DE GRUYTER OPEN

ICAME Journal, Volume 39, 2015, DOI: 10.1515/icame-2015-0004

Word frequency and collocation: Using children's literature in adult learning

Ed Thomas, Kanda University of International Studies, Japan

Abstract

This study involved the creation of a corpus of children's literature spanning 5.5 million words. Using concordance software, the corpus was able to show the most frequent words and collocations. These will be of interest both to literary researchers in the genre of children's literature and also teachers and applied linguists working with adult students of English.

1 Introduction

Modern computer technology offers us a view of language structure that was not available to researchers before (Sinclair 2004: 1). The growth of computer-aided corpus linguistics has seen a surge of interest in collocation (Shin and Nation 2008: 339) – something that distinguishes native from non-native language users (Schmitt 2000: 79).

Separate to that, literature and naturally-occurring text is once again becoming a primary source of input for language teachers and course designers (Khatib 2011: 201). While audiolingual and communicative language teaching (CLT) models replaced older text-reliant grammar translation methods in the last century, computer corpus linguistics is once again bringing text to the classroom as an authentic tool for learning. On a personal level, I have gainfully used children's stories in adult teaching as a way of inspiring students to develop a range of imaginative and linguistic skills. These include narrative building, character development, reading and writing skills (such as use of narrative verb tenses, or descriptive adjectives) as well as pronunciation aid through storytelling. Others too believe children's literature can play "a crucial role in the education of EFL adult [students]" as "an initial step in developing literary competence" (Ho 2000: 268). To date, a sizable corpus of children's text is yet to be compiled for pedagogical purposes. The aim of this study was to build such a corpus and extract a basic word frequency count and a set of common collocations.

Words are collocates if, in a given sample of language, they are found together more often than their individual frequencies would predict (Jones and Sinclair 1974: 19). This tendency to co-occur can be measured statistically with a computer.

2 Theoretical background

J. R. Firth stated we best know the meaning of a word not by examining it in isolation but by the company it keeps (Firth 1957: 11). This emphasis on collocational properties within meaning is something now widely accepted by linguists (Nation 2001: 56; Milton 2009: 149). A 'neo-Firthian' school has emerged, looking at words in combinations, phrases, formulaic sequences and chunks (Halliday 1973; Leech 1974; Sinclair 1991; Nattinger and DeCarrico 1992; Louw 1993; Weinert 1995; Stubbs 1995; Schmitt 2004; Hoey 2005; Almela Sanchez 2006; Biber 2009). The advent of corpus techniques has done much to bolster the position, as "Analyses of corpora by computer have now revealed detailed and hitherto unsuspected patterns of idiomaticity... and in so doing have provided descriptive substantiation of the insights that Firth expressed over fifty years earlier" (Widdowson 2007: 410).

Hoey's recent theory of lexical priming marks "the pervasiveness of collocation" as its starting point (Hoey 2005: 1). His examination of the statistical distribution of words leads him to a theory that every word is lexically 'primed' to co-occur with others, and as a word is acquired through encounters with it in speech and writing it becomes "cumulatively loaded" with context and meaning (Hoey 2005: 8). Other researchers, too, owe their impetus to Firth (Halliday 1966; Leech 1974; Sinclair 1991; Nattinger and DeCarrico 1992; Louw 1993; Weinert 1995; Stubbs 1995; Hopper 1998; Schmitt 2004). It is becoming accepted that priming within language (certain words showing greater attraction to each other, such as *coffee* to *mug* more so than *tea* to *mug*, for example) is a well-established and widely-tested phenomenon of the human mind (Pace-Sigge 2013: 167).

3 Previous studies relating to the classroom

3.1 Shin and Nation (2008)

Shin and Nation (2008) analysed a spoken corpus to aid the teaching of collocations in ELT. Motivating them was the belief that learning common collocations will develop fluency and native-like selection. There is always more than one possible way to say something, but only one or two ways will sound natural to a

native speaker (Pawley and Syder 1983). They cite examples from Korean students of English who used *lying story* instead of *tall story* or *artificial teeth* instead of *false teeth*. Both *lying story* and *artificial teeth* are grammatically and semantically correct, but they fail to sound native.

The authors used a computer to search the 10 million-word spoken section of the British National Corpus (BNC). Using only content words (nouns, adjectives, verbs and adverbs) as their pivot (search) words, they found the most frequent collocations to include *you know*, *I think*, *a bit*, *always / never used to*, *as well*, *a lot of*, *[number] pounds*, *thank you*, *[number] years* and *in fact*. They noted the "here-and-now" nature of spoken language in their findings, with many interjections and amplifiers (*you know*, *a bit*, *come on* and so on) in the high-frequency list.

3.2 Baker and Freebody (1989)

Corpus-based studies relating specifically to children's literature are small in number. Indeed, "The consideration of literature written for children from a linguistic perspective is a comparatively new field of study" (Knowles and Malmkjaer 1996: 1). Baker and Freebody uploaded text from 163 primary school readers and carried out an analysis. Of particular interest was the appearance of *little* in the highest frequency range of words: ranked number 18 in the corpus, and the only two-syllable word in the top 20. Its relative length (compared to words such as *the*, *and*, *a*, *to*, *I*) together with its "grapho-phonetic irregularity" thus invited detailed examination.

Another feature of children's books their corpus showed was the relative occurrence of *boys* and *girls*; *boys* appeared more frequently than *girls* by a ratio of about 3:2. An analysis of verbs showed *boys* to be more energetic in their interactions with others (*shout*, *hurt*, *work* being amongst the frequent collocations with masculine nouns and pronouns) while *girls* were seen to *like*, *play with*, *talk to*, *walk with*, *hold onto* and *kiss*. Their analysis suggested girls are more emotional and less physical in their portrayal. In terms of common adjectival collocation (besides *little*), girls were *young*, *pretty* and *dancing*, while boys were *brave*, *kind*, *sad* and *naughty.* Fathers did things like *paint*, *pump*, *fix*, *drive*, *pull*, *start*, *shout* and *let*, while mothers *baked*, *dressed*, *hugged*, *kissed*, *packed*, *picked*, *set*, *splashed* and *thanked*. They thus argued a certain social theory was "embedded" in the literature, with computer corpus methods bringing this to light (Baker and Freebody 1989: 135).

