Investigating the Lexical Profile of a Popular Children's Storybook Series

Aaron C. SPONSELLER
Institute for the Promotion of Global Education
Hiroshima University

The Family Reading Project (FRP), initiated in 2016, trains Japanese parents to read English language storybooks to their elementary schoolers (Ferguson, Sponseller, & Yamada, 2017). One challenge the FRP has faced is selecting storybooks that are not only interesting in terms of content but lexically within the reach of the participating parents. At present, no comprehensive corpus of children's texts or principled manner for measuring the difficulty of a text for children exists. The primary aim of this research was to begin developing and refining a principled approach to analyzing candidate texts for inclusion in the FRP. A principled approach is needed if the FRP wishes to make lexically-informed book selections for future iterations of the program. Fifty-five titles from the popular children's storybook series The Berenstain Bears were digitized in order to create a miniature corpus. The lexical profiles for the series overall as well as individual titles that were generated and investigated are discussed, and implications for evaluating potential texts for use in the FRP are briefly discussed.

BACKGROUND

The Family Reading Project (FRP) Underway in Osaka

The FRP began at an elementary school in Osaka in January, 2016 (Ferguson, Sponseller, & Yamada, 2017). The project involves preparing parents of 6-7 year olds to engage their children in shared reading of English language children's books. This project was born from my interest in family literacy programs I was involved with in the United States.

'Family Literacy' is something of an umbrella term covering a wide variety of educational objectives across very diverse communities. Most typically these programs are found in low socioeconomic status (SES) communities, frequently with substantial populations of immigrants. In practice, these programs can range from simply offering survival English courses to local members of the community, all the way to civic ESL or even ESP. This project is not a full family literacy program such as those found in the US and other nations with substantial immigrant populations; instead, it focuses entirely on encouraging Japanese parents to read with their children in English in the home.

Positive English outcomes for the children are, of course, the endgame. While most studies on the benefits of parent-child dialogic reading have been conducted in English L1 contexts, there is also evidence of Spanish L1 mothers reading to their children in English (Dever & Burts, 2002; Jordan et al., 2000; Shanahan et al., 1995) with successful outcomes. Similar outcomes have been shown in Taiwan (Wu & Honig, 2010) and Singapore (Yeo, Ong, & Ng, 2014). In their meta analysis of 33 articles stretching all the way back to 1960, Bus, Ijzendoorn, and Pellegrini (1995) concluded that shared book reading has positive effects.

The research conducted on these programs has come primarily from the US so far. Many of the studies actually look at low SES or at-risk communities where English is the L1 (Barbour, 1998; Bus, Ijzendoorn, & Pellegrini, 1995; Dever & Burts, 2002; Dickinson & De Temple, 1998; Hindman & Morrison, 2012; Jacobs et al., 2011; Sénéchal & LeFevre, 2002; Sénéchal, LeFevre, Thomas, & Daley, 1998; Weigel, Martin, & Bennett, 2005, 2006). Some studies have been conducted within family literacy programs and sampled at-risk populations for whom English was not the L1. These are typically immigrant communities in rural or inner-city regions (Jordan et al., 2000; Shanahan, Mulhern, & Rodriguez-Brown, 1995).

The Japanese context seems ideal for implementing a family reading program. Overall, the population is well-educated and literate in their L1. Many Japanese mothers are housewives and might be in a position to choose to spend time engaging in an FRP, and parents of young children in today's Japan almost certainly went through six years of English education between junior and senior high school. This does not mean they can have fluent conversations in English, but reading children's stories might be within reach.

The FRP has had several successes so far. Parents have realized they are not supposed to explicitly teach their children to read, but to enjoy the storytelling together and foster their children's interest in English stories and words. Several children have certainly expressed such interest. The primary challenge the FRP has faced thus far is selecting appropriate texts. Some texts have been too long. Some texts appear simple but actually contain vocabulary that is unknown to many of the parents. Onomatopoeia is also very common in children's storybooks, but these are the kinds of words that are a.) not likely taught in English classes in Japan, and b.) not at all intuitive without very clear contextual clues. The FRP has supported these parents through face-to-face training sessions, distribution of glosses for most books, and video recordings of instructors reading the books hosted on a private YouTube channel. However, a primary goal as the FRP continues to grow is to establish a method of identifying ideal texts that may rely more upon the lexical knowledge these parents already possess.

