*Concordia Working Papers in Applied Linguistics*, 4, 2013


# Lexical Analysis of the Dr. Seuss Corpus

**Jordan Foster**
*Concordia University*

**Craig Mackie**
*Concordia University*

## Abstract

For decades, the work of Dr. Seuss has been both a cultural mainstay and a frequently used tool in L1 and L2 learning. In this paper, an examination of Seuss' work from the standpoint of literary analysis, imaginative writing and linguistic pedagogy is combined with insights from a corpus analysis in order to explore how these texts may facilitate or hinder the acquisition of language. The aim of this study is to establish the corpus of Dr. Seuss' works for the first time in order to analyse these texts from a vocabulary standpoint. In doing so, the pedagogical implications and appropriateness of the use of Dr. Seuss books will be discussed in the context of teaching L2 English and early L1 English literacy. The results of the Dr. Seuss corpus analysis will shed light on trends in Geisel's children's writing, provide the basis for future studies using the Seuss corpus and facilitate discussion on the pedagogical implications this kind of analysis may have for the teaching and learning of both L1 and L2 English.

The work of Theodor Geisel, better known as Dr. Seuss, has been a cultural mainstay in North America and around the world (via translations) for well over sixty years. Geisel was responsible for the creation of some of children's literature's best-known characters and his books are often some of the very first read to children or read by children themselves. A continued interest in Geisel's work from the standpoints of

literary analysis, imaginative writing and linguistic pedagogy continues the relevance of his work, years after the author's death. At the same time, sophisticated methods of linguistic analysis continue to develop, with corpus analysis in particular presenting itself as a powerful tool for exploring the way certain texts facilitate or hinder the acquisition of language. With this in mind, the purpose of this study is to begin to gather all of Geisel's works intended for children together for the first time and then analyse the corpus of his work from a vocabulary standpoint. In doing so, the pedagogical implications and appropriateness of the use of Dr. Seuss books will be discussed in the context of teaching English as second language and early L1 English literacy. Other corpora of children's literature have been compiled for the purposes of investigating the make-up of the vocabulary in children's literature, but none have focussed solely on the works of a single author. The results of the Dr. Seuss corpus analysis will shed light on trends in Geisel's children's writing, provide the basis for future studies using the Seuss corpus and lead to discussions on the pedagogical implications this kind of analysis may have for the teaching and learning of both L1 and L2 English.

## BACKGROUND

### Dr. Seuss

Dr. Seuss' works and imagery are omnipresent in modern North American culture partly owing to the very poignancy of the messages presented in his stories, whether they are overtly manifested or covertly transmitted (Menand, 2002). Lange (2007) discusses the widespread implications of the dominant loci, or themes, present in Dr. Seuss' works. Concepts such as activism, acceptance, independence, perseverance, possibility and imagination, are identified as being central to the worldview in Dr. Seuss' rhetoric; themes which must have been refreshing for a post-World War II audience able to appreciate the fanciful yet challenging themes presented (Menand, 2000). Menand also expands on these ideas by looking at the political subversiveness of the messages in Dr. Seuss' works. Dr. Seuss' works pushed the envelope during the cold war period and brought controversial messages to the larger public consciousness. The story *The Lorax* is discussed as being targeted for its overt environmental message while stories like *The Butter Battle Book* drew criticism for being too soft on communism. Despite the deliberately silly

and irreverent character of Seuss' works, the themes and messages in Dr. Seuss' works are used to explore deeper philosophical issues in Held's (2011*) Dr. Seuss and Philosophy: Oh, the Thinks You Can Think!*. The widespread familiarity with Seuss' stories is used as a rationale for exploring deeper philosophical issues, predicated on the very accessibility of such themes when presented in Seussian format. Essays from the book use well-known Dr. Seuss stories and characters as allegories that are used to open up an exploration of the profound and humanistic themes earlier presented by Dr. Seuss.

Beyond an analysis of Dr. Seuss' central themes, there is an extensive literature available that charts the unique use of vocabulary, both authentic and invented, which characterizes Seuss' work. Schroth (1978) examines many of the features of Dr. Seuss' vocabulary use trends, by looking at both Dr. Seuss' word use and word invention. Many examples are given of how Dr. Seuss' use of superlatives, repetition and generous alliteration create a fun, rhythmic tone for the reader. Dr. Seuss' use of invented collocations, open juncture (e.g., '*Bad-animal-catching-machine'*) and creatively-employed, highly productive bound morphemes (e.g., '*gladdish*') help create invented vocabulary that is logical and easily understood for emerging English readers. Schroth also discusses Dr. Seuss' use of onomatopoeia as being a device used to communicate real information through invented imaginative vocabulary. Kies (1990), in his discussion of the use of phonemes in children's literature to evoke reader responses, gives the example of Dr. Seuss' use of voiceless stops (/p/, /t/, and /k/) in authentic vocabulary in *Hop on Pop* to give the reader a sense of abrupt movement. In Crystal's (1996) discussion of how children's literature does not often reflect emerging child speak in the form of language play, Dr. Seuss is cited as a classic example of a children's author who bucks this trend and employs language play extensively.

The richness of Dr. Seuss' content and writing style has inspired many researchers to use his work in studies looking at both first and second language English acquisition. In a study looking at strategies used for improving L1 English graphophonemic awareness, Dr. Seuss books were used as a treatment for helping first grade inner city students improve their letter-sound correspondences (Jenkins, Vadasy, Firebaugh & Profilet, 2000). The program, which focused on the pronunciation of segmentals, used *Hop on Pop* by Dr. Seuss as a treatment tool for its use of rhyme and segmental repetition. In a study looking at, among other things, French L1 English learners' initial aspirated /h/ production, Horst, White and Bell (2010) had students read Green Eggs and Ham as a prompting tool for

students to produce and become aware of the target form. Notestine and Tanner (2007) show how the use of Dr. Seuss texts can be effective in helping adult learners of English in a foreign language setting acquire various phonetic forms. Dr. Seuss works were chosen to target the acquisition of English supra-segmentals because they are interesting, inviting, playful, culturally relevant, contain a unique poetic rhythm and contain many instances of intonation, linking, blending and consonant clusters carried in meaningful text.

