Learning from Analysis of Japanese EFL Texts

George R. S. Weir and Toshiaki Ozasa

1. Introduction

Japan has a long tradition of teaching English as a foreign language (EFL). A common feature of EFL courses is reliance on specific textbooks as a basis for graded teaching, and periods in Japanese EFL history are marked by the introduction of different textbook series. These sets of textbooks share the common goal of taking students from beginners through to 'able' English language users, so one would expect to find common characteristics across such series. As part of an on-going research programme in which Japanese EFL textbooks from different historical periods are compared and contrasted, we have recently focussed our efforts on using textual analysis tools to highlight distinctive characteristics of such textbooks. The present paper introduces one such analysis tool and describes some of the results from its application to three textbook series from distinct periods in Japanese EFL history. In so doing, we aim to encourage the use of textual analysis and seek to expose salient features of EFL texts which would likely remain hidden without such analytical techniques.

2. Textual analysis

With the growing availability of text collections (corpora), many educators realise the potential for employing such resources in order to support teaching. An obvious application lies in language teaching, wherein available textual resources may serve as examples or illustrations of language use (e.g., Granger, et.al, 2002; Aston et al, 2004). In addition, local collections of texts, e.g., in the form of student submissions, are on the increase and this raises the need for tools that support the exploration and analysis of these text corpora. Tools such as Wordsmith (Scott, 1998) and AntConc (Anthony, 2005) offer approachable means whereby non-computer specialists may analyse their own data collections, and more ambitious facilities are available in systems such as NooJ (Silberztein, 2005), and GATE (Bontcheva, et. al., 2004). In the following, we introduce an alternative set of tools for textual analysis (developed by the first author) and thereafter describe the application of these tools in contrasting the content of historical Japanese EFL texts.

The Posit Text Profiling Toolset (Weir, 2007) comprises several software modules that work together to provide a broad range of textual analysis facilities. Built as an extensible set of Unix shell scripts and Perl programs, the system provides a means of generating frequency data, includes Part-of-Speech (POS) tagging, and can accommodate large text corpora with ease. In its initial version, the toolset is command line driven and depends for its flexibility upon users gaining a good understanding of the component scripts and available command options. In addition, for ease of use, a version with graphical interface has recently been developed (Baillie & Weir, 2008). In both cases, output from the toolset takes the form of multiple files that store a wide variety of results from the textual analysis.

A key feature of the Posit tools is part-of-speech profiling on any specified text, with word occurrence information

TABLE 1: Summary POS Profile output for 'Emma'

Input filename		emma.txt	
Total words (tokens)		159826	
Total unique words (types)		7364	
Type/Token Ratio (TTR)		21.7037	
Number of sentences		8585	
Average sentence length (ASL)		18.6169	
Number of characters		914519	
Average word length (AWL)		5.72197	
POS TYPES		**POS TOKENS**	
noun_types	4268	nouns	69060
verb_types	2603	verbs	67678
adjective_types	1346	prepositions	38600
adverb_types	487	personal pronouns	31192
preposition_types	65	determiners	26178
personal_pronoun_types	23	adverbs	25432
determiner_types	18	adjectives	25086
possessive_pronoun_types	7	possessive pronouns	9582
interjection_types	5	interjections	516
particle_types	0	particles	0

FIGURE 1: Posit Toolset with graphical user interface

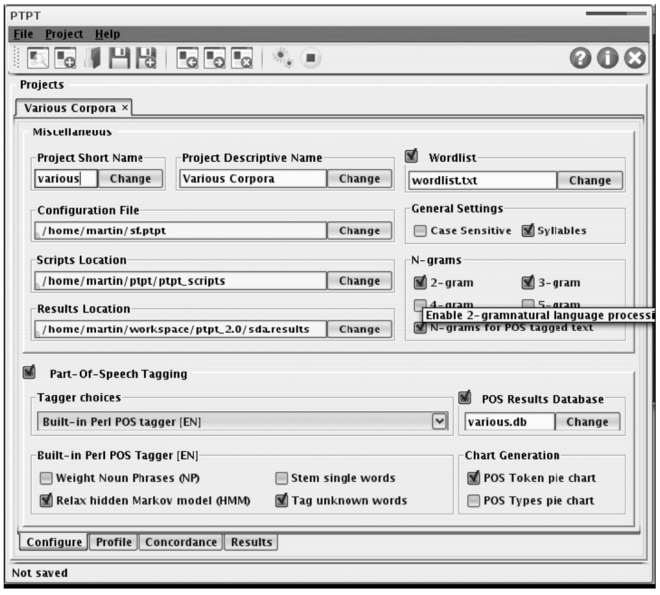


TABLE 2: Example 4-gram frequency data

Frequency	4-gram
50	i do not know
26	a great deal of
20	i am sure i
19	it would have been
18	mr and mrs weston
18	it would be a
18	i do not think
16	i have no doubt
16	i am sure you
15	and i am sure

TABLE 3: Example 2-gram frequency data

Frequency	2-gram
608	to be
566	of the
449	it was
446	in the
395	i am
334	she had
331	she was
308	had been
301	it is
299	mr knightley

detailed by raw frequency and by part-of-speech frequency. Totals are given for word tokens, word types, part-of-speech types, and part-of-speech tokens. As well as the summary level of detail on parts-of-speech, the system provides detailed frequency data on specific sub-types within parts-of-speech, such as common nouns, superlative adjectives, etc.[1] As an illustration, Table 1 shows the summary level data from analysis of Jane Austen's novel Emma.

