Reading in a Foreign Language October 2010, Volume 22, No. 2
ISSN 1539-0578 pp. 355–361

An improper assumption? The treatment of proper nouns in text coverage counts

Dale Brown
Nanzan University
Japan

The calculation of text coverage, that is the proportion of a text covered by a vocabulary of a given size, has become a standardized procedure in recent years. Such calculations provide important information for research and pedagogy about issues such as the goals of vocabulary learning and the ease or difficulty of particular texts. Chujo and Utiyama (2005) have pointed out that it is important to ensure that methodological issues involved in such calculations are properly addressed, and their research delineates the vocabulary size, sample size and text length needed to secure reliable results. However, one methodological issue that remains open to debate is how to deal with proper nouns.

In studies of text coverage it has become standard practice to assume that proper nouns cause no problems for second language (L2) readers. Recent studies include comments such as, "assuming that proper nouns are easily understood," (Nation, 2006, p. 70); "if we . . . assume that the proper nouns in the discourse are known," (Schmitt, 2008, p. 330); and "if we assume that proper nouns . . . have a minimal learning burden" (Webb & Rodgers, 2009a, p. 345). In fairness, there is occasional recognition that proper nouns can be problematic for learners and that it is essential that learners are able to recognize them (Webb & Rodgers, 2009b). Nevertheless, the general tendency is clear: These papers acknowledge the assumption, but proceed with minimal discussion or explanation of the reasons for making the assumption. Webb (2010), for example, explains the inclusion of proper nouns in his coverage figures simply by referring to Nation (2006) and his own previous papers (Webb & Rodgers, 2009a, 2009b), which shows how the assumption has become part of the procedure of calculating text coverage. Chujo and Utiyama's (2005) paper emphasizes how standard this practice has become as they state that "all proper nouns . . . were excluded *since these are usually excluded* [italics added] from source data" (p. 5). In recent years then, proper nouns are excluded simply because they have historically been done so.

This paper questions the way proper nouns are dealt with in text coverage calculations. After first looking at the various ways proper nouns are actually treated, the paper examines in detail the assumption that proper nouns are unproblematic for L2 readers. The roots of the assumption are traced back and examined, before a number of other issues that cast doubt on it are raised. The paper concludes with some questions that arise as a result.

While the assumption that proper nouns are unproblematic is consistently maintained in text coverage studies, the actual treatment of proper nouns varies. Chujo and Utiyama (2005) exclude

proper nouns from their data by simply deleting them. Nation (2006) and Webb and Rodgers (2009a, 2009b), on the other hand, integrate an open-ended list of proper nouns into their data analyses by using the Range software (Nation & Heatley, 2002) to produce a profile of the vocabulary in a text. They thus leave the proper nouns in the data and calculate a coverage figure for them, which is then added to the base level(s) being considered. VocabProfile (Cobb), a web-based version of Range, gives two options for dealing with proper nouns. First, users can identify words as proper nouns and have the program count these items as 1K words, that is, the first thousand words of the General Service List (West, 1953). Second, users can opt to use an algorithm that treats all words with an initial capital letter appearing anywhere other than at the beginning of a sentence as 1K items, the assumption being that these are proper nouns (Cobb, 2010). While the differences between these approaches may seem minor, they do have an impact on the results. Table 1 illustrates the differing results obtained using the first three methods explained above. Of these approaches, only that taken by Nation and by Webb and Rodgers makes the position of proper nouns in the data visible. Even in these studies, however, the subsequent discussion focuses on the total coverage figure.

Table 1
Three approaches to the treatment of proper nouns in text coverage calculations, using a concocted example text of 100 words containing 85 words from the 1K level, 4 proper nouns and 11 other words.

