the jaltcalljournal

ISSN 1832-4215
Vol. 5, No.1 Pages 65–76

Constructing a blog corpus for Japanese learners of English

Patrick Foss
Kwansei Gakuin University
pfoss@kwansei.ac.jp

Researchers have directed increased attention to the building and analysing of written learner corpora – databases of written language produced by language learners – to address issues such as the words that non-native learners of English use in their writing, and how their word use differs from that of native speakers. This paper offers an initial look at a new written learner corpus, currently under construction, which is composed of lower/intermediate-level learner blogs. Preliminary data from the corpus regarding high frequency vocabulary use is compared to frequency lists from the British National Corpus in order to illustrate basic usage differences.

Introduction

What words do non-native learners of English use in writing? How does their word use differ from that of native speakers? Increasingly, researchers have been building and analysing written learner corpora, i.e. databases of written language produced by language learners, to answer questions like these. Concrete evidence concerning issues such as learner use of vocabulary can only lead to better-targeted materials and more efficient language learning.

Probably the best known and most researched written learner corpus is the International Corpus of Learner English (**ICLE**), a 2.5 million word collection of essays written by advanced English level university students of various language backgrounds (Granger, 2003). Many other major corpora are also composed of essays written by advanced university learners. Examples include the Uppsala Student English Corpus (Axelsson, 2000; Axelsson &

Hahn, 2001); the Polish Learner English Corpus (Lenko-Szymanska, 2002; Pravec, 2002); the Janus Pannonius University Corpus (Horváth, 1999); and the Montclair Electronic Language Database (Fitzpatrick & Seegmiller, 2004). Secondary school learner essays representing a range of proficiencies have been collected in the Japanese EFL Learner Corpus (Tono, 2004). Secondary-school writing of various types has been collected in the TeleNex Student Corpus (Tsui, 2004). Other written learner corpora which have been described in the literature include the Hong Kong University of Science and Technology Learner Corpus, the Chinese Learner English Corpus, and the Corpus of English by Japanese Learners (Nesselhauf, 2004; Pravec, 2002). There are also two very large learner corpora – the Cambridge Learner Corpus and the Longman Leaner Corpus – that are commercial in nature.

As Granger and Tribble (1998) have noted, written learner corpora can benefit researchers, educators and learners in two main ways. First, they can help researchers and educators confirm or reject assumptions about learner use of language, thus leading to more efficient classroom practices and better-targeted materials or learning aids (including computer software tools). For example, is it true, as many a teacher in Japan might attest, that Japanese university learners tend to use the word *enjoy* at the expense of other equally appropriate words or phrases? If so, materials focusing on alternatives – *have a good time, have fun, appreciate* – can be developed. Second, corpus data can be shown directly to the learners themselves, to help raise their awareness of particular strengths or weaknesses (see Millar & Lehtinen, 2008, for an example of a step-by-step approach).

However, despite these benefits and the seemingly wide variety of corpora available, there are several gaps in current learner corpora research. Nesselhauf (2004) has mentioned proficiency and medium as two of these gaps. Concerning proficiency, of the corpora for which most published research has been done, the majority represent advanced learners exclusively; beginners and intermediate-level learners have been the focus of much less attention. Concerning medium, academic essays predominate; other types of learner writing have not been collected to anywhere near the same extent. A third gap, mentioned by Granger (2004), involves longitudinal studies; few learner corpora consist of writing collected from the same learners at different points in time.

How can these gaps be addressed? This paper offers an initial look at a new written learner corpus, currently under construction, which is composed not of advanced-level essays but of lower/intermediate-level learner blogs. After a brief description of this corpus-in-progress and the rationale behind using blogs, preliminary data from the corpus regarding high frequency vocabulary use will be compared to that from the British National Corpus in order to illustrate basic usage differences.

The corpus

This written learner corpus – hereafter referred to as the Japanese Learner English Blog Corpus (JLEBC) – is being constructed at a private university in Japan. First and second-year students in one of the English programs at this school write regularly on individually-created blogs as part of their English coursework. This is a required component of the program. Students are graded according to production; grammatical mistakes and spelling errors are ignored. Dictionary use is allowed but not encouraged. As there are seven instructors and approximately thirty classes involved, there are a number of variables concerning how these blogs are assigned and produced. These include:

1. Number of blog entries required per semester. This typically ranges from 5–10.
2. Number of words required per entry. This ranges from approximately 100–250.
3. Use of class time. Some teachers use the first 15-20 minutes of class time for blog-writing. Other teachers assign blogs mostly for homework.
4. Choice of topic. Some blog topics are teacher-generated, others student-generated.
5. Availability of models. Some teachers write models on their own blogs for student reference.
6. Level of interaction. Most, but not all, class blogs are organized into 'blog circles,' and students are asked to read 1–3 other student blogs per week and contribute brief comments regarding them.
7. Type of blog. A variety of free blog providers are utilized.

Most students in the program have low/intermediate-level English skills as measured by institutional **TOEFL** exams and other tests.

During the 2007–2008 academic year, 654 students wrote a total of 1,625,741 words (tokens) in 8,858 blog entries, which is where the **JLEBC** stands as of this writing. The mean per student was 2,486 words spread over 13.5 blog entries, for an average of 184 words per entry. For inclusion in the corpus, all entries have been anonymized, copied into computer text files, and organized by 1) learner, 2) semester, 3) entry number, and 4) topic choice: teacher-generated or student-generated. Of the 8,858 blog entries, 6078 (69% of the total) were written in response to 60 different teacher-generated topics. The remaining 2780 entries (31% of the total) were written in response to student-generated topics. No spelling or grammatical errors have been corrected.

Rationale for using blogs

Blogs have been defined by Ward (2004) as websites which are "updated regularly and organized chronologically according to date, and in reverse order from most recent entry backwards" (p. 1). Carney (2009) has noted several key characteristics of blogs: their (potentially) broad audience; their ownership by individuals; their frequent updates; and their communicative features such as commenting and hyperlinks.

As far as the use of blogs in language or writing education is concerned, researchers have particularly noted the 'real world' nature of blogs and other forms of online writing and how learners respond to this authenticity in ways they may not respond to less authentic forms of writing such as academic essays (Blanton, 2005; Bloch, 2007; Lowe & Williams, 2004; Pinkman, 2005). Although using blogs in the classroom is not inherently authentic, the potential audience and interactive features of blogs "make them more likely, or at least more simply, used in authentic communicative ways" (Carney, 2009, p. 301). In a related vein, the personal nature of blogs can make them a more authentic method of self-expression than standard academic rhetorical forms (Farmer, 2004; Thorne & Payne, 2005). Blogs have also been noted for their immediacy, making them particularly suitable for peer review and collaboration (Lowe & Williams, 2004). The less formal, less structured nature of blog writing or online writing in general has furthermore been shown to benefit lower-level or insecure learners without a solid background in academic writing (Blanton, 2005; Bloch, 2007; Lam, 2000).

For all of these reasons, blogs seem to be an ideal medium for lower/intermediate level learners using general (i.e. non-academic) English. A corpus constructed from blogs

produced by these learners would also seem to be an ideal starting point for research into lower/intermediate level vocabulary use.

Methodology

Learner corpora are valuable sources of information in and of themselves. They are also commonly compared with other corpora, particularly native speaker corpora, in order to ascertain differences in language usage. For this exploratory study, statistical data on the overall size of the corpus and the most frequent words was obtained for the **JLEBC** using Wordsmith Tools 5.0 (Scott, 2008). Using keyword analysis, this word frequency list was then compared to a similar list generated from the World Edition of the British National Corpus (**BNC**) (Scott, 2004). Table 1 shows the size of each of corpus.

Table 1. Corpora size.

Name of corpus	Size (tokens)
Japanese Learner English Blog Corpus (JLEBC)	1,625,741
British National Corpus: World Edition (BNC)	99,465,296

The **BNC** was chosen as the control corpus for three reasons. First, the corpus was designed to be representative of general English language use (Aston & Burnard, 1998). Although the **JLEBC** technically consists of writing in only one genre (blogs), it is designed to be representative of the general English language usage of a group of learners on a wide variety of general topics. Second, the **BNC** is one of the largest native speaker corpora in existence; size is an important consideration where the authority of a particular corpus is concerned (Granger & Tribble, 1998). Finally, wordlists from the **BNC** are freely available (Scott, 2004; see also Leech, Rayson, & Wilson, 2001).

Of course, the **BNC** is not a perfect model of English usage, but then no single corpus is. Even describing a native speaker corpus as a model of any sort for learners is troubling for some. Ringbom (1998a) pointed out in a study concerning high frequency vocabulary in the **ICLE** that comparing a learner corpus with a native speaker corpus can be problematic, as frequently employed terms such as 'overuse' and 'underuse' "presuppose a norm" from which learners fall short (p. 191). Cook (1998) asked the question plainly: "Why should the attested language use of a native speaker community be a model for learners of English as an international language?" (p. 60). The chosen model for this study, the **BNC**, is composed primarily of texts written in British English by professional adult writers; is it fair to make the English found in this type of writing the norm for young Japanese university learners writing on blogs? Though blogs are technically a written medium, their personal nature and less-structured form also gives them certain qualities found usually in speech. Would it be more appropriate to compare them to a primarily spoken corpus? Or is it, to take Ringbom's (or Cook's) point to one possible conclusion, actually *in*appropriate to 'presuppose a norm' and measure them against a native speaker corpus at all?

Perhaps – to all of these questions. However, it is natural for learners to look for norms, for standards by which to measure their own production. It is also understandable if the language usage of any large group of native speakers is considered standard (if not *the* standard). Though national, regional, and cultural differences must be taken into account, if the language usage of native speakers taken in the aggregate cannot be considered standard

usage, then what can be? Significant differences between learner usage and native speaker usage of vocabulary, for example, could therefore be an indication that this vocabulary has not been learned to the necessary extent (in the case of underuse) or is being used at the expense of normal linguistic variation (in the case of overuse). These conclusions can be tested, and if found valid, materials can be developed and targeted instruction employed to help the learners move forward in an efficient manner.

Particularly where high frequency vocabulary is concerned, differences in learner and native speaker usage should be closely examined. As they are largely function or basic-content words, high frequency words by their very nature are the words found in virtually all types of written and spoken discourse. They are the glue that holds the language together; it is therefore necessary for learners to know and be able to use them appropriately. As Nation (2001) has written, "high frequency words are so important that anything that teachers and learners can do to make sure they are learned is worth doing" (p. 16). Caveats aside, it seems 'worth doing,' then, to make comparisons between the learner writing in the **JLEBC** and the native speaker language in the **BNC**, despite its limitations, and see what sort of differences exist, especially concerning use of high frequency vocabulary. Given the above concerns, however, for the purposes of this study the descriptive phrases 'used more' and 'used less' will be employed rather than the possibly value-laden terms 'overuse' and 'underuse.'

Before making these comparisons, a second issue – topic sensitivity – must also be briefly addressed. As Ringbom (1998b) has noted, topic sensitivity "will to some extent be present whenever word frequency patterns are established for texts with different content" (p. 48). Even a corpus as large and varied as the **BNC** has a number of words at very high levels of frequency – *London* and *British* among them – which likely would not be present in a differently constructed corpus. A small corpus of texts created by Japanese learners of roughly the same age, studying in the same program and sharing many of the same experiences, is bound to have even more of these words. For this initial study, content words with obvious connections to assigned blog topics (e.g. *school, English, university*) have been removed from the frequency lists. However, it is likely that other words with less obvious connections are present and influencing the results.

For this reason, and because the **JLEBC** is still incomplete, all results should be considered tentative. The planned inclusion of additional blog entries will affect frequency ratios, and additional statistical analysis will be necessary in the future to substantiate any claims regarding usage differences. Think of the words in the **JLEBC** like runners in a marathon, this paper a report on the race in progress. Final standings are likely to change.

Results

1. High frequency words used more in the JLEBC than in the BNC

Of the top 300 most frequent words in the **JLEBC**, which were used significantly more by learners than by native speakers in the **BNC**? Keyword analysis revealed several categories of words on this list that are not surprising – indeed might be expected from lower/intermediate-level writing. One of the most obvious is first-person pronouns. Table 2 (and all subsequent tables in this section) lists these words followed by the approximate number of occurrences per 10,000 words in each corpora and the log-likelihood (**LL**) measure of the difference (the higher the number, the bigger the difference; see Dunning, 1993 for more on log-likelihood).

Table 2. First-person pronouns.

Word	JLEBC	BNC	LL
I	615.0	73.6	243140.4
my	150.0	14.8	66139.4
we	97.3	30.2	14835.1
me	37.9	13.2	4826.0

Some of these differences are quite large. Not only learner language is likely having an effect here; the personal nature of blogs, noted previously, also lends itself to first-person pronoun use.

Second, this list is peppered with basic adjectives, the most notable of which are shown in Table 3.

Table 3. Basic adjectives.

