Project Title: A mobile, personal reading tutor for learners of Chinese

Grantee: City University of Hong Kong

Principal Investigator: LEE Sie-yuen, John
Department of Linguistics and Translation
City University of Hong Kong

Co-investigators: KIT Chunyu
Department of Linguistics and Translation
City University of Hong Kong

Jonathan WEBSTER
Department of Linguistics and Translation
City University of Hong Kong

Final Report

by

Principal Investigator

A Mobile, Personal Reading Tutor for Learners of Chinese

Principal Investigator: Dr. John S. Y. LEE
Department of Linguistics and Translation
City University of Hong Kong
jsylee@cityu.edu.hk

Abstract

Extensive, extra-curricular reading is important for learning foreign languages. Learners therefore need to venture beyond their textbooks to seek additional reading materials. However, it is often difficult to identify suitable materials with an appropriate number of new words to stretch vocabulary knowledge, but not to hinder comprehension.

Most existing systems require users to choose a level on a proficiency scale. These scales can be opaque for users, and often too coarse-grained to cater to individual needs. We present a personal and adaptive text retrieval method for language learning. A user can search for documents with the desired percentage of words that are new to himself or herself. To compute this percentage, the learner model estimates the user's vocabulary knowledge, and dynamically updates itself through user interactions.

This report describes our implementation of this method in a personal reading tutor on a mobile app, and presents empirical evaluations in the context of learning Chinese as a foreign language We investigate the performance of the personalized learner model in predicting new vocabulary, and the extent to which the model helps users retrieve texts at their desired difficulty levels.

Keywords: Computer-assisted language learning (CALL); Mobile learning; Chinese as a foreign language; Readability assessment; Natural language processing

1 Introduction

"Free voluntary reading" serves as a major source of reading competence and vocabulary development (Krashen, 2005). Since recreational reading, or reading for pleasure, plays such an important role in second language acquisition, learners benefit from reading a wide range of texts, beyond their textbooks and graded readers.

The web provides a convenient and large source of extra-curricular reading. However, search results may not be suitable for everyone as pedagogical material, because search engines do not typically cater to individual needs or accommodate the variability among learners. A text that has the right proportion of new vocabulary for the intermediate student might bring little benefit to an advanced student, who is already familiar with most or all of the words. Yet, the same text might overwhelm a beginner student, due to an excessive percentage of unfamiliar words. Our goal is to develop a personal reading tutor that is sensitive to vocabulary difficulty, in order to support independent, self-directed reading and learning.

We describe a personalized and adaptive text retrieval method for language learners, that is centered on the vocabulary profile of individual learners. It tailors search results according to the user's language proficiency level, preferred learning pace and learning interests; further, it adapts to the user's evolving proficiency. We describe the implementation of this method in a personal reading tutor on a mobile app, and presents empirical evaluations in the context of learning Chinese as a foreign language We investigate the performance of the personalized learner model in predicting new vocabulary, and the extent to which the model helps users retrieve texts at their desired difficulty levels.

The rest of this report is organized as follows. The next section introduces the conceptual framework of the project. Section 3 reviews the research literature. Section 4 gives details on our methodology.

Section 5 describes our data. Section 6 discusses our results. Section 7 concludes and makes recommendations.

2 Conceptual Framework of the Project

The personal reading tutor is intended to serve as a source of extra-curricular reading in order to support independent, self-directed learning. Rather than presenting a user with a fixed reading curriculum, the user is to build his or her individualized curriculum by searching for reading material from a pool of candidate documents.

We must first define what makes a text suitable for a language learner. According to the "i+1" concept (Krashen, 1981), the most suitable texts lie within the proximal zone ("+1") of the learner's proficiency level ("i"). In order to implement this concept, a system must determine the value of "i" with learner and text modeling, and the amount of "+1" according to the user's preferred learning pace. In the rest of this section, we lay out the main considerations in designing our system.

2.1 Learner modeling

To make user-specific recommendations, the system must be able to estimate the user's language proficiency. Proficiency encompasses multiple dimensions, including vocabulary, syntax and semantics. Given the significant correlation between proficiency and vocabulary level (Coniam, 1999), we follow existing CALL systems in focusing on vocabulary (Miltsakaki and Troutt, 2008; Brown and Eskenazi, 2004).

One possible approach is to simply ask the user to pick a vocabulary level on a scale, such as a grade level. It can be difficult, however, for the user to determine what his/her level should be. One cannot assume, for example, that a learner of English outside the U.S. to be familiar with the English curriculum at American schools. Despite emerging standards such as the *Common European Framework of Reference for Languages* (CEFR), there is no universal scheme for most languages. To mitigate this problem, the system can offer automatic assessment to help users choose his/her level. Assessment metric should be easy to interpret and correct. An example of such a transparent metric is to predict the words that the user knows or does not know — an approach that we pursued in our system.

2.2 Text modeling

The conventional approach is to automatically label each candidate document with a level on a difficulty scale, such as a grade level. Given an estimated proficiency of the user on the same scale, the system can then recommend reading materials at the matching level. Despite recent advances in language technology, readability assessment remains a challenging task; for example, the estimated difficulty of a text may be off by one or two grade levels in a mature system that has been deployed in the classroom (Heilman et al., 2010). Further, similar to the learner model, there is no universally understood scale. Labels such as "intermediate" or "Grade 6" is likely to be opaque for many users. The system should express the difficulty level of a document with a more intuitive, objective metric. We used the percentage of new vocabulary as the metric.

2.3 Adaptive modeling

In graded readers, the levels are fixed and typically limited in number. For example, there are only six levels in the widely used *Hanyu Shuiping Kaoshi* (HSK) scheme for learning Chinese as a foreign language (CFL). Since language skills increase gradually over time, these coarse-grained levels do not optimally cater to one's evolving needs. Indeed, it has been argued that readability measures for self-directed learning should be according to individual dimensions, not overall prediction for standard classroom teaching (Beinborn et al., 2012). Instead of placing the onus on the user to decide when to advance to the next level, the system should "grow" with the user; in other words, the learner model should continually update itself through user interaction, and automatically increasing text difficulty in small steps. Our simple learner model allows the user to participate in this process by informing the system of new words they learned, and adjust the learner model in a fine-grained manner.

2.4 Learning pace

The user's preferred pace of learning may vary depending on the usage scenario. Sometimes, the user might prefer fluent reading. In this case, the best reading material would be those that the user can understand without the disruption caused by looking up unknown words; this means at least 95% of the words should be familiar to the user (Hu and Nation, 2000; Schmitt et al., 2011). At another time, the same user might wish to maximize vocabulary acquisition, and prefer to tackle a text with many new words. If so, the system should retrieve texts with 10% to 31% new vocabulary, the proportion that is observed in a popular CFL textbook series (Liang and Song, 2009). These variations cannot be captured by a user proficiency model, no matter how accurate it is. Since learning pace is less predictable, a system should allow users to define their desired ratio of new vocabulary.

2.5 Learning interests

At both the document level and at the word level, the system should return reading materials that are interesting for the user. The former means that the system should include documents on wide-ranging themes, and tailor its recommendations to those on topics of personal interest to the user (Heilman et al., 2010; Hsu et al., 2013). The latter means the document should contain vocabulary items that the user wants to learn or to review. The ideal system should combine both, allowing users to choose topics and to explicitly specify target words of interest as part of their search queries.

3 Review of Literature of the Project

Recent advances in information technology have given rise to intelligent tutoring systems that can adapt to users. An adaptive system is one that can "adjust instruction based on learner abilities and/or preferences, at any particular point of the instruction process, with the goal of acting on identified learner characteristics and improving the efficiency and efficacy of learning" (Oxman and Wong, 2014). An adaptive system typically has three components: the domain model, the learner model, and the adaptation model (Brusilovsky, 2012; Knutov et al., 2009; Vandewaetere et al., 2011).

We now review the research literature on these three models, and then focus on a central task of the learner model, that of complex word identification (Section 3.4). Finally, we describe assessment scales in Chinese learning, which will serve as features in our CWI models (Section 3.5).

3.1 Domain model

The domain model represents the subject matter to be taught. It may encode properties of the documents themselves, such as their title and length. For systems that provide dictionary entries, these entries also form part of the domain model. To facilitate search, it may also include meta-information such as the text theme or category, as well as the difficulty level.

Most previous research has focused on automatic classification of documents into difficulty levels. Given an input document, a readability assessment model estimates its difficulty level, which can be a point on an ad hoc scale (Hsu et al., 2013); a holistic score on the overall difficulty level of the text (François and Fairon, 2012); or, most commonly, a grade level in a school system, such as that of the U.S. (Colleen Lennon and Hal Burdick, 2014; Vajjala and Meurers, 2012; Collins-Thompson and Callan, 2005; Pitler and Nenkova, 2008; Miltsakaki and Troutt, 2008).