3.3 Stubbs (1995)

Baker and Freebody's corpus comprised 83,838 words – small by today's standards. Stubbs (1995) used larger corpora and the bearings of newer thinking in lexical studies (Sinclair 1991; Halliday 1993; Weinert 1995) to place collocation and 'chunks' at the centre of an emerging theory of language learning.

He confined his search to just four pivot words – *large*, *small*, *big* and *little* – and used both a 2.3 million word corpus of contemporary English and data from the *Oxford English Dictionary* CD-ROM to investigate their collocates. He made insights into ambiguity and meaning, noting how word clusters take on connotations of their own which dictionary definitions are unable to describe consistently (Stubbs 1995: 381). While *little* and *small* may appear to be synonymous, he sees *Little Red Riding Hood* as an example where *small* cannot be substituted as an adjective without meaning changing in some way. He follows Baker and Freebody in emphasising how *little* "connotes cuteness" in a way that *small* does not. Apart from in certain metaphorical and pejorative phrases (*small fry*, *small beer*), *small* is usually about physical size. *Little*, however, carries more in terms of ideological message; it has a "cuddle factor" (Stubbs 1995: 383).

Stubbs finds other examples of common collocations using his corpora – *big toe* and *little finger*, but *large intestine* and *small intestine*. He notes how usage is often nonliteral: one's big toe might be small in size, one's little brother might be bigger than oneself (Stubbs 1995: 384). *Big* can carry with it positive connotations of being grown-up (*Big boys don't cry*), or negative ones of self-importance (*big fish*, *big mouth*, *big head*, *big guns*). *Large*, on the other hand, is usually confined to mean more-than-average in terms of quantity (*large amount*, *large majority*, *large part*, *large-scale*).

Learning a language therefore means learning such fixed and semi-fixed units which are not always coextensive with traditionally recognized syntactic units (Stubbs 1995: 386). This is a point made by Sinclair (1987, 1991) when he talks of the "idiom principle" within language. For neo-Firthians like Stubbs and Sinclair, they believe once a word is selected for use there is a high probability that other words and features of grammar are co-selected with it (Stubbs 1995: 386), and these linguistic relations must be learned when one learns a language. An L2 learner cannot simply translate from one language to another using a dictionary; so much more in terms of connotation and ideological message is conveyed through words and their collocates.

3.4 Knowles and Malmkjaer (1996)

The following year Knowles and Malmkjaer carried out a study of children's books both from a literary and linguistic point of view. They developed ideas of ideology, stereotyping and implicit messaging using research from the social sciences (for example Thompson 1990) and applied them in a genre-specific way. They argued that writers' choices aid the creation and maintenance of relations of power in society (Knowles and Malmkjaer 1996: 68) – be it men dominating women, adults dominating children, or any other relation of power. This is so whether the writer intends it or not (Knowles and Malmkjaer 1996: 68), and impressions are built up by whole texts, passages, clauses, phrases or just collocations of words. Chunks of language become "linguistic facts", and what the world is actually like is not so much an issue for writers (Knowles and Malmkjaer 1996: 69). To use their example, a girl is more likely to be described as *blonde* than a car – even though the car's colour may be very similar to a fair-haired person. These "linguistic facts" have become firmly established, so that everyone talks of brains and eggs being *addled*, but butter and bacon as *rancid* (Knowles and Malmkjaer 1996: 69). They borrow Louw's (1993) term "semantic prosody" for describing words' tendencies to appear together and call certain associations to mind. *Pretty* calls to mind smallness and femininity, bringing them to the claim: "It would not be unreasonable to suggest that in the normal course of events semantic prosodies are... learnt collocationally" (Knowles and Malmkjaer 1996: 70).

Children, when learning language, are unlikely to be explicitly told which associations are called to mind when seeing words together. Rather, they "gain this impression through exposure, gradually, as part of a developing base of implicit knowledge about the language system" (Knowles and Malmkjaer 1996: 70). It is the task of the writer to exploit this implicit knowledge in presenting passages to the reader and successfully conjuring emotions and reactions (Knowles and Malmkjaer 1996: 71). Consequently, linguistic "habits" and stereotypes are perpetuated, potentially having damaging effects on groups or individuals (Knowles and Malmkjaer 1996: 71). They consider the word *black* as an example. It carries negative connotations due to well-known collocations such as *black magic*, *Black Wednesday*, *black sheep, black cloud*, so that when it occurs alongside a word like *man*, a negative impression is conveyed (even if the colour of the man's skin is, or is very close to, black). Traditional male and female roles, stereotypes about family relations, power structures within society and moral codes are all present in children's literature, they argue, and these stereotypes and language habits are slowly engrained into children's minds as they acquire language.

3.5 Recent years

Unfortunately in the past two decades, not a great deal more has been done in this research area. Thompson and Sealy (2007) compiled a corpus of children's texts from the BNC for comparison with texts written for adults. Their aim was "to explore the issue of whether language deployed in writing for children can been seen to represent the world and human experience differently form the ways in which they represented in writing for adults" (Thompson and Sealy 2007: 3). However their claims were limited, with their corpus compiled of 30 children's texts amounting to 698,286 words.

A current teaching trend to emerge from corpus linguistics is "data-driven learning" or DDL (Johns 1991; Gavioli 2000; Braun 2007; Chambers 2007). Students are now being given access to corpora to discover patterns and rules within authentic language themselves. Leel (2011) took this framework and used a modern work of children's literature – J.K.Rowling's *Harry Potter and the Philosopher's Stone* – as his authentic language data. A concordancer allowed his students to see collocational properties of prepositions such as *about* and *around*, *in* and *on*, with which his students made regular mistakes. Leel concluded it was a useful exercise for them which yielded improvement in their literacy.

This "data-driven" approach to teaching is not without its pedagogical issues. Criticisms of Leel's study rest on the size and nature of his corpus: a single novel by a single author. To date, a significant corpus of children's work is yet to be compiled for analysis, hence the need for this study.

