DE GRUYTER OPEN

ICAME Journal, Volume 41, 2017, DOI: 10.1515/icame-2017-0007

Taking the corpus to the classroom: Using children's stories to compare learner and native text

Ed Thomas, Kanda University of International Studies, Japan

1 Introduction

The potential of children's literature in adult L2 learning has received some attention (Ho 2000; Massi and Benvenuto 2001; Webb and Macalister 2012), but much research remains to be done and applied to foreign language learning. Analyses of L2 writing have predominantly used English for academic purposes (EAP) as data – essays, research papers and reflective compositions (Johns 1997; Hyland 2005; Yoon 2008; Crossley and McNamara 2009; Lee and Chen 2009; Conor-Linton and Polio 2014). Despite narrative being considered the primary means by which humans create meaning from their experiences (Casanave 2005:17), the role of creative writing within adult L2 learning needs more attention. This study uses corpus methods to analyse learner text and compare it to native writing. Variation between native and non-native language use can then offer insights for teaching practice.

An analysis of a five-million-word corpus of children's stories (Thomas 2015) showed function words to be the most frequent (*the*, *and*, *to*, *of*, *a*), followed by verbs (*was*, *had*, *said*, *is*, *be*, *have*, *were*). Notably, the eighth most frequent content word was the adjective *little*, with common collocations including *little girl*, *little boy*, *a little*, *little while*, *little more*, *little way*, *little bit*, *little longe*r and *little later*. The most frequent content-carrying collocation was *it was* – reflecting the narrative nature of children's stories and often a lack of pronoun reference. Vocabulary range is also large for native authors. The aim of this study is to compare a body of learner text with native writing, and examine if imaginative L2 language reflects similar or different patterns. The research questions guiding the study are:

1. What are the most frequent words and collocations used by learners in their stories?
2. What range of vocabulary do learners display in their creative writing, and how does this compare to native writers' vocabulary?

2 Method

The Japanese 19- and 20-year-old students in this study were enrolled on a sophomore writing class at Kanda University of International Studies (KUIS). Two 20-member classes from 2014/15 and two 20-member classes from 2015/16 were used to pool the data. Fourteen students chose not to have their writing included in the study, making 66 texts available as a data set. The level of English for all participants was advanced, each having studied for more than a decade during school and university education in Japan, and many already having completed periods of study abroad in the US, UK, Canada, Australia and New Zealand.

The bulk of the writing course was taken up with EAP, with the focus on essay structure, argument and the presentation of research. However, I (their teacher) laid emphasis on more imaginative writing exercises throughout, particularly on narrative. The texts studied included children's tales such as *Jack and the Beanstalk*, *Alice in Wonderland* and *Little Red Riding Hood*. The content of such stories is well-known to non-native English users, as they have been translated into hundreds of languages around the world and have become "quintessential classics" of our culture (Hunt 2001: 37). They are also illustrated stories, offering scope for "multi-modal" activities gaining ground in language teaching (Leki, Cumming and Silva 2008, Flood, Heath and Lapp 2014). Imagery was central to driving imagination and language production during the composition stages; illustrations from books and also cinematic images from Tim Burton's *Alice in Wonderland* movie were used to elicit plots and descriptions. Schmidt's (1994) theory of noticing was also used in the reading stages to highlight formulaic language from the genre, such as 'one day', 'once upon a time' and 'happily ever after'.

Each student was assigned to write a 500-word fictional narrative, using the previously-mentioned texts as models. The narrative structure of these texts was seen to fall into a five-part pattern of: beginning, journey, conflict, resolution and ending (others like Massi and Benvenuto 2001 use terms such as orientation, complication and climax for these stages, but I use simplified language presented in the classroom). Students were encouraged to follow this five-part pattern, but they were given free rein to let their imagination flow, break conventions and write however they wanted. All of the student texts have been collected at www.iword2014.blogspot.jp. This bank of 66 stories totals 39,120 words, forming the learner corpus for this study. Both Laurence Anthony's *AntConc* and Paul Nation's *Range* software were used to analyse the data.

3 Results

Range showed that there were 3,152 word types in the learner corpus and that the type-token ratio (used by Laufer and Nation 1995 as a measure of lexical variation) was 12.4. The software showed that 82.5 per cent of the learner's words (33,633 in total) are contained within the most frequent 1,000 words of English[1]. Just 5.2 per cent of the learners' words are from the second 1,000 words, and 1.9 per cent are from the third 1,000-word set. The native corpus, on the other hand, contained 37,717 distinct word types giving it a much higher type-token ratio of 67.5.