Assessment can be performed with readability formulas such as the Flesch Reading Ease Readability and the New Dale-Chall Readability Formula. These formulas consider factors such as the number of words, the length of words, the number of sentences, sentence length, graded vocabulary lists, etc. More recent research has explored a large range of features, including statistical language model scores (Collins-Thompson and Callan, 2004), type-token ratio, as well as the frequency of content words, complex sentences, negation words, polysemy, pronouns and conjunctions (Sung et al., 2015; Chen et al., 2013). In a recent study on estimating the grade level of primary school textbook material for Chinese, the best model achieved 72.92% accuracy (Sung et al., 2015).

3.2 Learner model

The learner model captures user characteristics, chiefly their language proficiency and reading interests.

Reading interests. Reading interests encompass many aspects; most existing systems allow users to indicate their interests in two ways. At a more coarse-grained level, users can report reading preferences in a questionnaire (Hsu et al., 2013), for example by express how much they enjoy reading about a topic on a 5-point scale (Heilman et al., 2010). At a more fine-grained level, users can use search keywords to specify the vocabulary items they want to see in their reading materials (Miltsakaki and Troutt, 2008).

Language proficiency. The learner model typically expresses the user's language proficiency with respect to the same scale for text difficulty in the domain model, for example a grade level. Some systems directly ask the user to select a grade level (Miltsakaki and Troutt, 2008). Others estimate the level based on pre-test scores, for example the GEPT reading comprehension pre-test (Hsu et al., 2013) or entrance exam (Wu, 2014).

The model may also record the user's knowledge at the word level, i.e., specific words that the user knows or does not know. In REAP, at the beginning of a session, the instructor provides a list of words to be taught, and the students indicate whether they know each word in the list. The system then prioritizes texts that contain those words that are marked as "unknown" (Heilman et al., 2010).

3.3 Adaptation model

The adaptation model defines how the learner model modifies the domain model. It can be classified along two dimensions, in terms of the form of modification, and when it is applied.

Content vs. navigation adaptation. In "content adaptation", the adaptation model changes the content or presentation of the learning items. Unknown or target words can be underlined or otherwise highlighted with special fonts and colors, or given English glosses. Some systems identify these words from predictions by the learner model (Miltsakaki and Troutt, 2008; Brown and Eskenazi, 2004); others use graded vocabulary lists (Wu, 2014). The website *Guidelines for CLT Materials Development* (http://www.cltguides.com) uses a similar approach with the HSK lists. Some systems automatically display glosses for unknown words, so as to reduce the number of clicks needed (Ehara et al., 2012).

In "navigation adaptation", the adaptation model changes the sequence in which the learning items are presented (Brusilovsky and Peylo, 2003). It may change, for example, the sequencing of documents in the database (Hsu et al., 2013). It may also simply filter out documents outside the proficiency level or text category indicated by the learner model (Miltsakaki and Troutt, 2008), similar to the "graded reader" approach. Other systems use document rankings to adapt navigation in a softer manner. The ranking may be determined by a combination of article correlation, article difficulty and learner's ability (Wu, 2014). REAP uses a weighted average of scores based on text length, reading grade level, the number of unknown target words, and topic interests (Brown and Eskenazi, 2004; Heilman et al., 2007). It thus prefers texts with the optimal length and number of unknown target words, as chosen by the instructor.

Macro- vs. micro-adaptive instruction. In "macro-adaptive" instruction, the adaptation model is applied only once, typically with an initial test or self-assessment at the beginning of user interactions (Lee and Park, 2008). This paradigm is also known as "aptitude-treatment interactions" or "diagnosis-prescription" (Shute and Zapata-Rivera, 2012). The adaptation model decides on the navigation and content adaptation once and for all, assuming all learner characteristics to remain unchanged. Graded readers, and systems that ask users to choose their grade level, are examples of this approach.

In "micro-adaptive" instruction, the adaptation model is repeatedly applied. The system constantly monitors the learner to update learner characteristics and to re-run the adaptation model. The monitoring can be explicit, such as post-reading exercises to test whether the user has learned a target word, and tracking the number of times a target word has appeared in readings (Brown et al., 2005; Heilman et al., 2010); or user feedback on the interest and perceived difficulty of the document. It can also be implicit, such as the logging or user queries or dictionary look-up (Wu, 2014), to infer their (lack of) knowledge and/or interest in learning the word. Future recommendations may then prefer documents with unknown target words, target words that have been practiced fewer times, or those the user frequently looked up in dictionaries.

3.4 Complex word identification

In this section, we first review previous approaches in complex word identification (CWI) (Section 3.4.1). We then summarize current approaches for personalizing CWI (Section 3.4.2).

3.4.1 CWI shared tasks

Two shared tasks on CWI have been organized in recent years. The 2016 SemEval shared task focused on English CWI (Paetzold and Specia, 2016). The training data consisted of 200 sentences, with each target word annotated by 20 different non-native speakers of English. The test data consisted of 9,000 sentences, entirely annotated by a single annotator. In an analysis of the overall results, word frequencies were found to be the most reliable predictor of word complexity (Paetzold and Specia, 2016). The best team, which combined lexicon-based, threshold-based and machine learning voter subsystems, achieved a precision of 0.147 and recall of 0.769.

The 2018 shared task expanded to multiple genres, languages and user groups (Yimam et al., 2018). With the CWIG3G2 dataset (Yimam et al., 2017), the participating systems tackled texts from different genres, including news written by professional writers and by amateurs, and Wikipedia articles. The training set was labeled by both native and non-native speakers, and was expanded to include German and Spanish. Further, some systems were evaluated on their performance in cross-lingual CWI, being trained on English, German and Spanish and tested on French CWI.

This research is distinguished from the shared tasks in several ways. First, we do not address the effect of genres or cross-lingual complexity. Second, we evaluate our models on Chinese, a language that has received relatively less attention in CWI research. Most significantly, our CALL-oriented perspective demands a different research methodology. Since their training sets include annotated by multiple learners, systems in the shared tasks attempted to learn an aggregate, user-independent notion of word complexity, despite "the expected heterogeneity among non-native speakers with different language backgrounds and proficiency levels" (Paetzold and Specia, 2016). The significant variation among learners is also reflected in a Krippendorff's Alpha agreement of 0.244 among the annotators in the 2016 task (Paetzold and Specia, 2016). Both native and non-native speakers were involved in annotating the training sets in the 2018 shared task, and the absolute agreement between them is only 70% (Yimam et al., 2018). In contrast, this work attempts to model differences in vocabulary proficiency among learners. We develop personalized CWI models, and evaluate these models on a dataset that includes CFL learners with a wide range of vocabulary competencies.

3.4.2 Personalized CWI

To maximize their utility, CALL applications should ideally cater to language learners spanning a large range of language proficiency. A one-size-fits-all CWI model, therefore, would not adequately deliver materials that suit individual users. An early effort to optimize CWI on individual users was reported by Zeng et al. (2005), who showed that demographic features can help improve personalized CWI performance in the medical domain. Most recent approaches consist of two components: training set creation, and automatic classification based on the annotated training set.

Training Set Creation. Compare to those in the shared tasks (Section 3.4.1), training sets for personalized CWI tend to be limited in size. Since each user must do his/her own annotation, the training set must be kept reasonably small; yet, in order to be informative, it must include representative words that can discriminate between users of different proficiency levels. Previous work has explored the following approaches in creating training sets.

Graph-based Active Learning: This method constructs the training set with Error Bound Minimization (Gu and Han, 2012), a non-interactive graph-based active learning algorithm. The entire vocabulary is first organized as a multiple complete graph, where nodes correspond to words, and edge weights reflect the similarity between the frequency ranks of the words. The assumption is that the vocabulary knowledge of learners is similar for words with similar frequency ranks. The algorithm selects the k most informative nodes from the vocabulary graph in a non-interactive way, i.e., without using human labels during the learning process. This algorithm selects nodes that are globally important, based on the number of edges. Further, the nodes must not be heavily connected to previously sampled nodes.

This graph-based active learning approach was adopted by Ehara et al. (2014a) for English CWI with the values of k ranging from 10 to 50, and by Lee and Yeung (2018a) for Chinese CWI. We will also take this approach to create training sets.

Word sampling: An alternative method is to draw words from vocabulary lists at different levels of difficulty. Laufer and Nation (1999) proposed this "word sampling" approach with a ten-level proficiency scale, using 1000 words at each level. In a more coarse-grained implementation, Lee and Yeung (2018c) adopted a four-level scale in the context of English lexical simplification. The four levels corresponded to four graded vocabulary lists, based on rankings in the New General Service List, the TOEIC Service List, the New Academic Word List, and the Business Service List. They created a 40-word training set by drawing 10 words from each of these four lists.

