Finding Something to Read:
Intelligibility, Readability and Learner Chinese Texts

James C. Loach

INPAC and Dept. of Physics, Shanghai Jiao Tong University, Shanghai 200240, China
james.loach@gmail.com, +86 131-6279-0215

Abstract

This article presents a new way of thinking about the readability of texts intended for non-native (L2) language learners. Readability is conceived of as the degree to which a learner experiences a text extensively, in the sense of being able to read it fluently and enjoyably, and a new framework is developed for estimating and communicating it. Readability is expressed as a reading level-specific measure of the likelihood that a text will be experienced extensively, and is presented as a distribution across reading levels. This contrasts with the traditional model in which texts are assigned single, most-probable estimates of their reading level. The new framework is applied to the problem of measuring the readability of Chinese L2 texts, where it forms the basis for a practical, easily-implemented technique. In addition to its novel conceptual basis, the technique is unique in how it deals with word segmentation and how it accounts for incidental word learning in extended texts. The technique is applied to a variety of different L2 texts and its performance is evaluated. It is used to investigate the variation of readability within extended texts.

1. Introduction

Learning to read is a central activity in most second language (L2) learning programs, and for reasons that extend beyond the importance of literacy as a goal. Reading is also a learning technique, and can be immensely beneficial as a way to process large quantities of comprehensible input (Krashen, 1989). This can help learners to internalize words, in all their shades of meaning and subtle patterns of use, and help them learn new words incidentally (Nagy et al., 1985; Nation, 2015; Restrepo Ramos, 2015). It can help learners internalize grammar structures and promote a native speaker-like sense for what sounds right and wrong (Elley, 1991).

However, the benefits of reading are as well-established as the difficulties in realizing them (Coady, 1997). The problem, which can be acute at lower levels, is known as the *beginner's paradox* - a state in which the learner does not know enough of a language to experience the benefits of reading. This represents a vicious cycle that can only be broken when the student's reading level reaches the point where they can handle extended texts (Nuttall, 1982). The cycle can then become a virtuous one, in which reading begets and reinforces language knowledge, which allows ever more fluent and enjoyable reading.

The logical way to help students break the vicious cycle is to bring down the language level required to read extended texts. And this is indeed what has been done, through simplified texts aimed at learners. These texts often take the form of *graded readers*, which are series of books, graded by reading level, through which learners progress as their reading level improves. There is debate over the educational trade-offs involved in substituting artificial texts for authentic ones (Wallace, 1992), but there is compelling evidence for the benefits of reading these texts, especially when they are well-matched to the level of the reader (Day & Bamford, 1998).

The importance of correctly matching the reading level arises because the closeness of the match strongly influences how the reader interacts with the text. In particular, it affects the *reading intensity*, the extent to which a reader experiences a text as *intensive* or *extensive* (Krashen, 1989; Hu & Nation, 2000). Intensive reading, at one extreme of the intensity scale, is reading that is slow, requires frequent dictionary use and provides little pleasure to the reader. At the opposite end is extensive reading, where the language is easy enough for the content to be enjoyed; the vast majority of words are known and reading proceeds at a satisfying pace (Nagy et al., 1987; Grabe & Stoller, 2011; Nation, 2015). Extensive reading is the only reasonable way to consume extended texts, and graded readers are designed to be read in this way. Matching a text to a reader means ensuring that they can read it extensively. This task is non-trivial,

however, because of the highly non-linear relationship between text complexity and reading intensity. As will be discussed below, a scattering of unknown words in a previously extensive text can make it much more difficult to read.

The traditional way in which the matching occurs follows a familiar pattern in language learning, one in which the learning materials are divided into levels. At a particular time, a student will self-identify with a particular level, as a proxy for their ability, and use it to guide their selection of materials. Thus the matching problem becomes one of making accurate and consistent level assignments to texts and communicating them effectively.

Reading levels are assigned using human judgment, often including feedback from students, or computational methods. The latter encompass a wide variety of techniques, from antiquated formulae, such as the Flesch–Kincaid readability test (Kincaid et al., 1975), to modern machine learning techniques, in which learning algorithms are trained using samples of level-graded text (Collins-Thompson, 2014).

In this work, we reformulate the process of matching texts with readers in a way that aims to make it easier for students to find suitable texts for extensive reading. We do this by focusing explicitly on the reader's experience of a text, and on providing them with rich information on which to base their judgment. Instead of asking the question 'what is the level of this book?', we ask 'given the reader is at level X, how readable are they likely to find this book?' We answer the question with a notion of readability as a measure of reading intensity, which is estimated on a level-by-level basis and presented as a distribution across levels. Readability is estimated in a principled way, using a non-linear relationship between text complexity and reading intensity, which we infer from experimental data.

We demonstrate our ideas by using them to develop a new computational technique for estimating the readability of Chinese L2 texts. The principles governing readability are similar across all languages (DeFrancis, 1989), but must be accounted for differently according to the particular features of the language under study. Chinese presents especially interesting challenges due to various features of its unusually complex script. These features prevent techniques developed for alphabetic scripts from being trivially adapted, and there has been little dedicated research into the readability of Chinese.

Our technique is principally characterized by its novel conceptual framework and by its practicality. It is designed to be easy to implement and easy to understand, and therefore to be of practical use to students, teachers, and developers of Chinese L2 reading materials. The technique

is also novel in how it handles characteristic features of the Chinese language—such as the absence of marked word boundaries, the absence of inflections, and the strong meaning associations of Chinese characters—and in how it accounts for incidental learning of new vocabulary.

2. Readability

We identify the readability of a text with the subjective experience of the reader, as a measure of perceived reading intensity—the extent to which the reader enjoys the text extensively. This conception follows the spirit of Dale & Chall (1949), who saw readability as 'the sum total… of all those elements within a given piece of printed material that affects the success that a group of readers have with it… the extent to which they understand it, read it at an optimum speed, and find it interesting.'

There is a well-developed literature on methods for estimating text readability, and there are two broad approaches to the problem. The first is to measure it by studying the responses of a target audience, and the second is to predict it from measurable characteristics of the text. We refer to the first as *usability studies* and the second as *computational techniques*. In both approaches, determining the readability means choosing between labels that have between assigned to stationary, external reference points: groups of learners in the usability studies and reference texts in the computational techniques.

Usability studies can provide clear results when they are well-designed and use carefully chosen groups (Redish, 2000). In particular, they provide the only way to directly assess subjective experiences. However, they can also be expensive and time-consuming, and because of this, are relatively little used.