## Children's literature corpora

Although an entire corpus consisting solely of Dr. Seuss' works has not been attempted up to this point, many other corpora focussing on children's literature have been compiled with different objectives in mind. What follows is an investigation of other children's corpora that have been created with various objectives. Attention will be given to the objectives of the corpora, the scope and size of the content they contain, as well as the methods used and methodological issues encountered in gathering the works contained within the individual corpora. The building of the Dr. Seuss corpus is influenced by the methods and goals of these prior studies.

Between 2002 and 2005, a large scale Welsh corpus building project took place with the intention of creating a rich database of Welsh language children's literature (Powell & Forbes, 2005). The corpus was built by collecting works from four Welsh publishers to assemble a collection of over 3,000,000 words from Welsh children's literature. The works came from books intended for various age groups and represent various genres of children's literature. The books were painstakingly scanned, converted into electronic text, passed through optimal character recognition (OCR) software and then placed into Word documents. The texts were cleaned up for OCR inaccuracies, invented words, English and onomatopoeic words (e.g., '*ahhh*'). The authors estimate that there is a <1% rate of inaccuracies in the entire corpus. The methodology for building this children's corpus (and its inherent methodological issues) is similar to those used in other studies such as Göbel and Peetz's (2005) corpus assembly of German children's literature spanning the 19th and 20th centuries. It also reflects certain issues faced in the building of much larger scale corpora using scanned text, such as the assembly of the Corpus of Contemporary American English (COCA) which in part consists of millions of words of scanned children's books and magazines (Davies,

2009). Following the assembly of the Welsh children's corpus, the texts were analysed by teams of researchers and educators to categorize each individual text in the corpus by genre, interest age, author and national curriculum correlation. The project was successful in producing a large scale title database that is searchable in its entirety or by sub-category. Text versions of the whole corpus, word and lemma count data and a compiled a list of the most frequent words across the corpus using Wordsmith software is available by download. In future, deeper analysis of the corpus will provide more robust statistics for publishers, educators and researchers robust information on the nature of Welsh used in literacy materials designed for Welsh children.

A smaller English corpus, the Corpus-based Learning about Language In the Primary-school (CLLIP) was designed in the United States for the investigation of L1 English literature (Thompson & Sealey, 2007). Thirty children's imaginative fiction books, made up of 698,286 tokens, with intended audiences of 8-10 year olds were extracted from the 100 million-word British National Corpus (BNC) of written and spoken English. The books were taken from the BNC as they were already transferred into an electronic version and were tagged for part of speech. The compiled corpus was analysed for word frequency and counts using Oxford Wordsmith Tools and was passed through MonoConc Pro in order to gather concordancing information. The results of these analyses were used to compare the corpus of children's imaginative literature to a larger corpus of adult fiction as well as a corpus comprised of newspaper text in order to examine how the groups of text varied in terms of the distribution and nature of word and parts of speech use, discourse patterns and representation of self and the world. After a very thorough analysis of the texts, parts of speech and discourse markers, the authors found that although the discourse pattern in children's writing represent how humans relate to their world (real or imagined) much differently than in fiction for adults, fiction in general has very similar vocabulary use patterns regardless of age of the target audience.

The creation of the VP-Kids corpora is different from the above-mentioned studies in that it is comprised of spoken language from children, as opposed to written language intended for children (Roessingh, n.d). The VP-Kids corpus, which is available as an online research tool (www.lextutor.ca/vp/kids), is comprised of samples of spoken text of children from aged four to seven gathered from studies looking at child speech development. The compiled corpus of spoken English is then subdivided into ten 250-word family groupings based on

frequency. This division of words in the online research tool allows users to enter any text they choose to see how its vocabulary profile compares to the coverage provided by the VP-Kids corpus. Furthermore, the VP-Kids online tool provides small text samples of both native and non-native English speaking children to compare to the larger child speech corpus. This corpus and online research tool is valuable in that it allows educators and researchers the ability to conduct a myriad of studies looking at how any text (written or spoken, native or non-native, produced by or for children) compares in vocabulary frequency and range to the larger corpus of child speech assembled.

Various studies have looked at bridging the gap between the authentic language description provided by corpus linguistics and language teaching research (LTR) by examining linguistic features of specific genres, examining corpora designed explicitly for LTR purposes and looking at how corpus based research influences materials development (Keck, 2004). Sealey and Thompson (2007) qualitatively examined how corpus derived material can be used to help young L1 learners gain passive metalinguistic knowledge by examining the lexico-semantic relationship between word endings provided in corpus derived language samples. Corpus derived data has been looked at in its role in assisting the development of L2 learning resources with focuses on grammar (Barbieri & Eckhardt, 2007), on grammar and vocabulary (Reinhart, 2010) or on how the teaching of vocabulary and grammar can be integrated with corpus derived data (Mahlberg, 2006). Corpus driven data has also been examined as an evaluation tool for current L2 learning devices by using frequency counts to gauge the learner level appropriateness of text based resources (Dodigovic, 2005). Most research on the links between language description corpus linguistics and LTR is, however, restricted to specialised academic L2 contexts, and a call has been made for the broadening of the field to represent different learner proficiencies and goals, including young or beginner L1 and L2 English learners (Keck, 2004).

Previous studies looking at the creation of children's corpora inform the building of the Dr. Seuss corpus in terms of assembly methods, categorization of assembled materials and questions raised for the analysis of the distribution of vocabulary among the most frequent words in the English language in Dr. Seuss' writing. The previous studies will also provide benchmark results that can be compared to preliminary analysis of the Dr. Seuss corpus. The studies examined above have highlighted the relevance of Dr. Seuss' works for L1 and L2 English language educators,

shown the unique nature of Dr. Seuss' use of vocabulary and given glimpses into the issues raised and questions asked by creators of other children's corpora.

## STUDY

In line with other studies looking to bridge the gap between corpus research and possible pedagogical implications, the present study looks to initially investigate the following research questions for subsequent discussion:

1. What is the frequency coverage of Dr. Seuss writing compared to a general English corpus?
2. How do the most frequent words across the Dr. Seuss corpus reflect the most frequent words found in previous analyses of children's literature corpora?
3. What is the frequency coverage of Dr. Seuss writing compared to a children's spoken corpus?
4. What is the frequency of imaginary lexical inventions in Dr. Seuss writing?
5. How does the frequency coverage of selected Dr. Seuss texts compare with the larger Dr. Seuss corpus?

## METHODOLOGY

### Corpus Composition

The development of the Dr. Seuss corpus is an ongoing process. At the time of writing, the corpus contained over half of Dr. Seuss' works and covers most of his top 20 bestselling titles. A list of the titles used in the preliminary corpus analysis presented here are included in Appendix A. The ultimate goal of this project is to run a comprehensive analysis of all of Seuss' works using a corpus method. As there has yet to be another children's literature corpus built using a single author's works, the building of the Dr. Seuss corpus has presented unique challenges that have changed the nature of its investigation in certain ways. In compiling the texts, it has become clear that there are clear distinctions between certain works of Dr. Seuss, bringing the need for a sub-categorization of the corpus into Dr. Seuss sub-genres or categories. This led to the creation of four tentative categories which subdivided the corpus into Dr. Seuss

genres, as defined by their lexical density and complexity of writing. Potentially useful categories include: Beginner Books/Early Phonic Readers, Non-rhyming Prose and Standard Seussian Rhyme. These categories are still being refined as titles are added to the corpus. The criteria and procedure for the sub-categorization is also to be examined in later studies. The titles will be examined and classified in a similar fashion to the categorization of the Powell & Forbes' Welsh children's corpus literature (Powell & Forbes, 2005). Educational implications of the sub-categorization will be discussed following the presentation of results from the corpus analysis.

Theodor Geisel's works span 7 decades and are written under various pseudonyms (Dr. Seuss, Theo LeSieg and Rosetta Stone). Not including the works that he produced for an adult reader audience (often highly political and sometimes bawdy - Menand, 2002), he wrote over 70 titles for young readers. The initial challenge for the building of the corpus was to figure out how to assemble all of these works, some quite famous and some quite obscure, a process that led to a reliance on a variety of media from audio recordings to .pdf documents and, of course, on the availablibility of the texts themselves. These texts were then either transcribed, decoded with OCR software or translated using voice recognition software, and then added to the corpus. A second reader checked the documents for accuracy and completeness to ensure the reliability of the corpus.

## Corpus analysis

The preliminary and final analyses off the Dr. Seuss corpus will be done using tools found on the *Compleat LexTutor* website (www.lextutor.ca). Results from the analysis of the corpus using the different online tools through the website can be compared to results from previous analyses of other children's literature corpora.

For our first analysis, the Dr. Seuss corpus was run through the Web-VP BNC-20 tool found on the Compleat Lexical Tutor v. 6.2 website (Cobb, 1994; Heatley & Nation, 1994) in order to evaluate how the vocabulary of the Dr. Seuss corpus compares in terms of lexical frequency to a larger English corpus. This tool provides an analysis of the count of words in the Dr. Seuss corpus that fall into the 20 frequency groupings of roughly 1,000 word families from the 100 million-word BNC of written and spoken English. This tool also categorizes words from the corpus into an off-list category that allows the analyst to capture highly infrequent

and imaginative words. Proper names were added to the exceptions list prior to running analysis of the text, while proper names that are imaginative were added to the exception list if they are names that are central to the story (e.g., *Grinch*) but not if they occur less than three times throughout the entire corpus. Results from the *Web Frequency Indexer* v. 1.3 analysis, found on the Compleat Lexical Tutor v. 6.2 website (Cobb, 1994; Heatley & Nation, 1994), provide total occurrences of proper names in the corpus which will inform the decision to add names to the exception list or not.

For our second analysis, the Dr. Seuss corpus was run through the Web Frequency Indexer v 1.3 in order to compare the most frequent words across the Dr. Seuss corpus to the most frequent words found in previous analyses of children's literature corpora. This tool produced a list, in descending order, of the most commonly occurring words throughout the corpus. The results from this list were compared against results from the analysis of the *Corpus-based Learning about Language In the Primary-school* (CLLIP) corpus of children's imaginative fiction (Thompson & Sealey, 2007). Lists of the top ten most frequent nouns, adjectives and lexical verbs of the CLLIP were compared to the corresponding items from the Dr. Seuss corpus. Although the two corpora are similar in size, the results were compared in terms of ranking of frequency rather than number of overall types or tokens.

Our third analysis focused on how the vocabulary used in the Dr. Seuss corpus writing compares to a children's spoken corpus, the Dr. Seuss corpus was run through the VP-Kids v.9 tool on the Compleat Lexical Tutor v. 6.2 website (Cobb, 1994; Heatley & Nation, 1994). The results of this analysis show how the vocabulary from the corpus falls into 250-word coverage bands created from a children's spoken corpus. Proper names were added to the exceptions list prior to running analysis of the text. Imaginative proper names not occurring more than 4 times throughout the corpus were not added to the exceptions list.

For the fourth analysis, in order to determine the frequency and proportion of imaginative vocabulary in the Dr. Seuss corpus, the corpus was run through the Web-VP BNC-20 tool in order to produce a list of off-list words. Any real word appearing on the list, as well as any standard proper name, was added to the exception list prior to running the analysis in order to ensure that the off-list results reflect only Dr. Seuss imaginative vocabulary. To help determine whether a word is a true Dr. Seuss lexical invention rather than an onomatopoeia (e.g., '*ahh*') or random strings of letters or numbers that may appear in one of the texts, E.C. Latham's

(2000) collection of Seussian lexical inventions '*Who's Who & What's What in the books of Dr. Seuss*' was be consulted. Although counted as proper nouns in the analysis of the vocabulary of the overall corpus, imaginative names were kept as off-list words for the purpose of this inquiry.

Finally, to investigate the fifth research question two separate titles from the Dr. Seuss corpus, chosen by their differing genres, were compared to the larger corpus. This investigation is intended to be a preliminary motivator for future studies looking at breaking the Dr. Seuss corpus into sub-categories based on genre. A standard Seussian rhyme book *The Butter Battle Book* and *a Dr. Seuss beginner book*, in *Green Eggs and Ham* (chosen part as it was written with an overt corpus-research influence), were compared to the larger corpus using Web-VP BNC-20. The lexical coverage of the corpus at large will be compared to those of the two texts. This comparison allows for further conversation on the nature of the differing genres by highlighting what words are unique to each text.

## RESULTS

### RQ #1: What is the frequency coverage of Dr. Seuss writing compared to a general English corpus?

After removing all proper names (N-107) and removing inaccuracies in the corpus that appeared as offlist words in initial searches, the Web-VP BNC-20 analysed the 26,263 token corpus and showed that K1 words accounted for 84.42% coverage, K2 words for 5.71% coverage and off list words accounted for 2.35% coverage of the corpus. There was an occurrence of 95% coverage (95.64%) at the K5 band level while 98% coverage (98.00%) occurred at the K15 band level. The significance of 95% and 98% text coverage will be discussed in the discussion of the results. Words (at least 7) were found in all of the K-band levels except for the K20 band level. A more detailed look at the number of families, types and tokens at each of these levels is shown in Table 1.

**Table 1.** Selected K-Bands from the Seuss Corpus Analysis Using Web-VP BNC-20

| Freq. Level | Families | Types | Tokens | Coverage % |
|---|---|---|---|---|
| K1 Words | 766 | 1283 | 22171 | 84.42 |
| K2 Words | 349 | 528 | 1500 | 5.71 |
| K5 Words | 109 | 149 | 345 | 1.31 |
| K15 Words | 18 | 22 | 42 | 0.16 |
| Off-List | ? | 324 | 618 | 2.35 |
| Total | 1984+? | 3217 | 26263 | 1.00 |

**RQ #2: How do the most frequent words across the Dr. Seuss corpus reflect the most frequent words found in previous analyses of children's literature corpora?**

The results from the Web Frequency Indexer v 1.3 analysis of the Dr. Seuss corpus provided a list of the most frequent words throughout the Dr. Seuss corpus. Certain words were excluded from the Seuss list as they are highly frequent in multiple grammatical categories. The word '*like*' was not included in the verb list as it is also frequent as an adjective and the word '*fun*' was omitted from the adjectives list as it is also highly frequent as a noun. The words are extracted to match the results from the (CLLIP) corpus of children's imaginative fiction (Thompson & Sealey, 2007) and are displayed in Table 2.

**Table 2.** Most Frequent Lexical Verbs, Adjectives and Nouns from the CLLIP and Dr. Seuss Corpora.

| Lexical verbs | | | | Adjectives | | | | Nouns | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CLLIP | % | Seuss | % | CLLIP | % | Seuss | % | CLLIP | % | Seuss | % |
| said | 6.30 | said | 0.48 | old | 2.13 | little | 0.34 | time | 1.18 | king | 0.45 |
| see | 1.33 | go | 0.29 | good | 1.91 | right | 0.23 | way | 0.86 | cat | 0.36 |
| know | 1.27 | say | 0.27 | little | 1.78 | big | 0.22 | thing | 0.72 | day | 0.22 |
| go | 1.24 | look | 0.25 | other | 1.54 | good | 0.18 | head | 0.71 | things | 0.21 |
| get | 1.23 | see | 0.25 | long | 1.31 | old | 0.17 | eyes | 0.71 | fish | 0.17 |
| looked | 1.17 | know | 0.23 | small | 1.1 | new | 0.15 | face | 0.63 | house | 0.17 |
| got | 1.16 | get | 0.21 | big | 1.08 | great | 0.13 | door | 0.63 | cats | 0.16 |
| come | 1.01 | come | 0.20 | great | 1.06 | high | 0.12 | people | 0.62 | hat | 0.16 |
| going | 0.98 | saw | 0.17 | sure | 0.96 | royal | 0.12 | day | 0.62 | head | 0.13 |
| think | 0.94 | eat | 0.15 | right | 0.93 | long | 0.12 | man | 0.57 | zoo | 0.13 |

### RQ #3: What is the frequency coverage of Dr. Seuss writing compared to a children's spoken corpus?

After removing all proper names (N-107) and removing inaccuracies in the corpus that appeared as offlist words in initial searches, the VP-Kids v.9 was used to provide analysis of the corpus. The results show the first 250-word band, which represents the 75-80% of words children use when speaking their native language (Roessingh, n.d.), providing 69.70% coverage, a gradual decrease in the percentage of coverage from bands 2 through 10, off-list known words providing a 2.16% coverage off-list unknown words providing 5.23% coverage of the Dr. Seuss corpus. Before entering into a discussion of the results, it should be mentioned that items appearing in the off-list unknown words category seem at a glance to be more frequent than some items on the off-list known words list or even the higher bands. Words like this include '*snack*', '*sprain*' and '*dawn'*. The results of the VP-Kids v.9 analysis are below in Table 3.

**Table 3**. Results VP-Kids v.9 Analysis of the Dr. Seuss Corpus

| Freq. Level | Families | Types | Tokens | Coverage % |
|---|---|---|---|---|
| Kid250 – 1 | 300 | 519 | 18291 | 69.7 |
| Kid250 – 2 | 193 | 345 | 2259 | 8.61 |
| Kid250 – 3 | 170 | 274 | 1148 | 4.37 |
| Kid250 – 4 | 140 | 216 | 795 | 3.03 |
| Kid250 – 5 | 128 | 189 | 484 | 1.84 |
| Kid250 – 6 | 83 | 135 | 363 | 1.38 |
| Kid250 – 7 | 102 | 150 | 317 | 1.21 |
| Kid250 – 8 | 89 | 120 | 256 | 0.98 |
| Kid250 – 9 | 75 | 104 | 223 | 0.85 |
| Kid250 – 10 | 67 | 81 | 170 | 0.65 |
| Off-List known | 182 | 222 | 566 | 2.16 |
| Off-List unknown | ? | 878 | 1372 | 5.23 |
| Total | 1529+? | 3233 | 26244 | 1 |

### RQ #4: What is the frequency of imaginary lexical inventions in Dr. Seuss writing?

An analysis of the Dr. Seuss corpus, with massive adjustments made to the off-list exceptions list (inclusion of any previously determined offlist word that was an actual English word ('*sailboat*') or a string of letters used solely for phonetic purposes ('*ahhh*'), was done using the Web-VP BNC-20. As the exclusion list automatically places exclusion words into the K1-

band of words, coverage outside of the off-list words will not be of interest for this research question. This analysis showed that Seussian imaginative terms such as '*Oobleck*', *'kerchoo'* and '*bopulous*' account for 2.63% of the running words in the Dr. Seuss corpus.

### RQ #5: How does the frequency coverage of selected Dr. Seuss texts compare with the larger Dr. Seuss corpus?

Results of Web-VP BNC-20 analyses of the entire Dr. Seuss corpus, *The Butter Battle Book* and *Green Eggs and Ham* were compared. As the sizes of these comparable differ greatly, only percentage of K-band converges will be compared. The results of this cross comparison are shown below in table 4. The results of this comparison show a striking difference in the make-up of the 630 token beginner book, *Green Eggs and Ham* from the general corpus and the 1237 token standard Seussian rhyme book The *Butter Battle Book*. The early phonic reader is covered almost entirely by K1-band words (94.8 % coverage compared to 81.43% for *The Butter Battle Book*) and has no offlist or imaginative words whereas the standard Seussian text is much more representative of the corpus in general in terms of coverage and off-list words. The results of this cross comparison are shown below in table 4.

**Table 4**. Comparison of Lexical Coverage of the Entire Dr. Seuss Corpus, *The Butter Battle Book*, and *Green Eggs and Ham*.

| BNC Freq. Band | Dr. Seuss Corpus Token Coverage % | *Butter Battle Book* Token Coverage % | *Green Eggs and Ham* Token Coverage % |
|---|---|---|---|
| K1 Words | 84.42 | 81.43 | 94.80 |
| K2 Words | 5.71 | 6.97 | 1.36 |
| K3 Words | 2.76 | 2.35 | 2.48 |
| K4 Words | 1.44 | 1.54 | 0.87 |
| Off-List | 2.35 | 3.55 | 0 |

## DISCUSSION

The analysis of the Dr. Seuss corpus, using the different tools described above, has provided some interesting findings that reveal unique characteristics of Seussian writing, while at the same time showing evidence of the works in the corpus being fairly representative of children's literature in general. Other results from the preliminary results of the corpus analysis show that there is more work to be done in terms of refining the corpus and suggest directions for future research. The

following section will examine the results stemming from analyses done in order to answer the five research questions and will discuss implications for educators as well as implications for future research.

In examining the lexical coverage across K-bands from the BNC, the first 1000 most frequent words of the BNC provided 84.42% coverage of the Dr. Seuss corpus, with knowledge of the first 5000 words of the BNC necessary for 95% coverage of words and the first 15 000 for 98% coverage of the Seuss texts (although the entire corpus itself is made up of only approximately 2000 families). According to Nation (2001, 2006), readers of differing genres of texts need to be familiar with 95% of a text to be able to comprehend it and 98% to read for enjoyment and that a vocabulary of 8000-9000 words sufficient for the comprehension of most texts, including newspapers. With these figures in mind, it would appear that a much larger vocabulary is needed for full comprehension of Dr. Seuss texts. Although studies such as Schmitt, Jiang and Grabe (2011) have refuted the threshold levels proposed by Nation, this is still evidence of a corpus with a very rich vocabulary. Webb (2012) in his own analysis of the usefulness of children's literature found that children's literature is often as lexically dense as adult fiction and may not be appropriate for ESL students with limited vocabulary. The results of the Seuss corpus analysis reflect this finding, so teachers may need to proceed with caution in using Dr. Seuss texts in lieu of graded readers. This is surprising as the material is intended for a younger audience and is also comprised of many simple beginner books. The results provide evidence of the books being more suitable for extensive reading than intensive reading as intensive reading for fluency development is best achieved using texts with almost no unknown vocabulary (Nation, 2001). If used for extensive reading, the books will be a valuable resource as extensive reading is most successful when L2 learners are motivated to take advantage of available resources (Horst, 2009). The attractiveness of the texts will also be beneficial in fostering early L1 literacy as print motivation is one of the key principles in early L1 literacy (Ghoting & Martin-Díaz, 2006). Dr. Seuss books are colourfully illustrated and playful which will attract learners and the books are often available in class or school libraries. Having lexical information for individual books, especially the more popular titles that are more often found in schools, would help teachers better guide their students (English L1 or ESL) towards more suitable reading material.

In examining the most frequent words in different categories to a similar children's corpus, there was a fair amount of overlap in lexical verbs and adjectives. The words contained in the lexical verb list are

representative mostly of fiction rather than of writing for a children's audience as they show protagonists making their way through the imagined worlds (*came, went*) and their descriptions of their perceptions of it (*said, looked*) (Thompson & Sealey, 2007). The similarities in the lists of adjectives, on the other hand, may be representative of children's literature as they show a child's description of things or characters in their world as being old but not young (Thompson & Sealey, 2007). Although many of the items on the list were similar, the items in the Dr. Seuss lists appeared with much lower frequency than in the CLLIP corpus, the highest lexical verb and adjective items in the CLLIP, respectively, accounted for 6.3% and 2.13% of running text while the figures in the Dr. Seuss corpus were much smaller at 0.48% and 0.34% respectively. This is even more striking when considering that the Dr. Seuss corpus is only about 1/9 the size of the CLLIP corpus and contains many basic early readers with limited vocabulary, thus providing evidence of there being a much wider range of vocabulary found in the Seuss corpus than in the CLLIP corpus. This is reinforced by the Web-VP BNC-20 analysis of the Dr. Seuss corpus which showed it to contain a wide range of words through almost all 20 BNC K-bands. The results from the most frequent nouns also follow the same frequency percentage patterns as the other two categories (frequencies of most frequent nouns in the corpora were 1.18% against 0.45%) but differ in other ways. Only two words, head and day are found in both lists. The most frequent noun in the Dr. Seuss corpus, king, and other items such as *cat* and *fish* are representative of the stories in the Dr. Seuss corpus and show how, despite the fact the Dr. Seuss corpus has great lexical variation, certain key words are repeated quite often throughout stories. Knowing which words are most frequent in individual stories will inform teachers as to which stories contain what highly repeated words. Repetition of words in the input, especially if well-spaced (and Dr. Seuss books tend to be read by children more than once), is vital for L2 vocabulary acquisition (Nation, 2001). As the development of vocabulary is also a tenet of early L1 literacy development (Ghoting & Martin-Díaz, 2006), lexical breakdown by frequency for individual books could prove a valuable resource for both L1 and L2 teachers of English in order to help make informed decision choosing appropriate course material.

The results of the VP-Kids v.9 coverage analysis show that 92.62% of the Dr. Seuss corpus is covered by the first 2500 most frequent words found in children's spoken English, and that 69.70% is covered by the most frequent 250 words. Roessingh (n.d.) claims that children use the

first 250 frequent words for 75-80% of their speech, so the coverage of the in the Dr. Seuss falls short of that mark, but not significantly. Considering the richness of the vocabulary and the presence of Seussian imaginative terms, it is somewhat surprising that the corpus represents children's speech as well as it does. To see whether Dr. Seuss early readers would better represent the children's spoken corpus, two early readers were passed through the VP-Kids tool. Interestingly, the 10 250 word K-bands provided worse coverage (90.33%) and the first 250 words only accounted for 59.5% of the words in the two stories. This raises questions as to the merit of comparing children's literature to a children's spoken corpus would be as the simplest stories in the corpus do not reflect native children's speech as well as the entire corpus which contains more challenging works and offlist words. The pedagogical implications of these results for the teaching of L1 and L2 English are still inconclusive at this point and require further investigation.

In terms of the proportion of imaginary words present throughout the Dr. Seuss corpus, the results show a lower percentage than maybe would have been expected from the works of an author famous for invented lexical terms. One of the explanations for this would be the occurrence of imaginary terms such as *grinch-ish-ly* which are recognised as three separate entries due to the separation of the components by hyphen. A word like this then does not appear in the list of imaginative Seussian terms as per the methodology prescribed. This is counterbalanced, however, by the inclusion of proper names from stories like Nizzard which account for a large proportion of the imaginative lexical terms. Perhaps a better set of criteria and means of controlling the Web-VP BNC-20 output would better serve this analysis. Despite, the limitations of the search for these terms, the list of terms created provides a great deal of food for thought. A subdivision of these terms may be useful as they are created in different ways, for instance by adding affixes to real words or proper names (*lightninged, grinch-ish-ly*), by creating minimal pairs for a rhyme scheme (*curtain/jertain, mustard/flustard*) or by straight lexical invention which could be of use for phonetic practice (*thwerll, zooskie*). Having these terms indexed by story and by category would give teachers control over how they may choose to use these terms with students. Certain terms, like the minimal pairs or the phonetically challenging terms, could be used for tackling pronunciation issues L2 students may be encountering. Words created imaginatively through the use of affixes could be used by teachers to help L1 and L2 students alike gain morphological awareness of word parts and affixes which can be a very

successful strategy in helping learners guess new words from context and subsequently learn new vocabulary (Nation, 2001). Although teachers might be wary of using texts with non-words with students who are emerging L1 and L2 English readers the lexical inventions can be used pedagogically in various ways and a breakdown of individual texts for non-words could steer teachers clear of titles that contain higher proportions of the items than others, should they choose.

Finally, in comparing two different types of texts from the Dr. Seuss corpus to the entire collection of texts, some striking differences are observable. There is a clear difference in the coverage provided by the BNC K-bands of a standard Seussian text, *The Butter Battle Book,* compared to a Dr. Seuss early reader, *Green Eggs and Ham*. This is to be expected as the latter was written in response to a challenge to write a successful children's book using less than 50 highly frequent word families from a children's corpus and are therefore well covered by the K1-band and contain no offlist words. This further supports the idea that subdividing the corpus into smaller Dr. Seuss genres would be appropriate for drawing any true concordancing or lexical frequency statistics. This is further supported by the findings from the previous research questions that show that a more fine-toothed analysis of the corpus, genre by genre and even title by title, would provide the most robust and relevant pedagogical information for researchers and L1 and L2 teachers. This will prove to be a valuable endeavour as examining an entire decontextualized corpus may not reveal the richness it contains and may in some ways actually be pedagogically misleading (Flowerdew, 2009).

