Keyword Analysis of Short Stories for Young Language Learners

Gabriella Lőcsey
University of Pécs, Hungary
locseyg@gmail.com

1. Introduction

As in recent years a great deal of attention has been given to discourse analysis, the goal of this study is to find out to what extent discourse analysis can be applied in foreign language education. Researchers identify lexical patterns with the help of corpus-based analysis. These lexical patterns determine both the content and the structure of texts (Fischer-Starke, 2010). The identification of these patterns used to be carried out intuitively, but nowadays this is done by computer programs. While researchers used to concentrate on finding and analyzing keywords to recognize and explain essential cultural and societal concepts, an alternative way of keyword analysis has emerged as a result of development of modern technology. *WordSmith Tools* (1996-2008), Scott's software, made it possible to use a quantitative approach to calculating keywords in a sample of texts or a corpus and to compare them to a reference corpus. This software allows for any word to become key if it occurs frequently enough in the examined text when compared to a reference corpus (Baker, 2004). These keywords indicate the content; this way they are also important from the educational perspective as well. According to Nation (2006), the knowledge of more than 95% of the written and spoken texts is needed for learners to be able to understand the meaning of a text. The more keywords of a text a language learner knows, the better the understanding of the content of the text. Analyzing keywords of various texts on different language levels might provide a keyword list that could be applied and would be advisable to use in foreign language education to make the instruction more effective. In this pilot study, texts of short stories for young learners were analyzed to find out what keywords and how many word-families the learners need to know to cope with tales.

2. Theoretical Background

2.1. Keyword and concepts

What does keyword mean? According to Pierre Guiraud, (1954) who first used this concept, keyword ('mots-clés') means simply a statistically significant lexical item. Enkvist (1964) and Gray (1964) referred to keywords as style markers whose frequencies differ significantly from their frequencies in a norm. Style markers are contextually bound linguistic elements. "Repetition is the notion underlying both style markers and hence keywords, but not all repetition, only repetition that statistically deviates from the pattern formed by that item in another context" (Culpeper, 2009, p.33).

Stubbs (2010) introduced three different senses of the term 'keywords'. Sense 1 is cultural and as Wierzbicka (1997, p.156) states, keyword is a 'focal point around which entire cultural domains are organized'. Firth (1957) and Williams (1967/1983) are the best examples of sense 1, as they found the most salient words intuitively describe culture in the finest way. However, Williams' work was grounded in a cultural studies: his list cannot be a good ground to define a 'general theory' (Stubbs, 2010, p.25). Stubbs (2010) argued that Williams had no theory of how the vocabulary of a language was structured and what the connection between the words, texts and text-types was. Sense 2 comes from comparative quantitative corpus analysis and it is statistical. 'Keywords are words which are significantly more frequent in a sample of text than would be expected, given their frequency in a large general reference corpus' (Bondi, 2010 p.25). Mike Scott's keyword software, *WordSmith Tools* (Scott, 1998) made it easy and fast to carry out corpus analysis. Baker (2004) highlighted the benefit of quantitative keywords and claimed that they guide the researcher to discover salient concepts in a text since the lexical dissimilarities of various texts are in focus. Sense 3 relates to Francis' approach and it is corpus-driven as it scrutinizes alterations in vocabulary and grammar use. Francis (1993) examined how people express their shared values and how the meaning is conveyed by a changeable lexico-grammar pattern.

The advent of computer programs such as *WordSmith Tools* (Scott,1999) and *W Matrix* (Rayson 2005, 2008) has facilitated both grammatical and semantical analyses of texts and allowed researchers to identify keywords easily and rapidly (Culpeper, 2009). Any word can become a keyword if it appears frequently enough in the examined text when compared to a reference corpus. Scott distinguishes three types of keywords: proper nouns, content keywords and high frequency words. Content words are words that people would recognize as key and they are markers of 'aboutness' of a certain text. High-frequency words are for instance *because*, *shall* or *already* which may be style markers (Baker, 2004). However, keywords do not need to consist of single words. Keyword lists consist of words of two-, three- and four- word 'clusters' (Scott, 1999) or lexical bundles (Biber, Johansson, Leech, Conrad and Finegan 1999, p.990).