	1K words	Proper nouns	Beyond 1K words	Text coverage at the 1K level
Proper nouns deleted from data (*Chujo & Utiyama, 2005*)	85	0	11	85/96*100 = 88.5%
Proper nouns coverage figure calculated separately then added to base level (*Nation, 2006; Webb & Rodgers, 2009a, 2009b*)	85	4	11	(85/100*100) + (4/100*100) = 85% + 4% = 89%
Proper nouns counted as 1K words (an option in VocabProfile)	89	0	11	89/100*100 = 89%

Though most recent text coverage studies simply follow the assumption that proper nouns cause no problems for learners, earlier studies did give some reasoning for the assumption. Nation and Wang (1999) explained:

> A list was made of all the proper nouns used in the graded readers. These were isolated from other words because it was considered that most of these words did not require prior vocabulary learning. That is, proper nouns could be easily understood from context and should not be counted as unknown vocabulary. Most of the proper nouns were first names (John, Colin, Julie, Carol), family names (Bligh, Jones) and places (Staines, Hollywood, London). (p. 358)

Going further back, Hirsh and Nation (1992) make a similar point, along with one other:

> There are strong reasons for considering proper nouns as words that do not require previous learning. First, the text reveals what we need to know about them as the story progresses. Who *Alice*, the *Dormouse* and the *Mad Hatter* are is revealed by the story.

> We are not expected to know this before coming to the story. Second, their form (an initial capital letter) and their function clearly signal they are proper nouns. (p. 691)

Both of these points are reasonable to an extent. The proper nouns in a text do not require previous learning. However, previous experience with proper nouns is undoubtedly of help, even basic knowledge such as that John and Colin are male while Julie and Carol are female. Proper nouns do still place some learning burden on the reader and readers must often be willing to defer their desire to know what the items refer to until later. Hirsh and Nation's (1992) first point also depends somewhat on the second, that proper nouns can easily be recognized as such. Allerton (1987) demonstrates the range and variability in the types and forms of proper nouns, including instances of proper nouns that are not usually capitalized as well as non-proper nouns that are frequently capitalized (adjectives related to country names, such as *English*, being perhaps the most prominent). Kobeleva (2008) has investigated learners' recognition of proper nouns in listening and found that proper nouns are often missed and mistaken for common expressions and vice versa: For example, in one of her experiments learners often identified unknown words as proper nouns. In reading, the initial capital letter, in most cases, and the less transitory nature of the input no doubt make recognition easier. Experienced readers may immediately recognize a proper noun when they see one. With less experienced readers, however, this may not always be the case.

In addition to these points, two further issues cast doubt on the assumption. First, while proper nouns do not have meanings in the conventional sense, they do have connotations and associations (Allerton, 1987; Van Langendonck, 2007). *Paris* is not just a city; it is a city that many people associate with fashion and culture. *Everest* is not just any mountain. Allerton describes a gradation of proper nouns in terms of meaning. This is indicated by the fact that some proper nouns, such as country names, are often translated into other languages (e.g., the English *Germany* becomes *Allemagne* in French and *Saksa* in Finnish), while others, such as personal names, are not.

A related problem is what Allerton calls mixed proper nouns, that is multi-word proper nouns which include a common noun element such as *President Obama* or the *Suez Canal*. Because of the capital letter on the common noun element, learners may consider it to be simply part of the name, and the meaning of it may be missed entirely. Particularly problematic may be words used in personal titles such as *Captain*, *Governor* and *Saint* as the status they confer may go unnoticed by the learner. Also potentially problematic for learners are proper nouns that happen to have the same form as regular words. Surnames are a good example. Of the 100 most common surnames in the 2000 US census, 24 appear as words in Nation's (2006) word lists, 18 of which are in the first 1,000 word families (e.g., *Cook*, *White* and *Young*).[1]

The second issue is that unfamiliar proper nouns can interrupt the flow of reading. Unfamiliar proper nouns may cause problems for two reading sub-processes: word recognition and phonological decoding. Fast and effortless word recognition is vital for reading, due to the limited nature of working memory (Koda, 2005; Grabe, 2009). If word recognition is not automated, precious working memory resources must be devoted to it, leaving insufficient resources available for the numerous other concurrent sub-processes involved in reading. We can thus suppose that when a learner comes across an unfamiliar proper noun, normal word

recognition can break down, causing working memory resources to be diverted to deal with the problem. Phonological decoding is also an important reading process (Koda, 2005; Grabe, 2009). The problem here is that English proper nouns are often difficult to decode because many are phonetically irregular and the spelling and pronunciation of names is less systematic than of ordinary words (Carney, 1994). Looking at Nation and Wang's (1999) examples in the quote given above, proper nouns such as *John* and *Staines* may cause phonological decoding difficulties for learners who have not encountered them previously. It seems likely then that proper nouns, through their combined impact on these two sub-processes, cause interruptions to the flow of reading.