The ten most frequent words learners used were *the*, *to*, *and*, *a*, *he*, *was*, *she*, *I*, *of*, and *it*. An abundance of function words and the definite article being the most frequent word is in line with native-speaker usage evidenced in other corpora (such as COBUILD and LOB – see Kuo 1999 – or Lee and Chen's 2009 learner corpus of Chinese undergraduate text). For these Japanese learners the most frequent word (the definite article) appeared 1,934 times out of 39,120 word tokens, giving it a 4.9 per cent coverage of the corpus as a whole. The indefinite article – the fourth most frequent word after *the*, *to* and *and* – appeared 1,076 times, providing 2.8 per cent coverage. Combined, the definite and indefinite articles account for 7.7 per cent of the learner corpus meaning, on average, an article appears every 13 words. In the native corpus of 5,484,937 words *the* appears 310,370 times (5.69 per cent coverage) and *a* is seen 124,583 times (2.27 per cent coverage). Combined, articles account for 434,953 or 7.93 per cent of the words.

Excluding function words, the ten most frequent content words (nouns, adjectives, verbs and adverbs) in the learner corpus are *was*, *had*, *said*, *is*, *were*, *time*, *day*, *are*, *can* and *went* – predominantly past tense verbs[2]. The most frequent collocation in the native corpus of children's stories is *it was* (Thomas 2015). *It was* is also a frequent collocation in the learner corpus, appearing 158 times. However *he was* out-ranks it as the most frequent collocation by learners (174 occurrences).

In native children's stories, the adjective *little* is used extremely frequently. Do learners employ little as often as native writers? The answer is no. In the native corpus, *little* is ranked the eighth most frequent content word. However, it is ranked 57 on the learner word list (see Appendix A), and is less frequent than *house*, *boy*, *mother*, *people*, *friends*, *girl*, *tree*, *forest*, *big* (ranked 38), *old*, *man* and *happy*. Writing a children's story does not seem to prompt a frequent use of *little* for L2 writers in the way it does for native storytellers. The most common noun collocation with *little* for natives is *girl*, but for learners it is *little bit*, fol-

lowed by *little later*, then *little while*. *Little girl* only appears four times across the 66 learner texts, and *little boy* only once.

4 Discussion

Lexical variation is striking between the learner and the native texts; the type-token ratio is 12.4 for learners but 67.5 for native writers. This clearly shows the larger range of words used by L1 writers. For learners to come closer to native writers in their texts, then, they should try to produce a greater number of distinct words rather than use the same ones over and over again.

The results showed article use to be extremely similar between the two sets of data. *The* and *a* accounted for 7.7 per cent and 7.9 per cent of the words in the respective learner and native corpora. But while the frequencies suggest similarity, a closer inspection points to dissimilarity. A look at concordance lines for *the* in the learner texts reveals many mistakes:

He looked for the way to get better, but he could not…

A few minutes later, he felt the change of body.

The boy had the way to run away.

…fins and scales had disappeared. Instead, I got the beautiful human's legs and smooth skin…

…she woke up and looked around. The shining sun came in the large blue sky.

…was working hard all day but he had the money shortage…

…come to the witch's castle! Muwhahaha!" After the happening, I couldn't have my confidence…

Japanese, like many Asian languages, operates without articles. Much has been written on article misuse by Asian learners (Butler 2002; Yoo 2009; Nickalls 2011), and results from this study add evidence to scholars' calls for better teaching of articles in Asian countries (Berry 1990). The ten most frequent content words in the learner corpus are *was*, *had*, *said*, *is*, *were*, *time*, *day*, *are*, *can* and *went* – predominantly past tense verbs. The native corpus of children's stories has *was*, *had*, *said*, *is*, *be*, *have*, and *were* appearing frequently, suggesting that learners in this study reflected native writers in verb usage. Using the concordancer, we can see learners correctly using verbs in both their main sense ('Koko was nine years old', 'his father had great inner strength', 'the other boys

were scared and they ran away') and auxiliary modes ('she was becoming more beautiful', 'they had promised to go to the forest', 'my parents were eaten by a shark'). So while articles seem to be a problem for these Japanese learners, verbs seem to be better used.

In native usage, the common collocation *it was* often sees *it* referring to nothing in particular:

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.

