Region - Educational Research and Reviews, 2024, Volume6, Issue9
DOI: 10.32629/rerr.v6i9.2666
ISSN Online: 2661-4634
ISSN Print: 2661-4626

Corpus creation and lexical profile analysis

Beini HE

Xingzhi Primary School, Shanghai 200436, China

Abstract: In the context of L2 teaching classroom, instructors' prime role in vocabulary instruction is to teach words and help specific learners make meaningful progress in lexical development. A great deal of planning and preparation needs to occur before teachers enter the classroom (Webb & Nation, 2017), among which the process of evaluating the appropriateness of texts, materials and assignments based on teachers and learners' current vocabulary level seems to be necessary [11]. The purpose of this paper is to demonstrate a detailed procedure in which instructors evaluate a certain text using the data of Compleat Lexical Tutor and determine whether it is a suitable language learning material for students who have mastered the most frequent 2,000 words, thereby conducting further lexical profile analysis.

Key words: L2 teaching classroom; vocabulary learning; lexical profile analysis; Compleat Lexical Tutor

1 Introduction

It is a common sense that in the L2 teaching classroom, texts, materials and assignments provided by teachers usually play an important role in language learning if they are appropriate resources that correspond to students' vocabulary size that teachers evaluate. However, a question here is that how teachers can evaluate the appropriateness of the texts and how they can assess the degree to which students will understand the vocabulary in the text. The answer is the result of Vocabulary Size Test of their students and data of RANGE Programme [4][5], which allows users to (a) determine the vocabulary size needed to understand the vocabulary in text, (b) create word lists based on the frequency of words occurrence and range of use in different types of discourse, (c) determine the number of encounters with words in a text [12]. To be simplified, what teachers need do is to make a comparison between the data and vocabulary size. The aim of this study is to give a detailed procedure for evaluating a certain text: 10 chapters from a fairy story: *Alice in Wonderland*, which is taken from BBC Learning English site, and determine whether it is suitable language learning material for students who have mastered the most frequent 2000 words. And effective ways of modifying the text to make it more suitable for students are then given.

2 Assessing students' understanding of text vocabulary

2.1 How much vocabulary is needed to understand a text

Text coverage is a measure that indicates how much vocabulary is necessary for adequate comprehension and whether learners are able to understand that discourse and guess words based on context. Research suggests that knowing 95% of the words in discourse may be enough for adequate comprehension and for learners to guess words from context [2][3]. Webb and Rodgers suggest that 95% coverage may be sufficient for comprehension of television and movies [9]. Nation suggests that knowing 98% of the words may be ideal for comprehension and for guessing from context [6][7]. Therefore, 95% is the ideal coverage for dealing with spoken text and 98% is the ideal coverage for dealing with written texts. I have chosen 10 chapters of the fairy story *Alice in Wonderland*, which is selected from BBC Learning English site (http://www.bbc.co.uk/learningenglish/english/features/drama) and used above standard to evaluate whether it is the suitable language learning materials.

2.2 Analyze text using the Compleat Lexical Tutor

Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	508 (77.56)	747 (74.33)	5,924 (85.42)	85.42
K-2 Words :	84 (12.82)	103 (10.25)	241 (3.48)	88.90
K-3 Words :	14 (2.14)	14 (1.39)	118 (1.70)	90.60
K-4 Words :	16 (2.44)	18 (1.79)	53 (0.76)	91.36
K-5 Words :	7 (1.07)	7 (0.70)	13 (0.19)	91.55
K-6 Words :	5 (0.76)	5 (0.50)	30 (0.43)	91.98
K-7 Words :	4 (0.61)	4 (0.40)	33 (0.48)	92.46
K-8 Words :	4 (0.61)	5 (0.50)	13 (0.19)	92.65
K-9 Words :	1 (0.15)	1 (0.10)	1 (0.01)	92.66
K-10 Words :	5 (0.76)	6 (0.60)	13 (0.19)	92.85
K-11 Words :	3 (0.46)	4 (0.40)	21 (0.30)	93.15
K-12 Words :	1 (0.15)	1 (0.10)	3 (0.04)	93.19
K-13 Words :	1 (0.15)	1 (0.10)	13 (0.19)	93.38
K-14 Words :				
K-15 Words :	1 (0.15)	1 (0.10)	8 (0.12)	93.50
K-16 Words :				
K-17 Words :				
K-18 Words :	1 (0.15)	1 (0.10)	1 (0.01)	93.51
K-19 Words :				
K-20 Words :				
K-21 Words :				
K-22 Words :				
K-23 Words :				
K-24 Words :				
K-25 Words :				
Off-List:	?	98 (9.75)	450 (6.49)	100.00
Total (unrounded)	655	1,005 (100)	6,935 (100)	≈100.00

