ReCALL (2022), 34: 1, 51–65
doi:10.1017/S0958344021000197

RESEARCH ARTICLE

An editable learner model for text recommendation for language learning

John S. Y. Lee
Department of Linguistics and Translation, City University of Hong Kong, Hong Kong SAR, China (jsylee@cityu.edu.hk)

Abstract

Extracurricular reading is important for learning foreign languages. Text recommendation systems typically classify users and documents into levels, and then match users with documents at the same level. Although this approach can be effective, it has two significant shortcomings. First, the levels assume a standard order of language acquisition and cannot be personalized to the users' learning patterns. Second, recommendation decisions are not transparent because the leveling algorithms can be difficult for users to interpret. We propose a novel method for text recommendation that addresses these two issues. To enhance personalization, an open, editable learner model estimates user knowledge of each word in the foreign language. The documents are ranked by new-word density (NWD) – that is, the percentage of words that are new to the user in the document. The system then recommends documents according to a user-specified target NWD. This design offers complete transparency as users can scrutinize recommendations by reviewing the NWD estimation of the learner model. This article describes an implementation of this method in a mobile app for learners of Chinese as a foreign language. Evaluation results show that users were able to manipulate the learner model and NWD parameters to adjust the difficulty of the recommended documents. In a survey, users reported satisfaction with both the concept and implementation of this text recommendation method.

Keywords: text recommendation; text difficulty; open learner model; personalization; new-word density

1. Introduction

Language learners benefit from extensive reading, beyond their textbooks, ideally with text that is authentic in the linguistic, cultural, and functional sense (Buendgens-Kosten, 2013). Vocabulary difficulty is an important criterion in identifying suitable reading materials. A text with an appropriate proportion of new vocabulary for an intermediate learner might bring little benefit to an advanced learner who is already familiar with most words. Yet the same text might overwhelm a beginner because of the excessive percentage of unfamiliar words.

Learners would thus benefit from automatic recommendation of reading materials whose vocabulary complexity is suitable for them. Most conventional tools adopt the "leveling" or "graded" approach, which employs a common scale to match users and documents at the same or a similar level, such as school grades (Collins-Thompson, Bennett, White, de la Chica & Sontag 2011). On the documents, the system performs automatic readability assessment to estimate their difficulty level. On the users, the system performs automatic proficiency assessment to determine their proficiency level. Although this approach can be effective, it lacks personalization and transparency.

Cite this article: Lee, J. S. Y. (2022). An editable learner model for text recommendation for language learning. *ReCALL* 34(1): 51–65. https://doi.org/10.1017/S0958344021000197

© The Author(s), 2021. Published by Cambridge University Press on behalf of European Association for Computer Assisted Language Learning.

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press

A personalized text recommendation system should cater for individual differences in language learning patterns and reading preferences. Using levels to quantify language proficiency can be problematic for several reasons. First, the scale may not be familiar to users. Second, discrete levels do not allow fine-grained adjustment of the preferred difficulty in reading materials. Furthermore, a common scale assumes a standard order of language acquisition. With no control over the scale's definition, users of different backgrounds cannot adapt the levels to suit their individual learning patterns.

A transparent text recommendation system should facilitate scrutiny of its inference process. To justify the recommendation of a document, it should provide supporting information other than its estimated difficulty level. Unfortunately, most machine learning algorithms for automatic readability and proficiency assessments are not designed for human interpretation, resulting in a "black-box" leveling approach that hampers transparency and user trust.

To enhance personalization and transparency, this article proposes an editable, open learner model (OLM) (Bull & Kay, 2010) and a text recommendation metric based on new-word density (NWD) – that is, the percentage of words in the document that are new for the user. In terms of personalization, users can set their target NWD to calibrate the desired difficulty of reading materials and update the OLM to improve NWD estimation accuracy. In terms of transparency, users can scrutinize recommendation decisions by examining which words are predicted by the OLM to be new vocabulary.

2. Scope and contributions

A text recommendation system searches for the most suitable reading materials for language learners. The suitability of a document is typically measured with two factors: its difficulty, with respect to the user's language proficiency, and its thematic area, with respect to the user's reading interests (Heilman, Collins-Thompson, Callan & Eskenazi, 2010).

This article focuses on the first factor. The "difficulty" of a document depends on many aspects, ranging from the vocabulary and grammatical knowledge of the user, to pragmatics, connotation, and semantic complexity of the document. Our scope will be limited to vocabulary knowledge. According to some studies in second language acquisition, vocabulary knowledge and grammatical knowledge have comparable effects on reading comprehension (Shiotsu & Weir, 2007; van Gelderen, Schoonen, Stoel, de Glopper & Hulstijn, 2007). However, others suggested vocabulary knowledge to be the better predictor (Brisbois, 1995; Haynes & Carr, 1990; Mecartty, 2000) and reported a significant correlation between vocabulary difficulty and text difficulty (François & Fairon, 2012; Heilman, Collins-Thompson, Callan & Eskenazi, 2007; Reynolds, 2016). While a number of computer-assisted language learning systems have addressed grammatical knowledge (e.g. Bull, Pain & Brna, 1995; Shahrour & Bull, 2008; Vajjala & Meurers, 2012), automatic syntactic parsing, semantic complexity, and pragmatics analysis do not yet reliably achieve high accuracy. We will therefore follow most existing text recommendation systems in focusing on vocabulary (e.g. Brown & Eskenazi, 2004; Hsieh, Wang, Su & Lee, 2012; Hsu, Hwang & Chang, 2013; Liang & Song, 2009; Miltsakaki, 2009; Wu, 2016).