In these and other cases, *it* acts as a dummy pronoun. The dummy copula chunk *it was* is extremely common amongst native storytellers (Thomas 2015), but one might expect L2 writers to avoid such usage due to its conceptually difficult and vague reference. On the contrary, the present study reveals that *it was* is often used correctly in a dummy fashion by L2 writers:

it was difficult to decide a strategy for them…

he realised it was 5pm…

it was dangerous for the mermaid to stay at her home…

It was a rainy day and William was walking in the…

One day she was walking in the forest alone and it was getting dark…

…rumour that a very beautiful deer lived in the forest. It was said that if people could find it…

It was time to leave, he gave me something.

The concordance lines see *it* referring to much more concrete objects too ('He believed it was the gnome and ran after it', 'He made a special weapon. It was a gun. It had a great power', 'Yes, it was a bear! They ran away desperately', 'It was a very quiet place. Pom felt happy', 'He was going to make an omelette containing fried rice. It was his favourite food in the world'), suggesting these learners have little problem switching between the two uses of *it* in their writing.

Native children's story writers employ *little* very frequently in collocational chunks. This study shows that learners do not. Lexical priming is a relatively new field within linguistics (Hoey 2005; Pace-Sigge 2013) but it is interesting to note that L2 writers do not seem aware of the "semantic prosody" surrounding certain culturally-loaded terms (Stubbs 1995), and employ a word like *little* very

differently to natives. Much native language within children's stories is concerned with little boys and little girls (think of *Alice in Wonderland*, *Little Red Riding Hood*, *Peter Pan*, J*ack and the Beanstalk* – these are characters constantly belittled by their authors) , but what characters do L2 writers invent and how are they described? While everyday girls and boys do appear from these L2 writers, we also see talking animals, mermaids, princes and princesses, witches, thieves, aliens, samurai, Aikido masters, ghosts, gods, and elves. More fantastical again, there are talking trees and fruits, raindrops, magical flowers, and cups who live together as a couple. Reading their stories, one is struck by the students' power of imagination and their desire to write non-real scenarios based in magical forests, under-the-sea realms, outer space and entirely other worlds. Girls are *beautiful*, *pretty* and *nice* while boys are *lonely*, *educated*, *injured*, *unkind* and *human*. Perhaps because the undergraduate authors of these L2 texts are closer in age to girls and boys than the likes of Lewis Carrol and J. M. Barrie, they feel less need to 'look down' and belittle the characters. This is an aspect of language and psychology which begs more research.

The escapism of these student narratives is obvious. Asked to write short fiction, these young authors leave behind their university studies, part-time jobs and family lives and go on imaginative journeys to places where normal rules are suspended. Tea cups talk, magical flowers cure illness, aliens seek human recipes for *onigiri*, princes and princesses fall in love in heavenly kingdoms, trees befriend lonely schoolchildren in spooky forests. Psychological analysis is beyond the scope of this linguistic research, but the lexis alone does say something about the ideas within these Japanese students' minds and their desire to suspend reality.

Much of the academic workload of these undergraduates is taken up with essay writing, factual research and presentation of real-world topics. One of the purposes of this imaginative teaching unit was to inject more fun into language learning. After the stories had been written, time was spent outside the classroom telling the stories – focusing on speaking skills and body language as much as the narrative (see Figure 1).

Figure 1: Storytelling class at Kanda University of International Studies, July 2015

At the end of the unit, students were asked to reflect on their experience. A small proportion of the 80 participants reported essay writing to be more useful for them compared to creative writing ("Essay will make us to think about things very deeply", "I want to think about many topics such as politics, nature, medical, biology and so on, because these knowledge would be useful in the future when I get a job"). However, the vast majority of participants reflected positively on the creative writing experience. Comments included:

> I prefer to make own stories because everyone writes freely, and we could listen other's interesting story. Everyone has good imagination skill.
>
> My imagination power was bigger than before. It could lead to good.
>
> I love making story from my imagination. It's fun to write story for me, and it's never same as other people.

> I prefer story because I can write with fun. Also I can study some words which are often used in daily life. When I write story I think English is fun.

Creative writing is not something that has traditionally featured within the Japanese education system, which has instead focused on grammar translation, reading comprehension, repeated spoken drills, practical communicative ability and exam preparation (Sasaki 2008). Asian learners of English are traditionally viewed as group-orientated and harmony-seeking rather than individual critical thinkers keen to express their own voices (Atkinson 1997; Fox 1994, Ramathan and Kaplan 1996; Stapleton 2002). The present study does much to offer an alternative to this view, and suggests that creative writing can take a more prominent role in Japanese education. On a lexico-grammar level, learners not only performed well in the writing task, but they also enjoyed the experience and developed their own ideas. I would agree with Massi and Benvenuto (2001:167) in saying that this type of writing activity can be a far more rewarding experience than simply writing a paper on an assigned topic, a chore that is often devoid of emotion.