2.2. Previous research

More and more studies have been conducted on keywords of various genres: romantic fiction (Tribble, 2000), political correctness in newspapers (Johnson, Culpeper and Suhr, 2003). Xiao and McEnery (2005) were interested in spoken and written discourse and Culpeper (2009) studied key part-of-speech and key semantic domains in addition to keywords in Shakespeare's Romeo and Juliet.

Fischer-Starke (2010), in her book, provides a detailed summary of previous research on keywords by highlighting the useful and successful combinations of quantitative data and qualitative analysis in stylistics in Toolan's (2004) and Culpeper's (2002) work. The aim of using mixed methods is to encode literary meanings in the data by looking at collocations of the keywords in their concordance lines.

2.3. Issues in keyword analysis

The selection of an appropriate reference corpus in keyword analysis is essential. Some researchers (Scott & Tribble, 2006) argue that a large sample is advisable, while

others (e.g., Xiao & McEnery, 2005) did not find the size of the datasets for comparison very important in making a keyword list. The scope of the reference corpus is not as important as its relevance to the corpus being examined. It is precisely the genre of the reference corpus that prominently influences the keywords identified.

An essential parameter set in a Keyword program is the minimum frequency cut-off point. It means that the program 'excludes the words that will be identified as unusual simply because they happen not to have occurred or to have occurred very infrequently in the dataset of the reference corpus' (Culpeper, 2009, p.36). The choice of minimum frequency of cut-off often depends on the size of the data set. In this study the Keynes cut-off point is 25, originally set in the Keyword analysis program.

The other important parameter is the test for statistical significance. This figure 'calculates the significance of the unusualness of the keyword' (Culpeper, 2009, p. 36). A text with a keyword ratio of 0.01 is an extremely high keyword text and it means that it contains many words distinct to this text. In contrast, a low keyword text with a ratio of 0.009 uses mainly general words.

3. The Study

3.1. The aims of the pilot study

The aim of this pilot study was twofold. It attempted to explore the keywords in a selected corpus of 48 online short stories; how many word-families short stories consisted of and how many of these word-families young learners need to be familiar with to cope with the text confidently.

3.2. Research questions

This study was conducted to provide answers to the following research questions:
1. What are the keywords of 48 short stories?
2. How many word families are there in the whole texts?
3. How many word-families do young learners need to be familiar with in order to read these short stories?

3.3. Data analysis and procedures

Forty-eight short stories were chosen as corpus data out of 60 short stories for young language learners provided by The British Council on its website (https://learnenglishkids.britishcouncil.org/en/shortstories?page=0%2C0%2C0%2C0%2C0%2C0). The corpus consisted of 10,087 words. These tales are recommended to children of three age ranges; 1-5, 6-8 and 9-11. The titles of the stories are listed in the Appendix in the order of their appearance on the website. The selected reference corpus was the 14-million-word mixed written-spoken, US-UK corpus developed by Paul Nation (2007), as a basis for the first 2k of the British National Corpus- Corpus of Contemporary American English lists (BNC-Coca list).

Four procedures were applied in order to answer the research questions. First, keyword analysis was performed with the help of the Keywords Extractor (Cobb, 2007)

which is a part of the *Compleat Lexical Tutor* (Cobb, 2007).' This program determines the defining lexis in a specialized text or corpus by comparing the frequency of its words to the frequency in a more general reference corpus' (http://www.lextutor.ca/cgi-bin/range/texts/index.pl). Then the concordance plots of some words were investigated by AntConc Computer Software (Version 3.4.3). Vocabulary Profilers (Cobb, 2007) were also applied to find out which frequency band each word of the texts belong to; and finally, a special version of Vocabulary Profilers, VP-kids (Cobb, 2007) was used to look at how many words the texts contain from 10 frequency bands of roughly 250 families. The Keynes cut-off point is 25 in this study which was originally set in the Keyword analysis program.