Within this scope, we aim to make two contributions. First, we propose a text recommendation framework based on NWD and an editable OLM (section 3), and show that it can offer greater personalization and transparency than the conventional approach (section 4). Second, we demonstrate the feasibility of user interaction with the model and positive user attitude toward the framework. A user study evaluates the ability of new users to issue search queries and to manipulate search parameters to adjust text difficulty (section 5), and a user survey measures satisfaction on the design and implementation of the model (section 6).

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press

Table 1. Comparison between the graded approach and proposed approach

Design goal	Component	Graded approach	Proposed approach
Transparency	Recommendation metric	Grade	New-word density (NWD)
	User proficiency estimation	Estimate user's grade	Estimate which words are new to user
	Text difficulty estimation	Estimate document's grade	Estimate document's NWD
Personalization	Learner model update	User to select own grade	User to indicate whether a word is known
	Document ranking	Recommend documents in selected grade	Recommend documents whose NWD is closest to user target

3. Design considerations

According to the "i+1" concept (Krashen, 1981), the most suitable materials for language learning lie within the proximal zone ("+1") of the learner's proficiency level ("i"). To determine whether a document lies within this zone, a text recommendation system must identify what content in the document can be understood by the user ("i") and what is new to the user ("+1").

A common approach is to label the user and all candidate documents on a scale that measures both user proficiency and text difficulty. The system then recommends documents whose level or score on the scale is similar to that of the user. A variety of scales, both discrete and continuous, have been used. Discrete scales, also referred to as "levels" or "stages," may involve school grades (e.g. Collins-Thompson & Callan, 2005); public examination grades (e.g. Hsu *et al.*, 2013), such as those in the General English Proficiency Test (Roever & Pan, 2008); stages in graded readers (Hill, 2008); and categories in language assessment schemes, such as the six categories in the Common European Framework of Reference for Languages. Continuous scales, such as Lexile measures (Lennon & Burdick, 2014) and readability formulas (e.g. Anderson, 1983; Coleman & Liau, 1975), are more fine-grained but still not transparent.

Conventional systems make recommendations according to the estimated grade or score, but do not typically reveal the basis of the estimation, for instance, by pointing to specific words or sentences in the document or relevant characteristics in the learner profile. In the rest of the article, we will use the term "graded approach" to refer to this approach. We argue that it suffers from two main shortcomings – the lack of personalization and of transparency – and propose a novel approach that mitigates them. Table 1 summarizes the differences between the graded approach and our proposal.

3.1 Personalization

Personalization allows users to modify the system's inference. One way to increase user control is through OLM, which can be broadly defined as "learner models that can be viewed or accessed in some way by the learner, or by other users" (Bull & Kay, 2010: 301). These models, which promote metacognitive activities, can facilitate learner reflection, planning, and self-monitoring. They also encourage learner independence by helping learners take control and responsibility over their learning (Bull & Kay, 2007). An *editable learner model* is an OLM that allows the learner to change its content, thus increasing the accuracy of the learner model data and the quality of adaptation (Bull & Kay, 2010). For text recommendation, such a model should enable users to adjust its estimation of their proficiency and of document difficulty.

In principle, the graded approach can offer some degree of personalization by allowing users to identify their grade and to modify the documents' grades. In practice, ambiguities in grade definitions often make proficiency self-assessment difficult for users. It is also unreasonable to ask users

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press

to review the grades for a large number of documents, which would in effect defeat the purpose of automatic text recommendation.

Further, the personalization is limited because grade definitions are fixed. All scales assume a standard order of vocabulary acquisition. Many are underpinned by vocabulary lists; for example, the list of 3,000 "familiar" words in the Dale–Chall readability formula or those from the widely recognized scheme *Hanyu Shuiping Kaoshi* (HSK), adopted by many tools for learning Chinese as a foreign language (e.g. Jin, Lu, Lin & Li, 2018; Liang & Song, 2009). However, learners may deviate from the standard because of their linguistic and professional backgrounds. For instance, a second language (L2) word is likely acquired at an earlier stage by a learner whose first language (L1) has a cognate word. A document full of legal jargon should not be recommended for general users, but may serve as excellent reading material for those with legal training. Yet users cannot change the grade definitions to reflect their learning patterns.

3.1.1 Learner model

We propose an editable learner model that directly estimates a user's knowledge of each L2 word; more precisely, the model predicts each L2 word to be either "known" or "unknown" to the user. This design increases learner control in two ways. First, the model has the flexibility to capture individual patterns in vocabulary acquisition without making any assumption on acquisition order. Second, users can edit the model with confidence, as it is easier for them to judge whether a word is new vocabulary than to grade themselves on an unfamiliar proficiency scale.

3.1.2 Recommendation metric

We quantify text difficulty by NWD (Holley, 1973). The NWD of a text is defined as the number of all unknown words in the text divided by the total number of words. This metric is user-centered in the sense that it quantifies text difficulty from the user's point of view: the same text may be labeled as 10% NWD for a high-proficiency user, but 30% NWD for a lower-proficiency user. In contrast, the graded approach is not user-centered as it labels a text with a fixed grade on a scale. The onus is placed on users to familiarize themselves with the scale and to interpret the grade in relation to their linguistic competence. The NWD metric also allows fine-grained calibration of the difficulty of reading materials. One can, for example, request 5% NWD or below in order to understand the text without glosses (Hu & Nation, 2000; Schmitt, Jiang & Grabe, 2011), or above 10% NWD to simulate the difficulty in a typical textbook (Liang & Song, 2009).