Classification. Following user annotation on the training set, the system performs classification to identify complex words. Previous work has explored the following approaches in training CWI classifiers.

Label Propagation: After selecting the nodes by active learning, the system uses Local and Global Consistency (Zhou et al., 2004), a label propagation algorithm, to train an independent, binary classifier for each user. The nodes corresponding to the words in the training set are already labelled; the labels are then propagated to the unlabeled nodes based on edge weights. The assumption is that two nodes connected by a heavily weighted edge should have similar labels, and more heavily weighted edges should propagate more labels. On a dataset of Japanese learners of English, the best model achieved 76.44% accuracy (Ehara et al., 2010, 2014a). For Chinese CWI, however, Lee and Yeung (2018a) reported that an SVM classifier outperformed this method.

Graded Lists: This approach requires a number of pre-defined, graded vocabulary lists, say from level 1 to N. Each list defines a CWI model: the level-i CWI model predicts all words at the level-i vocabulary list as "non-complex", and all other words as "complex". The system then needs to assign the user to one of the N models. It does so by calculating the precision and recall of each CWI model on the user's training set, and selects the model that produces the highest F-score (Lee and Yeung, 2018c).

Statistical classification: A number of standard statistical classifiers, such as logistic regression and Support Vector Machines (SVM), have been used in previous studies (Ehara et al., 2010, 2014a; Yimam et al., 2018). They explored a large range of features, including word frequency, number of syllables, word length, word embeddings, n-gram probabilities, part-of-speech and suffix length, as well as semantic features such as the number of synsets, hypernyms and hyponyms. One of the top performing teams in the 2018 shared task also incorporated frequency statistics from learner corpora (Kajiwara and Komachi, 2018). In the only previous study on Chinese CWI, classifiers were trained on word frequency and character frequency in both standard and learner corpora, and features based on graded vocabulary lists led to the best performance (Lee and Yeung, 2018a).

3.5 Assessment guidelines for Chinese as a foreign language

Our CWI approach exploits three assessment scales, authored by experts in Chinese language pedagogy. For Chinese as a foreign language (CFL), the two major scales are the *Hanyu Shuiping Kaoshi* (HSK) (Hanban, 2014) and the *Test of Chinese as a Foreign Language* (TOCFL) (Tseng, 2014). The HSK guidelines provide a character list and a vocabulary list for each of six difficulty levels, covering a total of 9,600 vocabulary items. The TOCFL guidelines also provide similar vocabulary lists, covering a total of 8,000 vocabulary items across seven difficulty levels. Both scales can be mapped to the *Common European Framework of Reference for Languages*.

The *Lexical Lists for Chinese Learning in Hong Kong*, published by the Hong Kong Education Bureau (EdB)[1], defines the norm in Chinese vocabulary proficiency for students in Hong Kong. It consists of 9,706 Chinese words that form part of the Chinese curriculum in primary schools. Of these, 4,914 words are labelled as "key stage 1", which means they should be acquired by the end of Grade 3; the remaining 4,914 words are labelled as "key stage 2", which means they should be acquired by the end of Grade 6. Since these labels are intended for native speakers, there is no established mapping from these lists

[1] https://www.edbchinese.hk/lexlist_ch/

to HSK and TOCFL, which aim at CFL learners. Even so, the difference in key stages may still help indicate the relative complexity between words, and hence their expected order of acquisition for non-native speakers.

Feature	HSK	TOCFL
1	Level 1	
2	Level 2	
3	Level 3	Band A1
4	Level 4	Band A2
5	Level 5	Band B1
6	Level 6	Band B2
7	-	Band C1

Table 1: The HSK+TOCFL feature, derived from the merged levels of the *Hanyu Shuiping Kaoshi* (HSK) and *Test of Chinese as a Foreign Language* (TOCFL) guielines, according to the mapping from Fachverband Chinesisch e.V.

4 Methodology

Following an overview of our system architecture (Section 4.1), we present the user interface of our mobile app (Section 4.2). Then, we describe our algorithm for complex word identification (Section 4.3).

4.1 System Architecture

Similar to existing CALL applications (Section 3), our system include a domain model, a learner model and an adaptation model.

4.1.1 Domain model

Similar to many existing systems, our domain model of a document includes its category, its title, English glosses for its words, as well as its text difficulty. In a departure from conventional approaches, we do not express text difficulty as a level on a fixed scale. Since these scales can be opaque to users (Section 2), we adopt a more transparent, intuitive metric for text difficulty — the percentage of new vocabulary in the document, i.e., the percentage of words that are unknown to the user. More formally, we define a user's vocabulary profile $V = \{w_1, \ldots, w_k\}$ as the set of all words that are known to him or her. We then construct the indicator function $unk(w, V)$, which returns 1 if the the word w is unknown to the user with profile V, and 0 if it is known. Given a document with L words, say $d = (w_1, \ldots, w_L)$, the function $td(d, V)$ computes its text difficulty in terms of percentage of new words:

$$td(d, V) = \frac{\sum_{i=1}^{L} unk(w_i, V)}{L} \quad (1)$$

As reflected in the definition of $td(d, V)$, a document's difficulty depends on the user. More precisely, it depends on V, his/her vocabulary profile, which is to be estimated by the learner model (Section 4.1.2).

4.1.2 Learner model

As shown in Table 4.1.2, our learner model includes four parameters. The first three concern reading interests and preferences, while the fourth models language proficiency.

Reading interests Our system allows the user to explicitly set the preferred category (Section 4.2.3). In contrast, REAP only considers user preferences as one of the factors for text recommendations (Heilman et al., 2010). Similar to Read-X (Miltsakaki and Troutt, 2008), users can use search keywords to directly retrieve documents that contain them (Section 4.2.1). By default, search results prioritize documents that include word from the Vocabulary List, a list where the user can save new vocabulary, words that need review, or whichever words that interest them. The equivalent in REAP is more limited: the user can indicate unknown words only within the vocabulary items in the list crafted by the instructor.

Parameter	Description	Example	Input method
Text difficulty	Maximum percentage of new words in document	$\leq$ 20% new words	Integer input
Text category	Theme of document	Fables	Checkboxes
Search keywords	Words to appear in document	*huli* 'fox'	String input; Vocabulary List
Vocabulary profile	Words known to user	20,000 words	Integer input; Vocabulary Assessment

Table 2: Parameters in document search queries in our system

Language proficiency In order to estimate a document's difficulty, our learner model keeps a vocabulary profile for each user. Users can either set this profile manually by indicating their vocabulary size, or let the system automatically estimate it on the basis of the Vocabulary Assessment (Section 4.2.5). This self-assessment is similar in format to that in REAP but differs in scope and purpose. REAP attempts to maximize the hits of the "unknown" words in recommended documents, in a way similar to our Vocabulary List; in contrast, our system uses the self-assessment as training data to estimate the user's entire vocabulary profile, $\hat{V}$, by performing complex word identification (Section 4.3).

The simplicity of this learner model brings several advantages. First, the learner model's predictions are transparent to users. In the reading environment, they can see which words are estimated to be known, or unknown (Section 4.2.2). It takes a single tap to correct the model, facilitating dynamic adaptivity of the learner model (Section 4.1.3).

Second, it provides a straightforward metric for learning pace (Section 2.4). Users can include their preferred percentage of new vocabulary as part of their search query. Those who prefer leisure reading can set a lower percentage, while those who wish to maximize vocabulary acquisition can set a higher percentage.

4.1.3 Adaptation model

Our adaptation model adapts both content and navigation, and provides micro-adaptive instruction.

Content adaptation Following previous approaches such as Toreador (Miltsakaki and Troutt, 2008), our reading environment highlights search keywords, unknown words, and words in the Vocabulary List with colors and underlines, in order to draw the user's attention (Section 4.2.2). However, there is no adaptation with reading aids; word segmentation and English glosses are always put at the users' disposal.