Word	JLEBC	BNC	LL
high	31.7	3.8	12197.3
good	38.9	8.1	9490.2
interesting	11.8	1.0	5800.6
happy	11.9	1.1	5341.8
hard	15.5	2.2	5266.5
beautiful	9.9	0.8	4729.6
bad	8.8	1.5	2589.8
fun	5.4	0.5	2503.9
difficult	10.1	2.2	2408.5
big	8.2	2.4	1285.8
important	10.3	3.9	1138.3

Why the heavy use of *high* by the first and second year university learners represented in this study? Using the concordance feature on Wordsmith Tools reveals that *high school* is still very much on their minds. There are also numerous examples of *high price,* as in "…it is very high price and I cannot buy it," possibly indicating difficulties with the adjective *expensive*.

Other categories of words used more include words which function as quantifiers or intensifiers (Table 4), words typically labelled as 'vague' (Table 5), and common ordinals and other words used for signalling purposes (Table 6). The substantial presence of this last set of words is perhaps due to the influence of the learners' instructors; many of these words are explicitly taught in the writing program at this university. Overuse of the verb *think* is mentioned often in the literature in regards to learner writing (e.g. Aijmer, 2002; Ringbom, 1998b); as Table 7 shows, the learners in this study used this word much more than native speakers as well.

Table 4. Quantifiers and intensifiers.

Word	JLEBC	BNC	LL
very	127.5	12.0	57798.2
many	48.1	8.9	13225.2
lot	25.2	2.8	10264.2
much	26.5	9.0	3491.9
especially	7.1	1.7	1486.8

Table 5. 'Vague' words.

Word	JLEBC	BNC	LL
things	18.4	4.1	4231.0
various	8.2	1.5	2245.6
people	26.5	11.7	2187.6
thing	10.6	3.4	1551.8
place	11.0	4.8	919.2

Table 6. Ordinals and 'signal words'.

Word	JLEBC	BNC	LL
first	31.2	12.1	3307.3
example	15.2	3.6	3251.8
second	16.1	4.1	3173.7
next	16.3	4.5	2893.5
finally	6.4	1.3	1637.1

Table 7. Think.

Word	JLEBC	BNC	LL
think	45.0	8.9	11589.5

A final category of words used more by the Japanese learners in this study concerns the different forms of *enjoy* and *play* (Table 8), which perhaps deserve special mention. In Japanese, the concepts of 'having fun' or 'having a good time' are almost invariably expressed by the verb *tanoshimu* or the adjective *tanoshii,* which are frequently translated as *enjoy* or *enjoyable,* respectively. As for *play,* the Japanese equivalent *asobu* is used in a much wider variety of situations than *play* is in English; the sentence *I went out with my friends,* for example, is commonly expressed in Japanese as *Tomodachi to asobimashita* (literally, 'I played with my friends'). Given the context, then (and, to some extent, the age of the subjects involved), it is perhaps not surprising that these two words are used so often by the Japanese learners represented here.

Table 8. Play and enjoy.

Word	JLEBC	BNC	LL
play	19.2	2.1	7888.8
enjoyed	8.6	0.5	5030.6
enjoy	9.3	0.7	4973.5
played	8.9	1.1	3335.7
playing	6.2	1.1	1830.8

2. Words used less in the JLEBC than in the BNC

Of the top 300 most frequent words in the **BNC**, which were used significantly less than might be expected by the Japanese learners in this study? Words used less are perhaps of even greater interest than words used more; as Granger and Tribble (1998) have written concerning the benefits of non-native learner corpus data, "perhaps the greatest gain comes from the way in which the **NNS** corpus shows what is absent in learner writing" (p. 205). Though the range of topics in the **JLEBC** is naturally more limited than in the **BNC**, it should be noted again that words at the highest levels of frequency are commonly found across topics. The absence or less frequent occurrence of these words in learner writing may indicate a lack of understanding or confidence where production of these words is concerned and therefore a need for further instruction.

Keyword analysis of the two corpora revealed several categories of words used significantly less by the learners in the **JLEBC**. Though one of these categories could be labelled 'adult' or 'professional' words (e.g., *local, national, public, system*) not likely to be used by university-age Japanese learners, the majority are simple function words, as should be expected given their high level of frequency. No one with teaching experience in Japan should be surprised to learn that articles and words commonly used as determiners headline this group. Table 9 (and all subsequent tables in this section) lists these words followed by the approximate number of occurrences per 10,000 words in each corpora and the log-likelihood (**LL**) measure of the difference.

Table 9. Articles and determiners.