Computational techniques seek to replicate the results of usability studies without the expense. They are techniques for mapping text features onto reading levels, and differ from one another in their choice of algorithm and how they are calibrated. They have a long history, and with recent advances in machine learning, are becoming increasingly sophisticated and effective.

There is also a related literature seeking to understand the nature of extensive reading and its benefits for learners. One of the principal concerns is identifying the conditions necessary for extensive reading, and in particular, how they relate to measurable properties of the text. In these studies, readability is most often identified with comprehension, as an easily-measured proxy, and the main text feature that is studied is lexical coverage, the fraction of words in a text that are known to a reader (Adolphs & Schmitt, 2003). The choice of lexical coverage reflects the

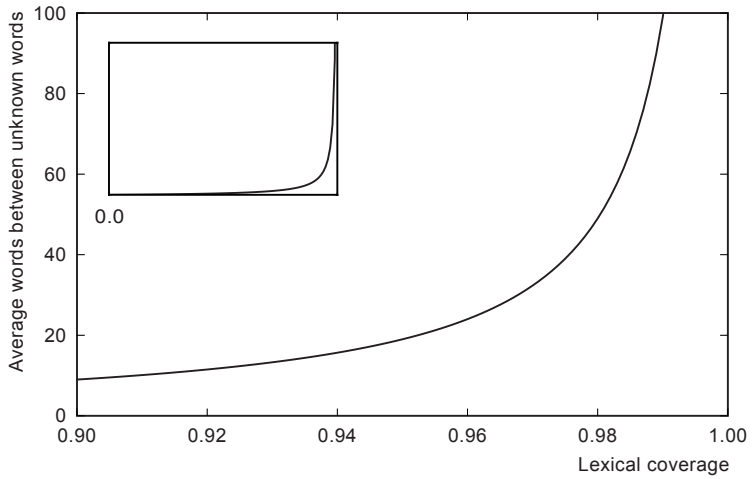

Figure 1 – Average words between unknown words in a text as a function of the lexical coverage. The insert shows the same curve over the full scale of lexical coverage.

overwhelming importance of vocabulary in reading comprehension, especially at lower language levels (Alderson, 2000; Read, 2000; Shen, 2005).

In this section, we reexamine experimental results from the literature on extensive reading and use them to suggest a way in which text features can be systematically related to readability as a measure of reading intensity. In so doing, we develop a new framework for the computational assessment of readability.

Research on how to identify texts suitable for extensive reading has focused on the relationship between reading comprehension and lexical coverage. The earliest experimental study by Laufer (1989) estimated that a lexical coverage of $l = 0.95$ is necessary for satisfactory comprehension. This number was based on a study of the relationship between lexical coverage and performance in a standardized comprehension test.

Hirsh and Nation (1992) discussed the reasons why readability might be associated with a high lexical coverage. They used a toy model to show how the average separation between unknown words, and therefore the average time between interruptions to the reader's flow, increases rapidly above $l \sim 0.98$. We illustrate their point in Figure 1, which shows the mean separation between words S as a function of lexical coverage. The shape of the curve is characteristic of a geometric progression, and $S(l)$ can written in that form explicitly

$$S = \frac{l}{1-l} = \sum_{k=0}^{\infty} l^k.$$

These arguments parallel those of West (1926), who, when considering the frequency of text interruptions that would interfere with children's enjoyment of a text, speculated that 'new words should not be introduced more frequently than one new in every fifty running words of the text' and identified $l < 0.90$ as an unacceptable impediment to fluent reading.

Hu and Nation (2000) studied the relationship between comprehension and lexical coverage at $l = 0.80, 0.90, 0.95$ and 1.00. They found a linear relationship, though there were wide variations in comprehension at each coverage. In particular, they found that a minority of subjects were able to achieve adequate comprehension at $l = 0.90$ and 0.95. Nevertheless, the authors recommend a coverage of $l = 0.98$.

Schmitt et al. (2011) performed a similar, though more sensitive study in the region $0.90 \leq l \leq 1.00$. They found a relationship that was linear and gentle (comprehension increased from 0.50 to 0.75 over the range under study). In their conclusions, they favored a coverage of $l \sim 0.98$ but described how the requirement might depend on the reading context.

In a study of L1 reading, Carver (1994) concluded that a lexical coverage of $l = 0.99$ was characteristic of a well-matched text, and $l = 0.98$ of a text that was too difficult.

These results do not represent a coherent picture. In particular, it is not obvious how to reconcile the linearity of the lexical coverage-comprehension curves with Hirsh and Nation's model, and with the apparent instinct among experimentalists that thresholds exist and are meaningful to quote. However, the contradictions only exist if readability is narrowly identified with comprehension. A richer notion of readability can easily allow for thresholds in the reader's experience of a text that *result* from incremental improvements in comprehension, or from other factors such as a reduction in the number of interruptions due to unknown words.

Consider a learner who can comfortably enjoy a graded reader at level X. When presented with a book at level $X - 1$, they might judge it 'pointlessly easy' (readability, $R = 1.0$), whereas if the book were at level $X + 1$, they might consider it 'far too difficult' ($R = 0.0$). However, if the student was asked to read the two books and take a comprehension test we might discover that their level $X - 1$ comprehension fell well short of 100%, perhaps because they didn't appreciate some nuances in word usage or lacked some higher-level reading skills. Likewise, with the more difficult book, they might recover a fair amount of meaning and score a comprehension well above 0%. In this way, a relatively small difference in comprehension can map to a dramatic change in perceived readability. This may be the threshold that experimenters look for when they vary the lexical coverage.

We can intuit what a general relationship between lexical coverage and readability might look like. There is some consensus in the literature that $l > 0.98$ is the threshold for extensive reading and we can identify this with maximum readability: $R(l > 0.98) \sim 1.0$. Below this, there is a transition region to very low readabilities. As $l \rightarrow 0.90$, corresponding to one unknown word in ten, we expect the readability to approach zero: $R(l \rightarrow 0.90) \rightarrow 0.0$. If the transition between these two regions is smooth, the curve is sigmoidal. Sigmoidal curves can be modeled using various analytical functions, and in the absence of more precise data, we choose an error function:

$$R(l) = \frac{1}{2}\mathrm{erf}\left(50 \cdot (l - 0.95)\right)$$

This function has a midpoint $R(l = 0.95) \sim 0.5$, and tails $R(l > 0.98) \sim 1.0$ and $R(l < 0.92) \sim 0.0$.