Research on the Use of Storybooks and Vocabulary Growth

Substantial research on the impact of parent-child dialogic reading (DR) in the L1 has consistently found the activity highly beneficial. Senechal, LeFevre, Hudson, and Lawson (1996) found that parents' familiarity with children's books predicted their child's receptive vocabulary knowledge, and that their child's knowledge of those books predicted both receptive and productive vocabulary. Among French speaking Canadians, Senechal (2006) has shown that storybook exposure at home was a hugely significant predictor of children's vocabulary level at kindergarten level. DR involves a greater amount of interaction between reader and child, and studies have shown DR training, often delivered through community or school-based intervention programs like the FRP, can have immense and long-lasting benefits on literacy overall (Huebner & Meltzoff, 2005; Huebner & Payne, 2010). Such studies on DR interventions exist outside of the English speaking world as well. A DR intervention aimed at preschoolers in rural Bangladesh demonstrated significant vocabulary gains for the treatment (DR) group over a control group that received prototypical storytelling lessons (Opel, Ameer, & Aboud, 2009). This study appears to have been a partial replication of Hargrave and Senechal (2000), who found nearly identical results in Ottawa, Canada.

Studies on the use of storybooks in L2 contexts exist as well. Elley's (1991; 2000) 'book flood' studies involved L2 learners reading illustrated kids' books for 1-3 years. These studies, conducted primarily in

developing countries, have demonstrated impressive gains in reading, speaking, writing, and vocabulary.

> After only six months on the project, reading comprehension and vocabulary tests produced gains which were three times as great as those of the control groups. The impact was consistent across both grade levels, and in both urban and rural schools... The "effect sizes" ... clustered around 1.00, which represents a very strong impact. The sample size in each grade was over 600 pupils. (Elley, 2000, p.242)

In the Japanese context, however, little has been done at the primary school level. Uchiyama (2011) looked at the style of reading English storybooks in the Japanese primary school classroom, and concluded that both traditional reading and "character imagery" (e.g. the teacher/reader dresses up as a story character and/or acts the part of the protagonist) work well in the primary classroom setting. Overall, the use of storybooks and DR for English language development remains underexplored in Japan.

The Case for Using a Series for Vocabulary

Book series that feature the same characters in each volume are potentially very good for inclusion in a program in which language acquisition is a primary goal. Ostensibly, once a character is known to the reader, their name no longer presents an obstacle to understanding the story. Moreover, the more volumes one has read in the series, the more likely it is that the personality, character, and world of the protagonist(s) will be understood by the reader. Such schema building should facilitate greater understanding of the stories over time. Beyond characters *per se*, the unique writing style of the authors might become more familiar to readers as they move through a greater number of volumes.

In the field of SLA and vocabulary acquisition, several researchers have investigated the lexical load of related texts or other learning materials. Some have looked at academic texts, such as economics books (Sutarsyah, Nation, & Kennedy, 1994). Rodgers and Webb (2011) looked at related TV programs such as crime, medical drama, etc., and Schmitt and Carter (2000) investigated newspaper articles touching on related topics. Related materials appear to have significantly less lexical load than unrelated materials.

Webb and Macalister (2013) compared material written for children, material written for L2 learners, and the Wellington Written Corpus to investigate the lexical load for extensive reading purposes. They broke their corpus of "text written for children" (p.300) into four sub-sections according to target age group. Their results indicated texts targeting 7-8 year olds required vocabulary knowledge around the 11K level, whereas texts targeting 11-13 year olds required knowledge around the 9K level. This seems counterintuitive. Moreover, it is problematic considering researchers have advocated that readers must know 95% or even 98% of the vocabulary in the text in order to comprehend what they are reading (Hu & Nation, 2000). Utilization of these BNC-derived bands, in which children's materials ostensibly contributed little, to the analysis of children's texts may well be a spurious choice. There are no word bands purpose-built for assessing such texts, however, so researchers have little alternative. It is clear that the construction of the corpus against which texts are analyzed must be considered carefully. Nation and Webb (2011) argue for the creation of specialized word lists, mentioning children's materials explicitly as an area in which such lists might be extremely useful. The research presented here does exactly that, though on a small scale. I have

created a corpus from a very particular series of texts, and then investigated at the lexical profile of the world in which those texts are set.

Research Questions

1. *What is the lexical profile of a selection of volumes in the popular children's story book series The Berenstain Bears?*
2. *How should the lexical profiles of individual texts within this series be interpreted when considering them for use in the Family Reading Program?*

METHOD

Materials

The book series analyzed in this study was *The Berenstain Bears*. This series was selected for analysis for several reasons. These books have been wildly popular in the USA for nearly 50 years. They are widely available, well-known, and present an ecologically valid choice from that standpoint. Kids of varying ages seem to enjoy them. The pages are absolutely filled with colorful pictures. The series has titles at many levels, including: *Picturebook, I-Can-Read, First Time Books, First Time Readers, Bright & Early, Beginner, Storybook Series, Cub Club,* and *Living Light.* Finally, at around $5 each, they are very economical. Taken together, this series seemed like a ideal candidate for analysis and, depending on the results of analysis, consideration for use in future iterations of the FRP. A total of 55 titles were analyzed in this study. See Table 1 for the individual titles listed by number of tokens (e.g. words). This fledgling corpus was comprised of 58,653 total words, and the average number of total words per book was 1,066.

Measures

The 14K bands, based on the BNC and developed by Nation (2006), were used to assess the lexical load of these texts. These bands use the word family as opposed to lemma as the unit of measurement. While this study has made it abundantly clear that the target audience consists of low-proficiency English L2 speakers and kids, and while Nation and Webb (2011) caution that "one of the most important principles to follow is that the unit of counting must match the purpose for which the lists are used" (p.146), the fact is that lemma frequency bands, which would have been preferable, do not exist at present, and certainly not for children's materials specifically.

Procedure

All 55 books were scanned/digitized and then run through optical character recognition (OCR) software to extract the textual content. Each scanned page was checked to ensure OCR errors were identified and fixed prior to inclusion in the analysis. These files were then saved in .txt file format. All .txt files were first run through AntWordProfiler (Anthony, 2014) individually to establish book-by-book lexical profiles. Finally, a master .txt file consisting of the text from all books analyzed in this study was created in order to examine the lexical profile of the series overall.

TABLE 1. Summary of Berenstain Bears Texts, Ordered by Total Token Count

The Berenstain Bears...	Series[1]	Total Tokens	Coverage Statistics		
			% Coverage at: K2 + Names & MWs	Reaches 95% at	Reaches 98% at
... *Bears in the Night*	BE	113	94.7	K3	K3
... *Ready Set Go*	FTR	307	93.5	K3	K13
... *Home Sweet Tree*	CC	347	91.4	K5	K7
... *On the Job*	FTR	430	95.4	NM[2]	K3
... *New Pup*	ICR	445	93	K5	K5
... *Big Road Race*	FTR	515	89.7	K6	K10
... *Wild Wild Honey*	MS	590	84.1	K5	NEVER
... *Picnic*	BB	593	96.1	NM	K4
... *Ghost of the Forest*	FTR	636	89.5	K4	K7
... *Big Election*	MS	642	91.4	K4	K6
... *Dinosaurs*	MS	660	84.6	K8	NEVER
... *Missing Honey*	FTR	704	88.2	K4	K7
... *New Baby*	PB	713	95.1	NM	K4
... *Attic Treasure*	FAM	725	90.9	K5	K8
... *Trouble with Chores*	FTB	743	92.1	K4	K7
... *Visit the Dentist*	FTB	815	87.9	K5	NEVER
... *Get in a Fight*	FTB	825	95.5	NM	K3
... *Sitter*	FTB	858	94.4	K3	K10
... *Go to the Doctor*	FTB	880	96	NM	K5
... *Go to School*	PB	916	95.2	NM	K6
... *Moving Day*	FTB	928	94.7	K3	K5
... *Messy Room*	FTB	1000	92.9	K3	K5
... *Baby Makes Five*	FTB	1006	92.2	K4	K7
... *Substitute Teacher*	FAM	1082	92.1	K3	K5
... *Go to Camp*	FTB	1096	92.4	K4	K6
... *Mama's New Job*	FTB	1104	92.7	K4	K7
... *Mansion Mystery*	HH	1112	87.6	K5	K10
... *Truth*	FTB	1116	93.4	K3	K6
... *No Girls Allowed*	FTB	1143	91.2	K5	NEVER
... *Big Blooper*	FTB	1148	94.1	K3	K5
... *Knight to Remember*	HH	1155	88.2	K5	K7
... *Forget their Manners*	FTB	1163	92	K3	K4
... *Trick or Treat*	FTB	1185	90.8	K4	K8
... *Too Much TV*	FTB	1199	95	NM	K6
... *In the Dark*	FTB	1210	93.3	K3	K7
... *Nursery Tales*	PB	1214	92.1	K4	K7
... *Green-Eyed Monster*	FTB	1220	95.6	NM	K3
... *Trouble with Money*	FTB	1289	94.9	K3	K7
... *Birds, Bees*	FTB	1315	96	NM	K4
... *Go Out for the Team*	FTB	1321	93.5	K3	K6
... *Too Much Birthday*	FTB	1321	95.8	NM	K4
... *Too Much Junk Food*	FTB	1340	93.4	K3	K6
... *Week at Grandma's*	FTB	1369	94.6	K3	K6
... *Bad Dream*	FTB	1371	91.3	K3	K8
... *Bad Habit*	FTB	1379	90.2	K5	K9
... *Get the Gimmies*	FTB	1429	92	K3	K7
... *Double Dare*	FTB	1465	91.5	K3	NEVER
... *Get Stage Fright*	FTB	1479	94.7	K3	K5
... *Trouble with Pets*	FTB	1486	93.4	K3	K5
... *Trouble with School*	FTB	1513	92.7	K4	K10
... *Learn about Strangers*	FTB	1543	95	NM	K6
... *Trouble with Friends*	FTB	1553	95	NM	K4
... *Meet Santa Bear*	FTB	1558	92.3	K4	K9
... *Papa's Day Surprise*	FTB	1679	94.7	K3	K4
... *In-Crowd*	FTB	1720	93.2	K3	K5