4 This corpus

4.1 Compilation

Hunston (2002) notes the four key features of a successful corpus: size, content, representativeness and permanence. It is generally agreed that "bigger is better" (Sinclair 1991; Flowerdew 1996), so I collected children's texts totalling 5,481,834 words. Every text was taken from the online *Children's Bookshelf* of Project Gutenberg, with a full list of titles in Appendix 4 below. They were all published within what literary scholars term the 'Golden Period' of children's literature 1863–1913 (see Hunt 1994: 59; Knowles and Malmkjaer 1996: 16).

Being out of copyright, the text was free to take. One might claim Hunston's criterion of permanence was not met in this body of work due to its dated nature. However, I argue many of these texts are still widely read and experienced today. *Alice in Wonderland*, for example, has been translated into hundreds of languages around the world and to date there have been 29 films and

nine television series made from it. Books like *Alice and Wonderland*, *Treasure Island* and *Peter Pan* are "quintessential classics" (Hunt 2001: 37) and surely rank among the permanent texts of any age. Fairy tales are from the oral and folk tradition and are therefore hard to date or assign authorship. But they were collected and printed in vast quantities during the 'Golden Age' of children's literature. Since then, they have become "ageless" and "inscribed on our minds", remaining with us from childhood throughout the rest of our lives (Zipes 1983: 1). I therefore believe there is a permanent factor to this corpus and much of the language in these texts is still current.

Regarding representativeness and balance of the corpus, both male and female authors' works were used, from a wide range of geographical locations and cultural / linguistic backgrounds. Poems were included as well as prose; novels as well as shorter fairy tales. There was also the text from a number of magazines and 'penny dreadfuls' which were popular at the time, such as *The Chatterbox*, *The Girl's Own Paper* and *St. Nicholas: Scribner's Illustrated Magazine for Girls and Boys*. Bearing in mind the types of text a child living 1863–1913 might have read (or had read to them), attempts were made at providing a representative balance of this language. Excluding editors of the magazines and the editing collectors of fairy and folk tales, some 47 authors were included. Two authors – Jules Verne and Johanna Spyri – appeared in translation, using the English versions of their work which appeared at the time. There were 65 single-author works, 21 editions of magazines, and two compendia of fairy tales, poems, stories, fables and nursery rhymes. Admitted, poetry is an entirely separate genre to novels and brings with it different language. But analyses were carried out on the corpus as a whole, rather than sub-corpora, as a child would have experienced language from all these genres at the same time. One night he / she might read *Kidnapped*, the next *The Owl and the Pussy-Cat*. A first language learner arguably experiences an "immersion pedagogy" while growing up (Gee 1994), and this has guided theories of language teaching in applied linguistics.

Laurence Anthony's program AntConc was chosen as the software to analyse the data. It is free to download and has powerful tools such as word frequency lists, a KWIC concordancer and collocation generators ranked according to raw frequency, mutual information or T-score.

4.2 Research questions

Compilation of the corpus was itself a goal of this study, and took some months to complete and yield results. I therefore limited myself to two very simple research questions as a preliminary use for this corpus. I hope in future to carry

out more detailed analyses and further investigations (such as a study of collocates with the adjectives *little* and *big*, following Stubbs, or a description of collocates relating to gender and age). For now, my research questions were simply:

1. What are the most frequent words in this corpus?
2. What are the most frequent collocations of these words?

From here, possible classroom uses of the data can be discussed and an outline of a basic pedagogical application put forward.

5 Results

As will be the case in any corpus of natural language, the most frequent words were function words – *the*, *and*, *to*, *of*, *a* and so on. With the aim of finding meaningful chunks of language to teach, they were ignored as search words in favour of content (NAVA) words.

Many of these words are *polysemous* – having more than one sense (e.g. *well* can be a place to store water, or otherwise an adverb imparting a sense of 'goodness'. Well, it can even be used as an exclamation, to resume a narrative or change the subject!). Some of the polysemous senses see non-content words entering into consideration as collocation pivots (*back* as a preposition, for example, rather than the place where your spine is). Similarly, many of the verbs on the list can act as auxiliaries, which carry little content (*was*, *had*, *is*, *were*, *are*, *do*, *did* acting as function words as in "They had eaten lunch" or "Did she finish her homework?", rather than main verbs as in "They had lunch" or "She did her homework"). Word classification is therefore an issue in any corpus study, as a computer cannot distinguish between function and content. Trying to take this into consideration, I found the 100 most frequent content words in this corpus comprised:

Verbs	-	44
Nouns	-	23
Adjectives	-	20
Adverbs	-	13

To answer the first research question, the frequency list for the most common content words in this 5.5million word corpus is listed in Appendix 1.

Approaching the second research question, these frequent content words were used as search pivots to find frequent collocations. An initial analysis yielded the results in Appendix 2. One would expect collocations involving

verbs to be ranked highly, considering this is a corpus of largely narrative work. However, because of the computer's inability to distinguish between function and content, many of these verbal collocations failed to convey any real meaning (such as *to be*, *had been*, *did not*, *was not* and *have been*). Due to their lack of content, I therefore excluded these from the list. After all, the motivation behind this study is to help language learners. Chunks like *to be* and *had been* are more a part of functional grammar than lexis and should have been mastered by students before higher level vocabulary learning and literacy improvement. Similarly, article + noun collocations were excluded. A more conclusive list of collocations I found is listed in Appendix 3.

6 Discussion

Looking at Appendix 1, the first thing to note is the highest frequency content word: *was*. This past tense copula verb does a lot of work in this body of text, which is not surprising considering its narrative nature. Indeed, the three most common words are all past tense verbs: *was*, *had* and *said*. Other common verbs are *do*, *see*, *go*, *come*, *know*, *make*, *think* and *take*. From a teaching point of view, good use of these verbs amongst students should be an aim.

Perhaps the next thing to note is the only non-verb in the top ten words: *little*. This occurs 15659 times in the 5.5million word corpus. Neither its synonym *small* nor its antonym *big* ranks in the top 100 words, suggesting *little* is given special status in these texts.