3.2 Transparency

A transparent system explains its decisions to users. Transparency promotes user trust (Bull & Kay, 2007) and leads to significant benefits in language learning and other subjects (e.g. Mitrovic & Martin, 2007; Shahrour & Bull, 2008). In text recommendation, the system needs to communicate why it deems a document suitable for the user.

The graded approach justifies its recommendation decision with a grade, without identifying the aspects that make the document difficult for the user. Further, the meaning of a grade tends to be opaque. Automatic proficiency and readability assessments can potentially remove the need for users to interpret the grade (e.g. Hsieh *et al.*, 2012). Most assessment models use machine learning algorithms, which typically score a document by combining weighted features based on statistical language models and readability formula scores, parse tree structures, type-token ratio, sentence and word length, etc. (e.g. Collins-Thompson & Callan, 2005; Francois & Fairon, 2012; Pitler & Nenkova, 2008; Sung, Lin, Dyson, Chang & Chen, 2015; Wu, 2016). These models essentially operate in a "black-box" fashion, as the internal working of their algorithms cannot be easily summarized for human consumption.

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press

Table 2. Parameters in the open learner model

Learner characteristic	Parameter	Input method
Vocabulary proficiency	Vocabulary set	Label words as "known" or "unknown"
Learning pace	Target text difficulty	Specify maximum new-word density
Reading interests	Document theme	Select document categories
	Search keywords	Enter search keywords

NWD can clearly specify which words in the document are predicted to be known or unknown to the user, according to the learner model. The explicitness enables scrutiny of recommendation decisions: users can easily verify the estimated NWD of a document by viewing the learner model and, if necessary, improve the estimation by correcting the model. Furthermore, NWD facilitates rapid development of multilingual text recommendation systems, since its computation is similar across languages and does not require handcrafted vocabulary lists.

4. System architecture

A personalized text recommendation system can be viewed as an instance of an adaptive system, which identifies learner characteristics and adjusts its content accordingly to improve learning efficiency (Brusilovsky, 2012; Oxman & Wong, 2014). An adaptive system is typically analyzed in terms of three models: the *domain model*, which defines what is to be learned; the *learner model*, which captures user characteristics and preferences; and the *adaptation model*, which adapts the content with respect to the learner model (Brusilovsky, 2012; Knutov, De Bra & Pechenizkiy, 2009). We present our proposed method within this framework and explain how it enhances personalization and transparency over existing approaches.

4.1 Domain model

The domain model defines the knowledge space as a set of knowledge elements – that is, the concepts to be learned. The system must label each pedagogical item with one or more of these elements, in a process known as "knowledge indexing."

For the graded approach, the knowledge element of each document can be viewed as its grade on the difficulty scale. Knowledge indexing thus requires grade estimation for each document through automatic readability assessment, which is prone to opacity (section 3.2). In contrast, our proposed method considers each L2 word to be a knowledge element. Thus, the knowledge elements of a document are simply the words it contains.

4.2 Learner model

Our learner model parameters are concerned with three user characteristics, namely the user's vocabulary proficiency; the preferred learning pace, defined by the target NWD; and reading interests (Table 2). The user can set all parameters as part of a personalized search query. Given our research focus (section 2), the rest of this section will concentrate on vocabulary proficiency and learning pace.

The NWD of a document is based on the user's "vocabulary set" – that is, the set of words known to the user. The learner model estimates this set by predicting *each* L2 word to be either "known" or "unknown" to the user. Although lexical knowledge can also be expressed on a five-point scale (Ehara, Sato, Oiwa & Nakagawa, 2012) or with a real-number score (Yimam *et al.*, 2018), the binary distinction enables a more transparent interpretation of NWD.