5 Conclusion

Narrative verb and pronoun use was good from the learners in this study. These are two salient features of native-writer work within the children's story genre, and students showed some mastery of them. However, two areas noted for improvement in this study are article use and vocabulary range.

The study is of course limited. The sample of learner text was small compared to the native corpus, making comparisons difficult. The type/token ratio as a measure of lexical variation is problematic, varying according to total text length. Perhaps a better reflection of vocabulary range is the D-value (Malvern and Richards, 2002). This is something my own future research can embrace as an alternative measurement. It should also be mentioned that the students worked through a drafting stage during the writing process – their teacher highlighted (but didn't correct) errors in their stories. A writing centre was also available at Kanda University, where students could take their text and discuss it with a native speaker. These two elements tightened up the learners' writing before final submission and helps account for its generally high quality.

To return to the research questions, do L2 writers reflect similar patterns of language to native writers of children's stories? Or are they very different? In terms of narrative verb usage, these L2 writers do seem to structure their stories in a native-like way. Also just like L1 writers, they employ function words (such

as *the*, *to*, *and*, *a*, *he*, *was*, *she*, *I*, *of*, and *it*) very frequently. Article use shows variation, however, and learners' vocabulary is much smaller in breadth. To improve these students' writing, I would advise vocabulary study and article practice. From a research angle, I will make efforts to increase the data pool of learner text from a variety of sources. Furthermore, a more modern native corpus of children's stories (containing, for example, the works of Roald Dahl and J K Rowling) should be available for purposes of comparison.

Research has shown that extensive reading leads to vocabulary learning (Nation and Ming-Tsu 1999; Webb and Macalister 2012), so students should constantly be encouraged to read. But a well-written composition makes effective use of vocabulary, and this need not be reflected in a wide range of vocabulary but a well-used one (Laufer and Nation 1995). While students' writing is limited, they should be encouraged to express themselves with whatever linguistic means they have. This is certainly something they enjoy doing, and something with huge potential for language learning too.

Notes

1. The first 1,000 words consists of around 4,000 types. Nation's sources for the lists are A General Service List of English Words by Michael West, 1953, and The Academic Word List by Coxhead, 1998.
2. Sometimes a verb carries content (as in, "they had lunch", "she did her homework") and sometimes not ("they had eaten lunch when...", "did you eat lunch?"). The software I used was not able to make this distinction, neither was tagging possible. So all verbs were classed as content words, despite the problem of polysemy and their functional potential.

References

Anthony, Laurence. 2011. *AntConc Version 3.2.4* (online). Tokyo: Waseda University. Available at: http://www.antlab.sci.waseda.ac.jp (accessed July 2015).

Atkinson, Dwight. 1997. A critical approach to thinking in TESOL. *TESOL Quarterly* 31: 71–94.

Berry, Roger. 1991. Re-articulating the article. *ELT Journal* 45: 252–259.

Butler, Yuko. 2002. Second language learners' theories on the use of English articles: An analysis of the metalinguistic knowledge used by Japanese students in acquiring the English article system. *Studies in Second Language Acquisition* 24: 451–480.

Conor-Linton, Jeff and Charlene Polio. 2014. Comparing perspectives on L2 writing: Multiple analyses of a common corpus. *Journal of Second Language Writing* 26: 1–9.

Crossley, Scott and Danielle McNamara. 2009. Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing* 18: 119–135.

Flood, James, Shirley Heath and Diane Lapp. 2014. *Handbook of research on teaching literacy through the communicative and visual arts*. New York: Rutledge.

Fox, Helen. 1994. *Listening to the world: Cultural issues in academic writing.* Urbana: National Council of Teachers of English.

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

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

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

Johns, Ann. 1997. *Text, role and context: Developing academic literacies*. Cambridge: Cambridge University Press.

Kuo, Chih-Hua. 1999. Can numbers talk? Basic data management of a corpus. *RELC Journal* 30 : 1–17.

Kwon, Sun-Hee. 2009. Lexical richness in L2 writing: How much vocabulary do L2 learners need to use? *English Teaching* 65: 155–174.

Laufer, Batia. 1991. How much lexis is necessary for reading comprehension? In P. Arnaud and H. Bejoint (eds.). *Vocabulary and applied linguistics*, 126–132. Basingstoke: Macmillan.