Figure 1. The distribution of the various levels of vocabulary in 10 chapters

The vocabulary size of students in class is known: the most frequent 2,000 words, but not yet mastered any other word levels, so the Compleat Lexical Tutor can be used to determine the vocabulary size needed to achieve 95% ~ 98% text coverage. These 10 texts are selected from a fairy story named *Alice of Wonderland* because each chapter is the right length for classroom reading. Figure1 is an analysis of the distribution of tokens, word types and word families in each of the different 1,000 word lists in the 10 texts. We should put much emphasis on the column of cumulative token because the percentage indicates whether the words in texts are within students' comprehension and how well the students understand the vocabulary in the text. Since the 10 texts are all from the same story, the proportion of the 4 items mentioned above in each text is almost the same. Some of the chapters will be selected and analyzed emphatically. The output clearly shows the relative importance of knowing the most common words. Over 85% of the words (5,924) are from the 1,000 word list, 3.48% of the words (241) are from the 2,000 word list. Thus, if the most frequent 2,000 word families are known to readers, they would know 6,165 of the 6,935 tokens in the text. Many researchers have taken the approach that proper nouns have a minimal learning burden and may be easily understood by readers (Webb & Nation). Thus we can add the proper nouns tokens in List 15 to the total number of tokens known to the readers who know the most frequent 2,000 words, then 6,173 of the tokens would be known leaving 762 unknown words. As mentioned above, if the percentage of tokens known to the learners reaches 95% ~ 98% coverage, the text will be suitable for classroom use without any assistance. However, in this case only 89.02% of the tokens would be known to learners with a vocabulary size of the most frequent 2,000 words. Thus this indicates that it is too difficult for learners to read and understand texts without any assistance from the teachers and dictionaries. Another way is to analyze the results of word families because similarity in forms and meanings of tokens from the same family may reduce the difficulty of word family learning. For example, the output shows that there are 33 tokens which occur in the 7th 1,000 word list, but these tokens only consist of 4 word families. From the perspective of the number of word families which may be unknown for learners with a vocabulary size of the most frequent 3,000 words, we find that there are 63 different word families. Therefore, if the vocabulary size of 2,000 word families is scored, there would be a maximum of 63 unknown word families in the text. The Compleat Lexical Tutor helps teachers know the vocabulary which may be unknown to their students.

2.3 Modify the text to make it more appropriate for students

Through the help of Compleat Lexical Tutor, it indicates that it is very difficult for learners to read the text without any assistance from the teacher and a dictionary. Thus it should be modified to make it more suitable for students to read. One way is to paraphrase into words known to the students, which means using the words within the most frequent 2,000 to substitute for those out of the list. We can select one text for example. Figure 2 is the output of distribution of tokens, word types and word families of the text *(Alice in Wonderland*: Chapter 2: The pool of tears) across each of the different 1,000 word lists. We can find that there are 20 tokens out of the most frequent 2,000 list plus 14 tokens classified as "Off-List". And Figure 3 shows those words. Firstly, it can be noted that a large proportion of "Off-List" words are mood particle, such as oh, ohhh, ouch and names such as Alice, they are already known by students and may be easily understood by readers. Now we can use some words to take place of them: fade → disappear, leather → animal skin, narrator → person who tells the story, gloves → clothing that wear on hand to keep warm, shrink → become small, splendidly → beautifully, duchess → a woman with the highest social rank, patter → hit, dodo → a large bird that no longer exist. We can clearly see that it is easy to replace the verb, adjective, adverbial and put new words to the text directly. However, in terms of noun, it's difficult to find an identical word that corresponds to the original one. What teachers can do is to give annotation to nouns or L1 translations when it is necessary, thereby allowing students to read the text relatively quickly by checking the relevant definitions. In general, replacement and annotation should be used together depending on different types of words.