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press

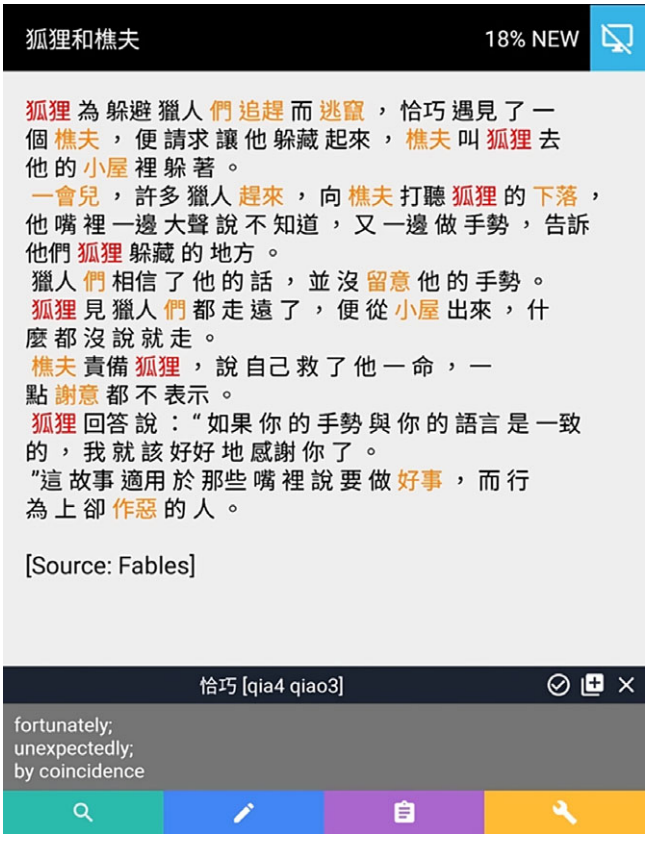


Figure 1. The reading environment

The learner model is thus an instance of the "overlay" model, which characterizes the user's knowledge level for each knowledge element (Brusilovsky, 2012). For an initial estimation of the learner model, new users rate their knowledge of 50 words as "known" or "unknown" (Figure 2). Using this self-assessment as training data, the system performs Complex Word Identification (Yimam *et al.*, 2018) to construct the user's vocabulary set. Our model, which automatically predicts each L2 word as either known or unknown to the user with an SVM classifier, performs at 78.0% accuracy (Lee & Yeung, 2018). New users may also opt out of self-assessment and manually choose a rough vocabulary set size. A size of N means the user knows the N most frequent words (up to $N = 40{,}000$), according to word frequency in Chinese Wikipedia.

As shown in Figure 1, the reading environment uses colors to visualize the learner model, displaying known words in black, new words in orange, and search keywords in red. If users see a known word wrongly predicted to be new vocabulary in their reading material, or vice versa, they can modify the word's status in the editable learner model by tapping on the word.

Learning pace refers to users' preferred difficulty for their reading materials (Table 2). In comparison to the graded approach, which requires users to dial up or down a grade, the target NWD parameter enables more fine-grained customization. Those who desire a faster pace can set a higher target percentage to retrieve more challenging materials, and those who prefer more leisurely or easy reading can set a lower target.

The maximum NWD is set by default at 20%, which means that documents in search results have at most 20% NWD. The relatively high default value is intended to encourage users to tackle

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press



Figure 2. Words to be labeled as new vocabulary (tap on the left icon) or known (right icon) during vocabulary self-assessment

challenging materials. Users can change the maximum NWD by adjusting the percentage with a slider (Figure 3).

4.3 Adaptation model

Given the learner model's diagnosis, the adaptation model adapts the pedagogical material to optimize learning outcomes (Knutov *et al.*, 2009). Content presentation, for example, can be adapted to highlight particular items. Our reading environment draws users' attention to the search keywords (highlighted in red) and the new vocabulary as predicted by the learner model (in orange). They can tap on these words to read the English gloss, which was taken from the CC-CEDICT project (https://cc-cedict.org), or update their status in the learner model (Figure 1).

Navigation prioritizes the most relevant learning items for the user according to the learner model. Text recommendation systems rank documents according to their suitability for language learning. The ranking metric can be a readability formula or a weighted score based on factors such as text length, reading grade level, and topic. Our app ranks documents in descending order in terms of NWD, based on word segmentation by the Stanford Chinese parser (Manning *et al.*, 2014). Search results include only documents whose NWD is below the user-specified maximum (Figure 4).

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press

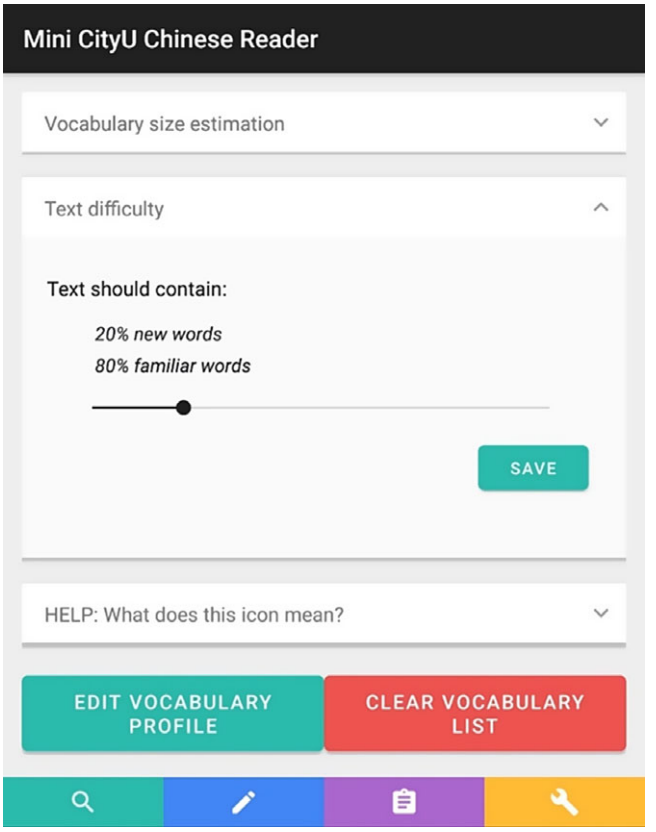


Figure 3. Specification of text difficulty in terms of the maximum new-word density (NWD)

5. User evaluation

The use of an editable learner model leads to more complicated user interaction, which may require a steeper learning curve. We conducted a user study with the twin goal of evaluating both the usability and performance of our proposed text recommendation method. For the first goal, we measure new users' ability to exploit personalized search by issuing complex queries involving multiple parameters. For the second goal, we measure whether these parameter variations lead to intended changes in the difficulty of the recommended documents.

5.1 Set-up

Text recommendation systems may be used by both language learners for personal reading and language teachers for class preparation. We therefore recruited both native and non-native speakers of Chinese for our evaluation, including 13 learners of Chinese as a foreign language (CFL) and five native speakers. The CFL learners represented a wide range of language backgrounds, with English, French, Indonesian, Korean, and Thai as L1, and also a variety of proficiency levels, with years of studies ranging from 1 to 15 years. The native speakers were undergraduate students at City University of Hong Kong.

