Available online at www.sciencedirect.com

ScienceDirect

System 41 (2013) 1043–1055

SYSTEM

www.elsevier.com/locate/system

Types of words identified as unknown by L2 learners when reading

Dale Brown*

Graduate School of Language and Culture, Osaka University, Machikaneyama-cho 1–8, Toyonaka-shi 560-0043, Japan

Received 20 February 2013; revised 23 July 2013; accepted 7 October 2013
Available online 25 October 2013

Abstract

In determining which words are likely to cause problems for learners in reading, the computer-based lexical profiling of texts has become routine. This study investigates the nature of items marked as unknown by two groups of learners ($n = 46$) when reading, with reference to the assumptions behind lexical profiling. The first assumption, that less frequent items are likely to be unknown, is supported by the results in that significantly more low frequency words were marked as unknown. The second assumption, regarding the use of the word family as the unit of counting for lexical profiling, is shown to be problematic. A significantly greater proportion of the higher frequency words marked were found to be inflected or derived forms. The third assumption, that few problems stem from the fact that computers can only recognise strings of characters, may be warranted. Relatively few of the higher frequency words that were marked occurred in the reading texts in ways likely to be unfamiliar to the participants. The study thus concludes that in using computer-based profiling of texts to judge which words cause problems for learners, the primary issue is the use of the word family as the unit of counting.

Keywords: Vocabulary; Reading; Text coverage; Lexical profiling; Word lists; Word families; Frequency effects; Word frequency

1. Introduction

In considering the types of words which cause problems for learners when reading in a second language, perhaps the commonsense response is that words which are less frequent in the language are the primary problem. This line of thinking can be easily operationalised today thanks to the development of lexical profiling. Lexical profiling involves examining the frequency level of each word in a text, allowing the less frequent items to be quickly identified. Lexical profiling can be conducted using freely available, simple tools such as the computer programs Range (Nation & Heatley, 2002) and AntWordProfiler (Anthony, 2013), with Cobb (2010) reporting that his online version of Range (www.lextutor.ca/vp/) has attracted pedagogical users from all over the world.

A body of research provides the basis for the lexical profiling of texts. One strand of this research focuses on the development of word frequency lists based on large corpora (Coxhead, 2000; Nation, 2004, 2006a). This research, building on earlier work culminating in West's (1953) *General Service List of English Words*, aims to identify the most important and useful words in the language. A second strand involves text coverage, that is the proportion of a text which is accounted for by a certain number of words. This research has found that texts of different genres vary, but the

* Tel.: +81668505897.
E-mail address: dbrown@lang.osaka-u.ac.jp.

0346-251X/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.system.2013.10.013

most frequent words always make up a large proportion of a text, with, for example, the top 2,000 word families accounting for between 80 and 90 percent of many texts (Nation, 2006b; Webb & Rodgers, 2009a, 2009b). This work reveals how much vocabulary is required for dealing with different texts and thus assists in establishing vocabulary learning goals. A third strand looks at the amount of text coverage required for successful comprehension of texts (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Schmitt et al., 2011). These studies have found that high levels of coverage are necessary, a figure of 98 percent being recommended.

Lexical profiling and the findings reported above provide teachers and materials writers with tremendously useful information. Words in a text that are of lower frequency can be quickly identified, and the overall difficulty of a text estimated. These are very useful procedures, but in applying them it must be recognised that several assumptions are being made. This study looks at the types of words which learners themselves identify as unknown in reading texts and in doing so examines these assumptions. Viewed through the practice of lexical profiling, there are four possible sources of difficulty for learners.

The first possibility is the central idea behind lexical profiling, that low frequency items are the main source of difficulty. This is based on the idea that more frequent words are learnt earlier, which Milton (2007, 2009) describes as "the frequency model of lexical learning" (2007, p. 48). The central place of frequency in language learning generally has been recognised in recent times (see Ellis (2002) and the responses to it), and with specific reference to vocabulary, Milton (2009) notes that Palmer (1917) had observed the frequency–difficulty link almost a century ago. Until recently the idea that vocabulary learning follows frequency had been widely accepted even though there was little empirical evidence for it. In recent years, however, the frequency model has been tested, with a number of studies (Aizawa, 2006; Brown, 2012; Milton, 2007) finding it to be robust for L2 learners in that knowledge of words is greatest at the highest frequency level and declines at subsequent levels. Milton (2009) suggests that the frequency effect "is so powerful that word difficulty features [such as degree of cognateness and word length], commonly accepted as influential in determining whether or not a word will be learned, fail to significantly impact on this effect" (p. 242).

A second possibility concerns gaps in learners' knowledge of high frequency items. As just discussed, frequency is a strong driver of learning when looking at the lexicon as a whole, but it is not the only factor and there are inevitably some gaps in learners' knowledge, while there are also some indications that the vocabulary knowledge of some individual learners may not follow the frequency model (Booth, 2013; Brown, 2012; Milton, 2007). Furthermore, Browne and Culligan (2008) report that many Japanese EFL learners, the participants in this study, display considerable gaps in their knowledge of high frequency vocabulary in comparison with their overall vocabulary knowledge. It must also be recognised that because high frequency items comprise such a large proportion of texts, even small gaps in learners' knowledge of these items could mean that many words encountered in a text are unknown. To illustrate, suppose that a text is 500 words long, and that the first 1,000 words in a frequency list make up 80 percent of the text, i.e. 400 of the words, while the third 1,000 make up 4 percent of the text, i.e. 20 words. A learner with a 5 percent gap in their knowledge of the first 1,000 words would, on average, lack knowledge of 20 out of the 500 words in the text. A gap of 5 percent in their knowledge of the third 1,000 words, however, amounts to just 1 unknown word. Thus gaps in learners' knowledge of the most frequent words could have a large impact on the overall level of vocabulary knowledge they bring to a text.

A third possibility arises as a result of the word family being adopted as the most appropriate unit in work on lexical profiling. A word family consists of a free-standing base form (e.g. *read*) and a range of affixed forms, including both inflected forms (e.g. *reads*, *reading*) and derived forms (e.g. *reader*, *readable*, *misread*) (Bauer & Nation, 1993). Word families can be defined more or less inclusively, that is allowing a greater or more restricted range of affixes. The word lists featuring in much of the recent work on lexical profiling contain word families at level 6 of Bauer and Nation's scheme. This level permits a range of affixes, but there remain affixes which are excluded and permitted affixes must operate in prescribed ways regarding regularity and transparency. The third possibility then is that the use of word families in the profiling of texts and estimation of difficulty may be problematic. Nation (2006b) states that "the assumption that lies behind the idea of word families is that when reading and listening, a learner who knows at least one of the members of a family well could understand other family members by using knowledge of the most common and regular of the English word-building devices" (p. 67). In support of this, Nation cites L1 research (Nagy et al., 1989) that shows the word family to be a psychologically real unit. Schmitt and Zimmerman's (2002) study of productive knowledge of derivations also offers some support for this position in that it found that learners could produce derived forms for some words which they apparently did not know. This would seem to indicate that word-building processes were being applied, although it should be noted that the learners were of advanced proficiency and