Word	JLEBC	BNC	LL
the	313.6	608.8	-29428.9
an	14.3	34.1	-2370.1
which	18.1	36.8	-1891.3
no	9.2	23.1	-1737.8
any	3.5	12.3	-1402.1
what	11.2	22.7	-1153.1
those	2.6	8.8	-963.5
such	4.6	10.8	-723.5
a	195.2	219.3	-452.5
that	90.0	106.0	-402.9

The Japanese language, of course, does not have articles, which might partly explain why the learners in this study used them to a lesser degree than the native speakers in the BNC.

Also striking is the number of prepositions used less in the JLEBC, of which Table 10 contains a sampling.

Table 10. Prepositions.

Word	JLEBC	BNC	LL
of	166.1	306.6	-12725.0
as	21.9	65.9	-6414.3
into	3.0	15.9	-2558.2
on	44.8	73.5	-2102.6
from	26.4	42.8	-1183.8
through	2.1	8.2	-1046.0
between	3.2	9.1	-828.4
by	36.7	51.6	-771.1

As with articles, Japanese also does not have prepositions in the English sense. *Between,* for example, can be translated as *~ (no) aida (ni),* which is technically a noun construction (literally 'the middle position').

Further notable for their underrepresentation are common verb modals (Table 11), other auxiliary verbs (Table 12), and verbs for reported speech (Table 13), suggesting possible difficulties with these forms for this level of Japanese learner.

Table 11. Verb modals.

Word	JLEBC	BNC	LL
would	8.3	23.1	-2042.1
might	1.2	6.0	-896.4
may	7.6	12.8	-411.1

Table 12. Auxiliary verbs.

Word	JLEBC	BNC	LL
be	27.8	65.5	-4477.3
been	6.6	26.2	-3376.1
had	25.6	41.5	-1148.8

Table 13. Verbs for reported speech.

Word	JLEBC	BNC	LL
said	5.9	19.7	-2134.4
says	0.8	4.0	-620.6
told	1.2	3.6	-327.4

Conclusion

This preliminary study has sought to point out basic usage differences concerning high-frequency vocabulary between the writing on blogs of a particular group of lower/intermediate-level Japanese learners and the writing of native speakers collected in the British National Corpus by examining the top 300 most frequent words in each corpus. That there are large numbers of basic words used more often by these learners than native speakers should not be surprising; learners of all types have fewer words to draw from to begin with and, in what Hasselgren (1994) famously termed 'the teddy-bear principle,' tend to use those words with which they feel most comfortable. However, discovering what these words are, specifically, can help local educators develop materials to wean learners away from these words in an efficient manner. Even more useful from a pedagogical point of view is discovering the words used less often by learners. Though production involves choice (see Corson, 1985, regarding motivation and language use) and any conclusions as to why learners avoid individual words must be considered educated guesses at best without specific information from the learners themselves, it seems reasonable to suggest that high frequency words which can be identified as underutilized may be words not fully understood by the learners in question. They therefore represent teaching opportunities – opportunities that often go overlooked, as it can be difficult on the spot to notice what is *absent* in learner speech or writing.

Hopefully one day materials based on the data in the **JLEBC** can be developed to help educators in Japan make vocabulary instruction better targeted and more useful for lower-intermediate level learners. In any case, as the **JLEBC** continues to grow, there will be further opportunities to study vocabulary use by this group and possibly even other groups of learners. While this corpus, at 1.6 million words, is large enough as a whole for some types of frequency analysis, a still larger one would be far more reliable. Furthermore, at this point in time, the subcorpora for the **JLEBC** are still too small to examine longitudinal issues effectively. How, for example, does vocabulary use by first and second year university students change as they progress through an **EFL** program? Questions like this await further data collection.

References

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 55–76). Amsterdam: Benjamins.

Aston, G., & Burnard, L. (1998). *The **BNC** handbook: Exploring the British National Corpus with **SARA***. Edinburgh: Edinburgh University Press.

Axelsson, M.W. (2000). **USE** – The Uppsala Student English Corpus: An instrument for needs analysis. ***ICAME** Journal, 24,* 155–157

Axelsson, M.W., & Hahn, A. (2001).The use of the progressive in Swedish and German advanced learner English – a corpus-based study. ***ICAME** Journal, 25,* 5–30.

Blanton, L.L. (2005). Student, interrupted: A tale of two would be writers. *Journal of Second Language Writing, 14,* 105–121.

Bloch, J. (2007). Abdullah's blogging: A generation 1.5 student enters the blogosphere. *Language Learning & Technology, 11* (2), 128–141. Retrieved January 14, 2008 from http://llt.msu.edu/vol11num2/bloch/default.html.