$R(l)$ can be made more general through a broader interpretation of the independent variable. We propose that lexical coverage be replaced with *lexical intelligibility*, which we define as the average probability that a token will be understood. In the simple case that words are either known or unknown, this is identical to the lexical coverage. However, it can also account for words that don't fit neatly into either category—words that have a finite probability of being guessed correctly. We can write

$$R(I) = \frac{1}{2}\mathrm{erf}\left(50 \cdot (I - 0.95)\right)$$

where I is the lexical intelligibility. This curve is plotted in Figure 2.

We can motivate the choice of an error function by considering how a set of readers with identical language knowledge might respond if we varied the lexical intelligibility of an ensemble of texts. Readability assessments might differ between readers because of variations in levels of interest in the subject matter (affecting their tolerance for interruptions from dictionary use), familiarity with the subject matter (affecting their ability to guess words), and general intelligence (affecting their ability to guess words and also employ higher-level reading strategies). For each reader-text combination i we could, in principle, determine a threshold intelligibility above which the text is more extensive than intensive, the set of I_i for which $R(I_i) = 0.5$. These I_i would form a distribution, and if this distribution were normal, then the probability of any reader-text combination being above the threshold would be given by an error function.

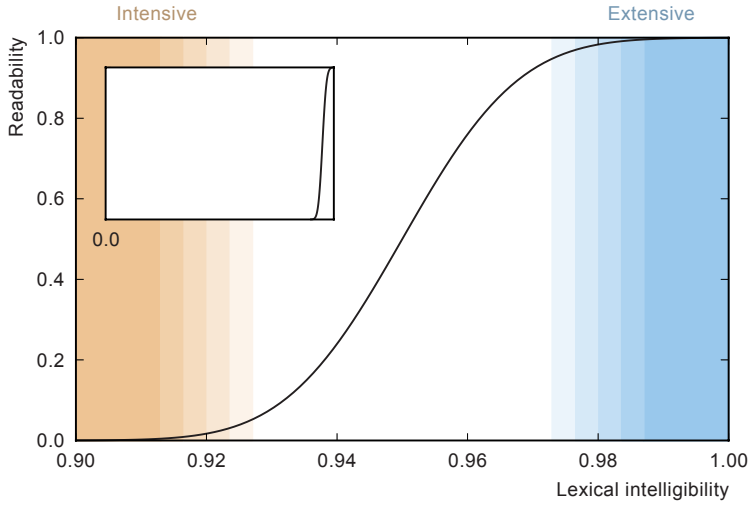

Figure 2– The readability curve representing the non-linear relationship between lexical intelligibility and readability. The insert shows the same curve over the full scale of lexical intelligibility. Lexical intelligibility is a measure of the probability that any given token in the text will be understood. Readability is a measure of reading intensity, the extent to which the text is experienced intensively or extensively.

The readability curve provides a mechanism for factoring the problem of estimating readability into two parts: a relatively simple determination of the lexical intelligibility, followed by a mapping onto readability using a measured readability function. This process is, of course, equivalent to any usability study or computational technique where the reference sets are properly calibrated in readability. But factoring the problem in this way permits a new style of computational technique that is conceptually and methodologically simple without being reductionist. The readability curve provides a straightforward and convenient way to parametrize much of the problem's complexity.

We now consider the question of how to present the output of this procedure to potential readers. This is not trivial, because both lexical intelligibility and the readability that follows from it, are reading-level specific parameters. As such, there is no single number that can be used to label a text. The solution is to express the readability as a function of level.

The readability, as we have defined it, is the answer to the question 'if I have mastered language level *X* then how readable will I find this book?' If potential readers span a range of levels, then the readability must also be quoted for different levels, and the most elegant way to do this is graphically, by plotting the readability as a function of level.

An example readability graphic is shown in Figure 3. The text it represents would constitute intensive reading for a student who had only mastered levels 3 or 4, would be a medium-difficulty text for someone

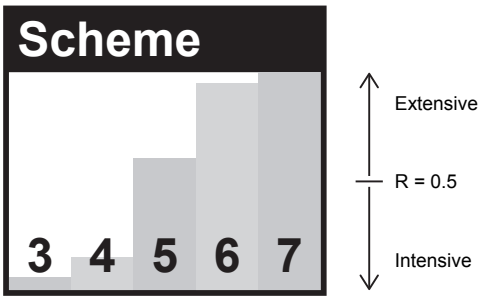

Figure 3 - A graphical way to communicate readability as a distribution across levels. 'Scheme' is the name of the level scheme whose levels '3', '4', '5', '6' and '7' appear along the x-axis. The readabilities at levels that are not shown in the graphic are easy to infer.

who'd mastered level 5, and would be easy extensive reading for those at higher levels. A student could use the graphic to factor in their personality—by learning what range of readabilities they can tolerate—and their intrinsic interest in the book, which will modulate this range.

Our conception of readability, and the way that we estimate it, leads very naturally to this style of presentation. When displayed as a graphic, the richer information content is easy and straightforward to understand, and may provide a more effective way for potential readers to judge the suitability of texts.

3. The readability of Chinese text

3.1. Reading Chinese

In most languages, the principal challenge in learning to read is mastering sufficient vocabulary (Nation, 2006). Chinese shares this difficulty, but adds another, and one that makes the task much more difficult than for an alphabetic language like English—the immense complexity of its script. Learning to read Chinese means learning several thousand distinct symbols (Kane, 2006). And while the task might not be quite as difficult as it sounds, it remains a formidable one.

The Chinese script interferes with many aspects of normal language learning, and not least by inhibiting incidental learning and extensive reading (Marton et al., 2010). The consequences for reading are so acute that even upper-intermediate learners, with useful levels of conversational Chinese, can struggle to find meaningful things to read. Basic L1 texts often use characters that are only known by advanced L2 learners, and the market for dedicated learner-oriented texts is extremely underdeveloped.

A practical technique for estimating the readability of Chinese L2 texts could help mitigate some of these problems (Zhang, 2015). In particular, it could help students and teachers navigate existing L2 materials, help publishers create new materials, and help developers organize and rank the reading materials they provide in their apps and on their websites.

3.2. Previous work

There have been a number of previous studies on the readability of Chinese. These have generally adapted techniques developed for English to account for the characteristic features of the Chinese language (which are discussed in more detail below).

The earliest of these studies followed the spirit of traditional English readability formulae. Yang (1971) created readability formulae using large numbers of independent variables, some of which were conventional, such as the presence of words in wordlists, and some of which were Chinese-specific, such as the average number of strokes in the characters. His work used L1 reading materials.