Some recent research seems to offer evidence of these problems. Kobeleva (2008) compared the listening comprehension of a news story in a names known condition (all proper nouns familiar before listening) and a names unknown condition (all proper nouns unfamiliar) and found significantly lower levels of comprehension in the latter condition along with an increase in the participants' rating of the difficulty of the task. Erten and Razi (2009) looked at the effects of nativizing a story on reading comprehension. A major part of the nativization process involved replacing the proper nouns in the story with ones from the home culture of the participants, in this case Turkish. For example, a character name was changed from *Frances* to *Özlem* and *Fifth Avenue* became *Kordonboyu*. Erten and Razi compared the reading comprehension of learners reading the story in its original version and in the nativized version, finding superior comprehension for those who read the nativized version. They attribute this superior comprehension to improved motivation and to reduction in the cognitive load. While this research was not focused on proper nouns, and nativization did involve other changes to the text, it is suggestive of the impact that proper nouns have.

I have argued that there are reasons to doubt the assumption that proper nouns are unproblematic for learners and can be treated as known items. However, it probably is true that learners deal with proper nouns more easily than other unknown words. The problem is that, as Allerton (1987) says, proper nouns "fall partly inside and partly outside the lexicon" (p. 62). As noted earlier, there is a gradation, from those at one extreme which are most definitely part of English, to those at the other, which most certainly are not. Thus some proper nouns will be unproblematic for learners, but some may cause considerable problems. If this argument is accepted, empirical research is needed into a number of questions.

Question 1: How good are learners at recognizing proper nouns as such?
As mentioned above, Kobeleva (2008) found that learners frequently mistake proper nouns for regular expressions and vice versa in listening. It seems likely that in reading, the answer to this question may vary with proficiency; it may also vary with the language background of the learners. Regarding first language processing, Coates (2006) has suggested, "the default interpretation for any linguistic string is a proper name" (p. 371). In other words, until we receive information otherwise, every linguistic string we encounter is assumed to be a proper noun. Does this process also operate in a second language? For learners, encountering unfamiliar linguistic strings is an all-too-familiar experience, so it may seem unlikely, but we do not have evidence either way.

Question 2: How do learners deal with proper nouns when reading?

Do unfamiliar proper nouns interrupt fluent reading, as I have suggested? What is the effect on comprehension of ensuring that all proper nouns are known or nativizing only the proper nouns in a text? Would it be wise for teachers to introduce the important proper nouns in a text before reading as is often done with key items of vocabulary?

Question 3: Should we reconsider the established figures concerning the vocabulary coverage believed to be necessary for reading?
The figures of 95% (Laufer, 1989), and 98% (Hu & Nation, 2000), have been widely quoted and have become benchmarks in assessing the readability of texts and calculating vocabulary learning goals. But how did these studies deal with proper nouns? Laufer's study makes no mention of them, making an evaluation of it impossible, but Hu and Nation were much clearer: The proper nouns in the text were left as they were, and the 98% figure includes the proper nouns as assumed known items. There are, however, in the light of the discussion so far, two issues with their study. First, the participants were of good proficiency: among the most proficient learners on a pre-sessional course at a university in an English-speaking country. Second, the single text used contained a small number of proper nouns. The proper nouns, by Hu and Nation's count, accounted for just 1.6% of the text, as opposed to typical levels of coverage for proper nouns in written text of 2–4% according to Nation (2006), or 4–5%, based on figures in Francis and Kucera (1982) and Johansson and Hofland (1989). Research with multiple texts containing more typical numbers of proper nouns and including less proficient learners would be helpful.