[1] Abbreviations are as follows: BE (Bright & Early), FTR (First Time Readers), CC (Cub Club), ICR (I Can Read), MS (Mini Storybook), BB (Beginner Books), PB (Pictureback), FTB (First Time Books), HH (Happy House, FAM (Family Time Books)

[2] "NM" indicates the text reached the threshold at the K2 + Names + MWs level.

ANALYSIS & RESULTS

Lexical Profile of the Series Overall

Several permutations for presenting the data in Table 1 were considered. The options included publication date, alphabetic order, the publisher-determined series assignment (*First Time Book*, etc.), and the band level at which the book reached 98% lexical coverage. None of the aforementioned options appeared to provide a sensible order for the texts, however. Therefore the texts are listed according to the total number of tokens they contain, from the fewest (*Bears in the Night*; *n* = 113) to the most (*In-Crowd*, *n* = 1720). The data on the texts in Table 1 shows very little pattern with regards to increasing token count and increasing difficulty. Most books reach 95% coverage at around the K3 band, and 98% around K6 or K7. The fact that the books in the sample came primarily from one series (*First Time Books*), makes discerning a pattern on the basis of publisher-designated series challenging.

Turning to Table 2, we begin to get a better idea of the lexical load of the series overall. As a series, the cumulative percentage of coverage at the K2 + names + marginal words level was just over 93%. The series hits 95% coverage at K3, and 98% coverage at K6. This is not that surprising given what we saw in Table 1 where the modes for 95% and 98% were K3 and K6/7, respectively.

This data presents questions, however. One question concerns which words are not present within the Berenstain Bears texts sampled that are within the lower K bands. This may be an area for future exploration. Also, names and marginals clearly comprise a critical percentage of the texts overall. Were this series to be introduced to the FRP it would be imperative that these words were recognized by the participants.

TABLE 2. Corpus Data for the Berenstain Bears Texts Analyzed (55 Texts Total)

Band	Total Tokens	Token %	Cumulative %	Unique Tokens
K1 Band	48784	83.17	83.17	1702
K2 Band	3110	5.3	88.47	820
Names	2397	4.09	92.56	113
Marginal Words	275	0.47	93.03	52
K3 Band	1527	2.6	95.63	517
K4 Band	585	1	96.63	224
K5 Band	574	0.98	97.61	191
K6 Band	322	0.55	98.16	120
K7 Band	219	0.37	98.53	93
K8 Band	104	0.18	98.71	46
K9 Band	82	0.14	98.85	50
K10 Band	108	0.18	99.03	43
K11 Band	59	0.1	99.13	24
K12 Band	29	0.05	99.18	18
K13 Band	28	0.05	99.23	22
K14 Band	14	0.02	99.25	8
(off list)	436	0.74	99.99	255
TOTAL	58653			4298

Lexical Profiles of Individual Texts

Investigating the lexical load of a given book is somewhat easier than analyzing a series; however, the lexical profile of any individual book is susceptible to misinterpretation if not explored critically. Tables 3,

TABLE 3. Lexical Profile for *The Berenstain Bears Ready, Set, Go!*

Band	Total Tokens	Token %	Cumulative %	Unique Tokens
K1 Band	260	84.69	84.69	77
K2 Band	13	4.23	88.92	8
Names	11	3.58	92.5	2
MWs	3	0.98	93.48	1
K3 Band	10	3.26	96.74	7
K4 Band	1	0.33	97.07	1
K5 Band	1	0.33	97.4	1
K13 Band	3	0.98	98.38	3
(off list)	5	1.63	100.01	5
TOTAL	307		105	