Looking at the collocations list (Appendix 3), *a little* jumps out – ranked the tenth highest collocation. *Little* therefore demands further attention in future research, but a brief investigation at this stage saw common chunks were: *a little girl*, *a little boy*, *a little while*, *a little more*, *a little way*, *a little bit*, *a little longer* and *a little later*. For teachers, these collocations would be useful building blocks to use. (Try substituting *small* into these phrases, however, and you sound rather un-native. The same holds for *big*).

The most common collocation in this corpus is *it was*. The concordancer helped put this phrase into more perspective, and we see that it often refers to very little (acting as a 'dummy' pronoun):

> *It was all very well to say "Drink me" but...*
> *It was high time to go....*
> *It was too late. The boat struck the bank full tilt.*

Similar issues apply for other common collocations such as *it is*, *there was*, *there is* and *there were*. These chunks perform essential functions within story-telling, but the pronouns often refer to nothing in particular. For example:

> *There was nothing else to do, so Alice soon began talking again.*
> *There was a large mushroom growing near her...*
> *There was the cat again, sitting on the branch of a tree.*

He was and *he had* are the collocations ranked second and third (8,258 and 7,352 occurrences). S*he was* is ranked 13th (3,960 occurrences), while *she had* does not feature in the list. Interpretation of these linguistic features must be left for future research, but for now it is useful to note the frequency of the verbs *be* and *have* within larger chunks of language.

7 Conclusion

Using corpora and the results of computer-based analysis within language learning syllabi is now widely held as a positive step (Braun 2005: 47), since language from a naturally occurring corpus better reflects linguistic reality compared to textbooks (Gavioli and Aston 2001: 238). A corpus can provide a "whole panoply" of activities for learning (Sinclair 2004: 297). Use of formulaic utterances is also a well-documented strategy for language learning success (Nattinger and DeCarrico 1992; Myles, Hooper and Mitchell 1998; Nunan 2001; Durrant and Schmitt 2009). Collocations and idioms are the building blocks of language (Murison-Bowie 1996: 183), and native-like selection involves the ability to select the preferred sequence from a number of grammatically acceptable variants (Weinert 1995: 184).

Our own use of lexical chunks may be subliminal, and the collocations we regularly use may not be obvious to us via introspection (Sinclair 1997: 29). But a corpus shows us the common word combinations we employ.

Teachers can use corpora as a base for material creation, test design, feedback and evaluation references (Braun 2005: 51). Or corpora can be used directly, by learners themselves. Tim Johns believes research is "too serious to be left to the researchers" (Johns 1991: 2), and students should be encouraged to "discover" foreign language and "learn how to learn" under a DDL framework (Johns 1991: 1). The corpus is not here a "surrogate teacher" but rather a "special type of informant" which can provide natural data (Johns 1991: 1).

The ability to comprehend language and the ability to produce language are very different things (DeKeyser 2007: 287). Using corpora in the classroom certainly involves students' passive reading skills and comprehension. But having

selected, say, frequent collocations from a body of text, teachers will want students to use them. If communication is to be successful, a relevant context has to be constructed by the discourse participants (Sperber and Wilson 1995: 52). This is where the traditional role of the teacher returns, providing communicative activities to practise new lexis. I would therefore agree with Braun that corpora provide pedagogic "enrichment", rather than a primary focus for learning (Braun 2005: 55). Successful use of corpora for learning and teaching will hinge on successful "pedagogic mediation" between the corpus materials and the corpus users (Braun 2005: 61).

The two research questions I set yielded data which brought me to two conclusions:

- Past tense verbs dominate the word frequency list of this corpus, although the adjective *little* ranks highly too.
- Many of the most frequent collocations in this corpus were not particularly interesting lexically (*it was*, *was a*, *I am*, *I have*). In fact, the *lack* of lexical content in the most frequent collocation – *it was* – is perhaps its most salient feature, showing how a much work the dummy pronoun does in a narrative setting.

Having created this corpus, I hope I and others can carry out more linguistic studies into children's literature and apply findings in new ways. I agree with Hoey (2005: 14) that a corpus can serve "as a kind of laboratory" for conducting experiments. Pedagogically too, I hope language teachers will see the value in using corpora to help language learners.

References

Almela Sanchez, Moises. 2006. *From words to lexical units: A corpus-driven account of collocation and idiomatic patterning in English and English-Spanish*. Frankfurt: Peter Lang.

Anthony, Laurence. 2011, online. *AntConc Version 3.2.4*. Tokyo: Waseda University. Retrieved from http://www.antlab.sci.waseda.ac.jp (accessed May 2013).

Baker, Carolyn and Peter Freebody. 1989. *Children's first school books*. Oxford: Basil Blackwell.

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14: 275–311.

Braun, Sabine. 2005. From pedagogically relevant corpora to authentic language learning contents. *ReCALL* 17: 47–64.

Braun, Sabine. 2007. Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL* 19: 307–328.

Chambers, Angela. 2007. Popularising corpus consultation by language learners and teachers. In E. Hidalgo, L. Quereda and J. Santana (eds.). *Corpora in the foreign language classroom*, 3–16. Amsterdam: Editions Rodopi.

DeKeyser, Robert. 2007. *Practice in second language: Perspectives from applied linguistics and cognitive psychology.* Cambridge: Cambridge University Press.

Durrant, Philip and Norbert Schmitt. 2009. To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching* 47: 157–177.

Firth, J. R. 1957. *Papers in linguistics 1934–51*. Oxford: Oxford University Press.

Flowerdew, John. 1996. Concordancing in language learning. In M. Pennington (ed.). *The power of CALL*, 87–101. Houston: Athelstan.

Gavioli, Laura. 2000. The learner as researcher: Introducing corpus concordancing in the classroom. In G. Aston (ed.). *Learning with corpora*, 108–137. Houston: Athelstan.

Gavioli, Laura and Guy Aston. 2001. Enriching reality: Language corpora in language pedagogy. *ELT Journal* 55: 238–246.

Gee, James. 1994. First language acquisition as a guide for theories of learning and pedagogy. *Linguistics and Education* 6: 331–354.

Halliday, Michael. 1966. Lexis as a linguistic level. In C. Bazell, J. Catford, M. Halliday and R. Robins (eds.). *In memory of J.R. Firth,* 148–162. London: Longman.

Halliday, Michael. 1973. *Explorations in the functions of language.* London: Edward Arnold.