After reading a manual about the app, the subjects were asked to independently perform a number of queries involving three parameters (Table 3): vocabulary size, NWD, and search keywords. Query #1, for example, involved a 40,000-word vocabulary set and 20% target NWD, whereas Query #4 required a search keyword.

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press



Figure 4. Search results rank documents by new-word density (NWD) up to the user-specified maximum

The app performed the document search in a text database containing 8,190 Chinese short essays downloaded from duanmeiwen.com. Two versions of the app, one in traditional characters and another in simplified characters, were made available to the subjects. After each query, the subjects reported the number of retrieved documents as shown on the search result page (Figure 4) and read the top five documents. Then, they rated the difficulty of each document on a 5-point Likert scale, from “very easy” (score 1) to “very difficult” (score 5).[1] For the fifth query, rather than using the app, the subjects performed the search on duanmeiwen.com.

5.2 Query execution accuracy

We first report on the subjects’ ability to correctly execute the queries. To determine whether the subjects issued the intended queries, we checked whether they reported the expected number of search results.

All 18 subjects correctly issued queries #2 and #3. For Query #1, one subject reported a wrong number of search results, likely due to an incorrect parameter value. For Query #4, 16 subjects reported the correct number, one subject reported an unexpected result, and one subject failed to report any. These results suggest that brief self-training is sufficient for most new users to use the search interface effectively.

[1]The native speakers were asked to rate the documents from the perspective of an intermediate CFL student.

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press

Table 3. Average difficulty ratings of the top five search results from queries in the user evaluation

Query	Vocabulary size	Target NWD	Keyword	Difficulty rating
#1	40,000 words	20%	None	3.55
#2	4,000 words	20%	None	2.44
#3	4,000 words	15%	None	2.27
#4	4,000 words	20%	*baba* "father"	2.28
#5 (web)	n/a	n/a	*baba* "father"	2.78

Note. NWD = new-word density.

5.3 Difficulty of recommended documents

We now analyze the subjects' perception of text difficulty with respect to different parameter values for vocabulary size and target NWD, and in comparison with web search. Table 3 presents the average rating given by the subjects to the top five documents recommended in each query.

5.3.1 Effect of vocabulary size

The five documents recommended in Query #1 received an average difficulty rating of 3.55. The 40,000-word vocabulary set included most common words. The 20% target NWD therefore led to a substantial number of rare words, which likely contributed to the relatively high difficulty rating. Query #2 maintained the target NWD at 20% but reduced the vocabulary size to 4,000 words, which had the effect of raising the NWD of all documents. As a result, those with more rare words were more likely to exceed 20% NWD and be rejected. The average difficulty rating of the retrieved documents decreased[2] to 2.44, confirming the intended effect of the vocabulary size parameter on text difficulty.

5.3.2 Effect of target NWD

Query #3 kept the vocabulary size the same as Query #2, at 4,000 words, but lowered the target NWD from 20% to 15%. This change led to a decrease in the average difficulty rating[3] from 2.44 to 2.27. The subjects perceived the recommended documents to be easier even with the relatively small drop (5%) in NWD, suggesting the effectiveness of NWD in fine-grained calibration of text difficulty.

5.3.3 Comparison with web search

The last two queries contrasted our text recommendation method (Query #4) with a search on duanmeiwen.com (Query #5), the source of the app's text database. Thus, both queries accessed similar pools of candidate documents and utilized the same search keyword. The only major difference was that text difficulty was not considered in Query #5, while Query #4 capped NWD at 20% for a vocabulary size of 4,000. The subjects gave a lower average score[4] to the materials returned by the app (2.28) than to those returned by web query (2.78). The lower rating suggests the ability of our text recommendation method to identify easier reading material, compared to general web search.

[2]The difference is statistically significant at $p < 0.00007$ in a two-tailed t-test.
[3]The difference approaches statistical significance at $p = 0.108$ in a two-tailed t-test.
[4]The difference is statistically significant at $p < 0.007$ in a two-tailed t-test.

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press

Table 4. User survey results of general design issues (left column) and app implementation (right column)

	General design issues		App implementation	
Topic	In general, it's a good idea for the text recommendation system to:	Score (*SD*)	In this app, it was easy for me to:	Score (*SD*)
Search methodology	(1a) rank search results by new-word density (NWD)	4.28 (0.67)	(1b) view the NWD in search results	4.56 (0.62)
	(2a) let users filter search results based on maximum NWD	4.39 (0.61)	(2b) set a maximum NWD in search results	4.67 (0.61)
Editable learner model	(3a) let users indicate whether they know a word or not	4.67 (0.69)	(3b) inform the app whether I know or do not know a word	4.71 (0.59)
Initial estimation of learner model	(4a) predict whether users know a word or not	3.67 (0.84)	(4b) complete the vocabulary assessment	4.61 (0.78)
	(5a) let users indicate their vocabulary size	4.28 (0.83)	(5b) specify my vocabulary size	4.28 (0.75)

6. User survey

To investigate users' opinion on the proposed text recommendation method, we conducted a survey that solicited both their attitude toward design issues in general and their experience with our app implementation in particular.

We administered this survey to the same 18 subjects as in the user evaluation (section 5.1). As shown in Table 4, the survey consisted of two parts, with the first concerned with general design issues and the second with the app implementation. Each part contained five statements on search methodology, the editable learner model, and the initial estimation of the model. The subjects rated each statement on a 5-point Likert scale, corresponding to "strongly disagree" (score 1), "disagree" (2), "neutral" (3), "agree" (4), and "strongly agree" (5).

