



How Accessible Are Unsimplified Novels for Advanced Learners of English?

August Aldred, *Manchester Metropolitan University, UK*

Acknowledgements

This study was made possible by the six wonderful people who agreed to take part, so my thanks are to them, first and foremost. I would also like to thank Dr. Huw Bell, who supervised this project and whose encouraging critique (and quibbling over my grammar) helped me further refine this paper to how it appears before you. Furthermore, I should thank those who constructively commented on the presentation form of this paper and Cheral Govind for her insightful comments on the first draft of this paper.

Abstract

Research into extensive reading, primarily by Nation (1992; 2000; 2006), suggests that an overall coverage of 98% is required to read a text with ease, further suggesting a vocabulary of 8000 words to achieve this. Through an analysis of advanced learner's (IELTS 6.5 or above in this study) interpretation of two texts, the researcher sought to uncover whether 'infrequent' vocabulary was the main issue in comprehending authentic texts, considering other factors such as polysemy and diction. VocabProfile's ability to predict the difficulty of texts in regards to frequency was analysed to answer the question, are 'difficult' words the same as 'infrequent' words?

In order to investigate the above factors, VocabProfile was used to analyse two extracts from vastly different novels – *Wuthering Heights* (Brontë, 1847) and *The Shadow of the Wind* (Ruiz Zafón, 2001) – in order to establish which 'infrequent' words occurred in the texts. Participants were then asked to read the texts and manually repeat this process by highlighting words they did not know, and to talk about their difficulties. The results suggest there are other difficulties beyond vocabulary for advanced readers and these are discussed further with some examination of why certain participants outperformed others.

Keywords: advanced learners; extensive reading; reading for pleasure; vocabulary; reading difficulties

Introduction

Hirsh & Nation (1992) suggested that 98% of all vocabulary in a text should be known for learners to be able to pleurably read it. The 2000 most common words in English are often

used as a starting point, as they cover a large portion of text – around 87% (Nation, 2006) to 90% (Hirsh & Nation, 1992). Hirsh & Nation further proposed that learners would need a vocabulary of approximately 5000 word families to reach 98% coverage. An example of a word family would be ‘run’, ‘runs’, ‘ran’, and ‘running’. Bauer & Nation (1993) define them as consisting of “a base word and all its derived and inflected forms” (p. 253).

Laufer (1989) had previously found that readers began to struggle with texts once the overall percentage of known words fell below 95%. Nevertheless, Laufer & Yano (2001) reported that advanced learners¹ in countries as diverse as Indonesia, Japan, Oman, and Israel possessed vocabularies of only 2000-3500 words (pp. 549-550). This was based on previous research by Laufer (2000) that looked at studies investigating passive vocabulary knowledge. Her investigations suggested vocabulary sizes to be between 1000 (in France) and 4000 (in China) (p. 48). When such estimates are considered alongside Hirsh & Nation (1992), we might reasonably assume that many learners would struggle with texts not aimed specifically at their level.

Ward (1999) highlights a flaw in the above suggestions for text coverage. Ward points towards the importance of the “exact position” (p. 309) of the word, which relates to the oft-overlooked point that readers can read comfortably, often by ignoring single words, until unknown words occur together. In fact, this seems to be how native speakers often deal with single unknown words (Block, 1992), regardless of their proficiency level.

Should learners encounter a text that falls below the coverage percentages suggested, they can choose from several reading strategies. Research suggests that more proficient readers tend to know which strategies to use and when to use them. (e.g. Anderson, 1991, pp. 468-469; Mokhtari & Sheorey, 2002, p. 7) Learners may assume that translation aids in comprehension of texts in English (e.g. Liao, 2006, p. 201) but translation may be a limited method depending on proficiency and resources. When reading for pleasure, regular reference to a dictionary or translation software would likely prove disruptive. Na & Nation (1985) attempted to explore what factors might affect the strategy of guessing from context, a strategy likely to be used in extensive reading. They stated that context plays a major role in helping readers to guess unknown words. It can be further reasoned from these findings that the larger the amount of unknown vocabulary, the less context is available to readers.

¹ The use of ‘advanced’ here is my choice based on the matching of reported learning hours in the articles and suggested learning hours to reach an ‘advanced’ level in frameworks like the CEFR.

Conversely, Bensoussan & Laufer (1984) concluded that guessing from context was only partially helpful to learners, suggesting that appropriate strategy use requires explicit instruction. Otherwise, students, regardless of proficiency level, fail to notice contextual clues or simply ignore unknown words.

When learners are ready to challenge themselves, there are graded readers available, aimed even at advanced users of English. Graded readers are organized by the number of word families contained, assuming that breadth of knowledge is more important than depth. A wide, but superficial vocabulary, is likely to be less useful in reading as language is very repetitive, with some words having several possible meanings. Rashidi & Khosravi (2010) found in their analysis of Iranian learners that deeper vocabularies not only contributed to stronger reading skills, but that a learner's depth of vocabulary knowledge was a strong point of reference when it came to predicting comprehension levels. According to the previously mentioned estimates of Laufer & Yano (2001), even advanced learners are unlikely to be working at the threshold of 4000 words/95% coverage, not least to mention the 8-9000 words suggested for 98% coverage needed to read novels (Nation, 2006). This comes even after significant periods of instruction of around 1500 hours (Laufer & Yano, 2001), which is close to estimates of time needed in class to reach C2 in the Common European Framework of Reference (CEFR) (EnglishProfile, n.d.).