Laufer, Batia and Paul Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics* 16: 307–322.

Lee, David and Sylvia Chen. 2009. Making a bigger deal of the small words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing* 18: 281–296.

Leki, Ilona, Alister Cumming and Tony Silva. 2008. *A synthesis of research on second language writing in English*. New York: Routledge.

Massi, Maria and Adriana Benvenuto. 2001. Using fairy tales to develop reading and writing skills. *CATESOL Journal* 13: 157–164.

Matsuoka, Warren and David Hirsh. 2010. Vocabulary learning through reading: Does an ELT course book provide good opportunities? *Reading in a Foreign Language* 22: 56–70.

Nation, Paul and Ming-Tsu. 1999. Graded readers and vocabulary. *Reading in a Foreign Language* 12: 355–381.

Nickalls, Richard. 2011. How definite are we about articles in English? A study of L2 learners' English article interlanguage during a university professional English course. P*roceedings from the Corpus Linguistics Conference* 92, 1–19. Available from University of Birmingham Centre for Corpus Research.

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

Ramathan, Vai. and Kaplan, Robert. 1996. Audience and voice in current composition texts: Some implications for ESL student writers. *Journal of Second Language Writing* 5: 21–34.

Sasaki, Miyuki. 2008. The 150-year history of English language assessment in Japanese education. *Language Testing* 25: 63–83.

Schmidt, Richard. 1994. Deconstructing consciousness in search of useful definitions for applied linguistics. *AILA Review* 11: 11–26.

Stapleton, Paul. 2002. Critical thinking in Japanese L2 writing: Rethinking tired constructs. *ELT Journal* 56: 250–257.

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

Thomas, Ed. 2015. Word frequency and collocation: Using children's literature in adult learning. *ICAME* 39: 85–110.

Webb, Stuart and John Macalister. 2012. Is text written for children useful for L2 extensive reading? *TESOL Quarterly* 47: 300–322.

Yoo, Isaiah. 2009. The English definite article: What ESL/EFL grammars say and what corpus findings show. *Journal of English for Academic Purposes* 8: 267–278.

Yoon, Hyunsook. 2008. More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning and Technology* 12 (2): 31–48.

Appendix A: Content words from the learner corpus

Rank (content)	Rank (overall)	Frequency	Word
1	6	998	was
2	20	295	had
3	21	292	said
4	28	224	is
5	30	201	were
6	32	188	time
7	34	169	day
8	37	156	are
9	38	154	one
10	46	128	can
11	47	128	didn't
12	49	125	went
13	51	121	could
14	53	116	have
15	54	114	go
16	57	110	be
17	58	110	house
18	60	104	boy
19	61	100	thought
20	62	91	couldn't
21	63	91	people
22	65	88	asked
23	66	88	want
24	67	87	do
25	68	87	mother
26	69	85	came
27	70	83	friends
28	72	80	know
29	73	80	next
30	74	80	other
31	75	79	found
32	79	77	looked
33	81	76	girl
34	82	76	many
35	83	75	tree
36	84	74	again
37	85	74	forest
38	86	70	big
39	87	70	got

40	88	69	here
41	89	69	'm
42	90	69	old
43	92	68	will
44	94	67	tried
45	96	65	find
46	97	65	like
47	99	64	felt
48	100	64	get
49	101	63	man
50	102	63	something
51	105	61	don't
52	106	61	wanted
53	107	60	started
54	108	60	suddenly
55	109	59	decided
56	110	59	happy
57	111	59	little
58	112	59	name
59	113	58	good
60	115	57	lived
61	118	54	human
62	119	54	life
63	121	53	surprised
64	122	52	became
65	123	52	home
66	124	52	long
67	127	51	now
68	129	50	been
69	130	50	see
70	131	49	did
71	132	49	heard
72	133	49	later
73	134	48	always
74	137	47	place
75	138	47	years
76	139	46	first
77	140	46	make
78	141	46	parents
79	143	45	friend
80	144	45	same
81	146	44	mermaid

82	147	44	once
83	148	44	world
84	149	42	dragon
85	150	42	dream
86	151	42	lot
87	152	41	answered
88	153	41	way
89	154	40	down
90	155	40	god
91	156	40	white
92	157	39	appeared
93	158	39	come
94	159	39	saw
95	160	38	away
96	161	38	beautiful
97	162	38	gave
98	163	38	magic
99	165	37	front
100	166	37	sad