By paraphrasing and diminishing the "Off-List" words, the percentage of cumulative tokens in the most frequent 2,000 list have swelled apparently, reaching 96.3%, as shown in Figure 4. In that case, the text would be suitable for classroom reading without any assistance and the remaining chapters can be modified using this method.

Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	183 (89.27)	219 (85.55)	570 (85.84)	85.84
K-2 Words :	13 (6.34)	15 (5.86)	23 (3.46)	89.30
K-3 Words :	3 (1.46)	3 (1.17)	13 (1.96)	91.26
K-4 Words :	2 (0.98)	2 (0.78)	3 (0.45)	91.71
K-5 Words :	1 (0.49)	1 (0.39)	1 (0.15)	91.86
K-6 Words :				
K-7 Words :	1 (0.49)	1 (0.39)	1 (0.15)	92.01
K-8 Words :				
K-9 Words :				
K-10 Words :	1 (0.49)	1 (0.39)	1 (0.15)	92.16
K-11 Words :				
K-12 Words :				
K-13 Words :				
K-14 Words :				
K-15 Words :	1 (0.49)	1 (0.39)	1 (0.15)	92.31
K-16 Words :				
K-17 Words :				
K-18 Words :				
K-19 Words :				
K-20 Words :				
K-21 Words :				
K-22 Words :				
K-23 Words :				
K-24 Words :				
K-25 Words :				
Off-List:	?	14 (5.47)	51 (7.68)	99.99
Total (unrounded)	205	256(100)	664 (100)	≈100.00

Figure 2. The distribution of the various levels of vocabulary in Chapter 2

BNC-COCA-3,000 types: [fams 3 : types 3 : tokens 13]
fade_[1] leather_[1] narrator_[11]
BNC-COCA-4,000 types: [fams 2 : types 2 : tokens 3]
gloves_[1] shrinking_[2]
BNC-COCA-5,000 types: [fams 1 : types 1 : tokens 1]
splendidly_[1]
BNC-COCA-7,000 types: [fams 1 : types 1 : tokens 1]
duchess_[1]
BNC-COCA-10,000 types: [fams 1 : types 1 : tokens 1]
pattering_[1]
BNC-COCA-15,000 types: [fams 1 : types 1 : tokens 1]
dodo_[1]
OFFLIST: [?: types 14 : tokens 51]
alice_[24] curiouser_[2] filledwith_[1] oh_[15] ohhh_[1] ouch_[1]

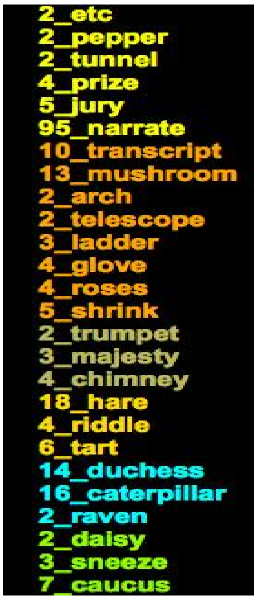
Figure 3. Tokens that out of the most frequent 2,000 list in Chapter 2

Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	187 (89.90)	226 (88.63)	599 (92.44)	92.44
K-2 Words :	14 (6.73)	16 (6.27)	25 (3.86)	96.30
K-3 Words :	3 (1.44)	3 (1.18)	13 (2.01)	98.31
K-4 Words :	1 (0.48)	1 (0.39)	1 (0.15)	98.46
K-5 Words :				
K-6 Words :				
K-7 Words :	1 (0.48)	1 (0.39)	1 (0.15)	98.61
K-8 Words :	1 (0.48)	1 (0.39)	1 (0.15)	98.76
K-9 Words :				
K-10 Words :				
K-11 Words :				
K-12 Words :				
K-13 Words :				
K-14 Words :				
K-15 Words :	1 (0.48)	1 (0.39)	1 (0.15)	98.91
K-16 Words :				
K-17 Words :				
K-18 Words :				
K-19 Words :				
K-20 Words :				

K-21 Words :				
K-22 Words :				
K-23 Words :				
K-24 Words :				
K-25 Words :				
Off-List:	?	7 (2.75)	7 (1.08)	99.99
Total (unrounded)	208	255 (100)	648 (100)	≈100.00