Carney, N. (2009). Blogging in foreign language education. In M. Thomas (Ed.), *Handbook of research on language acquisition technologies: Web 2.0 transformation of learning* (pp. 292–312). Hershey, PA: IGI Global.

Cook, G. (1998). The uses of reality: A reply to Ronald Carter. *ELT Journal, 52* (1), 57–63.

Corson, D. J. (1985). *The Lexical Bar.* Oxford: Pergamon Press.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19* (1), 61–74.

Farmer, J. (2004). Communication dynamics: Discussion boards, weblogs and the development of communities of inquiry in online learning environments. Retrieved January 14, 2008 from http://incsub.org/blog/?p=3.

Fitzpatrick, E., & Seegmiller, M.S. (2004). The Montclair electronic language database project. In U. Connor & T.A. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 223–238). Amsterdam: Rodopi.

Granger, S. (2003). The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly, 37* (3), 538–546.

Granger, S. (2004). Computer learner corpus research: current status and future prospects. In U. Connor & T. A. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 123–145). Amsterdam: Rodopi.

Granger, S., & Tribble, C. (1998). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In S. Granger (Ed.), *Learner English on computer* (pp. 199–209). London: Longman.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics, 4,* 237–258.

Horváth, J. (1999). Advanced writing in English as a foreign language. A corpus-based study of processes and products. (PhD Dissertation) Janus Pannonius University, Pécs, Hungary.

Lam, W. S. E. (2000). L2 literacy and the design of the self: A case study of a teenager writing on the Internet. *TESOL Quarterly, 34,* 457–482.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English.* Harlow: Longman. Frequency lists retrieved April 26, 2008 from http://ucrel.lancs.ac.uk/bncfreq/flists.html

Lenko-Szymanska, A. (2002). How to trace the growth in learners' active vocabulary: a corpus-based study. In B. Ketterman & G. Marko (Eds.), *Teaching and learning by doing corpus analysis: Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19–24, July 2000* (pp. 217–230). Amsterdam: Rodopi.

Lowe, C., & Williams, T. (2004). Moving to the public: Weblogs in the writing classroom. In L. Gurak, S. Antonijevic, L. Johnson, C. Ratliff & J. Reyman (Eds.), *Into the blogosphere.* Retrieved January 14, 2008 from http://blog.lib.umn.edu/blogosphere/

Millar, N., & Lehtinen, B. (2008). DIY local learner corpora: Bridging gaps between theory and practice. *The JALT CALL Journal, 4* (2), 61–72.

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

Nesselhauf, N. (2004). Learner corpora and their potential in language teaching. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125–156). Amsterdam: Benjamins.

Pinkman, K. (2005). Using blogs in the foreign language classroom: Encouraging learner independence. *The JALT CALL Journal, 1* (1), 12–24.

Pravec, N. (2002). Survey of learner corpora. *ICAME Journal, 26,* 81–114.

Ringbom, H. (1998a). High-frequency verbs in the ICLE corpus. In A. Renouf (Ed.), *Explorations in corpus linguistics* (pp. 191–200). Amsterdam: Rodopi.

Ringbom, H. (1998b). Vocabulary frequencies in advanced learner English: a cross-linguistic approach. In S. Granger (Ed.), *Learner English on computer* (pp. 41–52). London: Longman.

Scott, M. (2004). *BNC World Corpus wordlist* [Data file]. Retrieved from http://www.lexically.net/downloads/version4/downloading%20BNC.htm

Scott, M. (2008). WordSmith Tools (Version 5.0) [Software]. Oxford: Oxford University Press. Available from http://lexically.net/wordsmith/index.html.

Thorne, S. L., & Payne, S. (2005). Evolutionary trajectories, Internet-mediated expression, and language education. *CALICO Journal, 22* (3), 371–397.

Tono, Y. (2004). Multiple comparisons of IL, L1 and TL corpora: the case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In G. Aston, S. Bernardini & D. Stewart (Eds.), *Corpora and language learners* (pp. 45–66). Amsterdam: Benjamins.

Tsui, A. (2004). What teachers have always wanted to know-and how corpora can help. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 39–61). Amsterdam: Benjamins.

Ward, J. (2004). Blog assisted language learning (BALL): Push button publishing for the pupils. *TEFL Web Journal, 3*(1). Retrieved August 15, 2007 from http://www.teflweb-j.org/v3n1/blog_ward.pdf

Author biodata

Patrick Foss teaches in the School of Science and Technology at Kwansei Gakuin University in Japan.