As the importance of vocabulary to reading becomes ever more recognized, it seems likely that calculating text coverage will become a familiar procedure to more and more educators. Indeed, Cobb (2010) reports that a large number of teachers and researchers around the world use his VocabProfile software online. We must then ensure that the methodology in place is sound. We should not just follow a procedure because that is the way it has been done until now. We must carefully consider the assumptions we make. Cobb (2010) states, "it is difficult to explain to novice Vocabprofilers that proper nouns are not lexical items" (p. 187). I would suggest that this difficulty arises not because of any naivety among users of the software, but rather from their intuition that proper nouns are not as simple as researchers sometimes assume. This paper has suggested three questions for empirical research regarding how learners deal with proper nouns in reading and the consequences of this. My hope is that we may come to better understand the processes and become able to more confidently use the valuable information text coverage studies can provide.

Notes

1. The existence of mixed proper nouns and proper noun homonyms of regular words also means that the proportion of proper nouns is underestimated in computer-based calculations. Computers cannot "see" the common noun element of neither mixed proper nouns nor proper noun homonyms of regular words, and thus count both as regular words.

Acknowledgements

I would like to thank Rory Rosszell and the anonymous reviewer for their comments on earlier versions of this discussion.

References

Allerton, D. J. (1987). The linguistic and sociolinguistic status of proper names: What are they, and who do they belong to? *Journal of Pragmatics*, *12*, 61–92.

Carney, E. (1994). *A survey of English spelling*. London: Routledge.

Chujo, K., & Utiyama, M. (2005). Understanding the role of text length, sample size and vocabulary size in determining text coverage. *Reading in a Foreign Language*, *17*, 1–22.

Coates, R. (2006). Properhood. *Language*, *82*, 356–382.

Cobb, T. (n.d.). Web Vocabprofile [Computer software]. Accessed 15 June 2009 from http://www.lextutor.ca/vp/

Cobb, T. (2010). Learning about language and learners from computer programs. *Reading in a Foreign Language*, *22*, 181–200.

Erten, I. H., & Razi, S. (2009). The effects of cultural familiarity on reading comprehension. *Reading in a Foreign Language*, *21*, 60–77.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.

Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage*. Boston: Houghton Mifflin Company.

Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, *8*, 689–696.

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

Johansson, S., & Hofland, K. (1989). *Frequency analysis of English vocabulary and grammar*. Oxford: Clarendon Press.

Kobeleva, P. P. (2008). *The impact of unfamiliar proper names on ESL learners' listening comprehension*. Unpublished doctoral dissertation, Victoria University of Wellington, New Zealand.

Koda, K. (2005). *Insights into second language reading*. Cambridge: Cambridge University Press.

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 machines* (pp. 316–323). Clevedon: Multilingual Matters.

Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*, 59–82.

Nation, I. S. P., & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts [Computer software]. Available at Victoria University Web site, http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Nation, I.S.P., & Wang, K. (1999). Graded readers and vocabulary. *Reading in a Foreign Language*, *12*, 355–380.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*, 329–363.
Van Langendonck, W. (2007). *Theory and typology of proper names*. Berlin: Mouton de Gruyter.
Webb, S. (2010). Using glossaries to increase the lexical coverage of television programs. *Reading in a Foreign Language*, *22*, 201–221.
Webb, S., & Rodgers, M. P. H. (2009a). Vocabulary demands of television programs. *Language Learning*, *59*, 335–366.
Webb, S., & Rodgers, M. P. H. (2009b). The lexical coverage of movies. *Applied Linguistics*, *30*, 407–427.
West, M. (1953). *A general service list of English words*. London: Longman, Green and Co.

About the Author

Dale Brown teaches at Nanzan University in Nagoya, Japan. He is interested in vocabulary learning and teaching, the analysis of teaching materials and extensive reading. Email: dbrown@nanzan-u.ac.jp