Halliday, Michael. 1993. Quantitative studies and probabilities in grammar. In M. Hoey (ed.). *Data, description and discourse*, 1–25. London: Harper Collins.

Ho, L. 2000. Children's literature in adult education. *Children's Literature in Education* 31 (4): 259–271.

Hoey, Michae. 2005. *Lexical priming: A new theory of words and language*. London and New York: Routledge.

Hopper, Paul. 1998. Emergent grammar. In M. Tomasello (ed.). *The new psychology of language: Cognitive and functional approaches to language structure*, 155–175. Mahwah: Erlbaum.

Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunt, Peter. 1994. *An introduction to children's literature*. Oxford: Oxford University Press.

Hunt, Peter. 2001. The fundamentals of children's literature criticism. In J. Mikkenberg and L. Vallone (eds.). *The Oxford handbook of children's literature*, 35–51. Oxford: Oxford University Press.

Johns, Tim. 1991. Should you be persuaded – two samples of data-driven learning materials. In T. Johns and P. King (eds.). Classroom concordancing. *ELR Journal* 4: 1–16.

Jones, Susan and John Sinclair. 1974. English lexical collocations: A study in computational linguistics. *Cahiers de Lexicologie* 24: 15–61.

Khatib, Mohammad. 2011. Literature in EFL/ESL classroom. *English Language Teaching* 4: 201–208.

Knowles, Murray and Kirsten Malmkjaer. 1996. *Language and control in children's literature*. London and New York: Routledge.

Leech, Geoffrey. 1974. *Semantics*. Harmondsworth: Penguin.

Leel, Hsing-Chin. 2011. In defence of concordancing: An application of data-driven learning in Taiwan. *Procedia – Social and Behavioral Sciences* 12: 399–408.

Louw, Bill. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis and E. Tognini-Bonelli (eds.). *Text and Technology*, 157–176. Amsterdam: John Benjamins.

Milton, James. 2009. *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Murison-Bowie, Simon. 1996. Linguistic corpora and language teaching. *Annual Review of Applied Linguistics* 16: 182–199.

Myles, Florence, Janet Hooper and Rosamond Mitchell. 1998. Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning* 48: 323–363.

Nation, Paul. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nattinger, James and Jeanette DeCarrico. 1992. *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nunan, David. 2001. Second language acquisition. In R. Carter and D. Nunan (eds.). *The Cambridge guide to teaching English to speakers of other languages*, 87–93. Cambridge: Cambridge University Press.

Pace-Sigge, Michael. 2013. The concept of Lexical Priming in the context of language use. *ICAME* 37: 149–174.

Pawley, Andrew and Syder, Frances. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards and R. Schmidt (eds.). *Language and communication*, 191–226. London: Longman.

Schmitt, Norbert. 2000. *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, Norbert. 2004. *Formulaic sequences: Acquisition, processing and use*. Amsterdam: John Benjamins.

Shin, Dongkwang and Paul Nation. 2008. Beyond single words: The most frequent collocations in spoken English. *ELT Journal* 62 (4): 339–348.

Sinclair, John. 1987. *Looking up: An account of the COBUILD project in lexical computing and the development of the* Collins COBUILD English Language Dictionary. London: Collins ELT.

Sinclair, John. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sinclair, John. 1997. Corpus evidence in language description. In A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.). *Teaching and Language Corpora*, 27–39. London and New York: Longman.

Sinclair, John. 2004. *How to use corpora in language teaching*. Amsterdam and Philadelphia: John Benjamins.

Sperber, Dan and Deirdre Wilson. 1995. *Relevance: Communication and cognition*. Oxford: Blackwell.

Stubbs, Michael. 1995. Collocations and cultural connotations of common words. *Linguistics and Education* 7: 379–390.

Thompson, John. 1990. *Ideology and modern culture: Critical and social theory in the era of mass communication*. Cambridge: Polity Press.

Thompson, Paul and Alison Sealy. 2007. Through children's Eyes? Corpus evidence of the features of children's literature. *International Journal of Corpus Linguistics* 12 (1): 1–23.

Weinert, Regina. 1995. The role of formulaic language in second language acquisitions. *Applied Linguistics* 16 (2): 180–205.

Widdowson, Henry. 2007. Classic book review: Papers in linguistics by J.R.Firth. *Journal of Applied Linguistics* 17 (3): 402–413.

Zipes, Jack. 1983. *Fairy tales and the art of subversion*. London: Heinemann.

Appendices

Appendix 1

Rank (content)	**Rank (overall)**	**Raw Frequency**	**Word**
1	10	73331	was
2	17	40540	had
3	21	33949	said
4	27	28846	is
5	28	27582	be
6	33	22450	have
7	35	21038	were
8	53	15659	little
9	54	14247	are
10	57	13685	do
11	63	12244	like
12	65	11644	been
13	69	10664	see
14	70	10384	time
15	73	9936	did
16	76	9304	good
17	77	9120	well
18	78	9037	know
19	79	9006	go
20	81	8738	come
21	82	8717	came
22	85	7934	man
23	87	7698	back
24	88	7682	went
25	90	7577	made
26	91	7573	old
27	92	7560	great
28	94	7539	don
29	96	7493	never
30	98	7343	day
31	100	7261	just
32	101	7196	again

33	103	7102	other
34	104	7101	way
35	105	6919	long
36	106	6853	away
37	108	6733	much
38	109	6526	sir
39	110	6498	here
40	113	6426	get
41	115	6078	too
42	117	5844	think
43	118	5768	say
44	119	5707	got
45	120	5566	first
46	123	5532	such
47	124	5478	make
48	125	5438	thought
49	127	5224	take
50	129	5158	last
51	130	5153	tell
52	131	5092	head
53	132	5071	looked
54	134	5051	has
55	135	4982	Mr.
56	136	4952	am
57	137	4909	king
58	138	4886	while
59	139	4853	let
60	140	4753	put
61	141	4742	look
62	142	4740	eyes
63	143	4733	saw
64	145	4686	mother
65	147	4529	right
66	148	4514	quite
67	149	4482	once
68	151	4429	found
69	152	4398	ever
70	153	4381	hand
71	154	4369	water
72	155	4339	every
73	156	4337	still
74	157	4333	house
75	158	4300	going
76	159	4257	night
77	160	4224	boy
78	161	4220	took