Navigation adaptation Our model offers personalized document recommendations through the ranking algorithm. It calculates the percentage of new vocabulary in each document with respect to the user's vocabulary profile. Search results display this percentage for each document, and a document is ranked according to how close its percentage is to the user-specified target (Section 4.1.2). More formally, suppose M is the maximum percentage of new words set by the user, and V is the user's vocabulary profile. We define the distance score of a document d to be $dist(d, V, M)$:

$$dist(d, V, M) = M - td(d, V) \tag{2}$$

where $M \geq td(d, V)$. Given a database D with documents $\{d_i\}$, the search algorithm ranks the documents in increasing value of the distance score. In other words, the closer a document's percentage of new vocabulary is to the user-specified percentage, the higher it is ranked. In particular, the top-ranked document is:

$$top(V, M) = argmin_{d_i \in D} dist(d_i, V, M) \tag{3}$$

Micro-adaptive instruction As users' reading preferences evolve, the parameters in Table 4.1.2 need to be adjusted. Our system leaves it up to users to execute short-term or one-off changes. For example, they can adjust text difficulty to match the amount of effort they wish to put into the next reading, or

Figure 1: Search interface, with three buttons in the bottom for personalization options

use different search keywords or text categories to suit their mood. The Vocabulary List and vocabulary profile facilitate longer-term changes. As they encounter new words that they want to learn and review, users can insert them into the Vocabulary List, and the next search will prioritize texts that contain them (Section 4.2.1). Over time, as users gain in proficiency, text difficulty will become overestimated. When users see known words getting predicted as new (or vice versa) in a document, they can directly edit their status — from unknown to known, or in the opposite direction — with a single tap. The system then dynamically re-estimates the entire vocabulary profile (Section 4.3), and re-calculates text difficulty in the next search. Dictionary look-ups may provide some evidence of the user's lack of knowledge of a word, we do not wish to discourage the use of dictionaries. Another option is to administer vocabulary tests. Although they lead to be more accurate assessment, we prioritize convenience in order to encourage users to frequently update the learner model.

4.2 Mobile app description

This section describes the personal reading tutor's search interface (Section 4.2.1), read texts (Section 4.2.2). We then explain how users can personalize their search with respect to text categories (Section 4.2.3), text difficulty (Section 4.2.4), and their vocabulary profile (Section 4.2.5).

4.2.1 Search interface

The app has access to a text database of authentic documents. In the search interface on the main page of the app, the user can type in one or more Chinese keywords (e.g., *huli* 'fox' in Figure 1), and then tap the "Search" button to search for documents that contain all the keywords. This field can also be left blank if the search targets only one text category (Section 4.2.2).

By default, the app checks the box "Search for words in Vocabulary List", which instructs the system to include words in the Vocabulary List as search keywords. The user can store words that s/he wants to learn or review in this list. The words are displayed beneath the search field, with the most recently added ones placed first.

Figure 2 shows the search results. Documents whose titles contain the keywords are displayed on top, in red. These documents are sorted by the percentage that appears after each title. This percentage is the proportion of new vocabulary in the document, according to the Vocabulary Profile, which labels all words in the text database as "known" or "unknown" to the user. The percentage must meet a user-specified minimum (Section 4.2.4). They are followed by documents in which the keywords appear only in the body but not in the title; they are colored in blue.

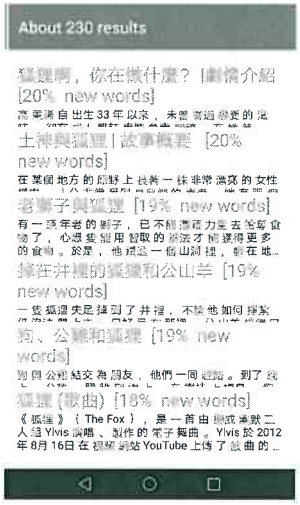

Figure 2: Search results.

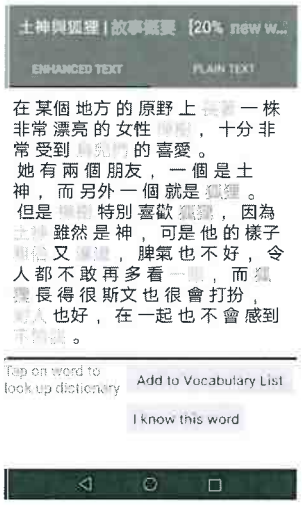

Figure 3: Reading environment.

4.2.2 Reading environment

After tapping on a search result, the user is taken to the "Enhanced Text" mode, as shown in Figure 3. In this mode, the text is word-segmented with the Stanford Chinese parser (Manning et al., 2014). The default font color is black. The search keywords are highlighted in red. New vocabulary is colored in orange to draw the user's attention. Further, words in the Vocabulary List are underlined and bolded.

The user can tap on any word to see its English gloss, as provided by the dictionary entry in CEDICT. Besides reading the gloss, the user can perform two actions on the word. First, the user can add the word to the Vocabulary List by tapping on "Add to Vocabulary List" (or remove it by tapping on "Remove from Vocabulary List"). Second, the user can edit the Vocabulary Profile. If a word is wrongly predicted as "unknown" and displayed in orange, the user can inform the system by tapping on "I know this word". Similarly, if a word wrongly predicted as "known", the user can inform the system by tapping on "I don't know this word". As part of the micro-adaptive algorithm (Section 3.3), the Vocabulary Profile would then be dynamically updated and the complex word identification algorithm is re-run (Section 6).

Alternatively, the user can read the "Plain Text" version, which does not provide word segmentation, English gloss, or access to Vocabulary List or Vocabulary Profile.

Figure 4: Selection of text categories.

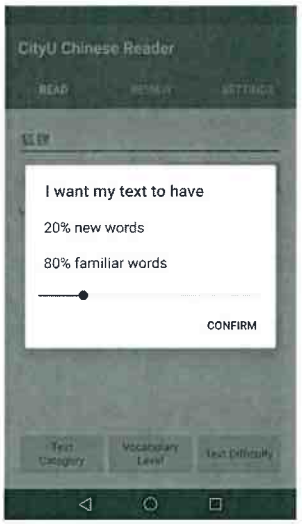

Figure 5: Selection of text difficulty in terms of percentage of new words.

4.2.3 Text category

The documents in our database belong to 24 categories, or themes (Section 5.1). To restrict search results to one or more text categories, the user can tap on the "Text Category" button on the main search page (Figure 1), which opens a window with checkboxes for the categories (Figure 4).

4.2.4 Text difficulty

The right button, "Text Difficulty", leads to a panel that states, by default, "I want my texts to have 20% new vocabulary and 80% familiar words". This means that the app returns documents in which at most 20% of the words are new with respect to the user's vocabulary profile. To retrieve texts that contain more (or less) new vocabulary, the user can use the slider to specify the desired percentage (Figure 5). The higher the percentage, the harder the documents.

4.2.5 Vocabulary level

The middle button, "Vocabulary Level", lets the user manage his or her vocabulary profile, on which text difficulty is based depends. With the "Custom" option, the user himself/herself chooses a particular vocabulary size. With the "Automatic" option, the system makes the estimation. We first describe the

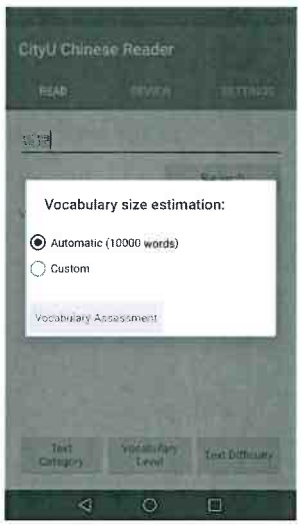

Figure 6: Selection of vocabulary level.

ranked vocabulary list, on which both options depend.

Ranked vocabulary list. The app maintains a list of all words in the text database, ranked from easy to difficult based on word frequency in Chinese Wikipedia and on two major assessment scales for CFL: the *Test of Chinese as a Foreign Language* (TOCFL) (**?**) and the *Hanyu Shuiping Kaoshi* (HSK) (Hanban, 2014) (Section 3.5). The list begins with the words in HSK; words within each level are ranked according to their frequency in Chinese Wikipedia; words in TOCFL are mapped to the equivalent HSK level according to the mapping from Fachverband Chinesisch e.V. [2]. All other words are appended to the end of the list, again ranked by frequency in Wikipedia.

Manual estimation. By default, the app assumes a vocabulary size of 10,000 Chinese words. This means that in the list described above, all words ranked in the top 10,000 are considered "known" in the vocabulary profile, and the rest as "unknown". As shown in Figure 6, the user may change the size manually by selecting the "Custom" button. The user can then use the slider to directly adjust the vocabulary size. All words whose status have been manually edited by the user, however, are not affected by this change. To provide a reference based on HSK and its mapping to CEFR, a vocabulary size of 2000 is labelled as "basic", 4000 as "independent", 10,000 as "advanced", and 20,000 as "proficient".

Automatic estimation. To fully utilize the learner model, the user may allow the app to automatically estimate his or her vocabulary size by checking the "Automatic" button. The Vocabulary Profile then runs the complex word identification algorithm to predicts its estimation on each word (Section 4.3). The more training data it has, the better its accuracy. To provide training data, the user can tap on the "Vocabulary Assessment" button to take a self-assessment. The self-assessment consists of 50 words, selected from a 9,000-word-list provided by the Hong Kong Education Bureau. The user is to annotate each word as "known" or "unknown" (Lee and Yeung, 2018b) (Figure 7).