6.1 General design issues

6.1.1 Search methodology

In the first part of the survey (left column in Table 4), the first two statements were concerned with recommendation methodology: the use of NWD as the ranking metric (1a) and the use of maximum NWD as the search criterion (2a). Consistent with the correlation between NWD and text difficulty observed in section 5.3, 16 out of 18 subjects "agreed" or "strongly agreed" with the use of NWD for ranking documents (score 4.28). Further, 17 out of 18 subjects "agreed" or "strongly agreed" with the use of maximum NWD as the search constraint (score 4.39). Hence, the subjects held favorable views on the two fundamental features of the proposed text recommendation method.

6.1.2 Editable learner model

Our design assumes users' willingness to interact with the editable learner model. Testing this assumption, statement (3a) revealed that a majority of the subjects "strongly agreed" that users should be able to manually update the model on their vocabulary knowledge (score 4.67).

6.1.3 Initial estimation of learner model

The last two statements consulted the subjects on the automatic method (4a) and manual method (5a) for initializing the learner model for a new user. With eight out of 18 subjects expressing reservation – either "neutral" or "disagree" – about automatic prediction of vocabulary knowledge,

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press

statement (4a) received the lowest score among the statements in the survey (score 3.67). This ambivalence may be attributable to system mistakes in automatic complex word identification (section 4.2), which likely needs improved accuracy to gain more acceptance among subjects. The subjects gave slightly higher scores (score 4.28) to statement (5a), indicating an overall preference to manual estimation of their vocabulary size.

6.2 App implementation

6.2.1 Search interface

In the second part of the survey (right column in Table 4), the first two statements polled the subjects on their experience with the search interface. In statement (1b), 17 out of 18 subjects "agreed" or "strongly agreed" that it was easy to work with the NWD-based search result (Figure 4) (score 4.56). Similarly, in statement (2b), 17 out of 18 subjects "agreed" or "strongly agreed" that it was easy to set the parameter for maximum NWD (score 4.67) (Figure 3). These relatively high scores, consistent with the high success rate in query execution (section 5.2), suggest the usability of the search interface.

6.2.2 Learner model update

The next statement (3b) gauged user opinion on interaction with the editable learner model – that is, tapping on a word to change its status from "unknown" to "known," or vice versa (Figure 1). This statement received the highest score (4.71) of all statements in the survey. As user updates are critical for the accurate NWD estimation, this result is promising for applying an editable learner model for text recommendation.

6.2.3 Initial estimation of learner model

The last two statements evaluated user experience with the automatic (4b) and manual (5b) methods for model initialization. Statement (4b) showed that most subjects were willing (score 4.61) to complete the 50-word vocabulary self-assessment (Figure 2). In comparison, in statement (5b), the subjects found it relatively difficult (score 4.28) to manually estimate their vocabulary size. All in all, while the subjects preferred the manual method in principle (cf. scores for statements (4a) and (5a)), they found the automatic method easier in practice.

7. Limitations

Results from the user evaluation (section 5) and survey (section 6) should be interpreted with a number of limitations in mind. First, the scope of the learner model is limited to vocabulary knowledge (section 2). An expanded model incorporating syntactic and semantic complexity would offer a more comprehensive evaluation of the effects of personalization and transparency.

Second, the learner model does not cover multiword expressions. Since language often comes in chunks or semi-preconstructed phrases (Sinclair, 1991), a model that accounts for multiword expressions can yield more accurate NWD estimation. It would, however, complicate user interaction due to the larger number of parameters. The learner model also does not distinguish between different senses of a word. Modeling word senses can improve NWD estimation accuracy, but would require development of a sense inventory for each word as well as automatic word sense disambiguation.

Third, the evaluation was relatively small scale. A larger number of subjects, reflecting greater variety in L1 background and L2 proficiency, would measure user experience more comprehensively among diverse user populations. Further, the evaluation, which focused on text recommendation quality and user perception, did not consider language learning outcomes. Although pedagogical benefits have been demonstrated in intelligent tutoring systems for various subjects,

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press

they have yet to be established for text recommendation specifically. A controlled experiment contrasting the learning outcomes of the proposed and graded approaches is needed to draw a conclusion.

8. Conclusions and future work

Text recommendation systems support self-regulated language learning by matching learners with appropriate reading materials. The conventional approach performs the matching through automatic readability assessment, which assigns each document to a difficulty level, and through automatic proficiency assessment, which measures the user on the same scale. This approach cannot be customized to the user's personal learning patterns, because the levels assume a standard order of language acquisition. It also lacks transparency as it reveals little of the basis of its recommendation decisions.

This article has presented a novel text recommendation method that addresses these limitations. This method increases personalization with an editable learner model, which enables users to update the status of their vocabulary knowledge in the model when it makes wrong predictions. Further, this method provides more transparency through document ranking by NWD, a metric that facilitates user scrutiny and fine-grained calibration of text difficulty. An evaluation showed that new users were able to issue queries with various personalization features to achieve the intended effects on text difficulty. According to a survey, users were receptive to the search methodology and interface and were willing to inspect and update the learner model.