TABLE 4. Lexical Profile Statistics for *The Berenstain Bears Big Road Race*

Band	Total Tokens	Token %	Cumulative %	Unique Tokens
K1 Band	399	77.48	77.48	148
K2 Band	45	8.74	86.22	24
Names	3	0.58	86.8	2
MWs	15	2.91	89.71	6
K3 Band	10	1.94	91.65	5
K4 Band	6	1.17	92.82	4
K5 Band	7	1.36	94.18	5
K6 Band	6	1.17	95.35	3
K7 Band	1	0.19	95.54	1
K10 Band	23	4.47	100	3
TOTAL	515			201

TABLE 5. Lexical Profile Statistics for *The Berenstain Bears and the Dinosaurs*

Band	Total Tokens	Token %	Cumulative %	Unique Tokens
K1 Band	495	75	75	177
K2 Band	43	6.52	81.52	29
Names	19	2.88	84.4	4
MWs	1	0.15	84.55	1
K3 Band	55	8.33	92.88	15
K4 Band	8	1.21	94.09	4
K6 Band	1	0.15	94.24	1
K7 Band	3	0.45	94.69	2
K8 Band	3	0.45	95.14	3
K9 Band	4	0.61	95.75	1
(off list)	28	4.24	100	14
TOTAL	660			251

TABLE 6. Lexical Profile Statistics for *The Berenstain Bears and the Green Eyed Monster*

Band	Total Tokens	Token %	Cumulative %	Unique Tokens
K1 Band	1018	83.44	83.44	288
K2 Band	90	7.38	90.82	44
Names	54	4.43	95.25	11
MWs	4	0.33	95.58	2
K3 Band	34	2.79	98.37	21
K4 Band	8	0.66	99.03	5
K6 Band	1	0.08	99.11	1
K7 Band	1	0.08	99.19	1
K8 Band	1	0.08	99.27	1
K9 Band	1	0.08	99.35	1
K10 Band	1	0.08	99.43	1
K11 Band	1	0.08	99.51	1
(off list)	6	0.49	100	5
TOTAL	1220			382

4, 5, 6, and 7 below present the lexical profile for five different individual titles in this series. These titles were selected for discussion because their lexical profiles each illustrate at least one trend noticeable among the 55 texts analyzed.

First, in *The Berenstain Bears Ready, Set, Go!* (Table 3), which was the second shortest book at only 307 total tokens, we see that 95% coverage is reached somewhere within the K3 band. According to the lexical profile, however, 98% coverage is reached at the K13 band, despite the fact that bands K6 through K12 are not represented by a single word in the text. Looking at the book itself, however, makes it clear that the lexical load presented in Table 3 is misleading. This entire book emphasizes the comparative and superlative forms. The word *spring* is within the K2 band, and the word *swing* is within the K3 band. However, the words *springy, springier, springiest*, and *swingy, swingier, swingiest* are six of the eight words that are classified as either K13 or off-list in the lexical profile. Given the repetitive nature of the comparative/superlative forms, as well as the pictorial support provided by the text, these words are very likely understandable by low-proficiency readers (e.g. children or their parents who are reading to/with them), particularly given the visual support provided by the pictures in the book.

The point is illustrated further in Table 4, where the total token to unique token ratio at the K10 band (23:3) indicates a few supposedly difficult words are prevent this text from being classified as an easy text for children. If we investigate these statistics with a critical eye, however, there is a single word causing the issue. In this case, that word is *putt.* The word *putt* (as in a race car going *putt, putt, putt)* accounts for 22 of the 23 tokens in the K10 band. That is roughly 4.2% of the total tokens in the text. Were we to move *putt* to the marginal words list this dramatically changes the outcome of 95% and 98% thresholds. Instead of 89.7% coverage at the K2 + Names + MWs level, the text would be at 93.9% (89.7 + 4.2 = 93.9). The book then reaches 95% coverage at K3 (93.9 + 1.9 = 95.8), and 98% at K5 (95.8 + 1.1 + 1.3 = 98.2). This provides a stark example of how a single infrequent word that appears repeatedly in a text can falsely appear to put the text out of reach of the target audience. The word *putt* was not moved to the MWs list because removing it from the K10 would compromise the integrity of the K10 list for future use against other texts. Looking for those ratios of high token total to low token type is a quick way of identifying vocabulary that is having a dramatic impact on the lexical profile of a text.