More recent approaches have used machine learning techniques (Jeng, 2001; Lau, 2006; Chen et al., 2013; Jiang et al., 2015). These works adapt well-established classification techniques to the Chinese language and produce promising results for L1 texts. Any of them might be adapted to L2 texts by using suitable training materials, and ideally, they would have provided useful reference points for assessing the performance of our technique. However, they are challenging to implement computationally and none are available as pre-packaged tools.

Studies on L2 texts have been limited to measurements of the correlations between text features and grade levels (Shen, 2005; Da, 2009). None of these analyses have matured into practical tools.

3.3. The algorithm

The aim of our algorithm is to assess the readability of Chinese text within the paradigm described in Section 2. If we assume that the readability curve in Figure 2 is a reasonable approximation for Chinese, our task becomes one of estimating the lexical intelligibility of Chinese text.

The algorithm needs to satisfy a number of requirements. In particular, it needs to deal with the *segmentation problem*, which arises because word boundaries are not marked in Chinese and because the concept of the word is not unambiguously well-defined (Packard, 2000). The algorithm should also account for the way in which Chinese characters often carry strong semantic associations that can help a reader guess the meanings of multiple-character words in which they appear. And, finally, we would also like the algorithm to account for words that are learned as the reader

progresses through the text. This is especially important for accurately assessing the readability of extended texts, where there may be a finite cast of names, places, etc., that the reader wouldn't know in advance, but which appear so frequently that they can be easily learned.

The design of our algorithm is based on the experimental finding that vocabulary knowledge is the dominant factor in determining readability at lower reading levels. As such, it is a vocabulary-based algorithm in which lexical intelligibility is estimated by comparing the characters and words in a text to those in reference word lists.

Our decision to base the method on word lists, rather than samples of the texts at different levels, was motivated by the difficulties we encountered when trying to use segmentation algorithms with limited quantities of reference material. An example is provided by the way in which the state-of-the-art NLPIR/ICTCLAS algorithm (Zhou & Zhang, 2003) segments sequences of characters differently depending on their grammatical contexts. Consider 下雨 (to rain), which is composed of the characters for 下 (to fall) and 雨 (rain). NLPIR/ICTCLAS will segment 下雨 as a word in 快要下雨了 ('it's going to rain soon') but segment 下 and 雨 separately in 像下雨的声音 ('like the sound of falling rain'). This may be logical, but it means that if the readability algorithm sees only the first context of 雨 in the reference texts but only the second context in the text being analyzed, then it might regard 雨 as being unintelligible, something which would clearly not be true. Such problems can be mitigated with large, varied and strictly-graded L2 reference texts, but adequate quantities of such texts do not exist in Chinese.

A further, practical issue with advanced segmentation algorithms is that they usually involve significant computational overhead. This acts contrary to our goal of creating a practical technique that can be widely implemented. An overview of Chinese segmentation algorithms is given by Huang and Xue (2012).

Our approach to segmentation is to avoid doing it explicitly. Rather, we search for strings of characters in the text that match words in a word list. This is not adequate as a general segmentation algorithm because there is often ambiguity as to which word a particular character belongs. However, this ambiguity is not important here because we are only interested in whether or not a match exists. It can be a problem when parts of unknown words are incorrectly identified as known, but we provide a tuning mechanism to account for this statistically.

We implement this approach in a procedure that assigns characters in the text to different classes depending on the probability that a reader at a given level will understand the words in which they appear. Each character is assigned a lexical intelligibility that depends on the class in

which it appears, and for characters being learned, also on the pattern of previous occurrences. These lexical intelligibilities are averaged to give the overall lexical intelligibility for the text, which can then be mapped onto a readability.

The classes, in order of decreasing priority, are:

- *Known* characters appear in words that are known to the reader. Words are considered known if they appear in the word list corresponding to the reader's level. These characters are assigned lexical intelligibilities of $I = 1.0$.
- *Guessable* characters are known to the reader from their presence in different words to those in which they appear in the text. They may help the reader guess the meaning of the words in which they do appear. They are assigned lexical intelligibilities of $I = \alpha$, where α is a tunable parameter.
- *Learnable* characters do not appear in words that are known to the reader. These characters have lexical intelligibilities that vary throughout the text, as successive appearances make it more likely that the reader will correctly identify them. The probability that a character will be correctly identified is identical with the lexical intelligibility, and for each character, is calculated on an occurrence-by-occurrence basis using a learning model that is inferred from experimental data.
- *Learned* characters are characters that have appeared so many times that they can be considered *known*. These are assigned lexical intelligibilities of $I = 1.0$.

Characters are assigned to the highest-priority class that applies. Punctuation marks and non-Chinese characters are retained when assigning characters to classes because they act to mark word boundaries and so make the procedure more accurate. However, they are not included when calculating lexical intelligibilities.

The class of *guessable* characters is designed to account for the way in which characters can suggest the meanings of words in which they appear, and also for peculiarities in the word lists. For example, if a reader knows the character 美 (beautiful) then it may help them guess or remember the meanings of words such as 美丽 (beautiful), 美术 (fine arts), and 完美 (perfect). And conversely, knowing these three words would help the reader understand the meaning of 美 when it appears alone as a word in phrases such as 她很美 (she is beautiful). 美丽, 美术 and 完美 appear in the word lists used in this study but 美 happens to be absent. This class helps to account for all this complexity statistically, by setting the lexical intelligibilities of the characters it contains to a tunable parameter α. For a given word list, α can be determined by using a training text with a known readability. If the training text is known to be

extensive, then the lexical intelligibility can be set to 0.98 (corresponding to $R \sim 1.0$) and the corresponding α can be found using an optimization procedure. If there are multiple reading levels, then the α's found for each level can be averaged to give a single overall α.

In the class of *learnable* characters, the lexical intelligibility (recall probability) of a particular occurrence of a particular character is calculated using a model based on experimental data from Teng (2015). Teng measured recall probabilities for nonsense words that had been introduced into an otherwise well-understood text based on how frequently they occurred. This was done separately for nouns, verbs and adjectives and in bins of numbers of occurrences. We averaged across the different word types and took the mid-point of each bin to give the data points shown in Figure 4. We used these points to derive a simple learning model, by fitting them to a linear slope with intercept constrained to the origin and not allowing the lexical intelligibility to exceed 1.0. We can write the model as

$$I(n) = \min(0.041\mathrm{n}, 1.0)$$

when n is the number of previous occurrences (see Figure 4). Once I reaches 1.0 the character is reclassified as *learned*. This learning model is clearly simplistic and could be improved by combining Teng's results with a more realistic model of human memory (for example, by integrating the work of Pavlik (2007)).

One consequence of modeling the learning of characters is that it allows a text of any difficulty to achieve a readability close to 1.0, provided that it is long enough. This behavior is not nonsensical per se—readability is an average measure and any text built from a finite subset of the language will naturally get easier as you work through it—but it is unrealistic when the rate of learning is allowed to be arbitrarily large. We model the finite mental capacity of the reader using a simple constraint: we do not allow characters to contribute to the overall lexical intelligibility if there are more than β unlearned characters in the preceding N_β characters. Throughout this work we use $\beta = 1$ and $N_\beta = 10$. A more sophisticated learning model might account for this behavior more naturally.

Figure 5 shows a passage of text marked-up according to the class of each character.

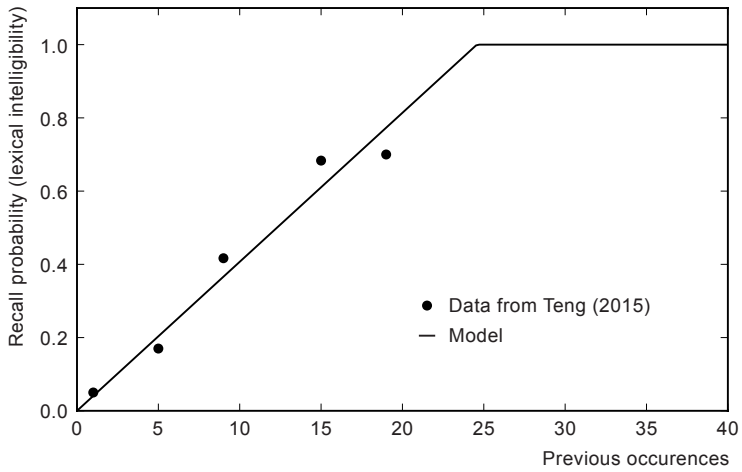

Figure 4 – Recall probability (equivalent to lexical intelligibility) as a function of how many times a character has previously occurred in the text.

(a) HSK 3

对一直住在地下的鼹鼠来说，这一天只是个开头。夏天慢慢地到了，
日子一天比一天长，也一天比一天有意思。他学会了游泳、划船，
爱上了水里的生活，还有风吹过芦苇丛发出的声音。

(b) HSK 5

对一直住在地下的鼹鼠来说，这一天只是个开头。夏天慢慢地到了，
日子一天比一天长，也一天比一天有意思。他学会了游泳、划船，
爱上了水里的生活，还有风吹过芦苇丛发出的声音。

Known Guessable Words
Learned Learning

Figure 5 – Example passage showing how characters are assigned to different classes. The text is an extract from The Wind in the Willows (Grahame, 1944) translated by X. Xu (personal communication, 2016). The text is part of a longer text that was analyzed at two reading levels: (a) HSK 3 and (b) HSK 5. 'Words' are multiple-character words identified in the relevant word list.

3.4. Performance

We used word lists corresponding to the six levels of the Hanyu Shuiping Kaoshi (HSK) Proficiency Test (汉语水平考试), a set of examinations administered by the Chinese National Office for Teaching Chinese as a Foreign Language (NOTCFL) ("Chinese National Office", 2016). To tune our algorithm and evaluate its performance we used text extracted from the HSK Standard Course textbook series (Liping, 2014a, 2014b, 2014c, 2014d, 2014e, 2015a, 2015b, 2015c). We omitted the HSK 1 level book because of the limited amount of Chinese text it contained, and the second HSK 6 level book, which had not been published. We omitted text that was not intended to match the level of the book, such as explanations of grammar points and cultural notes. The content of these books closely corresponds with the HSK levels, though proper nouns and other words are occasionally used, and with increasing frequency at higher levels.

We divided the texts into training data, containing the set texts from the beginning of each chapter, and testing data, containing the questions and exercises. There were a total of 220935 characters (including punctuation) in the testing data and 361201 in the testing data. Using the training data to tune the model, we found $\alpha = 0.900$ and the readabilities shown in Figure 6(a). Using this α, we calculated readabilities for the testing set shown in Figure 6(b). Tuning on the training set (not shown) yielded an almost identical α of 0.906.

The results in Figure 6 show that a single α gives consistent results across all HSK levels and between the training and testing sets. For their target levels, all but two books are predicted to be perfectly extensive. In the training set, the HSK level 5 book is slightly less than extensive (due to a particularly large number of unknown words in the set texts, but not in the examples and exercises). In the testing set, the HSK level 2 book is made slightly less extensive by the relatively high proportion of proper nouns (mostly people's names). Note that while the HSK 6 level book has the same HSK 6 readability in both the training and testing sets, the sets have very different HSK 5 readabilities. It is particular feature of this study's way of estimating and communicating readabilities that such a difference is readily apparent.

Table 1 shows a breakdown of how the readability of each text in the training set is calculated at the text's target readability. Note that guessable and learnable/learned characters make up relatively small fractions of the texts.

Figure 6 - Readabilities by HSK level for five volumes of HSK Standard Course textbook series (Liping, 2014a, 2014b, 2014c, 2014d, 2014e, 2015a, 2015b, 2015c). The training set for each book contains the set texts at the start of each chapter; the testing set contain the remaining text in each chapter.

Table 1 – Detailed analysis of the HSK Standard Course books at their target HSK levels. I and R are the lexical intelligibility and readability, respectively. f_{char} gives the fractions of characters in each class and I_{av} gives their average intelligibility. The 'learnable/ed' class is a combination of the learnable and learned classes.

Level	I	R	Known		Guessable		Learnable/ed	
			f_{char}	I_{av}	f_{char}	I_{av}	f_{char}	I_{av}
2	0.974	0.963	0.925	1.00	0.053	0.90	0.022	0.11
3	0.985	0.993	0.888	1.00	0.094	0.90	0.017	0.67
4	0.988	0.996	0.907	1.00	0.085	0.90	0.008	0.54
5	0.962	0.798	0.812	1.00	0.163	0.90	0.025	0.11
6	0.980	0.984	0.850	1.00	0.145	0.90	0.005	0.04

We applied the algorithm to two series of graded readers, by Mandarin Companion (Dickens, 2015a, 2015b; Irving, 2015; Jacobs, 2015; Verne, 2016; Wells, 2015) and Sinolingua (Shi, 2013a, 2013b, 2014a, 2014b, 2016a, 2016b). The Mandarin Companion books are divided in two levels that are differentiated by the numbers of different characters they use (around 300 characters for level 1, and 450 for level 2). The Sinolingua books are aimed at more advanced readers and are labelled by the number of different words they contain. The results are shown in Figure 7.