To select these 50 words, we used the complex word identification model that produced state-of-the-art results for English (Ehara et al., 2014b). First, the entire vocabulary is organized as a multiple complete graph. Nodes correspond to words and edge weights show how similar the frequency ranks of a word pair are. The assumption is that words with similar frequency ranks are known to learners whose familiar words are similar to each other. It is primarily based on word frequencies in BNC and COCA. The model then performs Error Bound Minimization (Gu and Han, 2012), a non-interactive graph-based active learning algorithm, to select the 50 most informative nodes from the vocabulary graph. It selects nodes that are globally important, based on the number of edges, and are not heavily connected to previously sampled nodes.

[2] http://www.fachverband-chinesisch.de/chinesischindeutschland/pruefungen/index.html

Figure 7: Vocabulary assessment.

4.3 Complex word identification

We constructed our training set from the *Lexical Lists for Chinese Learning in Hong Kong* (Section 3.5). Following Ehara et al. (2014a), we used the Graph-based Active Learning method to select the most informative nodes to generate a training sets with 50 words. While we could in principle use larger training sets, it would in practice be undesirable to burden users with a large amount of vocabulary annotation before allowing them to use the system.

Following user annotation of the training set, we trained a classifier to perform complex word identification (CWI) for each individual user. The classifier accepts any Chinese word as input; its output is "non-complex" if the user knows the word, and "complex" if the user does not. We compared the following approaches for classifier training.

Type	Native/non-native	Linguistic level	Feature
Frequency	Non-native	Word	LearnerFreq
		Character	LearnerFreq-char-max, LearnerFreq-char-min
	Native	Word	Freq
		Character	Freq-char-max, Freq-char-min
Assessment scale	Non-native	Word	HSK+TOCFL
		Character	HSK-char-max, HSK-char-min
	Native	Word	EdB

Table 3: Features used in our statistical classifiers contain both frequency statistics and difficulty scale, at both the word and character levels, derived from native and non-native resources.

4.3.1 Statistical Classification

Table 3 summarizes the features that we implemented. They can be divided into two main types, with some based on assessment scales and others on frequency statistics. Both feature types are defined at both the word and character levels, and derived from native and non-native sources. Features based on assessment scales include:

- **HSK+TOCFL**: The difficulty level of the word (1, 2, 3, 4, 5, 6 or 7), according to the HSK and TOCFL guidelines (Section 3.5). We use the mapping proposed by Fachverband Chinesisch e.V. [3] (Table 1). To resolve conflicts between the two scales, when a vocabulary item appeared in more than one difficulty level, we always selected the lowest level assigned to the word.

[3] http://www.fachverband-chinesisch.de/chinesischindeutschland/pruefungen/index.html

- **EdB**: The key stage (1 or 2) of the word according to the Chinese curriculum guidelines from Education Bureau of Hong Kong. For words not included in the guidelines, the EdB feature is "3".
- **HSK-char-max**: A Chinese word may contain multiple characters, potentially at various levels of difficulty. This feature takes the maximum level among the characters in the word (1, 2, 3, 4, 5, or 6), according to the HSK guidelines.
- **HSK-char-min**: Same as above, except that the minimum of the difficulty levels is taken.

Features that are based on frequency statistics include:

- **Freq**: The frequency of the word in a standard Chinese corpus.
- **Freq-char-max**: For each character in the word, we compute its frequency count in the standard Chinese corpus. This feature takes the maximum frequency.
- **Freq-char-min**: Same as above, except that the minimum is taken.
- **LearnerFreq**: The frequency of the word in a learner Chinese corpus.
- **LearnerFreq-char-max**: For each character in the word, we compute its frequency count in the learner Chinese corpus. This feature takes the maximum frequency.
- **LearnerFreq-char-min**: Same as above, except that the minimum is taken.

Similar to Ehara et al. (2014a), we used SVM and logistic regression as statistical classifiers.

4.3.2 Graded Lists

In addition to statistical classifiers, we implemented an approach based on graded vocabulary lists, similar to the one proposed by Lee and Yeung (2018c).

Graded Lists: This approach considers seven "typical" learners, whose vocabulary knowledge conforms exactly to one of the seven levels in the HSK and TOCFL guidelines (Section 3.5). More specifically, if the learner is at level N, then he or she knows precisely all those words in the corresponding vocabulary lists of HSK and TOCFL from level 1 up to level N, and no other word. The system calculates the precision and recall of each of the seven CWI models on the user's training set. It then assigns the user to the model that optimizes the F-measure.

Graded Lists Oracle: To benchmark the accuracy of the Graded Lists approach, we also evaluated its oracle version. The oracle takes the same approach as above, except that it has access to the test data, and assigns the user to the model that optimizes the accuracy on the test set.

5 Data collection and analysis

During development of the personal reading tutor, we retrieved Chinese web documents to compile a text database (Section 5.1), and invited learners of Chinese as a foreign language to evaluate our app (Section 5.2). This section gives details on these datasets.

5.1 Text data and processing

We retrieved Chinese documents from the web, including 10,258 documents from OPUS (http://opus.lingfil.uu.se/), a publicly available parallel corpus; 71,039 pages from the Chinese Internet Corpus (Sharoff, 2006); 232 fables from Project Gutenberg; 28,947 short essays from Duanmeiwen.com; 843,436 texts from Chinese Wikipedia (zh.wikipedia.org). In addition, our collection includes all Chinese documents from Common Crawl (http://commoncrawl.org/), an open repository of web crawl data. Some of these documents are very short, or are mixed with other languages. In the interest of text quality in the personal reading tutor, we filtered the text collection to keep only those with at least three sentences and mostly Chinese content. The database in the publicly available version of the app contains 788,982 documents (Table 4).

Table 4: Composition of text database in our app

Category	Source	Number of documents
Fables	Project Gutenberg	232
Short essays	Duanmeiwen.com	8,190
Web pages	Chinese Internet Corpus	49,548
Computer Science	Chinese Wikipedia	1,008
Information	Chinese Wikipedia	151
Geography	Chinese Wikipedia	174,119
Humanities	Chinese Wikipedia	62,983
History	Chinese Wikipedia	62,481
People	Chinese Wikipedia	81,540
Science	Chinese Wikipedia	72,801
Society	Chinese Wikipedia	60,574
Applied Science	Chinese Wikipedia	22,554
Social Science	Chinese Wikipedia	6,483
Technology	Chinese Wikipedia	11,890
Religion	Chinese Wikipedia	9,587
Other	Chinese Wikipedia	132,329
General	Chinese Wikipedia	18,686
Art	Chinese Wikipedia	3,929
Interdisciplinary	Chinese Wikipedia	1,515
Philosophy	Chinese Wikipedia	1,829
Leisure	Chinese Wikipedia	5,693
Language	Chinese Wikipedia	277
Science and Technology	Chinese Wikipedia	452
Literature	Chinese Wikipedia	121

We extracted the content tags from the Chinese Wikipedia documents, thereby automatic classifying them into the categories Computer Science, Information, Geography, Humanities, History, People, Science, Society, Applied Science, Social Science, Technology, Religion, Other, General, Art, Interdisciplinary, Philosophy, Leisure, Language, Science and Technology, and Literature. Further, we performed word segmentation on all documents with the Stanford Chinese segmenter (Manning et al., 2014). We used Solr (http://lucene.apache.org/solr/), a high performance search server that supports full-text search, for our database.

5.2 Learner data

We invited learners of Chinese as a foreign language to participate in a user study of the app interface (Section 5.2.1), to express their opinion on the design of the app (Section 5.2.2), and to evaluate our complex word identification model (Section 5.2.3),.

5.2.1 App interface study

While the personalization options help tailor search results, they also add complexity to the search interface. We evaluate user experience of the major functions of the app, and measure the extent to which they help users find texts with different levels of difficulty, compared with a simple search engine on the web. The participants in this user study were seven learners of Chinese, whose native languages include English, French, Korean, and Thai. Their years of CFL studies ranged from 7 to 13 years.

The subjects independently read an introductory manual, which guided them through a few search query scenarios. They were then shown Table 6.1 and asked to perform the five search queries. The purpose of the evaluation was two-fold. First, we evaluate whether they were able to correctly execute the search, which would indicate the quality of the design of the user interface. Second, we measure their perception of the effects of varying two of the search parameters (Table 4.1.2): vocabulary profile and text difficulty. For each query, the subjects read the five top-ranked texts returned from the text category "Short essays", and then rated their difficulty on a 5-point Likert scale, from "very easy" (score 1) to "very difficult" (score 5).

5.2.2 Survey

The same subjects also completed a survey, which consists of a number of statements on the general design of text search tools for language learners, and on the specific implementation of the our app (Table 6.1). They indicated their opinion on each statement on a 5-point Likert scale, from "strongly disagree" (score 1) to "strongly agree" (score 5).

5.2.3 Complex word identification dataset

To derive word frequency statistics for the Freq feature, we used a corpus of 9.2 million sentences from Chinese Wikipedia. For the LearnerFreq feature, we used the *Jinan Corpus of Learner Chinese* (JCLC) (Wang et al., 2015), which contains 6 million Chinese characters written by students from over 50 different native language backgrounds. We performed word segmentation on both corpora with the Stanford Chinese parser (Levy and Manning, 2003). We used the implementation of SVM and logistic regression (LR) classifiers in scikit-learn (Pedregosa et al., 2011), with all combinations of the features listed in Table 3.

As baselines for the proposed approaches, we implemented the **Majority baseline**. This baseline always predicts the label that is assigned to the majority of the words in the training set. The Majority baseline can be a very strong baseline for low-proficiency language learners, who have limited vocabulary knowledge.

As test set, we drew 550 words from the *Lexical Lists for Chinese Learning in Hong Kong* (Section 3.5), such that they did not overlap with the training set. We randomly selected words spanning different levels of difficulty, as measured by their frequency counts in Chinese Wikipedia.

We asked seven subjects, all CFL learners, to label each word in these datasets on a five-point scale: (1) Never seen the word before; (2) Probably seen the word before; (3) Absolutely seen the word before but do not know its meaning, or tried to learn the word before but forgot its meaning; (4) Probably know, or

able to guess, the word's meaning; and (5) Absolutely know the word's meaning. Following Ehara et al. (2010), we consider a word to be non-complex if it is scored five, and complex otherwise. The numbers of non-complex and complex words in the test set and training set for each learner are shown in Table 5. For analysis purposes, we divide the seven learners into two groups. The four who knew less than 150 of the 550 words were designated as "Low-Proficiency", and the other three as "High-Proficiency".

Following Ehara et al. (2014a), we asked the learners to score their knowledge of words in isolation, rather than in context such as in the CWI shared task. The context of a word provides clues for learners to guess its meaning, and thus affects how they score their knowledge of a word. Even if the learner is able to guess a word in one context, the same is not guaranteed in another context since the content and the density of new words in each text is different. Since the CWI model in our system is intended for text selection, we did not wish to assume any one particular context when determining the learner's knowledge of a word. CWI annotation in a context-free manner thus allowed us to more accurately judge the learners' ability to understand different reading materials.

Proficiency	**Training Set**		**Test Set**	
	non-complex	complex	non-complex	complex
Low	11	38	68	482
	7	43	78	472
	14	36	113	437
	14	36	146	404
High	13	37	188	362
	17	33	217	333
	22	28	296	254

Table 5: The seven subjects in our dataset (Section 5.2.3), divided into "Low" and "High" proficiency according to the number of words annotated as "non-complex" in the test set.

6 Results and discussion

We discuss our evaluation results on the app interface (Section 6.1), learners' opinion on the design of the app (Section 6.2), and the performance of our complex word identification model (Section 6.3).

6.1 App interface study

For four of the search queries, all subjects received the expected results. In query #1, one subject reported a different search result than the expected.[4] Given the limited self-training of the subjects, these results suggest that the user interface was easy to use. We now report the subjects' perception of the difficulty of the texts as they varied the parameters.

Vocabulary profile In search query #1 in Table 6.1, the vocabulary profile was set at 40,000 words. Since most words were known, there was practically no filtering even though text difficulty was capped

[4]The subject reported 7492 results instead of 7942 results, which could have been a handwriting error.

Search	Mode	Vocabulary Profile	Text Difficulty	Keyword	Average Score
# 1	App	40,000 words	20%	None	3.43
# 2	App	4,000 words	20%	None	2.07
# 3	App	4,000 words	15%	None	1.80
# 4	App	4,000 words	20%	*baba* 'father'	2.13
# 5	Web	None	None	*baba* 'father'	2.60

Table 6: Difficulty ratings for various search parameters, within the "short essay" category

General Design issues In general, it's a good idea for this kind of tool to ...	**Score**	**App implementation** In this app, it was easy for me to:	**Score**
Let users filter search results based on text category	5.00	Choose my preferred text category [Text Category checkboxes]	5.00
Let users store a list of words for review or search later	5.00	Save words in the Vocabulary List for review or search later [Vocabulary List]	4.86
Rank search results by percentage of new words	4.71	View the percentage of new vocabulary in search results [Search button]	4.86
Let users filter search results based on a maximum percentage of new words	4.71	Set a maximum percentage of new words in search results [Text Difficulty slider]	4.86
Let users tell the tool whether they know a word or not	4.71	Inform the app whether I know or do not know a word [Vocabulary Profile]	5.00
Try to predict whether users know a word or not	3.71	Complete the Vocabulary Assessment [Vocabulary Profile ("Automatic")]	4.86
Let users indicate their vocabulary size	4.42	Specify my Vocabulary Level [Vocabulary Profile ("Custom")]	4.71

Table 7: Statements in the survey and their average ratings

at 20% new words. The five documents retrieved in this setting received an average difficulty score of 3.43. In query #2, when the profile was reduced to 4,000 words, the average difficulty score decreased to 2.07 This suggests that the manual adjustment of the vocabulary profile, via the "Custom" option (Section 4.2.5), had the intended effect of retrieving easier documents. The "Automatic" option is to be evaluated in Section 6.3.

Text difficulty In search query #3, the vocabulary profile remained the same as #2, at 4,000 words (Table 6.1). The text difficulty, however, was lowered from 20% to 15%. As expected, the average difficulty score of the top five documents further dropped, from 2.07 to 1.08.

App vs. generic search Finally, queries #4 and #5 compare text retrieval in the app with a similar search on the website duanmeiwen.com, from which the documents in "Short essays" were drawn (Table 6.1). Both queries include the keyword *baba* 'father', and have a similar pool of candidate documents. The main difference lay in the capping of 20% new words (at a vocabulary size of 4,000 words) in #4, and the absence of such constraints in #5. On average, the subjects found the documents returned by the app to be easier (score 2.13) than those returned by the website (score 2.60).

6.2 Survey

The statements on the left column of Table 6.1 deal with general design of text retrieval tools for language learning. All subjects strongly agree (score 5.0) that users should be able to choose text categories, and that they would benefit from making a list of words for which they wish to review or search later. Most also liked (score 4.71) the idea of ranking and filtering search results by percentage of new words — a central feature in our personalized text retrieval method. They also mostly approve (score 4.71) of the

user indicating whether they know a word or not, one of the premises of our adaptive algorithm. They were less enthusiastic (score 3.71) about the system making predictions on their word knowledge; this was perhaps partly attributable to mistakes made in complex word identification, which is a challenging NLP task (Section 6.3). Instead, they slightly prefer (score 4.42) to manually indicate their vocabulary size.

The statements on the right column solicited the subjects' opinion on the app itself, asking whether they found it easy to perform various functions. The actual features were not specified in the survey, but are shown here to clarify that these statements allude to the selection of text categories, the Vocabulary List, the search results, setting the text difficulty parameter, working with the vocabulary profile and vocabulary assessment. Most subjects found the interface well implemented. The "Custom" option of the vocabulary profile received the lowest level of satisfaction, likely due to the difficulty of estimating one's vocabulary size. We now turn to an evaluation on complex word identification, which seeks to automate this process.

6.3 Complex word identification

We discuss the overall CWI performance of the various approaches (Section 6.3.1). We then provide more in-depth analysis on the best model, focusing on its performance for low- and high-proficiency learners (Section 6.3.2).

6.3.1 Overall results

Table 8 shows the CWI performance of the baseline and the various approaches discussed in Section 4.3. Excluding the Graded Lists Oracle (Section 4.3.2), the SVM trained with the feature set that combines non-native (HSK+TOCFL) and native (EdB) assessment scales achieved the best performance, followed by the feature set with non-native information (HSK+TOCFL) alone.

The Majority baseline predicted all words to be complex. Since our subjects had relatively low proficiency in Chinese, this baseline yielded a strong performance of 71.1% accuracy and 82.3% F-measure, outperforming both the SVM (60.9% and 68.0%) and logistic regression (49.8% and 33.9%) classifier trained on word frequency in Chinese Wikipedia (Freq).