We plan to pursue future work in two directions. First, we plan to expand the editable learner model to cover other dimensions of text difficulty (section 7). Since an expanded model may result in more complex user interaction, it would be worth considering less invasive methods for model updates (e.g. Hokamp, Mihalcea & Schuelke, 2014). Second, we intend to conduct a longitudinal study, involving subjects with a greater variety of L1 backgrounds and L2 proficiency, on the learning outcomes of the proposed text recommendation approach. Such a study would help assess its potential to serve in a personal user model for lifelong learning (Kay & Kummerfeld, 2019).

Acknowledgements. This work was supported by grants from the Standing Committee on Language Education and Research (EDB(LE)/P&R/EL/164/6) and from the Innovation and Technology Fund (ITS/389/17) in Hong Kong SAR, China. We thank Lam Xing, Zhenqun Yang, and Chak Yan Yeung for their assistance in app implementation.

Ethical statement. The experimental procedure in this research was approved by the Human Subjects Ethics Sub-Committee at City University of Hong Kong (Application No.: H000899).

References

Anderson, J. (1983) Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6): 490–496. http://www.jstor.org/stable/40031755

Brisbois, J. E. (1995) Connections between first- and second-language reading. *Journal of Reading Behavior*, 27(4): 565–584. https://doi.org/10.1080/10862969509547899

Brown, J. & Eskenazi, M. (2004) Retrieval of authentic documents for reader-specific lexical practice. In *Proceedings of the InSTIL/ICALL 2004 Symposium on Computer Assisted Learning*. Venice, Italy, 17–19 June.

Brusilovsky, P. (2012) Adaptive hypermedia for education and training. In Durlach, P. J. & Lesgold, A. M. (eds.), *Adaptive technologies for training and education*. New York: Cambridge University Press, 46–66. https://doi.org/10.1017/CBO9781139049580.006

Buendgens-Kosten, J. (2013) Authenticity in CALL: Three domains of 'realness'. *ReCALL*, 25(2): 272–285. https://doi.org/10.1017/S0958344013000037

Bull, S. & Kay, J. (2007) Student models that invite the learner in: The SMILI Open Learner Modelling Framework. *International Journal of Artificial Intelligence in Education*, 17(2): 89–120.

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press

Bull, S. & Kay, J. (2010) Open learner models. In Nkambou, R., Bourdeau, J. & Mizoguchi, R. (eds.), *Advances in intelligent tutoring systems: Studies in computational intelligence, Vol. 308*. Berlin: Springer, 301–322. https://doi.org/10.1007/978-3-642-14363-2_15

Bull, S., Pain, H. & Brna, P. (1995) Mr. Collins: A collaboratively constructed, inspectable student model for intelligent computer assisted language learning. *Instructional Science*, 23: 65–87. https://doi.org/10.1007/BF00890446

Coleman, M. & Liau, T. L. (1975) A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2): 283–284. https://doi.org/10.1037/h0076540

Collins-Thompson, K., Bennett, P. N., White, R. W., de la Chica, S. & Sontag, D. (2011) Personalizing web search results by reading level. In Berendt, B., de Vries, A., Fan, W., Macdonald, C., Ounis, I. & Ruthven, I. (eds.), *CIKM '11: Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. New York: Association for Computing Machinery, 403–412. https://doi.org/10.1145/2063576.2063639

Collins-Thompson, K. & Callan, J. (2005) Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13): 1448–1462. https://doi.org/10.1002/asi.20243

Ehara, Y., Sato, I., Oiwa, H. & Nakagawa, H. (2012) Mining words in the minds of second language learners: Learner-specific word difficulty. In Kay, M. & Boitet, C. (eds.), *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*. Mumbai, India, 8–15 December.

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

Haynes, M. & Carr, T. H. (1990) Writing system background and second language reading: A component skills analysis of English reading by native speaker-readers of Chinese. In Carr, T. H. & Levy, B. A. (eds.), *Reading and its development: Component skills approaches*. San Diego: Academic Press, 375–421.

Heilman, M. J., Collins-Thompson, K., Callan, J. & Eskenazi, M. (2007) Combining lexical and grammatical features to improve readability measures for first and second language texts. In Sidner, C., Schultz, T., Stone, M. & Zhai, C. (eds.), *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics: Proceedings of the main conference*. Stroudsburg: The Association for Computational Linguistics, 460–467.

Heilman, M., Collins-Thompson, K., Callan, J. & Eskenazi, M. (2010) Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education*, 20(1): 73–98.

Hill, D. R. (2008) Graded readers in English. *ELT Journal*, 62(2): 184–204. https://doi.org/10.1093/elt/ccn006

Hokamp, C., Mihalcea, R. & Schuelke, P. (2014) Modeling language proficiency using implicit feedback. In Calzolari, N., Choukri, K., Declerck, R., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. & Piperidis, S. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland, 26–31 May.

Holley, F. M. (1973) A study of vocabulary learning in context: The effect of new-word density in German reading materials. *Foreign Language Annals*, 6(3): 339–347. https://doi.org/10.1111/j.1944-9720.1973.tb02613.x

Hsieh, T.-C., Wang, T.-I., Su, C.-Y. & Lee, M.-C. (2012) A fuzzy logic-based personalized learning system for supporting adaptive English learning. *Educational Technology and Society*, 15(1): 273–288.

Hsu, C.-K., Hwang, G.-J. & Chang, C.-K. (2013) A personalized recommendation-based mobile learning approach to improving the reading performance of EFL students. *Computers & Education*, 63: 327–336. https://doi.org/10.1016/j.compedu.2012.12.004

Hu, M. H.-C. & Nation, P. (2000) Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1): 403–430.