3.4. Results and discussion

3.4.1. Keywords analysis of short story corpus

The keyword list (Table 1) shows all the words in the short story corpus that are at least 25 times more frequent in the text than in the reference corpus. The number preceding each word is the number of times more recurrent this word is in the short stories than it is in the corpus BNC-COCA list. For example, in the output the 27, 516 princess (Table 1) means that the word princess is 27,516 times more frequent in the short story corpus than it is in the reference corpus. This possibly means that 'princess' word has an important role (key) in the analyzed corpus.

Table 1 Keywords analysis of short story corpus

(1)	27,516.00	princess	(36)	79.00	tummy
(2)	18,827.00	okay	(37)	66.21	cave
(3)	17,379.00	clothes	(38)	63.24	monster
(4)	11,586.00	computer	(39)	62.97	surf
(5)	10,138.00	quinine	(40)	62.07	lantern
(6)	8,689.00	caliph	(41)	62.06	circus
(7)	5,793.00	especially	(42)	61.19	bean
(8)	4,055.00	splodge	(43)	59.66	Santa
(9)	3,258.50	Persia	(44)	54.59	giant
(10)	2,896.33	ghoul	(45)	54.31	clown
(11)	1,448.17	ogre	(46)	50.37	tower
(12)	579.27	snowman	(47)	49.51	eagle
(13)	452.56	harp	(48)	47.95	mouse
(14)	440.78	pyramid	(49)	47.78	lion
(15)	434.50	rainforest	(50)	47.10	torch

(16)	413.79	malaria	(51)	47.05	roar
(17)	386.20	hippo	(52)	44.56	jungle
(18)	382.55	ouch	(53)	44.54	jack
(19)	310.36	Ramadan	(54)	39.04	seed
(20)	296.23	dinosaur	(55)	38.79	emperor
(21)	271.56	lonely	(56)	37.78	snake
(22)	241.39	shush	(57)	36.82	elephant
(23)	206.90	tweet	(58)	33.55	asleep
(24)	169.27	dragon	(59)	31.48	shout
(25)	163.95	Greece	(60)	31.26	knight
(26)	160.92	porridge	(61)	30.94	clue
(27)	149.81	pigeon	(62)	29.48	castle
(28)	129.51	buzz	(63)	28.40	medicine
(29)	125.93	icon	(64)	27.50	fame
(30)	120.69	marathon	(65)	26.49	tunnel
(31)	117.43	tomb	(66)	26.16	animal
(32)	92.83	teddy	(67)	25.83	climb
(33)	91.06	planet	(68)	25.32	ocean
(34)	85.20	hooray	(69)	25,19	bounce
(35)	81.59	parrot			

With the help of the Concordance plot the frequency and the places of the words' occurrence in the corpus can be investigated. The more common a word is in the corpus, the darker it is in that part of the concordance plot. According to the Keyword list, 'princess' was the most frequent word; thus, a question emerged whether this word appeared in different parts of the corpus or can it be densely found in one particular section. As it can be clearly seen from the concordance plot (Figures 1-3) 'princess' appears 19 times in the texts (Figure 1) and this word is seemingly repeated in one or two stories. The first 20 most frequent keywords are generally content words and nouns and this is also the characteristic feature of the rest of the list. However, there are two exceptions in the first twenty words: *okay* (5 times in Figure 2) which 'serves as a routine compliant response' (Biber et al, 1999, p. 1090) and *ouch* (14 times in Figure 3) which is 'an interjection' and has 'an exclamatory function' (p.1083) expressing sudden physical pain.

The word 'okay' can be found in different part of the corpus. The title of the last short story is 'I'm too ill' and this must be the reason why 'ouch' appears so frequent at the end of the corpus.

Figure 1. Concordance plot of 'princess' (AntConc (Version 3.4.3) [Computer Software].

Figure 2. Concordance plot of 'okay' (AntConc (Version 3.4.3) [Computer Software].

Figure 3. . Concordance plot of 'ouch' (AntConc (Version 3.4.3) [Computer Software].