Consistent with the observations reported by Kajiwara and Komachi (2018), using word frequencies in a learner corpus led to more accurate results than frequencies in a standard corpus. A logistic regression (LR) classifier trained on the *Jinan Corpus of Learner Chinese* (JCLC) reached 76.3% accuracy and 83.4% F-measure, and the SVM achieved 76.6% accuracy and 83.7% F-measure, both above the Majority baseline. Despite the smaller size of the JCLC comparative to Chinese Wikipedia, word usage statistics from texts produced by language learners themselves appear to align more closely to their vocabulary knowledge.

The Graded Lists approach achieved 78.4% accuracy and 85.3% F-measure, outperforming all LR and SVM classifiers trained on word frequencies. This showed the effectiveness of the assessment scales in predicting a learner's vocabulary knowledge. Indeed, classifiers with features based on the assessment scales also proved more effective than those based on word frequencies. For the SVM classifier, the feature set with word difficulty levels in the HSK and TOCFL guidelines (HSK+TOCFL) pushed the accuracy up to 78.4%, tied with the Graded Lists approach, and F-measure up to 85.4%, slightly outperforming the Graded Lists approach. The best performance, at 79.2% accuracy and 85.4% F-measure, was obtained by adding the key stage information in the EdB lexical list (HSK+TOCFL+EdB). Notably, this model lay within 0.2% of the accuracy and 0.3% of the F-measure of the Graded Lists Oracle. Combining this model with other frequency-based features resulted in slight degradation in performance.

6.3.2 Low- vs. High-Proficiency Learners

We now examine variations in CWI performance with respect to learner proficiency, by dividing our subjects into a low-proficiency group and a high-proficiency group (Table 5):

Low-proficiency learners. The SVM classifier trained on the HSK+TOCFL+EdB feature set achieved the highest accuracy for low-proficiency learner, at 83.8%. Both precision and recall were high, at 87.8%

and 92.5% respectively. Since the word lists cover most of the words they know, the word difficulty scales seem particularly useful for capturing the limited vocabulary of these beginners.

High-proficiency learners. For more advanced learners, the word lists are not sufficiently comprehensive for modelling their larger vocabularies. The SVM classifier trained on the HSK+TOCFL+EdB feature set achieved a lower accuracy for this group of learners, at 73.0%, though still above the baseline. While the model maintained a high recall (91.4%), its precision was lower (70.1%) since it misclassified plenty of non-complex words as complex. Due to the relatively poor coverage of difficult words in the assessment scales, this feature set underestimated the learner's knowledge of more advanced vocabulary.

Augmenting the HSK+TOCFL+EdB feature set with frequency-based features lowered the overall results (Table 8). Among high-proficiency learners, however, these features may help the model cover a greater range of words. Indeed, the best model for the high-proficiency learners was the LR model trained on HSK+TOCFL+EdB, LearnerFreq and LearnerFreq-char-min features, with 75.4% accuracy. The frequency-based features helped predict more of the advanced vocabulary as "known", increasing precision from 70.1% to 74.5%, though at the cost of a decrease in recall, from 91.4% to 86.3%.

Method	Feature set	Accuracy	Precision	Recall	F-measure
Majority baseline	n/a	0.711	0.711	1.0	0.823
Logistic Regression	Freq	0.498	0.936	0.310	0.339
	LearnerFreq	0.763	0.807	0.877	0.834
	HSK+TOCFL	0.724	0.818	0.789	0.796
	HSK+TOCFL+EdB	0.727	0.815	0.811	0.804
	HSK+TOCFL+EdB+LearnerFreq	0.767	0.821	0.863	0.835
	HSK+TOCFL+EdB+LearnerFreq +LearnerFreq -char -min	0.765	0.836	0.831	0.829
SVM	Freq	0.609	0.754	0.770	0.680
	LearnerFreq	0.766	0.803	0.889	0.837
	HSK+TOCFL	0.784	0.783	0.946	*0.854*
	HSK+TOCFL+EdB	*0.792*	0.802	0.920	*0.854*
	HSK+TOCFL+EdB+LearnerFreq	0.765	0.787	0.903	0.835
	HSK+TOCFL+EdB+LearnerFreq +LearnerFreq -char -min	0.761	0.761	0.956	0.844
Graded Lists	HSK+TOCFL	0.784	0.790	0.934	0.853
Graded Lists Oracle	HSK+TOCFL	0.794	0.789	0.942	0.857

Table 8: Accuracy in complex word identification using various methods and feature sets.

7 Conclusions and Recommendations

The main contributions of this project are three-fold. First, we have developed a personalized and adaptive text retrieval method that ranks search results in terms of its ratio of new vocabulary for the user. Second, we have advanced the state-of-the-art in complex word identification (CWI), an important component in this method. CWI automatically predicts the user's vocabulary knowledge based on a small, annotated sample of their vocabulary. With previous studies focused on European languages, we disseminated the first study on CWI for Chinese as a foreign language (CFL). We show that an SVM classifier, trained on native and non-native assessment scales on Chinese, achieves the highest accuracy of 79.2% on a dataset consisting of seven CFL learners at different proficiency levels. Third, we implemented this method in a mobile app as a personal reading tutor. In a survey among CFL learners, most agree that such tutors should rank and filter search results by percentage of new words — a central feature in our method. Most found the app easy to use, with average scores above 4.7 out of 5 for all seven questions regarding

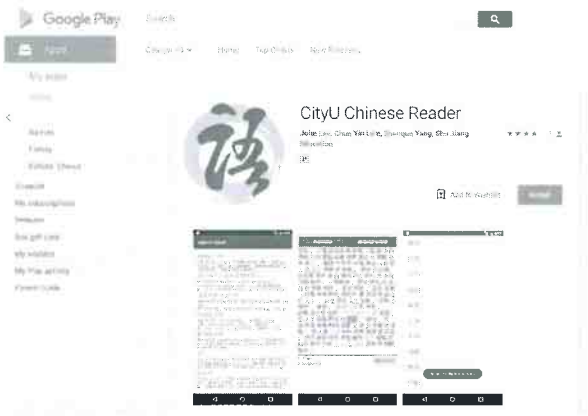

Figure 8: Access to app from Google Play.

functions of the app. With one exception, all subjects were able to perform all search queries correctly, providing corroborating evidence of the quality of the app interface.

Matching learners with appropriate texts remains a challenging task. It is hoped that this app will encourage active learning among language learners by providing individualized reading experience. To further improve and promote this new type of reading tutor, we recommend further research in three directions. First, we would like to raise CWI performance with active learning, by dynamically querying users on their vocabulary knowledge as they interact with the app; and with other linguistic features, such as semantic and n-gram patterns, for training the statistical classifier. Second, the tutor can benefit from a more comprehensive assessment of text difficulty, beyond vocabulary items, that takes into account syntactic complexity (Chinkina and Meurers, 2016). This would help the search algorithm return more pedagogically suitable results. Third, longitudinal studies on users of the app would evaluate the pedagogical benefits, not only in terms of raising language proficiency but also in self-directed learning. Fourth, gamification of the app, including graphics and motivations for the user, would make the tool even more attractive for CFL learners.

8 Appendix

8.1 Publications

We published three papers describing the research outcomes of this project:

- Lee, J., Lam, C. Y., and Jiang, S. (2016). A Reading Environment for Learners of Chinese as a Foreign Language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*.
- Lee, J. and Yeung, C. Y. (2018). Automatic Prediction of Vocabulary Knowledge for Learners of Chinese as a Foreign Language. In *Proceedings of the International Conference on Natural Language and Speech Processing* (ICNLSP).
- Yeung, C. Y. and Lee, J. (2018). A Personalized Text Retrieval System for Learners of Chinese as a Foreign Language. In *Proceedings of the 27th International Conference on Computational Linguistics* (COLING).

8.2 Access to app

The app has been made available to the public at the Google Play app store (http://play.google.com). It can be accessed by searching for "CityU Chinese Reader" in the app store (Figure 8), or by directly entering the following URL on the browser: https://play.google.com/store/apps/details?id=com.cityu.lt.chinesereader. At the app store, click on the "Install" button to download and run the app on any Android device.

8.3 Dissemination and Promotion

For the dissemination and promotion of the app, we have pursued the following avenues:

Web presence We constructed a web page for the app (http://www2.lt.cityu.edu.hk/~jsylee/app.html), with a brief description of its functions and a link to the download site at Google Play.

Demos We gave five demonstrations of the app:

- At the 26th International Conference on Computational Linguistics (COLING), in December 2016;
- At the Quality Education Fund workshop at the Department of Linguistics and Translation, City University of Hong Kong, in February 2017;
- At the Chinese language course taught by Dr. Hui Wu at the Department of Chinese and History, City University of Hong Kong, in March 2017;
- At the 2nd International Conference on Natural Language and Speech Processing, in April 2018;
- At the 27th International Conference on Computational Linguistics (COLING), in August 2018;

User study Eleven learners of Chinese — from CityU and beyond — responded to our promotion efforts through the web page and demos, and completed a user study (reported in Section 5.2.1) and a survey (reported in Section 5.2.2) on the app.