In addition to word and part-of-speech analysis, the Posit toolset can also return frequency data on the presence of multiword sequences. Such sequences are termed 'n-grams,' where n has a value that indicates the number of words in the sequence. Thereby, 2-grams are sequences of word pairs, 3-grams are sequences of word triples, etc. The result of 4-gram frequency analysis on the text of the novel 'Emma' gives the 'top ten' results shown in Table 2, below.

The configuration screen for the graphical version of the Posit Text Profiling Toolset is illustrated in Figure 1.

Using the n-gram facility of the Vocabulary Profiler, we can readily contrast the 4-gram result with the 'top ten' 2-gram result from the same text (Table 3).

2.1 Other Posit features

A range of features has been added to the core Posit functionality with the graphical interface development. The principal additions are

i. results database,

ii. optional POS tagging and support for multiple taggers and pre-tagged text,

iii. concurrent profile execution, and

iv. concordance.

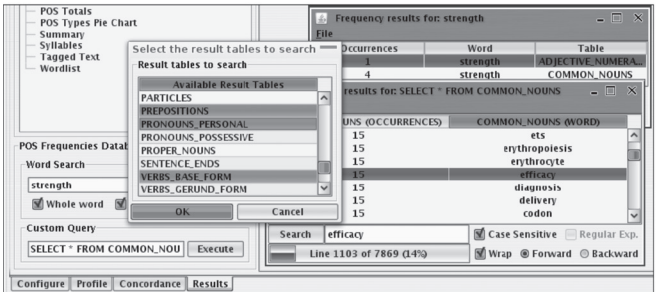
2.1.1 Results database

Inclusion of a relational database for storing the results of word/POS tag frequency analyses affords a powerful new addition to the Posit system. Through this facility a user may perform searches across numerous results files and cross reference words to determine the grammatical types under which they have been categorised and in what contexts they are used within the test corpora (Figure 2).

2.1.2 Optional POS tagging

Since most script features are configurable, the GUI also allows the user to configure the POS tagging. As well as

FIGURE 2: Database search facility



turning POS tagging off altogether, thereby accommodating pre-tagged texts, the user may opt to change from one POS tagger to another. The application comes with two POS taggers but, through the addition of wrapper scripts, also allows for the use of external POS taggers.

2.1.3 Concurrent profile execution

This useful facility allows the user to perform simultaneous independent analyses on two or more sets of texts and manage the profiles and results through separately specified project windows (Figure 3). Each profile will have its own set of configuration, profile, concordance and results windows. Although processed independently, the concurrent availability of separate sets of results will facilitate ease of visual comparison across the analysed texts.

2.1.4 Concordance

The concordance feature adds a common and useful textual analysis component that was absent from the original Posit tools. Through the concordance a user can select a keyword and the desired word span on either side of the keyword. The system will then display all occurrences of the keyword in the contexts provided by the surrounding number of adjacent words. The concordance feature is illustrated in Figure 4. With the addition of a concordance, the Posit Toolset becomes one of the most versatile and complete textual analysis tools available.

Concordance searching is an interactive feature that is performed on the original corpora; it is not a batch job carried out by the scripts. Since the concordance facility did not fit with any of the existing tabs in terms of similar functionality it is provided in a separate tab. Concordance results are displayed in a file viewer similar to that of the Results tab. This also allows the user to have many concordance result windows open simultaneously for comparison purposes. The concordance results can also be saved as an HTML file for subsequent viewing as a 'webpage' within the Posit tool or through a Web browser.

FIGURE 3: Managing multiple projects

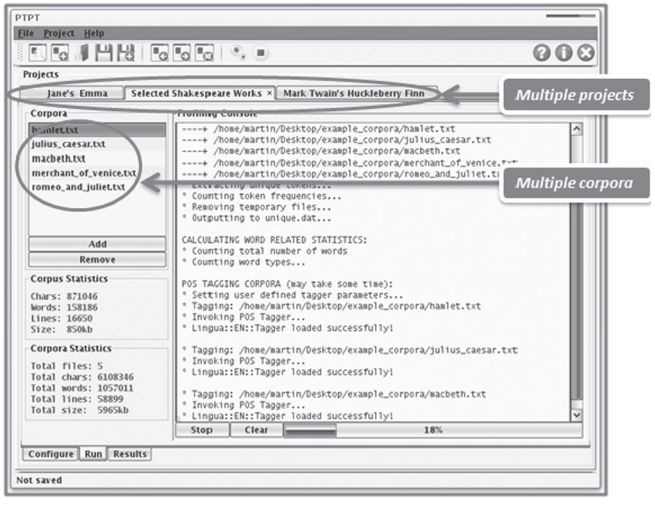
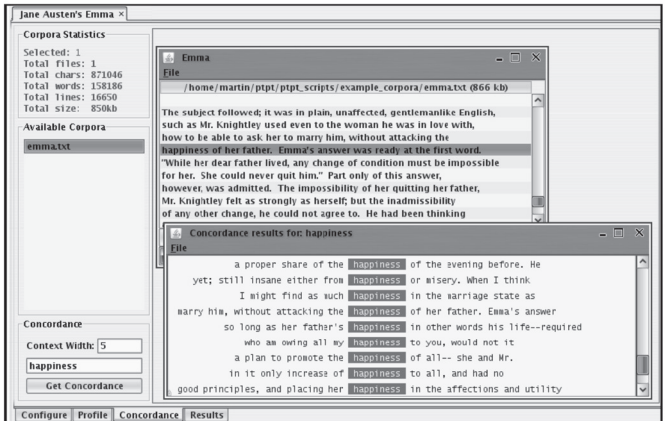