79	164	4191	things
80	165	4151	many
81	167	4089	even
82	168	4072	asked
83	170	3979	white
84	171	3962	own
85	172	3947	round
86	173	3928	home
87	174	3868	place
88	175	3865	nothing
89	176	3861	face
90	177	3797	another
91	178	3787	people
92	179	3751	began
93	180	3736	young
94	182	3717	always
95	183	3715	thing
96	184	3712	most
97	187	3603	heard
98	188	3598	told
99	189	3596	cried
100	190	3579	give

Appendix 2

Rank	**Collocation**	**Raw Frequency**	**T-score**
1	it was	12,735	105.96905
2	to be	9,151	89.28990
3	he was	8,258	81.73596
4	he had	7,352	80.46944
5	was a	6,807	62.34047
6	it is	5,948	78.22217
7	had been	5,159	70.67682
8	there was	4,995	69.23218
9	said the	4,972	43.43640
10	a little	4,885	64.89444
11	I am	4,209	66.67374
12	he said	4,109	67.99710
13	I was	4,055	47.38148
14	they were	4,018	62.25898
15	did not	3,988	62.31378
16	she was	3,960	56.57640
16	was the	3,566	-9.71436

17	I have	3,354	53.88177
18	the other	3,296	51.48479
19	to see	3,222	52.85129
20	to do	3,059	51.34715
21	the king	2,822	48.41814
22	was not	2,793	45.28896
23	like a	2,717	46.87238
24	a great	2,705	49.29649
25	have been	2,697	51.01442
26	the little	2,657	34.99894
27	would be	2,645	49.90734
28	that was	2,611	34.43586
29	I had	2,608	39.60764
30	is a	2,585	38.05081
31	the old	2,559	42.34596
32	began to	2,530	48.44936
33	you are	2,524	58.38478
34	such a	2,471	47.19122
35	at last	2,458	49.11210
36	do you	2,436	54.08873
37	to go	2,428	57.56863
38	going to	2,423	47.07252
39	she said	2,371	48.60414
40	a good	2,344	44.67478
41	would have	2,311	46.72992
42	the first	2,283	42.52635
43	to have	2,248	49.68558
44	you know	2,218	48.63568
45	to make	2,175	43.85190
46	to get	2,075	44.92761
47	is the	2,073	9.70075
48	had a	2,012	24.46664
49	will be	1,863	41.76028
50	the water	1,861	38.85288
51	I know	1,821	44.68065
52	the house1	,780	38.06770
53	at once	1,779	41.89339
54	there is	1,733	41.60256
55	be a	1,764	27.17580
56	a long	1,728	38.28299
57	they are	1,716	41.48998
58	came to	1,680	35.60881
59	that is	1,651	40.77140
60	who had	1,644	38.36814
61	I think	1,640	45.75029
62	the way	1,608	32.62283

63	you have	1,605	44.30496
64	was so	1,599	32.16739
65	to say	1,598	39.66129
66	he is	1,596	33.48953
67	to take	1,593	37.38569
68	the great	1,583	29.27396
69	and was	1,580	-18.31343
70	had not	1,576	34.89302
71	he did	1,564	40.23234
72	have a	1,552	26.44786
73	a man	1,551	35.57675
74	there were	1,549	38.96420
75	't know	1,547	38.27084
76	the time	1,536	32.94812
77	was in	1,527	12.90283
78	his head	1,517	38.05487
79	must be	1,501	37.98994
80	had to	1,501	12.29610
81	the man	1,488	28.82459
82	tell you	1,464	40.04689
83	I do	1,464	38.28576
84	back to	1,456	33.44090
85	who was	1,422	33.81068
86	said I	1,400	36.84376
87	his own	1,393	36.38324
88	the most	1,386	31.69530
89	which was	1,377	31.48752
90	and said	1,361	15.33160
91	was no	1,343	30.91116
92	you see	1,334	39.37876
93	this is	1,333	37.35392
94	was very	1,312	31.13794
95	come to	1,294	45.74243
96	the boy	1,293	30.07676
97	said to	1,269	12.54657
98	is not	1,235	30.81130
99	to come	1,222	45.74243
100	we were	1,221	33.29239

Appendix 3

Rank	**Collocation**	**Raw Frequency**	**T-score**
1	it was	12,735	105.96905
2	he was	8,258	81.73596
3	he had	7,352	80.46944

4	was a	6,807	62.34047
5	it is	5,948	78.22217
6	there was	4,995	69.23218
7	said the	4,972	43.43640
8	a little	4,885	64.89444
9	I am	4,209	66.67374
10	he said	4,109	67.99710
11	I was	4,055	47.38148
12	they were	4,018	62.25898
13	she was	3,960	56.57640
14	was the	3,566	-9.71436
15	I have	3,354	53.88177
16	the other	3,296	51.48479
17	to see	3,222	52.85129
18	to do	3,059	51.34715
19	the king	2,822	48.41814
20	like a	2,717	46.87238
21	a great	2,705	49.29649
22	the little	2,657	34.99894
23	I had	2,608	39.60764
24	is a	2,585	38.05081
25	the old	2,559	42.34596
26	began to	2,530	48.44936
27	you are	2,524	58.38478
28	such a	2,471	47.19122
29	at last	2,458	49.11210
30	to go	2,428	57.56863
31	she said	2,371	48.60414
32	a good	2,344	44.67478
33	the first	2,283	42.52635
34	to have	2,248	49.68558
35	you know	2,218	48.63568
36	to make	2,175	43.85190
37	to get	2,075	44.92761
38	is the	2,073	9.70075
39	had a	2,012	24.46664
40	the water	1,861	38.85288
41	I know	1,821	44.68065
42	the house	1,780	38.06770
43	at once	1,779	41.89339
44	there is	1,733	41.60256
45	be a	1,764	27.17580
46	a long	1,728	38.28299
47	they are	1,716	41.48998
48	came to	1,680	35.60881
49	I think	1,640	45.75029