Based solely on estimates of vocabulary knowledge, learners are therefore seen to be at a kind of impasse. If they continue with graded readers, their language knowledge has no room to grow, but they may feel unprepared to tackle texts written for readers who are highly proficient or native speakers. Graded readers, ubiquitous as they are, have garnered much attention, but fewer studies have looked at what comes after these readers. Uden, Schmitt & Schmitt (2014) found that learners can transition directly on to unsimplified novels, though they are likely to have less knowledge of the vocabulary. Motivation and enjoyment were also seen as potentially important factors. McQuillan (2016) argued for a pathway in which learners would work their way up to unsimplified novels through a series of books such as *Harry Potter* or *The Hunger Games*, maintaining that background knowledge acquired in the first instalment might aid the learner as they progressed.

Though there has often been less research on advanced learners' reading, this has been changing. For example, Abosnan (2016, p. 142) argues that zealous focus on understanding

single words in Libyan universities neglects learning to comprehend, with teachers providing the information for students. Extensive reading is one of many ways advanced learners can improve their language skills. Krashen (1989; 2004) has used his input hypothesis as a springboard to explore vocabulary gains from extensive reading. Other research shows that it can also lead to gains in other skills, like spelling (Schmitt, 1998; Pigada & Schmitt, 2006), and can be motivating and enjoyable (Leung, 2002; Powell, 2005).

What stands out in the research into reading difficulties so far is a disregard for the readers' own experiences. It is presumed that vocabulary is at the root of learners' struggles, as lower vocabulary knowledge seems to indicate lower comprehension (Laufer, 1992), but there must be something beyond that which makes reading novels remain difficult, even for learners with vocabularies that extend outside the 2000 most common words, which Hwang (as cited in Hirsh & Nation, 1992) states offer great coverage for fiction texts. This study aims to offer some insight into advanced learners' personal reading experiences as well as to explore any additional factors of impediment.

The following research questions formed the basis of this investigation:

RQ1 Are 'infrequent' words the same as 'difficult' words?

RQ2 How do advanced learners explain their perceived vocabulary difficulties in an authentic text?

RQ3 How useful is VocabProfile (Cobb, 2002) as an indicator of difficulty?

Methodology

Six participants (four males and two females) took part, each with a different first language and from different fields of study to try to account for various English education experiences and specialisms. All participants, except P, who is currently employed in the UK, were current, international undergraduate students. For the purposes of this study, advanced learners were defined as those with an IELTS score of 6.5 or above, as this is roughly equivalent to the 'advanced' levels of the CEFR (IELTS, n.d.). The use of the IELTS framework to decide whether participants were 'advanced' users of English was not a perfect method as it is open to some risks, namely that all participants had not taken the exam at the same time, nor in the same country, and some had taken other qualifications that could be roughly translated into an IELTS score. This casts some doubt over the validity of the scores.

Given that all participants had been required to take IELTS or similar exams, mostly for the purpose of studying in the UK, meant, however, that it was a convenient benchmark by which to compare and contrast the participants' proficiencies. Table 1 describes each of the participants in further detail. Henceforth, each participant will be referred to by the pseudonyms listed here, which are simply the first letter of each of their first languages.

Table 1: *Biographical data*

Participant	Age	Sex	First language (L1)	IELTS Score	Field of study	Frequency of reading novels in L1	Frequency of reading novels in English
J	21	Male	Japanese	7.5	Politics & International Relations	Once a month	Twice a month
P	36	Female	Polish	8	Law	Once a month	Once a month
C	26	Male	Cantonese	7	Banking & Finance	Once/twice a week	Once/twice a month
M	24	Male	Mandarin Chinese	8.5	Japanese Studies	One novel every 5/6 years	One novel every year
R	19	Female	Romanian	7.5	Anthropology & Archaeology	Reads about eight novels per year	Reads about two novels per year
G	19	Male	Gujarati ²	7.5	Aerospace Engineering	N/A	Once every month/two months

² Gujarati is a language spoken primarily in the north-western state of Gujarat, India

Participant recruitment was carried out primarily through advertising the study in student or language exchange groups on social media, with two of the participants being previous acquaintances of the researcher. In line with good ethical practice, all participants were briefed orally on what would happen in the study, what the aims were, and what would happen with the data. They also received this information in the form of an information sheet, including points of contact with both the researcher and the project supervisor. Participants were given the option to leave the study at any time they wished, with no negative consequences, and all participants gave their oral consent to be recorded during the post-reading interview. The reading tasks and interviews were held in public spaces, such as the university and, in one case, at the participant's workplace.

Participants were presented with two extracts from English-language novels, both of which were taken from the very beginning of the novels to replicate the natural process of reading. The novels selected were *Wuthering Heights* (Brontë, 1847) and *The Shadow of the Wind* (Ruiz Zafón, 2005). Given the vastly different eras of publication of these novels, there were several stylistic and grammatical differences, which were elements considered useful in eliciting from participants what was difficult about the extracts beyond unknown vocabulary. To challenge previous estimates, these extracts were also selected as the likely amount of known vocabulary coverage was below both 95% and 98%. 81.96% of words were in the first 2000 most common words in English in *Wuthering Heights* and 91.46% in *The Shadow of the Wind*. If Laufer & Yano's estimates (2001) about advanced learners were correct, it could be assumed that the learners' vocabularies might be insufficient, and, according to Hirsh & Nation (1992), these texts therefore would prove difficult.

Texts were analysed with VocabProfile to establish how many 'infrequent' words occurred in the extracts. For the purposes of this study, 'infrequent' words were those outside the most common 2000 words in English. Participants were asked to carry out a similar process while reading by highlighting words they did not know. A dictionary of their own choice was available throughout the study and participants were informed that they could make notes on words they did not know or the events of the extract. These notes could be in English or in their first language. Following the reading, participants were asked to talk about their comprehension of the texts through a series of questions, including a retelling of the text (see appendix). These conversations were recorded and then transcribed.

Finally, the highlighted texts were compared to the data from VocabProfile to assess how accurate the tool was in predicting ‘difficult’ words. Data from the post-reading interviews was also analysed for shared difficulties and exploration of each participant’s own explanation of what made these extracts challenging.

Results

Using VocabProfile’s output and the manual data from participants, coverage of known vocabulary for each participant was calculated. To analyse what impediments were shared, specific difficulties mentioned by participants in the post-reading interview were quantified and these results are displayed in table 3. Sometimes, participants were unaware of misinterpretations of the text that they had made, but these were noted by the researcher. The causes of these misinterpretations are listed in table 4.

Table 2: *Mean participant coverage for each text*

	<i>Wuthering Heights (WH)</i>	<i>The Shadow of the Wind (SotW)</i>
Coverage with 2000 most common words	81.96%	91.46%
Mean participant coverage	96.8%	99.7%
Standard Deviation	2.16	0.87

Table 3: *Number of participants and their self-reported issues when reading*

	<i>Wuthering Heights</i>	<i>The Shadow of the Wind</i>
Unknown vocabulary	2	0
Unknown vocabulary grouped together	3	0
Style, including organisation	4	2
Figurative language	0	2
Vocabulary used in an unfamiliar or unmodern way	3	1
Insufficient background knowledge or context	4	1

Categorizing the reported difficulties was not always straightforward. In hindsight, a more rigorous process should have perhaps been employed, particularly in regards to splitting categories that held two types of possible impediments. At the time of the study, participants' reports engineered the categorization as it was their personal experiences that were being reported. In a similar vein, the issues reported in table 4 were largely subjective observation, but it was nonetheless felt that it was worth reporting additional difficulties.

Table 4: *Misinterpretations unnoticed by participants*

	<i>Wuthering Heights</i>	<i>The Shadow of the Wind</i>
Completely unknown vocabulary	1	0
Had apparent incorrect knowledge of word meaning	2	0
Misread the text	1	0

At first glance, there's some evidence for Laufer's argument (1997) in the "had apparent incorrect knowledge" category, in that she states an error learners often commit is applying the only meaning they know of a word. An incorrect interpretation of a word does not always mean the learner only knows one meaning; there may be deeper thought processes leading to that conclusion. For instance, a participant arrived at the correct meaning of "capital" in "capital fellow" by inferring that capital, as in a city in a country, meant "the top".

In regards to research question one, the quantitative difference between what VocabProfile selected as 'infrequent', and what participants highlighted as 'difficult', suggests 'infrequent' words are not always 'difficult' (see table 5). That being said, highlighted 'difficult' words were almost always 'infrequent'. It appears, however, that completely unknown, single words are perhaps easier to deal with than other perceived difficulties mentioned by readers here, such as style, figurative language, grouped unknown words, and lacking background knowledge. This is assumed on the basis that participants could usually guess the meaning of single words they had highlighted as unknown.

Table 5: Number of ‘infrequent’ words in extracts, excluding proper nouns. Bracketed numbers refer to highlighted words that were not ‘infrequent’ according to VocabProfile

	<i>Wuthering Heights</i>	<i>The Shadow of the Wind</i>
VocabProfile	76	26
Participant J	26 (5)	6 (2)
Participant P	9 (1)	0
Participant C	34 (3)	8 (2)
Participant M	5 (1)	0
Participant R	22 (1)	3
Participant G	7	0

Given that this study engaged with six very different readers, each participant’s responses will now be explored further to answer research question two: How do advanced learners explain their perceived vocabulary difficulties in an authentic text?

Individual Participants (Known Vocabulary Coverage)

Participant J (WH: 95%, SotW: 98.5%)

Wuthering Heights posed something of a challenge to this participant for a variety of reasons. Vocabulary was one issue, but in the post-reading interview the diction of the text was also mentioned, namely that the roundabout style in which the author wrote was frustrating as J felt he did not understand what was happening. This is not uncommon in texts from this period, in which much of the plot is woven into long-winded and seemingly unimportant soliloquies and conversations.

J also provided some very interesting interpretations of the text itself. He assumed that some prior knowledge of the Bible was required to understand parts of the extract, regarding the phrase “Go to the Deuce”. Intriguingly, “Deuce” was a word he highlighted as unknown, yet he could interpret it as a religious reference, perhaps because of its likeness to “go to the devil”. Furthermore, until he read that Mr. Lockwood had arrived by horse, he had assumed the two characters to already be inside a room. This he speculated from the fact that Mr. Lockwood describes Heathcliff’s features in detail and he felt that this would be only possible if the two were sitting quite close together. This second point of confusion may have also

stemmed from a vocabulary issue in that he did not know the phrasal verb “rode up”, missing information on Mr. Lockwood’s mode of arrival.

There were fewer unknown words in *The Shadow of the Wind*, but even looking these words up did not help him. This was, he said, because the words were often part of figurative speech. For example, he searched for the word “wreath”, but was unable to fully comprehend what the author meant when he described dawn as a “wreath of liquid copper”. Issues in comprehension in this text were largely linked to fragmented understanding of similar metaphorical language and the organization of the text which jumps between timelines; all elements VocabProfile was unable to predict.

In general, J could extract many important points from the first extract, but was clearly not confident with the text, judging from his language (“I’m not sure”, “I’m just guessing”) and seemed heavily distracted by the number of unknown words, despite his ability to guess some words successfully from context when asked. Based on J’s responses, it appears that too many unknown words can cause reading to become burdensome as readers may be distracted by the amount of unknown vocabulary whether it is the main impediment or not. It is not clear at this stage how many words may be needed to prove distracting, but it is likely to differ between readers in regards to their confidence dealing with the unknown.

Participant P (WH: 98.5%, SotW: 100%)

This participant was one of the stronger performers. It is difficult to attribute this to one, clear reason. We may consider factors such as exposure and types of exposure, with P working a full-time job in an Anglophone environment. Laufer & Hulstijn (2001) argued with their Involvement Load Hypothesis that vocabulary is acquired and retained better when language users are more involved in a task. Jing & Jianbin (2009) provided evidence that this is true not only of reading and writing tasks, but also listening tasks. Arguably, as a lawyer, the stakes are higher for P, and she is more involved in the task of understanding English on a day-to-day basis, leading her to have possibly acquired, and retained, a larger vocabulary. By and large, the Involvement Load Hypothesis is to be applied to explicit learning in an implied class-setting, but aspects of it are applicable to P’s experiences, such as the possibility of needing to negotiate the meaning of more words, which has been seen to lead to better retention (Newton, 1995). This of course raises some questions about the size of each

participant's vocabulary, which was not a component of this study, but could be considered in future investigations.

P read *Wuthering Heights* very quickly and displayed a confidence in tackling the text despite some unknown words appearing. In the post-reading interview, P stated that the unknown words could not be guessed from context and took nothing away from the image she had in mind whilst reading, so she felt able to skip those words. This echoes findings from Freebody & Anderson's (1983) analysis of native English speakers' reading behaviours. In this investigation, they found it takes a great number of unknown words to affect comprehension and unknown words, when deemed unimportant, are simply skipped and absorbed into the context when a summary is required. In short, P's reading strategies seem to mirror a native English user. In her post-reading interview for the second extract, she also talked further about being able to imagine the story. Tomlinson (2011) wrote extensively about the benefits of teaching learners to use top down, deeper processing strategies like visualization to aid in better recall and comprehension. Again, this suggests P was perhaps tapping into higher level strategies than other participants, allowing her to bypass the difficulty of any vocabulary.

Interestingly, one of the more frustrating parts of this extract for P was when Heathcliff's "fingers sheltered themselves, with a jealous resolution." Despite understanding every single word, she could not picture what kind of action this could be describing, which perhaps impeded her main reading strategy. This also mirrored J's struggles with "wreath"; a word he knew but could not envision in a particular context. VocabProfile did not pick out any of the words in the quoted phrase above as being 'infrequent', suggesting most learners in this group ought to know them, but in this sequence, they are perhaps beyond comprehension given that they are used unusually.

Participant C (WH: 94%, SotW: 98%)

One might assume that as the most frequent reader of all participants, and with a long residence in the UK, that C might have dealt with this task quite easily, however, C highlighted the most unknown words of any participant in both extracts. C was quite well acquainted with the researcher beforehand, so it was known that he had a majority of Cantonese-speaking friends as well as a desire to return to Hong Kong, so there are perhaps questions to be raised about how acculturation has manifested itself in C's English education

and how this may have affected language learning and acquisition, however, these are outside the scope of this paper.

Of all who took part, only C read more often in his first language than in English, but there has been some observation that a certain level of proficiency is needed for higher level strategies to transfer from the L1 to the L2 (Kong, 2006), and C had the lowest proficiency amongst participants in the study. Alongside J, who also performed less well on the task, C was reliant on one reading strategy: translation. The strategy of translation seemed to be his preferred one also because he read a lot of texts online and it was easy to paste English text into online translation sites. In a similar vein, J, who used a Japanese dictionary, was partial to glossing in Japanese. All other participants used either no dictionary or an English-language dictionary. This apparent lack of confidence in guessing from context or to employ other strategies was a shared attribute between these two participants, suggesting perhaps that the confidence and ability to deal with unknown words in multiple ways is an important aspect of reading ability. Notably, C pointed to grouped unknown words as problematic, stating, “If I don’t know half of the vocabulary I can’t really guess”. This lends support to Ward’s (1999) caveat that the position of words must be considered. Intriguingly, he reported that looking up these words only helped him gain about an 80% understanding of the text, implying his struggles lay in other factors too. When questioned about this, his only other grievance was diction.

Participant M (WH: 99%, SotW:100%)

After the task, when he decided to guess the meaning of the six words he marked as unknown in *Wuthering Heights*, M got all but one correct. As with other participants in the study, M cited style as a possible impediment and stated that it took him reading a few paragraphs into the extract before he began to fully comprehend. The fact he struggled in this way despite having such an extensive knowledge of the text’s vocabulary lends further support to the idea that vocabulary is not the only factor at play.

M chose not to use a dictionary in the task and revealed this is his usual approach to reading in English. He showed some metacognitive strategies in his reasoning, such as when he guessed the meaning of the word “penetralium” correctly. He based his assumptions on the seemingly shared suffix with “museum”. Though this does not seem to be a shared suffix

etymologically, it showed an interesting cognitive process in the participant, which coincidentally led to the correct answer.

Extramural English could have also played a part in exposing learners to usually ‘infrequent’ vocabulary. In the case of M, his seemingly wide and deep vocabulary may have been boosted by his consumption of a televised version of Sherlock Holmes – a character famed for his loquacious speaking, and based on material written at a period quite close to Brontë’s work, which could have led to schema raising.

In *The Shadow of the Wind*, M marked no unknown words, but did illustrate his troubles with comprehending the order and organisation of the text in regards to timelines. This issue was not the only one he shared with participant J; they both also pointed out the difficulty in fully understanding the metaphorical language in the text, going as so far as to give the same example about the description of dawn. M, however, seemed to have grasped the meaning and was commenting merely on how much imagination such language required to get the full experience the author intended. This links back to points about the role of imagination discussed in regards to participant P. Like P, he also seemed to employ several strategies for reading that made use of both top down and bottom up processing.

Participant R (WH: 96%, SotW: 99.4%)

R’s reading of *Wuthering Heights* gave further credence to Ward’s argument (1999) about the location of vocabulary. When asked what made the extract difficult to read, R referred directly to a part of the text in which unknown vocabulary material was grouped together; around two to three unknown words per sentence. Tellingly, she missed the introduction of the servant, Joseph, who was introduced in this particular passage.

Choosing only to use a dictionary for some of the unknown words, R explained that she chose only to look up words she thought she might know the meaning to. This seems counter-intuitive, but guessing and verification have been identified as important reading strategies, as together they lead away from incorrect guesses (Laufer & Yano, 2001). Like participant M, R said that when reading in English usually, she would guess words and only look them up if they were absolutely crucial to understanding a sentence or idea.

It was noted that participants R and P had some of the most succinct summaries of the two extracts. They also appeared to have quite eloquent vocabularies, judging from some of their selected words and phrases e.g. using “encounter” instead of “meeting.” This led the

researcher to observe that an instrument much like the one used in this study could also prove useful in examining the active vocabularies of learners.

Participant G (WH: 98.8%, SotW: 100%)

Once the interview began, G disclosed that he had studied English Literature at A Level in his home country and this was perhaps an advantageous factor that attributed to his strong performance. Indeed, his analysis was quite in depth, paying attention to things like the use of darkness as a representation of emotion in *The Shadow of the Wind*. Like M and P, G displayed multiple reading strategies and hinted at others, such as “reading between the lines.”

Like R, he had no issues until about half way through the first extract. Unlike R, however, the unknown words were not grouped together. Despite finding several of the same details, G’s analysis was noticeably more well-rounded, suggesting that grouped unknown vocabulary was an issue. The two participants also shared the same strategy of reading to the end of a text before checking any vocabulary, and only when necessary; preferring to either guess or simply skip less important words. This strategy seemed to work better for G, as R needed to skip bigger chunks of text with grouped unknown words.

As with the other participants, G referred to the style of *Wuthering Heights* as a problem. He described how he needed to read the extract a couple of times to understand it fully. In comparison, he found the second extract straightforward and thus more enjoyable. For G, there were no lexical or structural struggles in the second text.

Discussion

Considering each participant’s coverage rates, it appears that around 94% is the threshold before texts become difficult (see participant C), which supports Laufer’s argument (1989) that readers will struggle once this coverage rate is below 95%. Furthermore, better performers on the tasks had knowledge coverage rates between 98% and 100%, supporting Hirsh & Nation’s claims (1992). What is evident, however, is that obstacles continue to exist and these are not always vocabulary-related. Amongst the participants, there were differences in the extent of comprehension, even amongst the better performers and despite working at similar coverage rates. Schmitt, Jiang & Grabe’s study (2011) into the relationship between comprehension and coverage rates seems to suggest the same, with some participants with 100% coverage knowledge but not full comprehension scores (p. 36). This suggests that there are other impeding factors that make texts less or more accessible to an advanced learner. Of

course, to support this assumption, these learners' vocabularies would need to be tested, which could also allow further exploration of how large an advanced learner's vocabulary needs to be to reach these coverage rates. This study has functioned purely on the basis of coverage rates, but we have no indication of what kind of vocabularies participants were working with, other than previously-made estimates (Laufer & Yano, 2001). Indeed, if those estimates are correct, then there is perhaps a challenge to be made against Nation's suggestion (2006) that more than 8000 word families are required to reach knowledge coverage rates of 98% and above, as Laufer & Yano estimated much smaller working vocabularies of advanced learners.

Overall, *The Shadow of the Wind* was easier for every participant. The three participants who found unknown words in *The Shadow of the Wind* shared difficulties with the words "wreath", "stifle" and "glint", but other words selected beyond these were harder to predict. Although the highlighted words were those predicted by VocabProfile to be 'infrequent', there seemed to be no pattern amongst the participants in what was difficult beyond those three words. Although we can predict what may be difficult based on frequency, it is impossible to guess what will be difficult for each individual reader. This is evident in how J and P both selected "thorns" as an unknown word despite their very different proficiencies and the fact this word is not infrequent according to VocabProfile (it appears in the second 1000 most common words). As the 2000 most common words seemed not to be a perfect indicator of difficulty, a more modern text like *The Shadow of the Wind*, which contained more of these words, was harder to judge for difficulty using VocabProfile alone.

Through the manual method employed by the participants, it became clear that they sometimes underestimated their own skills, which contrasted with Laufer & Yano's findings (2001) that EFL learners tend to overestimate their understanding. These phenomena were also evident in Lotherington-Woloszyn's study (1993), in which learners underestimated their ability to deal with unsimplified texts *and* overestimated themselves when it came to graded readers. The results of the abovementioned study suggest that the degree to which a text is simplified alters perceptions of difficulty more than comprehensibility. This supports the idea that participants like J and C, who highlighted more unknown words, were seemingly overwhelmed by the 'difficulty' of the text, and not focusing on what they understood. It may also point to the influence of style on readability and why the classic text in this study was

viewed as more difficult despite all participants offering similar – and generally correct – summaries about the plot and characters.

There was also a noted tendency for participants to look up words they thought they might know, apparently lacking confidence in their knowledge. Laufer & Yano (2001) suggest that evaluation of vocabulary knowledge can vary from nationality to nationality, with more ‘traditional’ societies perceived as being modest. In this study, those learners would have been those from China, Japan, and Hong Kong. Though the Japanese and Hong Kong Chinese participants selected a larger number of unknown words, this seemed to stem more from genuine lack of knowledge than modesty, and the Chinese participant was extremely confident in his evaluation, as well as able to guess the unknown words as he predicted he could. This may indicate that nationality is not always a good indicator of evaluating abilities.

Shang (2010) provides some evidence for a correlation between usage of multiple reading strategies and self-efficacy in reading. Similarly, Chamot’s findings (1993) seemed to uncover a possible relationship between these two elements. Comparing participants like C and J, who relied on one strategy, to others like M or P, who utilized visualization and compensation strategies, there does appear to be a case to be made for more confident readers to be those with more strategies at their disposal.

Participants’ first languages could have influenced comprehension depending on how widely their L1 differed to English. We must remember, though, that concepts of difficulty or difference in language are subjective to the individual. For example, Romanian may seem closer to English, but Mandarin Chinese shares grammatical similarities with English whereas Romanian uses a gender and declination system. Grabe (1991) has argued that, in any case, it seems less clear that the differences between L1 and L2 are as pronounced in advanced learners (p. 388). Additionally, Upton & Lee-Thompson (2001) looked at the ways native speakers of Chinese and Japanese used their first language while reading. They observed that not only do L2 users of English actively use their L1 to comprehend texts, but through a think-aloud task found that the advanced learners used their respective L1s much less than the intermediate students did. Good comprehension seemed to spring more from effective strategy use than overreliance on the first language.

In regards to comprehension of the texts, all participants could give satisfactory retellings of the second extract. The retellings of *Wuthering Heights* were generally correct in substance,

but in parts ranged from containing insightful and subtle detail to complete confusion. It seems clear, however, that advanced learners can pick up on several in-text clues that allow them to follow plots, even in more complex texts. Both texts elicited appreciation for the language used and some participants felt motivated to pursue reading *The Shadow of the Wind* after reading the extract, suggesting they derived pleasure from the task. This is congruous with Uden et al.'s findings (2014) that the book itself and motivation to read can affect the learner's attitude and performance.

Several of the participants identified themselves as infrequent readers in both their first language and in English i.e. they didn't read daily, or even weekly, for pleasure. Given the variety of results presented here, it seems that it is possible to be a relatively skilled reader without reading on a consistent basis (see participant M). Of course, this neglects any prior reading experience of the participants. Their responses to how often they read only related to fiction texts, so this does not take into account the amount of non-fiction reading they take part in – probably a great deal as university students. Extramural input is another factor to consider, with media such as music and television offering further exposure to words that may occur in fictional texts. For example, participant C knew the meaning of the relatively infrequent word “legacy” from an Eminem track of the same name. Participants J and P reported reading about the same amount in both their L1 and English, yet their reading comprehension and vocabularies differed greatly, so there may have been other sources of exposure that affected their acquisition.

Limitations

Although efforts were made to create a varied sample, it remains a small sample, and so it is difficult to apply these findings in a wider context. Furthermore, the participants were all of a similar age. The researcher also noted a personal tendency to move forward in the interview too quickly sometimes, and to give away minor information in the interview, which could have altered the data in some way. Due to these factors, there may have been missed opportunities to learn more about the reading experience, which could be solved by adding more structure in the interview process. Furthermore, the background of participants' education was not fully explored, such as whether they took classes in their home country in their L1 or in English. This could potentially have been another variable, as was suggested by G's performance.

If this study were to be replicated, it may prove fruitful to be more selective about the texts used. As explained earlier in this paper, *Wuthering Heights* and *The Shadow of the Wind* were chosen for their very different publication dates and styles, but to further explore the role of vocabulary and reading strategies, texts that are perhaps not as wildly different should be used to eliminate the possibility that *Wuthering Heights* is difficult simply because it's an old and turgid text that can prove difficult even for native speakers. Further combinations may draw on books that are closer in publication date but that vary in style or genre.

Previous comments on this study have alluded to the fact that learners lacked schemata for these texts. Carrell (1983) cites a number of studies into schema-based reading. She refers to a 1981 study of her own in which the schemata of stories were purposefully changed to something unexpected and that this caused problems in recall and comprehension. Details beyond this are not shared, but it is difficult to conclude with certainty that other variables were not at play. That is not to say that having a schema in place plays no part in comprehension of a text in the L2. It is simply argued here that it should be considered no smaller or larger than other variables presented in this research. Carrell cites another study, by Kintsch & Greene (1978), in which said researchers aimed to prove the importance of having a cultural schema in place by comparing comprehension rates amongst American students of European stories and American Indian stories, namely those of the Apache tribe. They found that the students had better comprehension of the European stories, claiming a greater familiarity in structure, but they were criticized for not including an American Indian group of readers, claiming the Apaches to be bi-cultural and already familiar with Western structures. If this was indeed true in 1978, then it would be safe to assume that many L2 users of English in 2018 are also 'bi-cultural', if not through the way they live their day-to-day lives, then through the spread of multimedia across the globe, including the internet. If L2 media is consumed enough, previously missing schemata may be acquired. If true, this would put speakers of a multitude of languages on a more level playing field in terms of possible comprehension when considered simply through the lens of schema-based reading.

Laufer (1997) claimed that L2 readers often mistakenly think they recognize words, overestimating their knowledge, whereas Lotherington-Woloszyn (1993) found evidence of the opposite. The conflicting reports of students' knowledge and what the researchers in these studies actually found remind us of the possible risk in validity in self-report studies, which suggests that instruments such as standardised tests should be employed alongside

qualitative methods like those used here for more accurate understanding of participants' levels of vocabulary knowledge. This is also a relevant point given that there has been much reference towards previous research on vocabulary coverage estimates without conducting any vocabulary testing on participants in this study, which was done in the interest of avoiding participant fatigue (readers took between 30 and 90 minutes to complete the task as it stands). Where this study has sought to fill the gap of more qualitative research into reader experiences, it has perhaps failed to incorporate pertinent quantitative information that could support conjecture made here.

A clear weakness of VocabProfile is its inability to deal with aspects like phrasal verbs, polysemy, and metaphorical language. These were some of the other issues raised by participants as points of difficulty. Alderson's findings (2007) suggest, however, that we ourselves are unable to give accurate or consensual judgements on word frequencies, meaning that we must still rely on instruments, such as concordancers like VocabProfile, for more valid estimates.

Conclusion

In *Wuthering Heights*, style and a belief in lacking schema were the most cited reasons for finding the extract difficult. When discussing 'infrequent' or 'difficult' vocabulary, it mattered more if such words were close together, as most of the participants otherwise felt capable of guessing the meaning. Participants could highlight 'infrequency' as something that made vocabulary 'difficult', suggesting that there is a link between 'infrequency' and perceived 'difficulty'. VocabProfile's output gives further evidence to this link. However, it should be remembered that 'infrequency' here also meant for participants words used in a different sense, suggesting some participants' vocabularies were not deep enough. In *The Shadow of the Wind*, style once again seems to have been an issue, though overall this text was well received by all participants and posed no real difficulties.

To answer research question three, VocabProfile is quite useful as an indicator of difficulty insofar as that 'infrequent' usually means 'difficult'. It cannot, of course, tell us how each reader will fare with a text, nor can it deal with more advanced uses of words in English, such as metaphorical language, which to the software may appear to be a string of 'frequent' words.



Researchers such as Nation and Laufer base their predictive assumptions of text difficulty on overall coverage, which fails to consider where words occur and in what context.

Unsimplified novels seem accessible enough to advanced learners, but learners must be equipped with enough motivation and self-belief in their knowledge to employ the correct reading strategies to progress through the novel comfortably.

Further research might analyse a larger group of learners and consider further the genre of text. Laufer (1997) admits that estimated coverage rates cannot be used to blanket all learners (p. 24), so efforts should be made to find specific thresholds for each level. For example, Laufer (1989) hints at a possibly lower threshold of coverage that, when combined with successful guessing, leads to the 95%+ coverage required e.g. 80% knowledge and 15% correct guessing. Laufer (1992; 1997) also accepts the role other aspects of texts play in comprehension, such as background knowledge, which have been explored in this paper. The gravity of such elements could be investigated further by measuring learners' vocabularies and then comparing how they deal with texts that contain elements mentioned in this paper e.g. unfamiliar style. Given the frequency with which text organisation was mentioned here, it seems a worthwhile endeavour to perhaps also explore the effects of teaching learners how to deal with different diction, such as non-linear timelines.

Biodata

August Aldred is a recent Master's graduate and aspiring language teacher, with experience teaching both English and Japanese as a private tutor. Her research interests within the field lie broadly in reading in L2, and in identity issues related to language and culture, both amongst L2 learners and heritage language users.

References

- Abosnan, S. H. (2016). *The teaching of reading English in a foreign language in Libyan universities: methods and models*. Unpublished PhD thesis, University of Glasgow.
- Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics*, 28(3), 383-409.
- Anderson, N. J. (1991). Individual differences in strategy use in second language reading and testing. *The Modern Language Journal*, 75(4), 460-472.