Table 5 provides another example of a text where the lexical profile makes it appear that the text is challenging when it is likely less challenging than it appears. The text of interest is *The Berenstain Bears and the Dinosaurs*. Given the word *dinosaurs* is in the title, it should come as no surprise that there were eight dinosaur names used in the book for a total of 22 tokens. These were all off-list words comprising roughly 3.3% of the text. Do words such as *stegosaurus* and *tyrannosaur* really put this book out of reach? This is doubtful, particularly when considering the audience consists of children and the fact that the entire book is built around these key off-list words. The title of the book itself, the cover art, and multiple illustrations of each dinosaur provide ample support for these off-list words.

Looking at a couple of books with higher token counts, it seems that they are less likely to have skewed lexical profiles due to one or two infrequent words used repeatedly. In Table 6, we see a text of 1220 tokens that hits 98% in the K3 band. Several more difficult bands are represented, but often by only one token and with total/type ratios that indicate nothing is being repeated excessively. Results for the longest book in the study (1720 tokens), found in Table 7, are similar. The book appears more difficult, reaching 98% coverage

TABLE 7. Lexical Profile Statistics for *The Berenstain Bears and the In-Crowd*

Band	Total Tokens	Token %	Cumulative %	Unique Tokens
K1 Band	1388	80.7	80.7	355
K2 Band	122	7.09	87.79	66
Names	91	5.29	93.08	21
MWs	3	0.17	93.25	2
K3 Band	50	2.91	96.16	29
K4 Band	13	0.76	96.92	11
K5 Band	20	1.16	98.08	13
K6 Band	7	0.41	98.49	3
K7 Band	4	0.23	98.72	3
K8 Band	1	0.06	98.78	1
K10 Band	3	0.17	98.95	2
K11 Band	2	0.12	99.07	2
K12 Band	4	0.23	99.3	2
K13 Band	1	0.06	99.36	1
(off list)	11	0.64	100	10
TOTAL	1720			521

at K5 and again without any token total/type ratios within the respective bands that indicate the lexical profile for the text is misleading.

DISCUSSION

Pedagogical Implications

The larger body of research surrounding the use of series or materials from similar genres to increase the likelihood of incidental vocabulary acquisition hints at a promising future for using a series of children's books in the FRP. However, it appears we might not be able make the blanket assumption that a children's book series lends itself to vocabulary acquisition the same way a TV series within a specific genre does. Looking back at Table 1 and all the various titles in the series clearly indicates that each book touches on a topic in itself. The characters are common between texts, but what those characters do in a given story ranges from visiting the dentist, to learning about dinosaurs, to going trick-or-treating. Shared characters alone might not be enough to make the argument for using a series such as this in the FRP at this time. It seems prudent to consider each text individually for now, perhaps selecting those texts which match a theme of the children's preference or what their curriculum at school might be covering at the moment.

It is critical to understand that supposedly easy texts with lower token counts might have lexical profiles that put them out of reach for learners. The lower token count means they are easier in terms of length. However, as the number of tokens in the text decreases each token will then represent an increasingly larger portion of the text. A critical eye is needed to ensure words such as *putt* or *springy/springier/springiest* or *Tyrannosaurus Rex*, which are often central to the very theme of the text, do not become the reason for eliminating a text from consideration in a program like the FRP.

Several limitations must be acknowledged. It was not possible to obtain and analyze the entire *Berenstain Bear* series. Also, the use of the 14K word bands is problematic. These bands use the word family as the unit of measurement. This is inappropriate for this study which targets low-proficiency L2 English speakers and their productive use of the language. "For productive use… the lemma or type is the best unit

of counting because knowing how to use one word in the family does not mean you can accurately use other members of the family" (Nation & Webb, 2011, p.136). Lemma-based frequency lists are not available to my knowledge. Even if they were/are, however, they would likely still suffer from the second issue that the 14K bands present: Some highly salient words might not be frequent according to the bands, but might actually be commonly understood by native English speaking children and low-proficiency NNSs. As an example, animals, foods, and dinosaurs are relatively common elements in children's books but might be much less frequent in larger corpora such as the BNC. The BNC is not optimal for the assessment of children's texts for the purposes of projects such as the FRP. Future work in this area should utilize a corpus or corpora derived entirely from children's texts. Finally, the 95% and 98% thresholds so common in studies related to lexical load (Hu & Nation, 2000) might or might not be appropriate for the purposes of the FRP. These parents will be sharing texts with their children, and therefore it makes sense that total mastery of all the vocabulary in a given text might be desirable. More research is needed in this area.