Figure 4. The distribution of the various levels of vocabulary after modified in Chapter 2

However, another way is teaching the words that are out of the 2,000 list in advance. In that case, it will decrease the learning burden of comprehending the whole text. The method is suitable for the text where unknown words are encountered multiple times. Figure 5 shows the number of occurrences of unknown words in 10 texts. For example, the words "narrate", "transcript", "mushroom", "duchess", "caterpillar" have a high number of occurrences. The greater the number of encounters, the more likely learning is to occur. It has been shown that repetition similarly affects the incidental learning, which means focusing on a specific topic several times will probably increase repetition, so that students will make a smaller contribution with each subsequent repetition and gradually add them to lexical knowledge [1][8][10].



2_etc
2_pepper
2_tunnel
4_prize
5_jury
95_narrate
10_transcript
13_mushroom
2_arch
2_telescope
3_ladder
4_glove
4_roses
5_shrink
2_trumpet
3_majesty
4_chimney
18_hare
4_riddle
6_tart
14_duchess
16_caterpillar
2_raven
2_daisy
3_sneeze
7_caucus

Figure 5. The encounter times of the unknown words in the 10 texts

3 Conclusion

Determining whether texts are appropriate for specific learners can be a difficult and important task for teachers. The procedure for evaluating a text can be: 1) measuring vocabulary size using a test such as the Vocabulary Levels Test or Vocabulary Size Test; 2) analyzing text with the quick and easy Compleat Lexical Tutor, simply eliminating the text from consideration may often be the best solution if learners do not have sufficient vocabulary to understand the words in the text; 3) analyzing the text's distribution of tokens, word types and word families in each of the different 1,000 word lists. This will involve going through not in the list and reclassifying hyphenated words and proper words, and then calculating the cumulative coverage. The cumulative coverage in relation to vocabulary size will provide some indication of whether students can understand a text and to what degree they can understand the vocabulary in the text. Comparing with the standard of 95% and 98%, teachers can make choices based on the following: 1) If there are a small number of unknown word tokens, teachers can use other common words to paraphrase, but it is limited to verb, adjective, adverbial. It is better to give glossing or L1 translation to nouns because they are not easy to be replaced. 2) If there are a large number of unknown words, teachers may pre-teach those items that appear several times in the text thus incidental learning will occur when learners read these materials. Webb and Nation also found other methods that necessary for teachers to do: ensuring the dictionaries are available and giving learners time to look up the dictionaries, simplifying the text or choosing a more appropriate one.

Conflicts of interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Durrant P, Schmitt, N. 2010. Adult learners' retention of collocations from exposure. *Second language Research*, 26(2):163-188.

[2] Laufer B. 1989. What percentage of text lexis is essential for comprehension? In C. Lauren.&M.Nordman (Eds), *Special Language: From Humans Thinking to Thinking Machine*s, 316-323. Clevedon, UK: Multilingual Matters.

[3] Liu N, Nation ISP. 1985. Factors affecting guessing vocabulary in context. *RELC Journal*,16: 33-42.

[4] Nation I.S.P, Beglar D. 2007. A vocabulary size test. *The Language Teacher*, 31(7): 9-13.

[5] Nation I.S.P, Healthy A. 2002. Range: A program for the analysis of vocabulary in texts. http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx

[6] Nation I.S.P. 2001. *Learning Vocabulary in Another Language*. UK: Cambridge University Press.

[7] Nation I.S.P. 2006. How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1): 59-82.

[8] Sonbul S, Schmitt N. 2013. Explicit and implicit lexical knowledge: acquisition of collocations under different input conditions. *Language Leaning*, 63(1): 121-159.

[9] Webb S, Rodgers MPH. 2009a. Vocabulary demands of television programs. *Language Learning*, 59(2): 335-366.

[10] Webb S, Newton J, Chang ACS. 2013. Incidental learning of collocation. *Language Learning*, 63(1): 91-120.

[11] Webb S, Nation I.S.P. 2017. How vocabulary is learned. *Conditions Contributing to Vocabulary Learning*, 65-67.

[12] Webb S, Nation I.S.P. 2008. Evaluating the vocabulary load of written text. *TESOLANZ Journal*, 16:1-10.