Jin, T., Lu, X., Lin, Y. & Li, B. (2018) *Chi-Editor: An online Chinese text evaluation and adaptation system*. Guangzhou: LanguageData (languagedata.net/editor).

Kay, J. & Kummerfeld, B. (2019) From data to personal user models for life-long, life-wide learners. *British Journal of Educational Technology*, 50(6), 2871–2884. https://doi.org/10.1111/bjet.12878

Knutov, E., De Bra, P. & Pechenizkiy, M. (2009) AH 12 years later: A comprehensive survey of adaptive hypermedia methods and techniques. *New Review of Hypermedia and Multimedia*, 15(1): 5–38. https://doi.org/10.1080/13614560902801608

Krashen, S. D. (1981) The "fundamental pedagogical principle" in second language teaching. *Studia Linguistica*, 35(1–2): 50–70. https://doi.org/10.1111/j.1467-9582.1981.tb00701.x

Lee, J. & Yeung, C. Y. (2018) Automatic prediction of vocabulary knowledge for learners of Chinese as a foreign language. In *Proceedings of the 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*. Algiers, Algeria, 25–26 April. https://doi.org/10.1109/ICNLSP.2018.8374392

Lennon, C. & Burdick, H. (2014) *The Lexile framework as an approach for reading measurement and success*. Durham: MetaMetrics.

Liang, S. & Song, J. (2009) Construction of an approach for counting Chinese graded words and characters — A tool for assessing difficulty level of word in Chinese language teaching materials writing system. *Modern Educational Technology*, 7: 86–89.

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. & McClosky, D. (2014) The Stanford CoreNLP natural language processing toolkit. In Bontcheva, K. & Zhu, J. (eds.), *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Stroudsburg: Association for Computational Linguistics, 55–60. https://doi.org/10.3115/v1/P14-5010

Mecartty, F. H. (2000) Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning*, 11(2): 323–348.

Miltsakaki, E. (2009) Matching readers' preferences and reading skills with appropriate web texts. In Kreutel, J. (ed.), *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*. Stroudsburg: Association for Computational Linguistics, 49–52. https://doi.org/10.3115/1609049.1609062

Mitrovic, A. & Martin, B. (2007) Evaluating the effect of open student models on self-assessment. *International Journal of Artificial Intelligence in Education*, 17(2): 121–144.

Oxman, S. & Wong, W. (2014) *White paper: Adaptive learning systems*. Downers Grove: DV X/DeVry Education Group and Integrated Education Solution.

Pitler, E. & Nenkova, A. (2008) Revisiting readability: A unified framework for predicting text quality. In *EMNLP '08: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics, 186–195. https://doi.org/10.3115/1613715.1613742

Reynolds, R. (2016) Insights from Russian second language readability classification: Complexity-dependent training requirements, and feature evaluation of multiple categories. In Tetreault, J., Burstein, J., Leacock, C. & Yannakoudakis, H. (eds.), *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. Stroudsburg: Association for Computational Linguistics, 289–300. https://doi.org/10.18653/v1/W16-0534

Roever, C. & Pan, Y.-C. (2008) Test review: GEPT: General English Proficiency Test. *Language Testing*, 25(3): 403–408. https://doi.org/10.1177/0265532208090159

Schmitt, N., Jiang, X. & Grabe, W. (2011) The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1): 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Shahrour, G. & Bull, S. (2008) Does "notice" prompt noticing? Raising awareness in language learning with an open learner model. In Nejdl, W., Kay, J., Pu, P. & Herder, E. (eds.), *Adaptive hypermedia and adaptive web-based systems: 5th international conference, AH 2008, Hannover, Germany, July 29 – August 1, 2008: Proceedings*. Berlin: Springer, 173–182. https://doi.org/10.1007/978-3-540-70987-9_20

Shiotsu, T. & Weir, C. J. (2007) The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, 24(1): 99–128. https://doi.org/10.1177/0265532207071513

Sinclair, J. (1991) *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sung, Y.-T., Lin, W.-C., Dyson, S. B., Chang, K.-E. & Chen, Y.-C. (2015) Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, 99(2): 371–391. https://doi.org/10.1111/modl.12213

Vajjala, S. & Meurers, D. (2012) On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*. Stroudsburg: Association for Computational Linguistics, 163–173.

van Gelderen, A., Schoonen, R., Stoel, R. D., de Glopper, K. & Hulstijn, J. (2007) Development of adolescent reading comprehension in language 1 and language 2: A longitudinal analysis of constituent components. *Journal of Educational Psychology*, 99(3): 477–491. https://doi.org/10.1037/0022-0663.99.3.477

Wu, T.-T. (2016) A learning log analysis of an English-reading e-book system combined with a guidance mechanism. *Interactive Learning Environments*, 24(8): 1938–1956. https://doi.org/10.1080/10494820.2015.1070272

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G. H., Specia, L., Štajner, S., Tack, A. & Zampieri, M. (2018) A report on the Complex Word Identification Shared Task 2018. In Tetreault, J., Burstein, J., Kochmar, E., Leacock, C. & Yannakoudakis, H. (eds.), *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Stroudsburg: Association for Computational Linguistics, 66–78.

About the author

John S. Y. Lee is an associate professor at the Department of Linguistics and Translation at City University of Hong Kong. He obtained his PhD in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology.

Author ORCiD. John S. Y. Lee, https://orcid.org/0000-0003-2505-2678

https://doi.org/10.1017/S0958344021000197 Published online by Cambridge University Press