CONCLUSION

Learning the steps and processes involved in creating a small corpus, and considering how to analyze them at both the series level and individual book level, is necessary inasmuch as it informs text selection in the FRP. The process has revealed that while much can be gleaned from examining corpus-level data, going through texts individually is critical in order to understand their particularities. As these examples have demonstrated, obtaining a lexical profile is merely the first step. The next step requires looking at the words or bands that give the text the appearance of difficulty. A single word (*putt*), or a few words critical to the content (e.g. *dinosaurs*) or grammar (e.g. comparative/superlative) might frequently be causing a book to appear more challenging than it actually is. Developing a principled approach to analyzing the lexical load of children's books and determining how those texts meet the proficiency of the readers, or how the readers' proficiency can be strategically increased to meet the demands of the texts is an ongoing goal.

REFERENCES

Anthony, L. (2014). AntWordProfiler (Version 1.4.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Barbour, A. C. (1998). Home Literacy Bags Promote Family Involvement. *Childhood Education, 75*(2), 71-75. http://doi.org/10.1080/00094056.1999.10521986

Bus, A. G., Ijzendoorn, M. H. van, & Pellegrini, A. D. (1995). Joint Book Reading Makes for Success in Learning to Read: A Meta-Analysis on Intergenerational Transmission of Literacy. *Review of Educational Research, 65*(1), 1-21. http://doi.org/10.2307/1170476

Dever, M., & Burts, D. (2002). An Evaluation of Family Literacy Bags as a Vehicle for Parent Involvement. *Early Child Development and Care, 172*(4), 359-370. http://doi.org/10.1080/03004430212721

Dickinson, D. K., & De Temple, J. (1998). Putting parents in the picture: Maternal reports of preschoolers' literacy as a predictor of early reading. *Early Childhood Research Quarterly, 13*(2), 241-261. http://doi.org/10.1016/S0885-2006(99)80037-4

Elley, W. B. (1991). Acquiring literacy in a second language: The effect of book- based programs. *Language Learning, 41*, 375-410. doi:10.1111/j.1467- 1770.1991.tb00611.x

Elley, W. B. (2000). The potential of book floods for raising literacy levels. *International Review of Education, 46*(3-4), 233-255.

Ferguson, P., Sponseller, A., & Yamada, A. (2017). Introducing the Family Reading Project. In P. Clements, A. Krause, & H. Brown (Eds.), *Transformation in language education*. Tokyo: JALT.

Hargrave, A. C., & Sénéchal, M. (2000). A book reading intervention with preschool children who have limited vocabularies: the benefits of regular reading and dialogic reading. *Early Childhood Research Quarterly, 15*(1), 75-90. http://doi.org/10.1016/S0885-2006(99)00038-1

Hindman, A. H., & Morrison, F. J. (2012). Differential contributions of three parenting dimensions to preschool literacy and social skills in a middle-income sample. *Merrill-Palmer Quarterly, 58*(2), 191-223.

Hu, M., & Nation, I.S.P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430.

Huebner, C. E., & Meltzoff, A. N. (2005). Intervention to change parent□child reading style: A comparison of instructional methods. *Journal of Applied Developmental Psychology, 26*(3), 296-313. http://doi.org/10.1016/j.appdev.2005.02.006

Huebner, C. E., & Payne, K. (2010). Home support for emergent literacy: Follow-up of a community-based implementation of dialogic reading. *Journal of Applied Developmental Psychology, 31*(3), 195-201.

Jacobs, G. M., Newland, L. A., Gapp, S. C., Cambetas Syed, D., Reisetter, M. F., & Wu, C.-H. (2011). Mothers' beliefs and involvement: Links with preschool literacy development. *Tarptautinis Psichologijos Žurnalas: Biopsichosocialinis Požiuris*, (9), 67-90.

Jordan, G. E., Snow, C. E., & Porche, M. V. (2000). Project EASE: The Effect of a Family Literacy Project on Kindergarten Students' Early Literacy Skills. *Reading Research Quarterly, 35*(4), 524-546.

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

Opel, A., Ameer, S. S., & Aboud, F. E. (2009). The effect of preschool dialogic reading on vocabulary among rural Bangladeshi children. *International Journal of Educational Research, 48*(1), 12-20.

Rodgers, M. P., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly,* 689-717.

Schmitt, N., & Carter, R. (2000). The lexical advantages of narrow reading for second language learners. *TESOL Journal, 9*(1), 4-9.

Sénéchal, M. (2006). Testing the home literacy model: Parent involvement in kindergarten is differentially related to grade 4 reading comprehension, fluency, spelling, and reading for pleasure. *Scientific Studies of Reading, 10*(1), 59-87.