Another finding of interest concerns the appearance of words such as quinine, caliph, Persia, ghoul, pyramid, malaria, Ramadan, Greece and marathon. No doubt that these words are mostly culturally dependent words and may not be a part of from the general knowledge of young language learners who are not part of the culture or tradition which they refer to. Short stories of different cultures have different keywords, as earlier Wierzbicka (1997) pointed out the relationship between keyword and culture. The selection of various expressions and traditional stories from the different parts of the world is welcome and can be a good start of awareness raising as it may make young learners and their social context more understanding and tolerant towards people from other cultures. Nowadays, it is essential because most of us live in a multi-cultural environment.

3.4.2. Lexical Vocabulary Profile

The lexical frequency profile method was developed by Laufer & Nation (1995). This procedure divides the words of texts according to which frequency band each word belongs to: first 1,000 most-frequent, second 1,000 most-frequent, the most-frequent 'academic' words not in either of the other two lists and the remainder or 'off list'. Vocabulary Profilers were adapted for Web by Tom Cobb (2007) (http://www.lextutor.ca/cgi-bin/range/texts/index.pl). This program shows the numbers and percentages of words and word families in the scrutinized English text coming from each of the three word lists and those which are not recognized.

Table 2 shows the number of word families, types, tokens and percentages at each level. The texts of these 48 short stories offered to young learners a total of 10,087 words. The first row of the Table 2 shows that 7,937 tokens belong to the K1 frequency band (first 1,000 most-frequent words) in the texts of the short stories. This amounts to 78.69% of the total words and it consists of 842 types. More than half (4,678) of these words are function words and 3,259 are content words. The second 1,000 most frequent words (K2 words) account for 913 tokens of 361 word types and 273 word families. Altogether 87.74% of the total words belong to K1 and K2 frequency band (K1+K2). Another finding of interest is that academic words also appear in these tales:

42 tokens of 16 different types and these types make up 14 word-families. The number of off-list words is extremely high, 1,195 words of 469 types. The appearance of off-list words has a higher percentage (11.82%) than the K2 words and academic words together (9.47%).

Table 2 Families, types, tokens at each word level in 48 short stories

Levels	Families	Types	Tokens	Percent
K1 Words (1-1000):	549	842	7,937	78.69%
Function:	...	...	(4,678)	(46.38%)
Content:	...	...	(3,259)	(32.31%)
K2 Words (1001-2000):	273	361	913	9.05%
K1+K2		...	...	(87.74%)
AWL Words (academic):	14	16	42	0.42%
Off-List Words:	?	469	1,195	11.85%
Total	836+?	1,688	10,087	100%

Table 3 shows the summary of the data analysis of Vocabulary Profilers. It shows that the texts are built of 10,087 tokens and 1,688 types. The lexical density of the texts, the ratio of the content and total words (0.54%) shows that slightly more than half of the words are content words in these short stories.

Table 3 Summary of the data analysis of Vocabulary Profilers

Words in text (tokens):	10,087
Different words (types):	1,688
Type-token ratio:	0.17
Tokens per type:	5.98
Lex density (content words/total)	0.54

3.4.3. Vocabulary Profilers for kids (VP-kids)

VP-kids matches the text of the short stories against 10 modified 250-word lists generalized from several studies of children's oral productions by Stemach and Williams (Cobb, 2007). Table 4 lists the classification of word families, types, tokens and the off-list words across the ten frequency levels. The most tokens are from the KID250-1 level (1 is the most frequent) and on the second most frequent level, KID250-2 there are 1,128 words. Furthermore, the numbers of word families, types and tokens are higher on the more frequent level. The number of words in each category gradually decreases except for the tokens of the first two levels where this reduction is radical. The most interesting outcome is that there are numerous off-list words (910 tokens, 409 types) in the texts and their number is close to the number of word tokens on the second level (1,128).

Table 4 Number of families, types and tokens

Freq. Level	Families	Types	Tokens	Coverage%	Cum%
Kid250 - 1:	182	338	6,331	62.78	62.78%
Kid250 - 2:	173	270	1,128	11.18	73.96%
Kid250 - 3:	137	191	517	5.13	79.09%
Kid250 - 4:	94	116	350	3.47	82.56%
Kid250 - 5:	71	95	268	2.66	85.22%
Kid250 - 6:	66	80	163	1.62	86.84%
Kid250 - 7:	58	67	141	1.40	88.24%
Kid250 - 8:	45	55	122	1.21	89.45%
Kid250 - 9:	45	48	115	1.14	90.59%
Kid250 - 10:	20	22	40	0.40	90.99%
Off-List known:	237	259	457	4.53	95.52%
Off-List unknown:	?	150	453	4.49	100.00%
Total	1,128+?	1,691	10,085	100%	100%

I looked for the first ten words of the keywords list in Table 2 in the 10 modified 250-word lists of Table 4 and the outcome is astonishing. Six out of ten, *quinine, caliph, especially, splodge, Persian and ghoul* are among the words on off-list but known. The other words such as *princess* can be found in the word group 6, *okay* in group 2, *clothes* in group 3 and the word *computer* in group 5.

4. Conclusion

The aim of this study was to analyze the keywords of 48 online short stories offered to early English learners in order to investigate how many word families can be found in the short stories altogether. Nation (2006) stated that 98% coverage is ideal, to entirely comprehend written and spoken texts. According to Table 4, in the text of these short stories the 98% coverage can be reached only if young learners also have the knowledge of some of the off-list words besides the other words on the frequency list. The different themes of these short stories are also thought-provoking and the high occurrence of culturally dependent keywords raises the issue of how much time a word needs to step from the off-list a bit higher at least to KID250 -10. Further research is necessary because of the limitations of the study. Only 48 stories were analyzed in this study and the analyses only offer insights into these short texts. It would be good to find out about many more stories. Keyword analysis can be of great help in language teaching as it can have various meaningful results. It can help teachers and learners comprehend and focus on the actual content and can give a kind of guidance on what words or word-families teachers need to teach to make young language learners understand this level of written and spoken English.

5. References

Anthony, L. 2014. AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net

Baker, P. 2004. Querying keywords: Questions of difference, frequency and sense in keywords analysis. *Journal of English Linguistics*, *32*(4), 346-359

Biber, D., Johansson, S., Leech, G., Conrad, S. &Finegan, E., (1999). *Longman grammar of spoken and written English*. London: Longman.

Bondi, M. 2010. Perspectives on keywords and keyness: An Introduction. In *Keyness in Texts,* M. Bondi & M. Scott (Eds), 1-21.Amsterdam: Benjamins.

Cobb, T. 2007. Keywords Extractor v.2 [computer program].

http://www.lextutor.ca/cgi-bin/range/texts/index.pl. Cobb, T. 2007. VP-Kids v.9 [computer program].

http://www.lextutor.ca/cgi-bin/range/texts/index.pl.

Cobb, T. Web VP Classic v.4 [computer program]. Accessed 3 Jan 2016 from

http://www.lextutor.ca/cgi-bin/range/texts/index.pl.

Culpeper, J. 2002. Computers, language and characterization: An analysis of six characters in *Romeo and Juliet*. In U. Melander-Marttala, C. Ostman & M. Kytö (Eds.), *Conversation in Life and in Literature: Papers from the ASLA Symposium*, *Association Suédoise de Linguistique Appliqué (ASLA),* 15. (pp.11-30). Uppsala: Universitetstryckeriet. (available at: *h*ttp://www.lexically.net/wordsmith/corpus_linguistics_links/papers_using_wordsmith.htm)

Culpeper, J. 2009. Keyness. Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics 14*(1), 29- 59.

Enkvist, N. E. 1964. "On defining style". In N. E. Enkvist, J. Spencer & M. Gregory (Eds.), *Linguistics and style*, (pp. 1–56). Oxford: Oxford University Press.

Firth, J. 1957. *Papers in Linguistics* 1934-1951. London: Oxford University Press.

Francis, G. 1993. A corpus-driven approach to grammar: principles, methods and examples.

In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds), *Text and Technology* (pp.137-56). Amsterdam: Benjamins.

Fischer-Starke, B. 2010. *Corpus Linguistics in Literary Analysis: Jane Austen and her Contemporaries.* London: Continuum. DOI: 101515/cast-2011-0010.

Gray, H. 1964. *Anatomy* or the London Telephone Directory of 1960.

Guiraud, P. 1954 [1970]. *Les Caractéres statistiques du vocabulaire*. (pp. 64–67) reprinted In: P. Guiraud & P. Kuentz (Eds.), *La Stylistique Lectures*, (pp.222–224). Paris: Klincksieck.

Johnson, S., Culpeper, J.& Suhr, S. (2003) From 'politically correct councillors' to 'Blairite nonsense': discourses of 'Political Correctness' in three British newspapers. *Discourse and Society* 14(1). 29-47.

Nation, I.S.P. 2006. How large a Vocabulary is needed For Reading and Listening? *The Canadian Modern Language Review, La Revue canadienne des langues vivantes, 63*(1), 59-82.

Nation, I.S.P. & Beglar, D. 2007. A vocabulary size test. *The Language Teacher, 31*(7), 9-13. Rayson, P. 2005. *WMatrix: A Web-based Corpus Processing Environment*. Computing Department, Lancaster: Lancaster University. http://www.comp.lancs.ac.uk/ucrel/wmatrix/. Rayson, P. 2008. From key words to key semantic domains". *International Journal of Corpus Linguistics*, *13* (4), 519–549.

Scott, M. 1998. *WordSmithTools*. Version 3. Oxford: Oxford University Press.

Scott, M. R. & Tribble C. 2006. *Keywords and corpus analysis in language education*. Amsterdam & Philadelphia: John Benjamins.
Stubbs, M. 1996. *Text and corpus analysis*. Oxford: Blackwell.
Stubbs, M. 2010. Three concepts of keywords. In *Keyness in Texts*, M. Bondi & M. Scott (Eds), 21-43. Amsterdam: Benjamins.
Toolan, M. 2004.Values are descriptions; Or, from literature to Linguistics and back again by way of keywords*. Belgian Journal of English Language and Literatures (BELL New Series 2*): 11-30.
Wierzbicka, A. 1997. *Understanding cultures through their key-words*. Oxford: Oxford University Press.
Williams, R. 1976/1983. *Keywords. A Vocabulary of Culture and Society*. London: Fontana.
Xiao, R. & McEnery, T. 2005. Two approaches to genre analysis: Three genres in Modern American English. *Journal of English Linguistics*, *33*(1), 62–82.

6. Appendix

Titles of 48 stories

1. My Secret Team
2. A Dog's life
3. Our colorful world
4. Ratty robs a bank
5. The Great Race
6. Why Anansi has thin legs?
7. The princess and the dragon
8. Dinosaur Dig
9. Spycat
10. The Lucky Envelope
11. The Mummy
12. Eric the Engine
13. The voyage of the animal
14. The Lucky Seed
15. My Favourite Day: Chinese New Year
16. Superhero High
17. The First Marathon
18. Santa's Little Helper
19. The story of quinine
20. No dogs
21. The lion and the mouse
22. The treasure map
23. The bird king
24. Teddy's adventure
25. My favourite day - Eid al Fitr
26. Ali and the magic carpet
27. Monster shopping trip
28. The hungry dragon
29. The cold planet
30. Buzz and Bob's big

31. The lazy bear
32. Jack and the beanstalk
33. George and the dragon
34. Goldilocks and the three bears
35. Pyramids in Paris
36. My favourite clothes
37. What's that noise?
38. The greedy hippo
39. The animal shelter
40. The haunted house
41. Twin's week
42. ABC Zoo
43. Circus escape
44. Planet Earth
45. I couldn't believe my eyes
46. The snowman
47. The lantern(A Ramadan story)
48. I'm too ill