- Bauer, L. & Nation, P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253-279.
- Bensoussan, M. & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, 7(1), 15-32.
- Block, E. L. (1992). See how they read: Comprehension monitoring of L1 and L2 readers. *Tesol Quarterly*, 319-343.
- Brontë, E. (1847). *Wuthering heights*. Thomas Cautley Newby.
- Carrell, P. L. (1983). Some issues in studying the role of schemata, or background knowledge, in second language comprehension. *Reading in a Foreign Language*, 1(2), 81-92.
- Chamot, A. U. (1993). Learning strategies in Japanese foreign language instruction. Final report.
- Cobb, T. (2002). Web vocabprofile. Retrieved from <https://www.lexutor.ca/vp/eng/>.
- EnglishProfile (n.d.). Introductory guide to the Common European Framework of Reference (CEFR) for English language teachers. In: *English Profile*. Retrieved from <http://www.englishprofile.org/images/pdf/GuideToCEFR.pdf>.
- Freebody, P. & Anderson, R. C. (1983). Effects on text comprehension of differing proportions and locations of difficult vocabulary. *Journal of Reading Behavior*, 15(3), 19-39.
- Grabe, W. (1991). Current developments in second language reading research. *Tesol Quarterly*, 25(3), 375-406.
- Hirsh, D. & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure?. *Reading in a Foreign Language*, 8, 689-689.
- IELTS (n.d.). Common European Framework How should the CEFR be used by recognising institutions wishing to set language ability requirements?. In: *IELTS*. Retrieved from <https://www.ielts.org/ielts-for-organisations/common-european-framework>.
- Jing, L. & Jianbin, H. (2009). An empirical study of the involvement load hypothesis in incidental vocabulary acquisition in EFL listening. *Polyglossia*, 16, 1-11.



- Kintsch, W. & Greene, E. (1978). The role of culture-specific schemata in the comprehension and recall of stories. *Discourse Processes*, 1(1), 1-13.
- Kong, A. (2006). Connections between L1 and L2 readings: Reading strategies used by four Chinese adult readers. *The Reading Matrix*, 6(2).
- Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal*, 73(4), 440-464.
- Krashen, S. D. (2004). *The power of reading: Insights from the research: Insights from the research*. ABC-CLIO.
- Laufer, B. & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, 22(1), 1-26.
- Laufer, B. & Yano, Y. (2001). Understanding unfamiliar words in a text: Do L2 learners understand how much they don't understand?. *Reading in a Foreign Language*, 13(2), 549-566.
- Laufer, B. (1989). What percentage of text-lexis is essential for comprehension. *Special language: From humans thinking to thinking machines*, 316-323.
- Laufer, B. (1992). How much lexis is necessary for reading comprehension?. In: *Vocabulary and Applied Linguistics* (pp. 126-132). London: Palgrave Macmillan.
- Laufer, B. (1997). The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. *Second Language Vocabulary Acquisition: A Rationale for Pedagogy*, 1, 20-34.
- Laufer, B. (2000). Task effect on instructed vocabulary learning: the hypothesis of 'involvement'. Conference paper presented at *AILA*, Tokyo.
- Lotherington-Woloszyn, H. (1993). Do simplified texts simplify language comprehension for ESL learners?. Retrieved from <https://eric.ed.gov/?id=ED371583>.
- McQuillan, J. (2016). What can readers read after graded readers?. *Reading in a Foreign Language*, 28(1), 63.
- Mokhtari, K. & Sheorey, R. (2002). Measuring ESL students' awareness of reading strategies. *Journal of Developmental Education*, 25(3), 2.



- Na, L. & Nation, I. S. P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal*, 16(1), 33-42.
- Nation, I. (2006). How large a vocabulary is needed for reading and listening?. *Canadian Modern Language Review*, 63(1), 59-82.
- Newton, J. (1995). Task-based interaction and incidental vocabulary learning: A case study. *Second Language Research*, 11(2), 159-176.
- Pigada, M. & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language*, 18(1), 1.
- Powell, S. (2005). Extensive reading and its role in Japanese high schools. *The Reading Matrix*, 5(2).
- Rashidi, N. & Khosravi, N. (2010). Assessing the Role of Depth and Breadth of Vocabulary Knowledge in Reading Comprehension of Iranian EFL Learners. *Journal of Pan-Pacific Association of Applied Linguistics*, 14(1), 81-108.
- Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning*, 48(2), 281-317.
- Schmitt, N., Jiang, X. & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1), 26-43.
- Tomlinson, B. (2011). 14 Seeing what they mean: helping L2 readers to visualise. *Materials Development in Language Teaching*, 357.
- Uden, J., Schmitt, D. & Schmitt, N. (2014). Jumping from the highest graded readers to ungraded novels: Four case studies. *Reading in a Foreign Language*, 26(1), 1.
- Upton, T. A. & Lee-Thompson, L. C. (2001). The role of the first language in second language reading. *Studies in Second Language Acquisition*, 23(4), 469-495.
- Ward, J. (1999). How large a vocabulary do EAP engineering students need?. *Reading in a Foreign Language*, 12(2), 309-324.
- Zafón, C. R. (2005). *The shadow of the wind*. London: Penguin.



Appendix: Interview Questions

Pre-reading questions

How old are you?

What is your major?

What is your current IELTS score?

How often do you read novels in your native language?

How often do you read novels in English?

Post-reading questions

Could you tell me what happened in the extract you read?

Did you find it hard to understand this extract? If so, what made it hard to understand?

What was difficult about the words you highlighted?

Did not understanding those words stop you understanding the story?

Did you feel that looking up the words helped you understand the story better?

Do you think you could read this book by yourself?

What did you learn about the characters?

What did you learn about the setting?

What do you think happens next?

Do you have anything else to add?