50	the way	1,608	32.62283
51	was so	1,599	32.16739
52	to say	1,598	39.66129
53	he is	1,596	33.48953
54	to take	1,593	37.38569
55	the great	1,583	29.27396
56	he did	1,564	40.23234
57	have a	1,552	26.44786
58	a man	1,551	35.57675
59	there were	1,549	38.96420
60	't know	1,547	38.27084
61	the time	1,536	32.94812
62	was in	1,527	12.90283
63	his head	1,517	38.05487
64	the man	1,488	28.82459
65	tell you	1,464	40.04689
66	I do	1,464	38.28576
67	back to	1,456	33.44090
68	said I	1,400	36.84376
69	his own	1,393	36.38324
70	the most	1,386	31.69530
71	and said	1,361	15.33160
72	was no	1,343	30.91116
73	you see	1,334	39.37876
74	this is	1,333	37.35392
75	come to	1,294	45.74243
76	the boy	1,293	30.07676
77	said to	1,269	12.54657
78	is not	1,235	30.81130
79	to come	1,222	45.74243
80	we were	1,221	33.29239

Appendix 4: Texts compiled for this corpus

Alcott, L., *Little Women*, 1869

Alger Jr, H., *Ragged Dick*, 1868

Anstey, F., *Vice Versa*, 1882

Baines Reed, T., *The Fifth Form at St Dominic's*, 1881

Bannerman, H., *Little Black Sambo*, 1899

Barrie, J.M., *Peter Pan and Wendy*, 1911

Baum, L., *The Wonderful Wizard of Oz*, 1900

Brazil, A., *The Fortunes of Philippa*, 1906

Brooke, L., *Johnny Crow's Garden*, 1903

Bruce, M., *A Little Bush Maid*, 1910

Burnett, F., *The Secret Garden*, 1911, *Little Lord Fauntleroy*, 1886, *A Little Princess*, 1905

Carroll, L., *Alice's Adventures in Wonderland*, 1865, *Through the Looking Glass*, 1871

Conan Doyle, A., *The Lost World*, 1912

Coolidge, S., *What Katy Did*, 1872

de la Mare, W., *Songs of Childhood*, 1902

Dickens, C., *Holiday Romance*, 1868

Dodge, M. (Ed.), *St Nicholas: Scribner's Illustrated Magazine for Girls and Boys*, 13 editions, 1877–1878

Ewing, J., *The Brownies and Other Tales*, 1870

Garis, H., *Uncle Wiggily's Adventures*, 1912

Grahame, K., *The Wind in the Willows*, 1908

Haggard, H., *King Solomon's Mines*, 1885

Harris, J., *Uncle Remus and Brer Rabbit*, 1907

Henty, G., *Out on the Pampas*, 1868

Hope, L., *The Bobbsey Twins*, 1904

Ingelow, J., *Mopsa the Fairy*, 1869

Jefferies, R., *Bevis: The Story of a Boy*, 1882

Kingsley, C., *Water Babies*, 1863

Kipling, R., *The Jungle Book*, 1894, *The Second Jungle Book*, 1895, *Stalky and Co*, 1899

Lang, A. (Ed.), *The Blue Fairy Book*, 1889

London, J., *The Call of the Wild*, 1903, *White Fang*, 1906

MacDonald, G., *At the Back of the North Wind*, 1871, *The Princess and Curdie*, 1883, *The Princess and the Goblin*, 1872

Mabie, H., Hale, E., Forbush, W., (Eds.), *The Young Folks Treasury*, 1909

Malet, L., *Little Peter*, 1888

Molesworth, M., *The Cuckoo Clock*, 1877

Nesbit, E., *Five Children and It*, 1902, *The Enchanted Castle*, 1907, *The Magic City*, 1910, *The Railway Children*, 1906

Potter, B., *Peter Rabbit*, 1902, *The Tale of Mr Tod*, 1912

Porter, E., *Pollyanna*, 1913

Pyle, H., *The Merry Adventures of Robin Hood*, 1883

Ransome, A., *The Child's Book of the Seasons*, 1906

Sewell, A. *Black Beauty*, 1877

Spyri, J. *Heidi*, 1884

Stevenson, R.L., *Treasure Island*, 1883, *Kidnapped*, 1886

Stratton-Porter, G., *The Girl of Limberlost*, 1909

Twain, M., *The Adventures of Tom Sawyer*, 1876, *The Adventures of Huckleberry Finn*, 1884, *A Connecticut Yankee in King Arthur's Court*, 1889, *The Prince and The Pauper*, 1881

Verne, J., *A Journey to the Centre of the Earth*, 1864, *Twenty Thousand Leagues Under the Sea*, 1870

Webster, J., *Daddy-Long-Legs*, 1912

Wiggin, K., *Rebecca of Sunnybrook Farm*, 1903

Wilde, O., *The Happy Prince and Other Tales*, 1888

Upton, B., *The Adventures of Two Dutch Girls and a Golliwogg*, 1895

The Young Folks Treasury, containing:

Nursery rhymes:	Hush-a-bye, Baby, on the Tree-top; Rock-a-bye, Baby thy Cradle is Green; Bye, Baby Bunting; Hush Thee, my Babby; Sleep, Baby, Sleep; This Little Pig Went to Market
Nursery tales:	The Three Bears; Cinderella; The Three Brothers; The Wren and the Bear; Chicken-Licken; The Fox and the Cat; The Rats and their Son-in-Law; The Mouse and the Sausage; Johnny and the Golden Goose; Titty Mouse and Tatty Mouse; Teeny Tiny; The Spider and the Flea; The Little Shepherd Boy; The Three Spinners; The Cat and the Mouse in Partnership; The Sweet Soup; The Straw, the Coal, and the Bean; Why the Bear Has a Stumpy Tail; The Three Little Pigs
Children's poems:	The Three Children; The Owl and the Pussy-Cat; Kindness to Animals; How Doth the Little Busy Bee; Suppose; Twinkle, Twinkle; Pretty Cow; The Three Little Kittens; The Land of Counterpane; There was a Little Girl; The Boy who never Told a Lie; Foreign Children; The Unseen Playmate; I saw Three Ships; A Was an Ant;