FIGURE 4: Concordance feature



3. Analysis of Japanese EFL textbooks

The present study contrasted three EFL textbook series used in Japan at different historical periods. This forms part of an on-going programme of diachronic analyses (e.g., Ozasa & Erikawa, 2004; Weir & Ozasa, 2007) and exploration of textual analysis techniques (e.g., Weir & Anagnostou, 2007). The first textbook series, Barnes' New National Readers, was published in 1883–84 and is taken to represent the 'early' period of ESL teaching in Japan. The second textbook series, Okakura, Yoshisaburo, The Globe Readers, was published in 1907 and is taken to represent the 'middle'

period of ESL teaching in Japan. The third textbook series, Jack and Betty: English Step by Step, was published in 1948 and represents 'recent' ESL teaching in Japan.

Our comparative analysis focused on part-of-speech profile and vocabulary—specifically, single word frequency and n-gram frequency (for n=2 through n=4). Our aim was to explore the POS and vocabulary profiles for each textbook series. The data for this work was derived using the Posit Text Profiling Toolset (described above). As well as comparing the word contents across these three textbook series, we also consider contrasts with data from the Brown corpus (Kucera and Francis, 1967). We also compared the degree of hapax legomena[2] for single and multiword vocabulary. Finally, we considered the textbooks' coverage of words from Dolch's high-frequency lists, as recommended for reading practice.

As a basis for our comparative study, we derived statistical data on each of the three textbook series listed above. To this end, each series was digitized and treated as a single text corpus. The three resultant corpora were analysed independently using the Posit Toolset to produce extensive word and n-gram frequency data. General statistics for the three textbook corpora are shown in Table 4, below.

The following data dimensions were considered in our contrastive analysis of the textbooks: (1) POS profile, (2) single word frequency; (3) 2-gram frequency; (4) 3-gram frequency, (5) 4-gram frequency. In addition, we compared the proportion of hapax legomena for each of the word frequency dimensions across the three textbook series. Finally, we considered the coverage of Dolch high frequency words. Each of these features is detailed below.

TABLE 4: General Statistics for Textbook Corpora

	National	Globe	JandB
Total tokens	190470	67231	51557
Total types	12154	6869	4923
Type/Token Ratio	15.67	9.78	10.47
No. of sentences	10572	4484	5392
Ave. sent. length	18	15	10
Ave. word length	5.5	5.5	5.4

3.1 Data results

3.1.1 Part-of-speech profiles

Percentage values for the contribution of each part-of-speech are shown in Table 5, below.

Already, we can note that each of the three textbook series has the same ranking for part-of-speech distribution. This is an interesting result that could not have been readily predicted.

Using the data from Table 5, we can contrast the distribution for parts-of-speech by graßphing each set of percentages. The result, shown in Figure 5, below, indicates a strong similarity in profile across the three textbook corpora.

TABLE 5: POS Percentages for Textbook Corpora

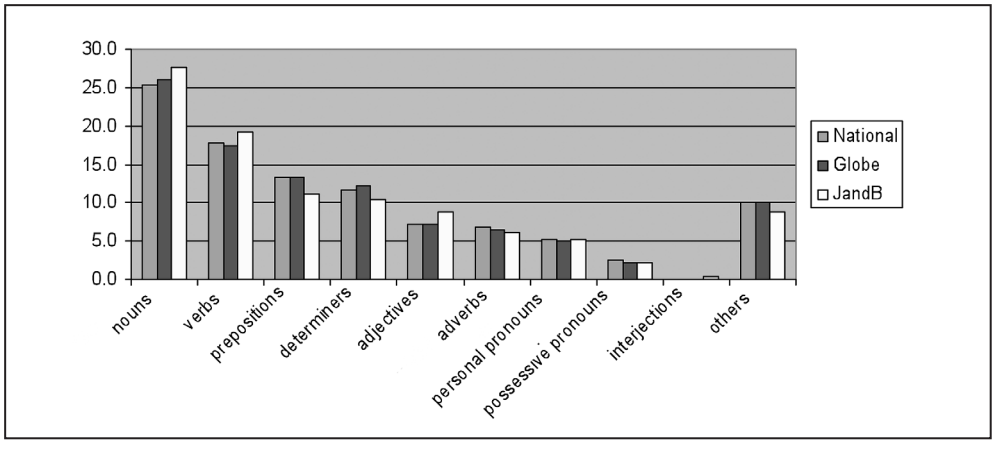
	National	Globe	JandB
nouns	25.3	26.0	27.7
verbs	17.9	17.4	19.3
prepositions	13.2	13.3	11.1
determiners	11.7	12.2	10.5
adjectives	7.2	7.2	8.8
adverbs	6.8	6.4	6.1
personal pronouns	5.3	5.0	5.2
possessive pronouns	2.6	2.2	2.2
interjections	0.0	0.1	0.3
others	10.0	10.1	8.8
total	100	100	100

In order to compare POS profiles as a dimension of English 'quality,' we employed a non-textbook corpus as a reference point. To accommodate the likely influence of American English in the composition of Japanese textbooks the Brown corpus was included as a source of American English.

A POS profile for the Brown corpus was produced by applying the Posit POS Profiling facility. This produced the POS profile shown in Table 6, below. We can better compare the POS distributions across the textbook series and the Brown corpus in Figure 6.

From this comparison, several conclusions seem appropriate. Firstly, as indicated earlier, each of the textbook corpora have remarkably similar POS profiles. Secondly,

FIGURE 5: POS Profiles for Textbook Corpora



each textbook POS profile is also markedly similar to that of the Brown corpus. What inferences may be drawn from these observations? Perhaps the textbooks share an underlying similarity in their choice of language forms. This may not be unreasonable, given their similar pedagogical objectives. Beyond this, their resemblance in POS profile to the Brown corpus suggests that they may also reflect a degree of English 'naturalness.' Certainly, their profiles appear consonant with American English characteristics as represented by the Brown corpus. We should perhaps note that National was originally published in the U.S. for American pupils and imported to Japan for use by Japanese learners of English. This may go some way toward accounting for simi-

TABLE 6: POS Percentages for the Brown Corpus

nouns	29.5
verbs	16.4
prepositions	14.0
determiners	11.4
adjectives	7.9
adverbs	4.7
personal pronouns	4.5
possessive pronouns	1.7
interjections	0.0
others	9.9
total	100

FIGURE 6: Textbook POS profiles compared with the Brown Corpus

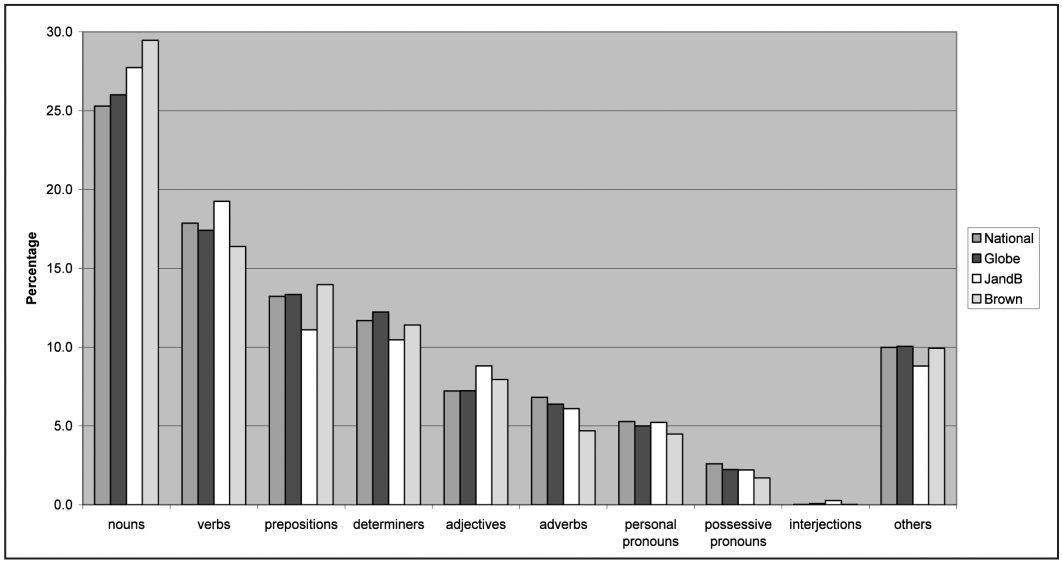


FIGURE 7: Single word frequency comparison

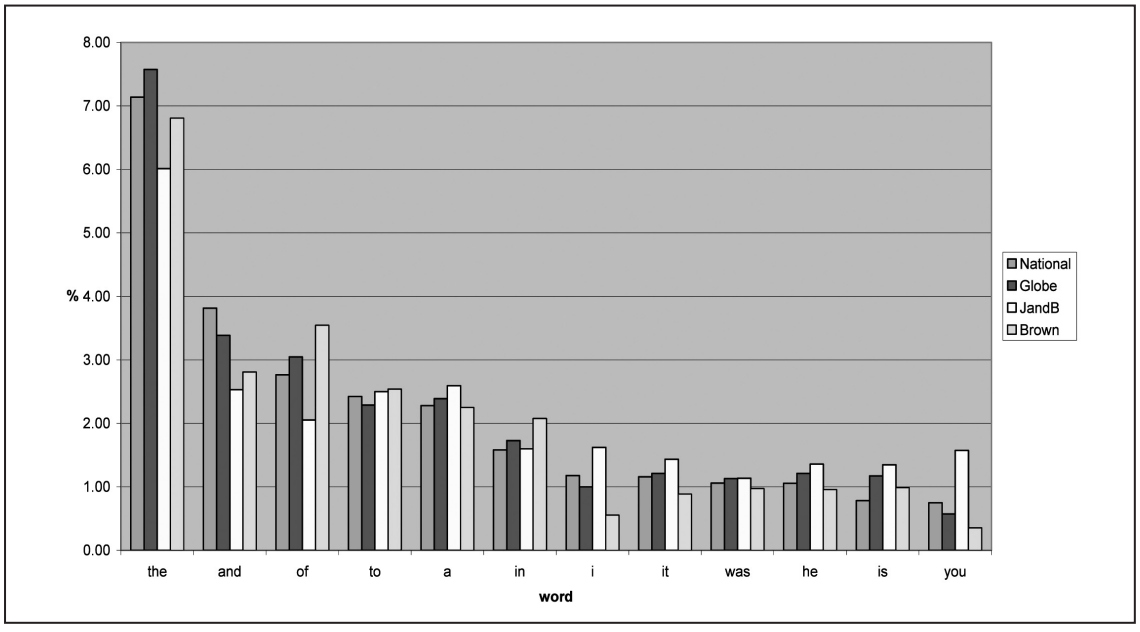


TABLE 7: Top ten words for each textbook corpus

National	Globe	JandB
the	the	the
and	and	a
of	of	and
to	a	to
a	to	of
in	in	i
i	it	in
it	he	you
was	is	it
he	was	he

larity between National and the Brown corpus. Yet, despite their Japanese origin, both Globe and Jack and Betty also exhibit similar POS profiles to the Brown corpus.

3.1.2 Single word frequency

In order to compare the word frequency and n-gram frequency dimensions, for each textbook series we selected the top ten items in each dimension and noted their frequency of occurrence as a percentage of the respective corpus. Thereby, we may compare not only item rank across textbook corpora but also item 'weight' relative to its own textbook corpus. One result of word frequency analysis for the textbook corpora is shown below (Table 7). This lists the top ten word occurrences by frequency for each textbook series. The absolute frequency value is shown for each word.[3]

This tabular comparison reveals that a compact set of only 12 distinct words accounts for the top ten occurrences by frequency across all three textbook series. This indicates a high degree of similarity in high frequency words.

Figure 4, above, allows a comparison across the frequency results in terms of percentage 'weight' of each word in its corpus. Additionally, Figure 1 includes a similar contrast with the Brown corpus. A visual comparison using Figure 1, suggests that National and Globe are close in terms of single word weighting, whereas Jack and Betty appears to diverge from these two textbook series. Finally, the Brown corpus results show a marked degree of deviation from the textbook series.

3.1.3 N-gram frequency

When considering the top ten word pairs (2-grams) by frequency for each of the textbook corpora, we find an intersecting list of 14 word pairs. These are shown in rank order, in Table 8, below.

The results of the frequency comparison of 2-gram occurrences are shown in Figure 8, below.

FIGURE 8: 2-gram frequency comparison

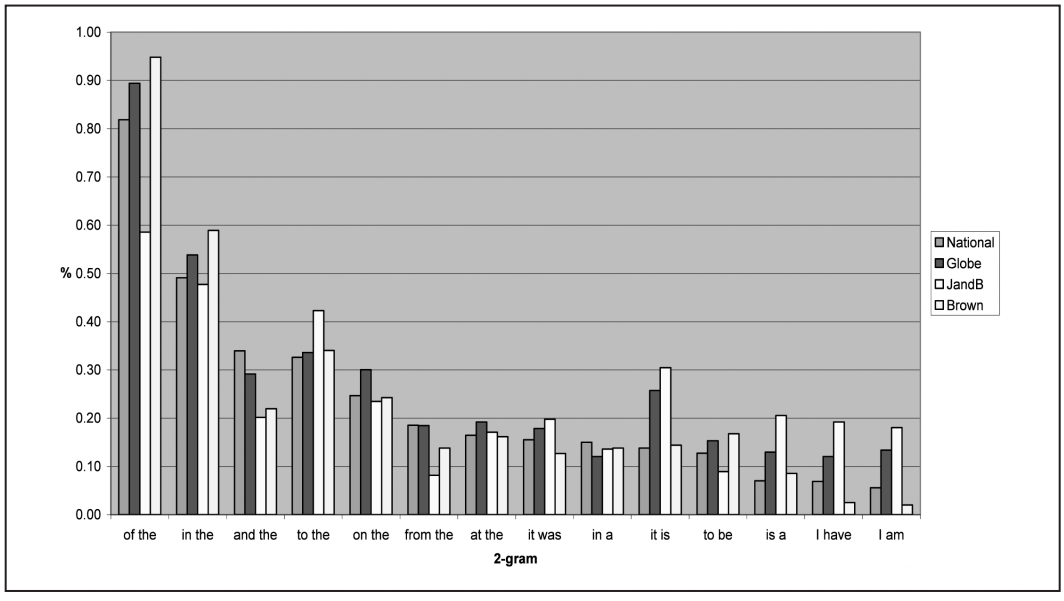


TABLE 8: Top 2-grams for each textbook corpus

National	Globe	JandB
of the	of the	of the
in the	in the	in the
and the	to the	to the
to the	on the	it is
on the	and the	on the
from the	it is	is a
at the	at the	and the
it was	from the	it was
in a	it was	i have
it is	to be	i am

As with the single word comparisons (Figure 7), Figure 8 suggests greater similarity between National and Globe, while Jack and Betty and the Brown corpus exhibit greater divergence. Likewise, a contrast in terms of 3-gram frequencies reveals the top ten three word sequences for each of the textbook series. This results in an intersecting list of 21 word triplets. These are shown in rank order in Table 9, below. Note that the presence of apostrophes in the original texts serve to differentiate some expressions, e.g., 'I do not' and the elision form 'I don't' appear as separate word triplets.

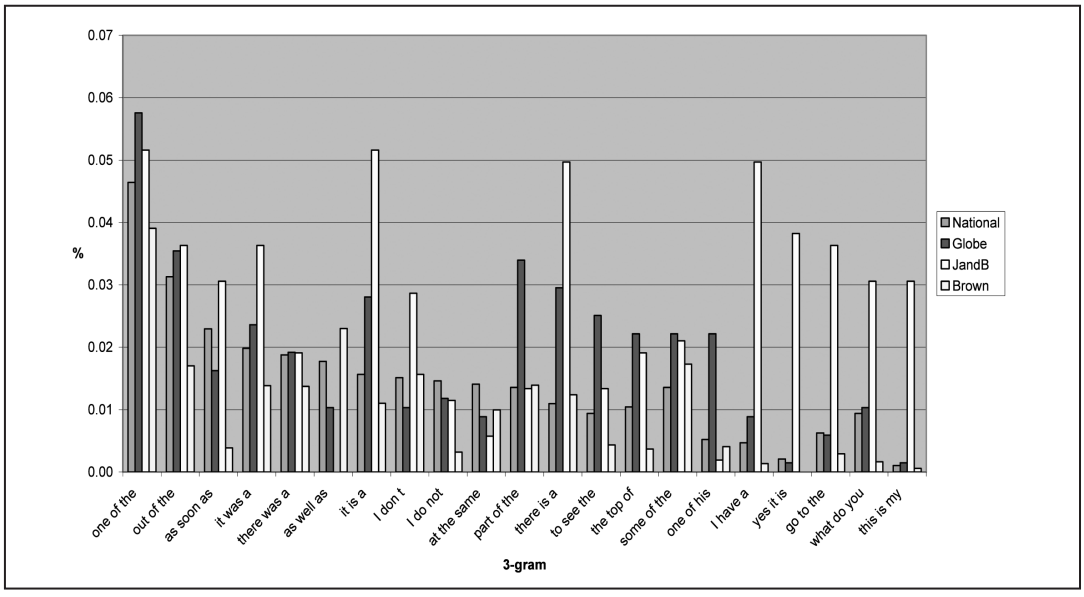
The results of the frequency comparison of 3-gram occurrences across all three textbook series and the Brown corpus are shown in Figure 9, below. This appears to show particular disparity on the part of Jack and Betty.

Analysis of 4-grams from our textbook corpora revealed a list of 29 separate four word sequences. Only one 4-gram was common in the top ten between corpora - the expression 'one of the most'. This was ranked 3rd in National and 4th in Globe. In Jack and Betty, this 4-gram is ranked 52nd. This hints at further divergence between Jack and Betty from the other two textbook corpora, although one might see the patterns for Globe and Jack and Betty as indicative of a focus upon grammar rather than vocabulary. Variation in usage is likely to be greater as we consider larger multiword expressions. This is evident in our comparison of 4-gram frequencies across the textbook series and the Brown corpus (Figure 10).

3.1.4 Hapax legomena

Conducting an exhaustive comparison of single or multiword units is beyond the scope of the present work. Our comparisons of 'top ten' items give some indication of the spread of most frequent items across the textbook corpora, and in contrast with the Brown corpus. Significantly, the highest frequency items account for relatively small proportions of the original texts. This suggests that vocabulary may not be the principal agenda for the textbook authors. A similar point can be made in a different fashion.

FIGURE 9: 3-gram word frequency comparison



The proportion of hapax legomena for each corpus may afford insight on the degree of vocabulary focus for each textbook series. Table 10 lists the percentages of hapax legomena for single words, 2-grams, 3-grams and 4-grams, across the textbook corpora. Clearly, the proportion of hapax legomena increases rapidly with the size of n-gram. For single words, National and Jack and Betty have around 45% and 48% hapax legomena, while Globe exceeds both of these with ~50%. This rises to ~98%, 99% and 97%, respectively for National, Globe and Jack and Betty, when we consider 4-gram occurrences.

One might assume that careful textbook design would address vocabulary coverage, as well as grammatical considerations. Furthermore, one might expect more frequent use of those words that were considered more important than others. Yet, the figures from Table 10 show that almost half of the individual words in each of the textbook series are used only once. This may suggest that only a small proportion of words were considered important enough to merit frequent use (or, indeed, that little consideration was given to word usage aside from meeting the needs of grammatical construction).

We may draw similar tentative conclusions regarding the data on multiword hapax legomena. There is decreasing indication of attention to multiword usage as the multiword sequence size increases. This leads to the small proportion of ~1–3% of 4-grams that are used more than once in each textbook series.

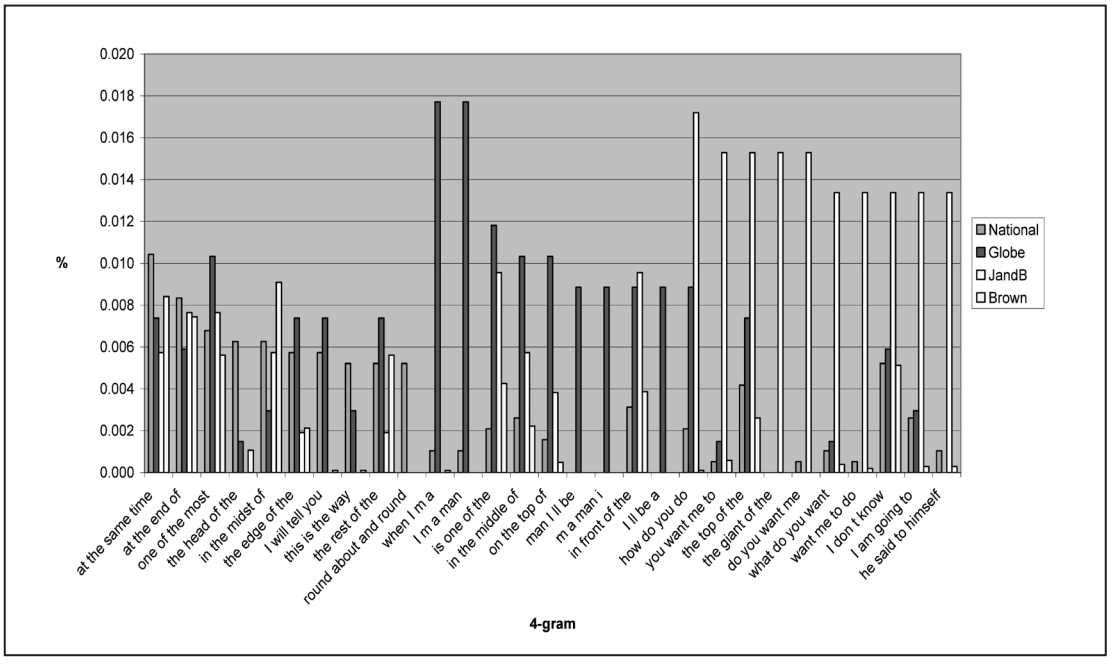
3.1.5 Dolch comparison

Edward Dolch compiled two word lists based upon their frequency of occurrence in children's books (Dolch, 1948) on the reasonable assumption that learners must be able to recognise such words in order to achieve reading fluency. His primary list contains 220 words (excluding nouns). Nouns are covered in a second list of 95 words. These Dolch lists are still in common use across schools in the United States and the United Kingdom, as a basis for gauging student progress in reading. In the context of our ESL textbook comparisons, the Dolch lists allow us to consider the degree of fit between the vocabulary of the textbooks and the content of the Dolch lists.

The first factor we consider is the presence or absence of the Dolch words in the textbook corpora. Indeed, we may regard the degree of Dolch word presence as a crude indicator of suitability of the textbook as a reading tuition, whereby, absence of a Dolch word represents a significant lack in terms of learner guidance.

Figure 11 shows a comparison of the top twenty Dolch words against our three textbook corpora. Values indicate the comparative ranking of these words (which, in turn, is indicative of relative frequency). The top twenty

FIGURE 10: 4-gram word frequency comparison



Dolch words have different rankings across our three textbook corpora but are all well represented in every textbook series.

Analysis of the word contents for each of the textbook corpora reveals some differences in their coverage of Dolch words. Firstly, with regard to the main Dolch list of 220 words, we find that the National textbook series contains all of these Dolch words. The Globe series contains all the Dolch words with two exceptions, the words 'ate' and 'slow.' Finally, the Jack and Betty textbooks contain all but one of the Dolch words. The missing item in this case is the word 'drink.'

For Dolch's list of high frequency nouns, we find greater variation in omission across the textbook corpora. Once again, the National textbook series is best in terms of Dolch word coverage. National omits only one word, 'birthday.' The Globe textbook series omits 10 words from the noun list and the Jack and Betty series omits 7 words from this list. The missing Dolch nouns for these textbooks are listed in Table 11.

4. Conclusions

This application of textual analysis to these three sets of Japanese EFL textbooks allows us to consider a range of comparisons across the textbook series. The comparative data allows us to consider the degree of similarity or divergence between these book series at the level of individual and multiword usage. For instance, Jack and Betty appears to diverge more from the other two textbook series in terms of n-gram frequencies. The high incidence of hapax legomena suggests a lack of focus upon multiword sequences on the part of the textbook authors and may indicate greater emphasis upon grammar over vocabulary. Finally, Dolch word analysis indicates good coverage of the main high frequency reading words but lower success in terms of high frequency nouns, especially in the case of the Globe series.

More broadly, our analysis of these textbooks provides a case study for the application of computer-based text analysis tools. There are other features in our textbook

TABLE 10: Percentages of hapax legomena

	National	Globe	JandB
1-gram	44.80	50.48	48.38
2-grams	76.68	79.84	76.49
3-grams	92.15	94.23	91.29
4-grams	97.99	98.76	96.62

FIGURE 11: Top twenty Dolch words

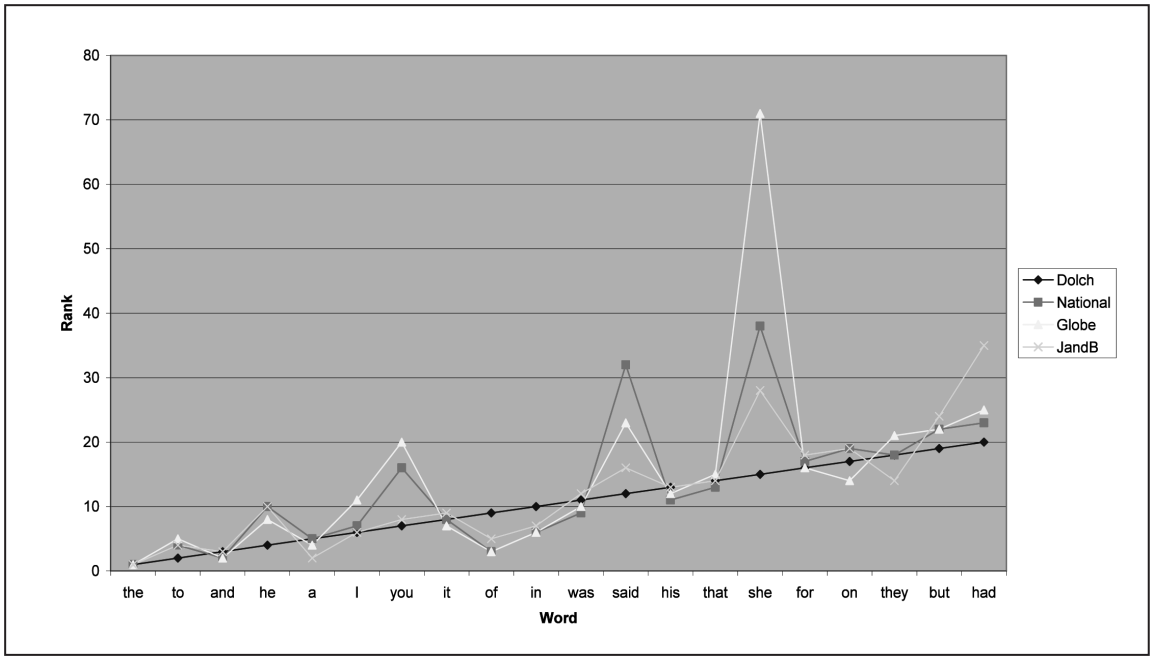


TABLE 11: Missing Dolch nouns

National	Globe	JandB
birthday	chair	chicken
	chicken	cow
	doll	duck
	egg	kitty
	kitty	rabbit
	pig	santa
	rabbit	seed
	santa	
	squirrel	
	toy	

series that could be elucidated through analysis tools like Posit. For instance, our hypothesis that grammar rather than vocabulary was prominent in the textbooks' design may be further explored by analysis of the grammatical structures employed across the textbooks. We propose that tools such as Posit have considerable potential as a means of learning from available text corpora, whether in the realm of EFL or elsewhere.

REFERENCES

Anthony, L. (2005). *AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. Proceedings of IWLeL 2004:* An Interactive Workshop on Language e-Learning.

Aston, G., Bernardini, S., & Stewart, D. (2004). *Corpora and Language Learners.* Amsterdam: John Benjamins.

Baillie, M. & Weir, G.R.S. (2008). The Posit toolset with graphical user interface. *Proceedings of the 4th ICTATLL Workshop,* Colombo, Sri Lanka, September 2008.

Bontcheva, K., Tablan, V., Maynard, D., & Cunningham, H. (2004). Evolving GATE to meet new challenges in language engineering. *Natural Language Engineering, 10,* 349–373.

Dolch, E. W. (1948). *Problems in Reading.* Champaign, IL: Garrard Press.

Granger, S., