoff S., Vadlapudi R. and Volodina E. (2014). Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation* **48**(1), 121–163.

Kireyev K. and Landauer T.K. (2011). Word maturity: Computational modeling of word knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Association for Computational Linguistics, Stroudsburg, PA, pp. 299–308.

Klare G.R. (1963). *The Measurement of Readability*. Ames, IA: Iowa State University Press.

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

Landauer T.K. and 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.

Landauer T.K., Foltz P.W. and Laham D. (1998). An introduction to latent semantic analysis. *Discourse Processes* **25**(2–3), 259–284.

Le Q. and Mikolov T. (2014). Distributed representations of sentences and documents. In *Proceedings of International Conference on Machine Learning*, Beijing, China, pp. 1188–1196.

Lee L.H., Chang L.P. and Tseng Y.H. (2016). Developing learner corpus annotation for Chinese grammatical errors. In *Proceedings of the 20th International Conference on Asian Language Processing*, Tainan, Taiwan, pp. 254–257

Lee L.S. and Chen B. (2005). Spoken document understanding and organization. *IEEE Signal Processing Magazine* **22**(5), 42–60.

Lété B., Sprenger-Charolles L. and Colé P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods Instruments, & Computers* **36**(1), 156–166.

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

Lu X. and Chen B. (2019). Computational and corpus approaches to Chinese language learning: An introduction. In Lu X. and Chen,B. (eds), *Computational and Corpus Approaches to Chinese Language Learning*, Singapore: Springer, pp. 6–14.

McConnell C.R. (1982). Readability formulas as applied to college economics textbooks. *Journal of Reading* **26**(1), 14–17.

McLaughlin G.H. (1969). SMOG grading-a new readability formula. *Journal of Reading* **12**(8), 639–646.

McNamara D.S., Louwerse M.M. and 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. and Graesser A.C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes* **47**(4), 292–330.

McNamara D.S., Crossley S.A. and Roscoe R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods* **45**(2), 499–515.

McNemar Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**(2), 153–157.

Mikolov T., Chen K., Corrado G. and Dean J. (2013). Efficient estimation of word representations in vector space. In *Proceeding of the International Conference on Learning Representations (ICLR)*, Scottsdale, Arizona. pp. 1–12. Available at https://arxiv.org/abs/1301.3781.

Miltsakaki E. and Troutt A. (2007). Read-x: Automatic evaluation of reading difficulty of web text. In *Proceeding of the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, Vol. 2007, No. 1, Quebec, Canada, pp. 7280–7286.

Miltsakaki E. and Troutt A. (2008). Real-time web text classification and analysis of reading difficulty. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, Stroudsburg, PA, pp. 89–97.

Ministry of Education (2014). General curriculum guidelines of 12-year basic education. Available at https://www.naer.edu.tw/ezfiles/0/1000/img/67/151760851.pdf (accessed January 2019).

Nan I. (2009). Available at https://trans.nani.com.tw/NaniTeacher/ (accessed March 2018).

Nolen-Hoeksema S., Fredrickson B.L., Loftus G. and Wagenaar W.A. (2009). *Atkinson and Hilgard's Introduction to Psychology*. Boston: Cengage Learning.

Pennington J., Socher R. and Manning C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543.

Petersen S.E. and Ostendorf M. (2009). A machine learning approach to reading level assessment. *Computer Speech & Language* **23**(1), 89–106.

Powers R.D., Sumner W.A. and Kearl B.E. (1958). A recalculation of four adult readability formulas. *Journal of Educational Psychology* **49**(2), 99–105.

Razek J.R. and Cone R.E. (1981). Readability of business communication textbooks-an empirical study. *Journal of Business Communication* **18**(2), 33–40.

Redish J. (2000). Readability formulas have even more limitations than Klare discusses. *ACM Journal of Computer Documentation (JCD)* **24**(3), 132–137.

Ribeiro M.T., Singh S. and Guestrin C. (2016). Model-Agnostic Interpretability of Machine Learning. In *Proceedings of 2016 ICML workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York, NY, pp. 91–95.

Rubenstein H. and Aborn M. (1958). Learning, prediction, and readability. *Journal of Applied Psychology* **42**(1), 28–32.

Salton G. and Buckley C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management* **24**(5), 513–523.

Samek W., Wiegand T. and Müller K.R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ITU Journal: ICT Discoveries Special Issue The Impact of AI on Communication Networks and Services* **1**(1), pp. 1–10.

Schloerke B. (2011). *GGally: Extension to ggplot2*. R package version 3.2.5. Available at https://github.com/ggobi/ggally (accessed January 2017).

Schvaneveldt R.W., Durso F.T. and Dearholt D.W. (1989). Network structures in proximity data. *Psychology of Learning and Motivation* **24**, 249–284.

Schvaneveldt R.W., Durso F.T. and Dearholt D.W. (2017). *Pathfinder network*. Available at http://interlinkinc.net/ (accessed January 2017)

Sherman L.A. (1893). *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*. Boston: Ginn and Company.

Spache G. (1953). A new readability formula for primary-grade reading materials. *The Elementary School Journal* **53**(7), 410–413.

Sparck Jones K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* **28**(1), 11–21.

Stenner A.J. (1996). Measuring reading comprehension with the Lexile framework. Available at https://www.wou.edu/~brownbr/Classes/SpEd_625_F_15/4_Infrmal_Asmt_Resources/1_Reading/8_Readability/Lexiles-Quantiles/1_Reading_Lexiles/Measurg_Reading_wi_Lexiles.pdf (accessed January 2017).

Sticht T.G. (1973). Research toward the design, development and evaluation of a job-functional literacy training program for the United States Army. *Literacy Discussion* **4**(3), 339–369.

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

Sung Y.T., Chen J.L., Cha J.H., Tseng H.C., Chang T.H. and 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.

Sung Y.T., Lin W.C., Dyson S.B., Chang K.E. and Chen Y.C. (2015b). Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal* **99**(2), 371–391.

Sung Y.T., Chang T.H., Lin W.C., Hsieh K.S. and Chang K.E. (2016a). CRIE: An automated analyzer for Chinese texts. *Behavior Research Methods* **48**(4):1238–1251.

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

Sung Y.T., Liao C.N., Chang T.H., Chen C.L. and Chang K.E. (2016b). The effect of online summary assessment and feedback system on the summary writing on 6th graders: The LSA-based technique. *Computers & Education* **95**, 1–18.
Tanaka-Ishii K., Tezuka S. and Terada H. (2010). Sorting texts by readability. *Computational Linguistics* **36**(2), 203–227.
Taylor M.C. and Wahlstrom M.W. (1999). Readability as applied to an ABE assessment instrument. Available at http://files.eric.ed.gov/fulltext/ED349461.pdf#page=30 (accessed January 2017).
Thorndike E.L. and Lorge I. (1944). *The Teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University. Bureau of Publications.
Truran M., Georg G., Cavazza M. and Zhou D. (2010). Assessing the readability of clinical documents in a document engineering environment. In *Proceedings of the 10th ACM Symposium on Document Engineering*, ACM, New York, NY.
Tseng H.C., Sung Y.T., Chen B. and Lee W.E. (2016). Classification of text readability based on representation learning techniques. In *Proceedings of the 26th Annual Meeting of the Society for Text & Discourse*, Kassel, Germany, pp. 1–6.
Vajjala S. and Meurers D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications using NLP*, Montreal, Canada, pp. 163–173.
Vajjala S. and Meurers D. (2014). Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, pp. 288–297.
Yan X., Song D. and Li X. (2006). Concept-based document readability in domain-specific information retrieval. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, ACM, New York, NY, pp. 540–549.
Zeno S.M., Ivens S.H., Millard R.T. and Duvvuri R. (1995). *The Educator's Word Frequency Guide*. New York: Touchstone Applied Science Associates, Inc. My Book.
Zhao J. and Kan M.Y. (2010). Domain-specific iterative readability computation. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, ACM, New York, NY, pp. 205–214.

Appendix

Table A1. Chinese readability features used in this study[a]

Feature	*Definition (all calculated per text)*
Lexical level	
Characters[a]	Total number of characters
Words	Total number of words
Adverbs	Total number of adverbs
Verbs	Total number of verbs
Low-stroke-count characters	Total number of characters with 1–10 strokes
Intermediate-stroke-count characters	Total number of characters with 11–20 strokes
High-stroke-count characters	Total number of characters with more than 20 strokes
Two-character words	Total number of two-character words
Three-character words	Total number of three-character words
Difficult words	Total number of words listed in Academia Sinica database of 3000 difficult words
Type–token ratio	Degree of diverse words
Semantic level	
Content words	Total number of content words
Average logarithmic frequency of content words	Logarithm of the average frequency of content words, according to Education Ministry word frequency list

https://doi.org/10.1017/S1351324919000093 Published online by Cambridge University Press

Table A1. (Continued)

Feature	*Definition (all calculated per text)*
Sentences with complex semantic categories	Total number of sentences with complex semantic categories
Noun word density	Ratio of content words to total words
Negatives	Total number of negation words
Syntactic level	
Average sentence length	Total number of words divided by the total number of sentences.
Modifiers per NP	Number of adjectives and adverbs before the head noun in noun phrases
NP ratio	Proportion of noun phrases to total sentence number
Cohesion level	
Pronouns	Total number of pronouns
Conjunctions	Total number of conjunctions
Positive conjunctions	Total number of positive conjunctions
Negative conjunctions	Total number of negative conjunctions
Personal pronouns	Total number of personal pronouns

[a] All "characters" in these readability features refer only to Mandarin Chinese characters.

Cite this article: Tseng H-C, Chen B, Chang T-H and Sung Y-T. Integrating LSA-based hierarchical conceptual space and machine learning methods for leveling the readability of domain-specific texts. *Natural Language Engineering* **25**, 331–361. https://doi.org/10.1017/S1351324919000093