	The Table and the Chair; Precocious Piggy; A Boy's Song; Buttercups and Daisies; The Violet; If ever I See; The Little Land; A Lobster Quadrille; Where Go the Boats; The Wind and the Moon; Where are you Going my Pretty Maid; The Lost Doll; Foreign Lands; Bed in Summer; Try Again; A Good Play; Good Night and Good Morning; The Wind; The Spider and the Fly; Let Dogs Delight to Bark and Bite; Child's Evening Hymn
Children's stories:	Hansel and Gretel; The Fair Catherine and Pif-Paf Poltrie; The Wolf and the Fox; Discreet Hans; Puss in Boots; The Elves and the Shoemaker; Hans in Luck; Master of All Masters; Belling the Cat; Little Red Riding-Hood; The Nail; Jack and the Beanstalk; How to Tell a True Princess; The Sleeping Beauty
Old-fashioned poems:	The Man in the Moon; Sage Counsel; Limericks by Edward Lear; More Limericks; The Dead Doll; Little Things; The Golden Rule; Do the Best You Can; The Voice of Spring; The Lark and the Rook; Thanksgiving Day; The Magpie's Nest; The Fairies of Caldon Low; The Land of Story Books; A Visit From St. Nicholas; Little Orphan Annie; The Chatterbox; The Voice of Spring; The History Lesson; Song of Life; The Good Time Coming; Windy Nights; The Wonderful World; Hark! Hark! The Lark; Jog On, Jog On; Sweet Story of Old; My Shadow; By Cool Siloam's Shady Rill; The Wind in a Frolic; The Graves of a Household; We Are Seven; The Better Land;The Juvenile Orator; The Fox and the Crow; The Use of Flowers; Contented John; The Old Man's Comforts, and How He Gained Them; The Frost; The Battle of Blenheim; The Chameleon; The Blackberry Girl; Mabel on Midsummer Day; Llewellyn and his Dog; The Snowbird's Song; For A 'That and A' That
Fables from Aesop:	The Goose that Laid Golden Eggs; The Boys and the Frogs; The Lion and the Mouse; The Fox and the Grapes; The Frog and the Ox; The Cat, the Monkey, and the Chestnuts; The Country Maid and Her Milkpail; The Ass in the Lion's Skin; The Tortoise and the Hare; The Vain Jackdaw; The Fox Without a Tail; The Wolf in Sheep's

	Clothing; The Crow and the Pitcher; The Man, his Son, and his Ass
Fables of India:	The Camel and the Pig; The Man and his Piece of Cloth; (adapted by P. V. She Sea, the Fox, and the Wolf; The Birds and the Lime; Ramaswami Raju) The Raven and the Cattle; Tinsel and Lightning; The Ass and the Watch-dog; The Lark and its Young Ones; The Two Gems
Scandinavian stories:	The Hardy Tin Soldier; The Fir Tree; The Darning-Needle; Thumbelina; The Tinder-Box; Boots and his Brothers; The Husband who was to Mind the House; Buttercup German stories: Seven at One Blow; One Eye, Two Eyes, Three Eyes; The Musicians of Bremen; The Fisherman and his Wife; Little Snow-White; The Goose Girl; The Golden Bird
French stories:	Beauty and the Beast; The White Cat; The Story of Pretty Goldilocks; Toads and Diamonds
English stories:	The History of Tom-Thumb; Jack the Giant Killer; The Three Sillies
Celtic stories:	King O'Toole and his Goose; The Haughty Princess; Jack and his Master; Hudden and Dudden and Donald O'Neary; Connla of the Golden Hair and the Fairy Maiden
Italian stories:	Pinocchio's Adventures in Wonderland
Japanese stories:	The Story of the Man who did not Wish to Die; The Accomplished and Lucky Teakettle; The Tongue-Cut Sparrow Battle of the Monkey and the Crab; Momotaro, or Little Peachling; Urashima Taro and the Turtle
East Indian stories:	The Son of Seven Queens; Who Killed the Otter's Babies; The Alligator and the Jackal; The Farmer and the Money Lender; Tit for Tat; Singh Rajah and the Cunning Little Jackals American Indian Stories:The White Stone Canoe; The Maiden who Loved a Fish; The Star Wife
Arabian stories:	The Story of Caliph Stork; Persevere and Prosper
Chinese stories:	The Most Frugal of Men; The Moon Cake; The Ladle that Fell from the Moon; The Young Head of the Family; A Dreadful Boar

Russian stories:	King Kojata; The Story of King Frost
Tales for tiny tots:	Tell Us a Tale; Little Red Hen; In Search of a Baby; Jock and I and the Others; Dolly Dimple; The Tale of Peter Rabbit; The Miller, His Son, and Their Ass; The Visit to Santa Claus Land; The Greedy Brownie; The Fairies' Passage; The World
Fanciful stories:	White Magic; The Brownies; The Story of Peter Pan; Sir Lark and King Sun; The Imps in the Heavenly Meadow; The Birthday Honours of the Fairy Queen

The Blue Fairy Book containing:

The Bronze Ring; Prince Hyacinth And The Dear Little Princess; East Of The Sun And West Of The Moon; The Yellow Dwarf; Little Red Riding-Hood; The Sleeping Beauty In The Wood; Cinderella, or, The Little Glass Slipper; Aladdin And The Wonderful Lamp; The Tale Of A Youth Who Set Out To Learn What Fear Was; Rumpelstiltzkin; Beauty And The Beast; The Master-Maid; Why The Sea Is Salt; The Master Cat, or, Puss In Boots; Felicia And The Pot Of Pinks; The White Cat; The Water-Lily; The Gold-Spinners; The Terrible Head; The Story Of Pretty Goldilocks; The History Of Whittington; The Wonderful Sheep; Little Thumb; The Forty Thieves; Hansel And Grettel; Snow-White And Rose-Red; The Goose-Girl; Toads And Diamonds; Prince Darling; Blue Beard; Trusty John The Brave Little Tailor; A Voyage To Lilliput; The Princess On The Glass Hill; The Story Of Prince Ahmed And The Fairy Paribanou; The History Of Jack The Giant-Killer; The Black Bull Of Norroway; The Red Etin