Outreach (on-going) We invited teachers at the following schools and organizations to use our app, and we plan to convene meetings for those interested in incorporating it as a pedagogical tool for Chinese language teaching:

- 地利亞修女紀念學校（百老匯）
- 天主教聖安德肋小學
- 香港中文大學校友會聯會張煊昌中學
- 樂善堂梁銶琚學校（分校）
- 油蔴地天主教小學
- 路德會梁鉅鏐小學
- 啟基學校
- 保良局王賜豪 (田心谷) 小学
- 大埔舊墟公立學校
- 景林天主教小學
- 華富邨寶血小學
- 香港教育工作者聯會黃楚標學校
- 上水宣道小學
- 鐘聲慈善社胡陳金枝中學
- 伊利沙伯中學舊生會中學
- 樂善堂李賢義裔群社

References

Beinborn, L., Zesch, T., and Gurevych, I. (2012). Towards Fine-grained Readability Measures for Self-directed Language Learning. In *Proc. SLTC Workshop on NLP for CALL*.

Brown, J. and Eskenazi, M. (2004). Retrieval of authentic documents for reader-specific lexical practice. In *Proc. InSTIL/ICALL Symposium*, Venice, Italy.

Brown, J. C., Frishkoff, G. A., and Eskenazi, M. (2005). Automatic Question Generation for Vocabulary Assessment. In *Proc. HLT-EMNLP*.

Brusilovsky, P. (2012). Adaptive hypermedia for education and training. In *Adaptive technologies for training and education*, pages 46–66. Cambridge University Press, New York, NY, USA.

Brusilovsky, P. and Peylo, C. (2003). Adaptive and intelligent web-based educational systems. *International Journal of Artificial Intelligence in Education*, 13:156–169.

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

Chinkina, M. and Meurers, D. (2016). Linguistically Aware Information Retrieval: Providing Input Enrichment for Second Language Learners. In *Proc. 11th Workshop on Innovative Use of NLP for Building Educational Applications*.

Colleen Lennon and Hal Burdick (2014). *The Lexile Framework as an Approach for Reading Measurement and Success*. MetaMetrics, Durham, NC.

Collins-Thompson, K. and Callan, J. (2004). A language-modelling approach to predicting reading difficulty. In *Proc NAACL-HLT*, Boston, MA.

Collins-Thompson, K. and Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13).

Coniam, D. (1999). Second Language Proficiency and Word Frequency in English. *Asian Journal of English Language Teaching*, 9:59–74.

Ehara, Y., Miyao, Y., Oiwa, H., Sato, I., and Nakagawa, H. (2014a). Formalizing word sampling for vocabulary prediction as graph-based active learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1374–1384.

Ehara, Y., Miyao, Y., Oiwa, H., Sato, I., and Nakagawa, H. (2014b). Formalizing Word Sampling for Vocabulary Prediction as Graph-based Active Learning. In *Proc. EMNLP*.

Ehara, Y., Sato, I., Oiwa, H., and Nakagawa, H. (2012). Mining words in the minds of second language learners: learner-specific word difficulty. In *Proc. COLING*.

Ehara, Y., Shimizu, N., Ninomiya, T., and Nakagawa, H. (2010). Personalized reading support for second-language web documents by collective intelligence. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 51–60. ACM.

François, T. and Fairon, C. (2012). An "AI Readability" Formula for French as a Foreign Language. In *Proc. EMNLP-CONLL*.

Gu, Q. and Han, J. (2012). Towards Active Learning on Graphs: An Error Bound Minimization Approach. In *Proc. IEEE 12th International Conference on Data Mining (ICDM)*.

Hanban (2014). *International Curriculum for Chinese Language and Education*. Beijing Language and Culture University Press, Beijing, China.

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

Heilman, M., Juffs, A., and Eskenazi, M. (2007). Choosing reading passages for vocabulary learning by topic to increase intrinsic motivation. In *Proc. International Conference on Artificial Intelligence in Education*.

Hsu, C.-K., Hwang, G.-J., and Chang, C.-K. (2013). A personalized recommendation-based mobile learning approach to improving the reading performance of EFL students. *Computers and Education*, 63:327–336.

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

Kajiwara, T. and Komachi, M. (2018). Complex word identification based on frequency in a learner corpus. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 195–199.

Knutov, E., Bra, P. D., and Pechenizkiy, M. (2009). AH 12 years later: A comprehensive survey of adaptive hypermedia methods and techniques. *New Review of Hypermedia and Multimedia*, 15:5–38.

Krashen, S. (2005). Free voluntary reading: New research, applications, and controversies. In Poedіosoedarmo, G., editor, *Innovative approaches to reading and writing instruction, Anthology Series 46*, pages 1–9, Singapore. SEAMEO Regional Language Centre.

Krashen, S. D. (1981). The fundamental pedagogical principle in second language teaching. *Studia Linguistica*, 35(1-2):50–70.

Laufer, B. and Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language testing*, 16(1):33–51.

Lee, J. and Park, O.-C. (2008). Adaptive instructional systems. In Spector, J., Merrill, M., van Merrienboer, J., and Driscoll, M., editors, *Handbook of research on educational communications and technology*, pages 469–484. Taylor and Francis Group, New York, NY, 3rd. edition.

Lee, J. and Yeung, C. Y. (2018a). Automatic prediction of vocabulary knowledge for learners of chinese as a foreign language. In *Proceedings of the 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*.

Lee, J. and Yeung, C. Y. (2018b). Automatic Prediction of Vocabulary Knowledge for Learners of Chinese as a Foreign Language. In *Proc. International Conference on Natural Language and Speech Processing (ICNLSP)*.

Lee, J. and Yeung, C. Y. (2018c). Personalizing lexical simplification. In *Proc. 27th International Conference on Computational Linguistics*.

Levy, R. and Manning, C. (2003). Is it harder to parse chinese, or the chinese treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 439–446. Association for Computational Linguistics.

Liang, S. and Song, J. (2009). Construction of an Approach for Counting Chinese Graded Words and Characters — A Tool fo Assessing Difficulty Level of Word in Chinese Language Teaching Materials Writing System. *Modern Education Technology*, 19(7):86–89.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proc. ACL System Demonstrations*, pages 55–60.

Miltsakaki, E. and Troutt, A. (2008). Real time web text classification and analysis of reading difficulty. In *Proc. Third Workshop on Innovative Use of NLP for Building Educational Applications*.

Oxman, S. and Wong, W. (2014). *White paper: adaptive learning systems*. Integrated Education Solution, Downers Grove, IL.

Paetzold, G. and Specia, L. (2016). Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: a unified framework for predicting text quality. In *Proc. EMNLP*.

Schmitt, N., Jiang, X., and Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *The Modern Language Journal*, 95(i):26–43.

Sharoff, S. (2006). Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.

Shute, V. and Zapata-Rivera, D. (2012). Adaptive educational systems. In Durlach, P. and Lesgold, A., editors, *Adaptive technologies for training and education*, pages 7–27. Cambridge University Press, New York, NY.

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

Tseng, W.-H. (2014). Huayu baqianci ciliang fenji yanjiu (Classification on Chinese 8 000 Vocabulary). *Huayu xuekan*, 6:22–33.

Vajjala, S. and Meurers, D. (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *Proc. 7th Workshop on Innovative Use of NLP for Building Educational Applications*.

Vandewaetere, M., Desmet, P., and Clarebout, G. (2011). The contribution of learner characteristics in the development of computer-based adaptive learning environments. *Computers in Human Behaviour*, 27:118–130.

Wang, M., Malmasi, S., and Huang, M. (2015). The jinan chinese learner corpus. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 118–123.

Wu, T.-T. (2014). English reading e-book system integrated with guidance mechanism. In *IEEE 14th International Conference on Advanced Learning Technologies*, pages 171–175, Los Alamitos, CA. IEEE Computer Society.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G. H., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*.

Yimam, S. M., Štajner, S., Riedl, M., and Biemann, C. (2017). CWIG3G2-complex word identification task across three text genres and two user groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 401–407.

Zeng, Q., Kim, E., Crowell, J., and Tse, T. (2005). A text corpora-based estimation of the familiarity of health terminology. In *International Symposium on Biological and Medical Data Analysis*, pages 184–192. Springer.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. (2004). Learning with local and global consistency. In *Advances in neural information processing systems*, pages 321–328.