Sénéchal, M., LeFevre, J.-A., Hudson, E., & Lawson, E. P. (1996). Knowledge of storybooks as a predictor of young children's vocabulary. *Journal of Educational Psychology, 88*(3), 520.

Sénéchal, M., LeFevre, J.-A., Thomas, E. M., & Daley, K. E. (1998). Differential Effects of Home Literacy Experiences on the Development of Oral and Written Language. *Reading Research Quarterly, 33*(1), 96-116.

Shanahan, T., Mulhern, M., & Rodriguez-Brown, F. (1995). Project FLAME: Lessons Learned from a Family Literacy Program for Linguistic Minority Families. *The Reading Teacher, 48*(7), 586-593.

Sutarsyah, C., Nation, P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based

case study. *RELC Journal, 25*(2), 34-50.

Uchiyama, T. (2011). Reading versus Telling of Stories in the Development of English Vocabulary and Comprehension in Young Second Language Learners. *Reading Improvement, 48*(4), 168-178.

Webb, S. (2011). Selecting television programs for language learning: Investigating television programs from the same genre. *International Journal of English Studies, 11*(1), 117-135.

Webb, S., & Macalister, J. (2013). Is Text Written for Children Useful for L2 Extensive Reading? *TESOL Quarterly, 47*(2), 300-322.

Nation, I. S., & Webb, S. A. (2011). *Researching and Analyzing Vocabulary*. Heinle, Cengage Learning.

Weigel, D. J., Martin, S. S., & Bennett, K. K. (2005). Ecological influences of the home and the child-care center on preschool-age children's literacy development. *Reading Research Quarterly, 40*(2), 204-233.

Weigel, D. J., Martin, S. S., & Bennett, K. K. (2006). Contributions of the home literacy environment to preschool-aged children's emerging literacy and language skills. *Early Child Development and Care*, *176*(3-4), 357-378. http://doi.org/10.1080/0300443050006374

Wu, C., & Honig, A. S. (2010). Taiwanese mothers' beliefs about reading aloud with preschoolers: findings from the parent reading belief inventory. *Early Child Development and Care*, *180*(5), 647-669. http://doi.org/10.1080/03004430802221449

Yeo, L. S., Ong, W. W., & Ng, C. M. (2014). The Home Literacy Environment and Preschool Children's Reading Skills and Interest. *Early Education and Development*, *25*(6), 791-814. http://doi.org/10.1080/10409289.2014.862147

ABSTRACT

Investigating the Lexical Profile of a Popular Children's Storybook Series

Aaron C. SPONSELLER
Institute for the Promotion of Global Education
Hiroshima University

The Family Reading Project (FRP), initiated in 2016, trains Japanese parents to read English language storybooks to their elementary schoolers (Ferguson, Sponseller, & Yamada, 2017). One challenge the FRP has faced is selecting storybooks that are not only interesting in terms of content but lexically within the reach of the participating parents. At present, no comprehensive corpus of children's texts or principled manner for measuring the difficulty of a text for children exists. The primary aim of this research was to begin developing and refining a principled approach to analyzing candidate texts for inclusion in the FRP. A principled approach is needed if the FRP wishes to make lexically-informed book selections for future iterations of the program. Fifty-five titles from the popular children's storybook series The Berenstain Bears were digitized in order to create a miniature corpus. The lexical profiles for the series overall as well as individual titles that were generated and investigated are discussed, and implications for evaluating potential texts for use in the FRP are briefly discussed.

要　約

子ども向け絵本シリーズの語彙使用レベルの研究

アーロン・スポンセラー
広島大学大学院教育学研究科グローバル教育推進室

2016 年に始まった Family Reading プロジェクト（FRP）は，日本人小学生の親を対象に，英語絵本の読み聞かせを練習する機会を提供するものである（Ferguson, Sponseller, & Yamada, 2017）。FRP で用いる絵本は，内容が面白いだけでなく，プロジェクトに参加する親の語彙に合わせて選ばなければならないが，現在，子ども向けテクストの包括的なコーパスや，その難易度を測定する方法は確立されていない。本研究の主目的は，FRP に用いるテクストの一貫した分析方法の開発とその改良にある。FRP が将来にわたって継続されていくためには，語彙に基づいて FRP で用いる本を選ぶ必要がある。そこで，子ども向け絵本シリーズとして有名な The Berenstain Bears から 55 作品を選んで電子化し，小規模コーパスを作成した。本論文では，そこからシリーズ全体と個別の作品の語彙使用レベルに関して調査し，今後の FRP において使用可能なテクストの評価について示唆を述べる。