## CONCLUSION

Despite the fact that Dr. Seuss' books contain very liberal use of grammatical forms in creating imaginative language and vocabulary, Theodor Geisel was quite conscious of the use of highly frequent words from children's literature (Menand, 2002). In response to a challenge to write a book for children in language they could understand, Geisel wrote *The Cat in the Hat* using a 225 word children's corpus made of word lists for early readers combined with 21 word choices of his own. Geisel, following the success of *The Cat in the Hat*, spent more time working on early beginner reader books and eventually wrote *Green Eggs and Ham* using a list of only 50 words, 49 being monosyllabic. These two books, informed by corpus research, transformed the landscape of children's literature and primary education (Menand, 2002).

This synchronistic finding has led our investigation toward the possibilities of a corpus approach being used to underline efficacious choices in literary primers and classroom exercises when introducing an author's works to second language or young first language readers. Thus, by means of a corpus analysis, particularly rich or complex instances of lexical, phonic and syntactical writing can be identified and highlighted in a manner amenable to teaching reading. This was confirmed by our experience of presenting these preliminary results at an ESL conference where local educators requested vocabulary lists of unconventional vocabulary to avoid confusion in their students, who might mistake the invented word for conventional vocabulary.

As the results of the preliminary analysis of the Dr. Seuss corpus have shown, Dr. Seuss' books are lexically rich and require a great deal of lexical knowledge for ease in comprehension. They also contain a variety of imaginative non-words and share some, but not all, of the characteristics of vocabulary seen in other children's literature. While the findings of this study are limited by the incomplete corpus and the limitations of the tools used on the Compleat LexTutor Website, this preliminary analysis does suggest that there is a great deal to be gained by further parsing the corpus along genre lines and pursuing a deeper analysis of the sub-genres and individual texts. Recommendations for future studies, upon full completion of the corpus, include a sub-division of the corpus, analysis of the subdivisions and individual texts and a complete inventory of invented imaginative terms in Dr. Seuss' works.

## ACKNOWLEDGEMENTS

The authors would like to thank Dr. Marlise Horst and Dr. Walcir Cardoso for their contributions and suggestions throughout the process of assembling the corpus as well as for providing direction in the discussion of the results. The authors would also like to thank Dr. Tom Cobb for the use of the online LexTutor corpus analysis tool and Eve Dewald for her assistance in assembling the corpus. We would also like to thank the COPAL editors and reviewers for their valuable input in the writing of this article. Finally, we thank our parents and teachers for sharing Seuss' works with us as children, a joy we were able to revisit as we spent time reading the books and examining them from new angles.

## REFERENCES


Barbieri, F., & Eckhardt, S. E. (2007). Applying corpus-based findings to form-focused instruction: The case of reported speech. *Language Teaching Research, 11*(3), 319-346.

Cobb,T. Web Vocabprofile [accessed 17 November 2012 from http://www.lextutor.ca/vp], an adaptation of Heatley & Nation's (1994) Range.

Cott, J. (1983). *Pipers at the gates of dawn: The wisdom of children's literature.* New York: Random House.

Crystal, D. (1996). Language play and linguistic intervention. *Child Language Teaching and Therapy, 12*(3), 328-344.

Davies, M. (2009). The 385+ million word corpus of contemporary American English (1990-2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*(2), 159-190.

Dodigovic, M. (2005). Vocabulary profiling with electronic corpora: A case study in computer assisted needs analysis. *Computer Assisted Language Learning, 18*(5), 443-455.

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics, 14*(3), 393-417.

Göbel, T., & Peetz, M. H. (2005). CCB: A corpus of children's books. Retrieved from http://peetz-intelligence.com/http://peetz-intelligence.com/

Ghoting, S. N., & Martin-Diaz, P. (2006). *Early literacy storytimes@ your library: Partnering with caregivers for success*. American Library Association

Heatley, A., & Nation, P. (1994). Range. Victoria University of Wellington, NZ. [Computer program, available at http://www.vuw.ac.nz/lals/.]

Held, J. M. (2011). *Dr. Seuss and philosophy: Oh, the thinks you can think!* Lanham: Rowman & Littlefield Pub Incorporated.

Horst, M. (2009). Developing definitional vocabulary knowledge and lexical access speed through extensive reading. In Z. H. Han & N. J. Anderson (Eds.), *L2 reading research and instruction: Crossing the boundaries*. Ann Arbor: University of Michigan.

Horst, M., White, J., & Bell, P. (2010). First and second language knowledge in the language classroom. *International Journal of Bilingualism, 14*(3), 331-349.

Jenkins, J. R., Vadasy, P. F., Firebaugh, M., & Profilet, C. (2000). Tutoring first-grade struggling readers in phonological reading skills. *Learning Disabilities Research & Practice, 15*(2), 75-84.

Keck, C.M., (2004). Corpus linguistics and language teaching research: Bridging the gap. *Language Teaching Research*, *8*(83), 83-109.

Kies, D. (1990). Three principles underlying iconicity in literature: The poetics of nonsense in children's and general literature. Retrieved from http://papyr.com/hypertextbooks/grammar/iconicity.htm.

Lathem, E. C. (2000). *Who's who & what's what in the books of Dr. Seuss.* Hanover, N.H: Dartmouth College.

Lange, K. N. (2009). Oh, the things you can find (If only you analyze): A Close textual analysis of Dr. Seuss' rhetoric for children (Doctoral dissertation). Kansas State University, Manhattan Kansas.

Mahlberg, M. (2006). Lexical cohesion: Corpus linguistic theory and its application in English language teaching. *International Journal of Corpus Linguistics*, *11*(3), 363-383.

Menand, L. (2002, December 23). Cat People. Retrieved from http://www.newyorker .com.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

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

Nilsen, D. L. (1977). Linguistic and non-linguistic images in Dr. Seuss: Or, how to read between the lines. *TESL Reporter*, *11*(1), 8-9.

Notestine, R., & Tanner, P. (2007). Oral interpretation of Dr. Seuss stories in the classroom. In K. Bradford-Watts (Ed.), JALT 2006 Conference Proceedings. Tokyo: JALT.

Powell, R., & Forbes, S. (2005). Ein Geirau NI: Corpus of children's literature in welsh. Retrieved from http://www.egni.org/.

Reinhardt, J. (2010). The potential of corpus-informed L2 pedagogy. *Studies in Hispanic & Lusophone Linguistics, 3*(1), 239-251.

Roessingh, H. (n.d.). Profiling the vocabulary of k – 2 learners. Retrieved from http://www.ucalgary.ca/languagelearning/profilingvocabulary+

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, *95*(1), 26-43.

Schroth, E. (1978). Dr. Seuss and language use. *The Reading Teacher*, 748-750.

Sealey, A., & Thompson, P. (2007). Corpus, concordance, classification: Young learners in the L1 classroom. *Language Awareness, 16*(3), 208-223.

Thompson, P., & Sealey, A. (2007). Through children's eyes?: Corpus evidence of the features of children's literature. *International Journal of Corpus Linguistics,12*(1), 1-23.

Webb, S., & Macalister, J. (2012). Is text written for children useful for L2 extensive reading? *TESOL Quarterly*. Published online 12 October 2012, DOI: 10.1002/tesq.70.

# APPENDIX A

## TITLES IN THE PRELIMINARY DR. SEUSS CORPUS

- *Green Eggs and Ham*
- *The Cat in the Hat*
- *One Fish Two Fish Red Fish Blue Fish*
- *Hop on Pop*
- *Oh, the Places You'll Go!*
- *The Cat in the Hat Comes Back*
- *How the Grinch Stole Christmas!*
- *I Can Read with My Eyes Shut!*
- *There's a Wocket in My Pocket!*
- *Yertle the Turtle and Other Stories*
- *Ten Apples Up on Top!*
- *Horton Hatches the Egg*
- *Happy Birthday to You!*
- *Mr. Brown Can Moo! Can You?: Dr. Seuss's Book of Wonderful Noises!*
- *Dr. Seuss's ABC*
- *And to Think That I Saw It on Mulberry Street*
- *Bartholomew and the Oobleck*
- *If I Ran the Circus*
- *The Butter Battle Book*
- *The King's Stilts*
- *If I Ran the Zoo*

Based on the data from our self-made corpus, this paper aimed to make a comparative analysis of the lexical bundles in the English introductions of Chinese graduate studentsâ€™ theses and those written by the graduate students in the international prestigious universities. The most frequently used four-word lexical bundles in the corpus were identified and classified structurally and functionally and corresponding analysis was made. Results indicated that lexical bundles in the English introductions of Chinese studentsâ€™ theses were structurally incomplete in comparison with those in the internat (In a bottom-up approach, the lexical and/ or form-focused corpus analysis comes first, and the discourse unit types emerge from the corpus patterns. See Biber et al., 2007, for discussion.)Â  Can a corpus be analyzed to identify the general patterns of discourse organization that are used to construct texts, and can individual texts be analyzed in terms of the general patterns that result from corpus analysis? Few studies have attempted to combine these two research perspectives.Â  One of the major methodological problems to be solved by any corpus-based analysis of discourse structure, then, is deciding on a unit of analysis. The use of corpus managers for analysis of large data files has been proposed more than once in translation studies by Baker who also published several empirical studies with examples of such analyses (e.g. Baker 1993, 1995, 2000). A similar proposal for interpreting studies was made in Shlesinger (1998).Â  Rationale Lexical density is one of the key quantitative corpus parameters (Stubbs, 2002:39). The parameter is based on the fact that languages are composed of content words which are the primary carriers of meaning (nouns, adjectives, verbs, etc.) and function words (auxiliary verbs, pronouns, conjunctions, etc.). u Text analytical tools. u A corpus-based lexical study â€“ Academic Word List (Coxhead, 2000). u What corpus linguistics is. OUHK - RIDCH 18th Seminar (April 2016) - Corpus linguistics as a research method. 2. What a corpus is. u A corpus is â€œa collection of pieces of language text in electronic formâ€ (Sinclair, 2004, p. 19).Â  u Purpose of your research u Criteria of the target corpus u Survey of existing corpora u Existing/self-built corpus. u Categories of texts in the existing corpus u Structure and size of the self-built corpus. OUHK - RIDCH 18th Seminar (April 2016) - Corpus linguistics as a research method. 12.Â  u derivations. u analyser, analysers, analysis, analyst, analysts, analytic, analytical, analytically. A corpus-based investigation of lexical pragmatics is in many respects a pilot project. There is no established paradigm,1 and since the interpretation of utterances in discourse may depart significantly from the linguistically encoded meaning, the analystâ€™s intuitions about the intended interpretation must play an important role.Â  (For further analysis and discussion, see the Corpus Analysis section of the AHRC Lexical Pragmatics website hosted by the Department of Phonetics & Linguistics, UCL.) 2. Lexical narrowing Lexical narrowing involves the use of a word or phrase to convey a more specific concept (with a narrower denotation) than the linguistically encoded â€˜literalâ€™ meaning. To illustrate, consider (1) and (2): (1) Mary is a working mother.