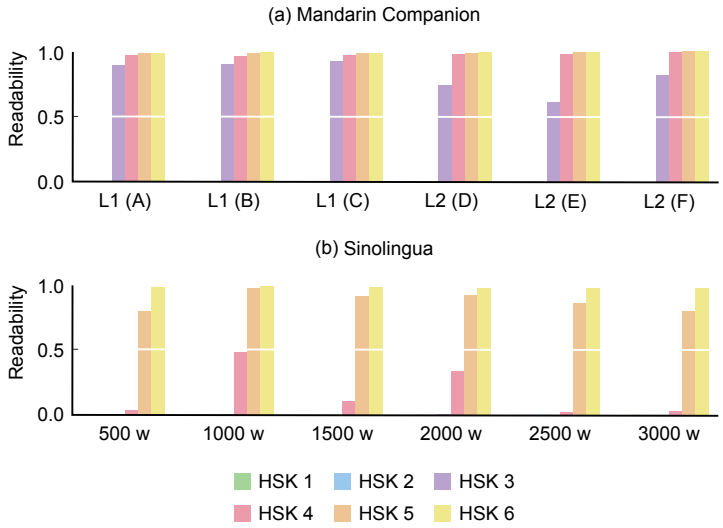

Figure 7 – Readabilities by HSK level for two series of graded readers. (a) shows six books from the Mandarin Companion series. The level 1 (L1) books are A: (Jacobs, 2015), B: (Irving, 2015), and C: (Blind, 2015); and the level 2 books are D: (Dickens, 2015a), E: (Dickens, 2015b) and F: (Verne, 2016). (b) shows the six books of the Sinolingua series, where w stands for words, as given in the titles (Shi, 2013a, 2013b, 2014a, 2014b, 2016a, 2016b). For the Sinolingua books, only the first 15000 characters in each book were analyzed.

With the Mandarin Companion books, there is good consistency between the books at each level and the higher-level books are indeed found to be slightly more difficult. The results for the Sinolingua books are more surprising, showing that the lexical difficulty of the books does not increase in the way that would be expected based on their titles. In addition (though not shown), the difficulties of the stories inside particular books are found to vary significantly. Manual inspection of the books accords with the results of the algorithm. In particular, the 1000 Word and 2000 Word books do appear to be simpler and easier to read that the 500 Word book.

We applied our algorithm to texts from the Chinese Reading World website (Shen & Tsai, 2010; "University of Iowa", 2016), which contains a broad range of authentic texts intended for extensive reading. Texts are organized into three difficulty levels (beginning, intermediate and advanced). They vary in length but are generally short. Figure 8 shows readabilities for 300 randomly-selected texts (out of a total of 900). There is a general decrease in readability from beginning to intermediate to advanced, but, within each level, there are significant variations. These variations are readably apparent on manual inspection of the texts.

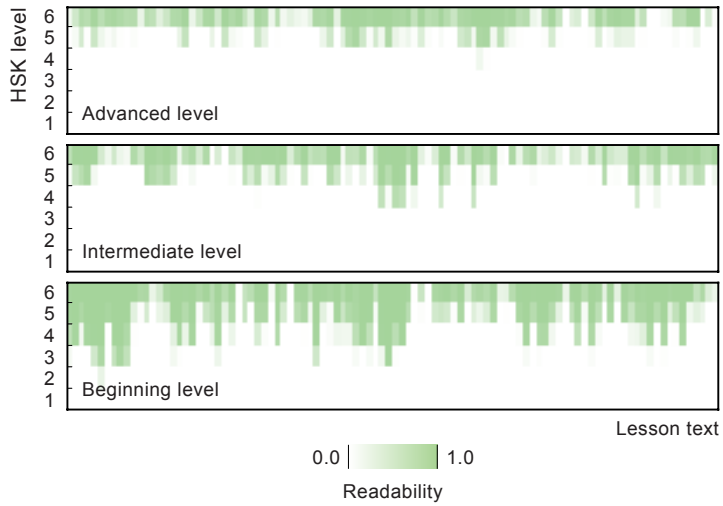

Figure 8 – Readability analysis of randomly selected articles from Chinese Reading World, shown in order of appearance, from left to right (Shen & Tsai, 2010; "University of Iowa", 2016.)

We also used our algorithm to explore how readability varies as a function of the reader's progress through extended texts. This contrasts with the studies above, in which readabilities are measured over complete texts. The top panel of Figure 9 shows the readability as a function of position for the six Mandarin Companion books of Figure 7. These readabilities are calculated up to and including a moving 250-character window centered on the x-axis value. For a single book, the lower panel also shows the contribution of the text in each window. Most of these books become easier as the reader progresses and learns vocabulary that is specific to the particular book. However, the pattern differs between books and they all contains sections of widely varying difficulty.

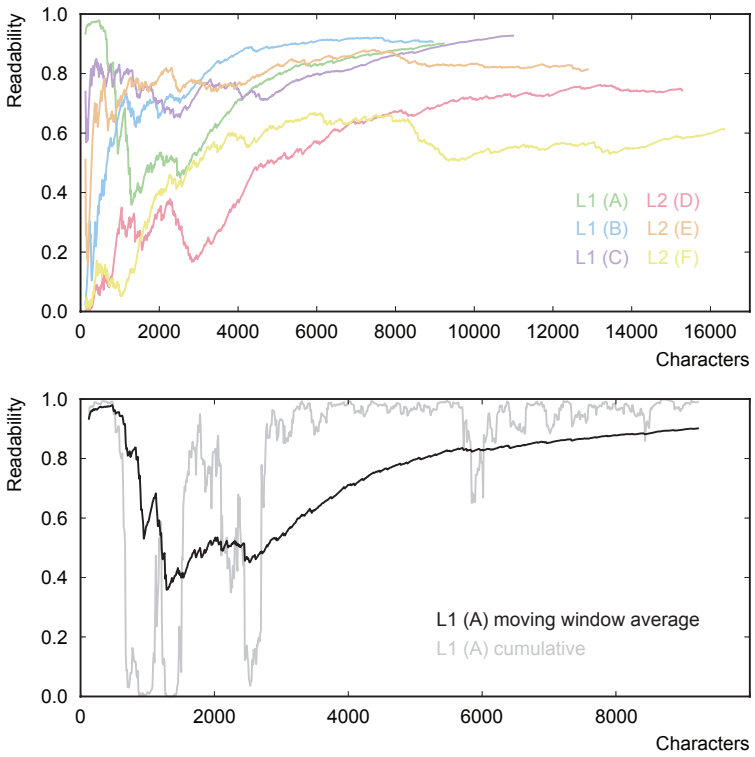

Figure 9 – The top panel shows readabilities as a function of position in the text for six books from the Mandarin Companion series. The level 1 (L1) books are A: (Jacobs, 2015), B: (Irving, 2015), and C: (Blind, 2015); and the level 2 books are D: (Dickens, 2015a), E: (Dickens, 2015b) and F: (Verne, 2016). Readabilities are quoted up to and including a 250 character moving window centered on the x-axis value. The lower panel shows the readability as a function of position for one of the texts (A) along with the average readability inside the moving window.

4. Discussion

In this study, we conceived of readability as a measure of reading intensity and developed a new framework for estimating it. Readability is derived from the lexical intelligibility using the non-linear readability function, and is a reading level-specific measure. We described how the readability of a text can be communicated graphically, as a distribution across reading levels.

The notion of the readability curve is consistent with experimental evidence and *a priori* consideration of the average time between interruptions as a function of lexical coverage. Nevertheless, we are unaware of any studies in which perceived readability is directly correlated with lexical coverage or other textual features. Readability curves may well differ in shape depending on language, style of text and reading context. They may also be affected by the how difficult it is to look up word, which can vary considerably depending on the reading medium, and by the presence of textual glossing. We recommend further study of these potential dependencies.

We used our conception of readability to devise a simple and effective algorithm for estimating the readability of Chinese L2 texts. The algorithm focusses on vocabulary as the key determinant of readability at lower reading levels, and contains a number of novel features. It is the only proven technique for measuring the readability of Chinese L2 texts.

The algorithm deals with the segmentation problem in a way that is computationally straightforward and helps make procedure suitable for widespread use. If required, it can be trivially adapted to work with modern segmentation algorithms.

The algorithm accounts explicitly for the learning of words during reading and can be used to study how readability varies throughout a text. We investigated readability as a function of position for several books and found that it can vary significantly and follow different patterns in different books. We also observed a general trend that accords with experience, that books tend to become easier to read as you progress through them. Our algorithm can be used to help control or remove undesirable variations in readability.

The algorithm was tuned and tested using text from HSK level-graded textbooks and shown to give sensible results when applied to graded readers. The number of quantitative studies was limited by the difficulty in obtaining materials that had had been reliably-graded by HSK level. The algorithm appears to be effective at the highest HSK level, but we might expect it to breakdown at around this point, when most common characters have been learned and higher-level reading skills are

becoming important. The algorithm does not analyze the grammatical structure of the text and will only be useful when the lexical complexity is broadly reflective of the grammatical complexity.

The algorithm can help users navigate existing texts. For example, the stories in the Sinolingua graded readers were found to vary considerably in difficulty, and much more so than the books themselves. The algorithm could be used to order these stories by readability into a proper, gently-sloping extensive reading program. Ordering them in this way would make them accessible to the largest number of students.

We observed similar variations in the electronic texts of Chinese Reading World. Like the Sinolingua books, these are high-quality texts whose usefulness is reduced by their unpredictable reading levels.

The most promising venue for application of the algorithm is within digital reading contexts, where it can be used to adapt the reader's experience based on their feedback. It provides a mechanism that allows a user to tell the software that a particular text is too difficult and, next time, be reliably presented with something easier. It could also be used to scan native texts for passages that can be read at particular levels. For example, one can imagine automated analysis and ranking of news articles, that could identify the set of articles that a student could plausibly read, and rank them in order of readability.

In addition to experimental investigation of the readability curve, future work should include application of this algorithm to other Chinese texts and improvements to the procedure for estimating lexical intelligibility. This could include a more sophisticated model for how words are learned incidentally during reading.

Acknowledgements

This work was supported by the Shanghai Key Lab for Particle Physics and Cosmology (SKLPPC), Grant No. 15DZ2272100.

References

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Carver, R.P. (1994). Percentage of Unknown Vocabulary Words in Text as a Function of the Relative Difficulty of the Text: Implications for Instruction. *Journal of Reading Behavior,* 26(4), 413-437.

Chen, Y. T., Chen, Y. H., & Cheng, Y. C. (2013). Assessing Chinese Readability Using Term Frequency and Lexical Chain. *Computational Linguistics and Chinese Language Processing*, *18*(2), 1-17.

Chinese National Office for Teaching Chinese as a Foreign Language (NOTCFL). Hanyu Shuiping Kaoshi (HSK) Proficiency Test. http://www.chinesetest.cn Accessed: 20.08.16.

Coady, J. (1997). L2 Vocabulary Acquisition Through Extensive Reading. In J. Coady & T. Huckin (eds.), *Second Language Vocabulary Acquisition* (pp. 225-237). Cambridge: Cambridge University Press.

Collins-Thompson, K. (2014). Computational Assessment of Text Readability: A Survey of Current and Future Research. *ITL-International Journal of Applied Linguistics*, *165*(2), 97-135.

Da, J. (2009). A Quantitative Approach to Measuring CFL Text Difficulty. *Paper presented at ACTFL 2009*. CA: San Diego.

Dale, E., & Chall, J. S. (1949). The Concept of Readability. *Elementary English*, *26*(1), 19-26.

Day, R. R., & Bamford, J. (1998). *Extensive Reading in the Second Language Classroom*. Cambridge: Cambridge University Press.

DeFrancis, J. (1989). *Visible Speech: The Diverse Oneness of Writing Systems*. Honolulu: University of Hawaii Press.

Dickens, C. (2015a). *Great Expectations: Part 1*: *Mandarin Companion Graded Readers Level* 1. Shanghai: Mind Spark Press.

Dickens, C. (2015b). *Great Expectations: Part 2*: *Mandarin Companion Graded Readers Level* 1. Shanghai: Mind Spark Press.

Elley, W. B. (1991). Acquiring Literacy in a Second Language: The Effect of Book-Based Programs. *Language learning*, *41*(3), 375-411.

Grabe, W., & Stoller, F. L. (2011). *Teaching and Researching Reading* (2nd ed.) New York, NY: Routledge.

Grahame, K. (1908). *The Wind in the Willows*. London: Methuen.

Hirsh, D., & Nation, P. (1992). What Vocabulary Size is Needed to Read Unsimplified Texts for Pleasure? *Reading in a Foreign Language*, *8*, 689-689.

Hu, M., & Nation, I.S.P. (2000). Vocabulary Density and Reading Comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

Huang, C. R., & Xue, N. (2012). Words Without Boundaries: Computational Approaches to Chinese Word Segmentation. *Language and Linguistics Compass*, *6*(8), 494-505.

Irving W. (2015). *The Sixty Year Dream: Mandarin Companion Graded Readers Level 1*. Shanghai: Mind Spark Press.

Jacobs, W. W. (2015). *The Monkey's Paw: Mandarin Companion Graded Readers Level 1*. Shanghai: Mind Spark Press.

Jeng, C. C. (2001). *Chinese Readability Analysis Using Artificial Neural Networks*. Northern Illinois University.

Jiang, Z., Sun, G., Gu, Q., Yu, L., & Chen, D. (2015). An Extended Graph-Based Label Propagation Method for Readability Assessment. In *Asia-Pacific Web Conference* (pp. 485-496). Springer International Publishing.

Kane, D. (2006). *The Chinese Language: Its History and Current Usage*. Tokyo: Tuttle Publishing.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel* (No. RBR-8-75). Naval Technical Training Command Millington TN Research Branch.

Krashen, S. (1989). We Acquire Vocabulary and Spelling by Reading: Additional Evidence for the Input Hypothesis. *The Modern Language Journal*, *73*(4), 440-464.

Lau, T. P. (2006). *Chinese Readability Analysis and its Applications on the Internet* (Doctoral dissertation, The Chinese University of Hong Kong).

Laufer, B. (1989). What Percentage of Text-Lexis is Essential for Comprehension? *Special language: From humans thinking to thinking machines*, *316323*.

Liping, J. (2014a). HSK Standard Course 1. Beijing: Beijing Language & Culture University Press.

Liping, J. (2014b). HSK Standard Course 2. Beijing: Beijing Language & Culture University Press.

Liping, J. (2014c). HSK Standard Course 3. Beijing: Beijing Language & Culture University Press.

Liping, J. (2014d). HSK Standard Course 4a. Beijing: Beijing Language & Culture University Press.

Liping, J. (2014e). HSK Standard Course 4b. Beijing: Beijing Language & Culture University Press.

Liping, J. (2015a). HSK Standard Course 5a. Beijing: Beijing Language & Culture University Press.

Liping, J. (2015b). HSK Standard Course 5b. Beijing: Beijing Language & Culture University Press.

Liping, J. (2015c). HSK Standard Course 6a. Beijing: Beijing Language & Culture University Press.

Marton F, Tse SK, Cheung WM, editors (2010). *On the Learning of Chinese*. Rotterdam: Sense Publishers.

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning Words from Context. *Reading Research Quarterly*, 233-253.

Nation, I. S. P. (2006). How Large a Vocabulary is Needed for Reading and Listening? *The Canadian Modern Language Review* 63, 59–82.

Nation, P. (2015). Principles Guiding Vocabulary Learning Through Extensive Reading. *Reading in a Foreign Language*, *27*(1), 136.

Nuttall, C. (1982). *Teaching Reading Skills in a Foreign Language*. London: Heinemann.

Packard, J. L. (2000). *The Morphology of Chinese: A Linguistic and Cognitive Approach*. Cambridge University Press.

Pavlik Jr, P. I. (2007). Timing is in Order: Modeling Order Effects in the Learning of Information. *In Order to Learn: How the Sequences of Topics Affect Learning*, 137-150.

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Redish, J. (2000). Readability Formulas have Even More Limitations than Klare Discusses. *ACM Journal of Computer Documentation (JCD)*, *24*(3), 132-137.

Restrepo Ramos, F. D. (2015). Incidental Vocabulary Learning in Second Language Acquisition: A Literature Review. *Profile Issues in Teachers Professional Development*, *17*(1), 157-166.

Schmitt, N., X. Jiang & W. Grabe (2011). The Percentage of Words Known in a Text and Reading Comprehension. *The Modern Language Journal* 95.1, 26–43.

Shen, H. H. (2005). Linguistic Complexity and Beginning-Level L2 Chinese Reading. *Journal of the Chinese Language Teachers Association*, *40*(3), 1.

Shen, H., & Tsai, C. H. (2010). A Web-Based Extensive Reading Program and its Assessment System. *Journal of the Chinese Language Teachers Association*, *45*(2), 19-47.

Shi, J. (2013a). *Graded Chinese Reader 500 words*. Beijing: Sinolingua.

Shi, J. (2013b). *Graded Chinese Reader 1500 words*. Beijing: Sinolingua.

Shi, J. (2014a). *Graded Chinese Reader 2000 words*. Beijing: Sinolingua.

Shi, J. (2014b). *Graded Chinese Reader 2500 words*. Beijing: Sinolingua.

Shi, J. (2016a). *Graded Chinese Reader 1000 words*. Beijing: Sinolingua.

Shi, J. (2016b). *Graded Chinese Reader 3000 words*. Beijing: Sinolingua.

Teng, F. (2014). Incidental Vocabulary Learning by Assessing Frequency of Word Occurrence in a Graded Reader: Love or Money. *LEARN Journal: Language Education and Acquisition Research Network*, 7(2), 36-50.

University of Iowa. Chinese Reading World. http://collections.uiowa.edu/chinese/ Accessed: 20.08.16

Verne, J. (2016) *Journey to the Center of the Earth: Mandarin Companion Graded Readers Level 2*. Shanghai: Mind Spark Press.

Wallace, C. (1992). *Reading*. Oxford: Oxford University Press.

Wells, H. G. (2015). *The Country of the Blind: Mandarin Companion Graded Readers Level 1*. Shanghai: Mind Spark Press.

West, M. (1926). *Learning to Read a Foreign Language: An Experimental Study*. London: Longman.

Yang, S. J. (1971). *A Readability for Chinese language* (Doctoral dissertation, Ph. D. Thesis for Mass Communication, University of Wisconsin).

Zhang, S. (2015). Teachers as Curators: Curating Authentic Online Content for Beginning and Intermediate CFL Learners. *Journal of Chinese Teaching and Research in the US*, 128.

Zhou, L., & Zhang, D. (2003). NLPIR: A Theoretical Framework for Applying Natural Language Processing to Information Retrieval. *Journal of the American Society for Information Science and Technology*, *54*(2), 115-123.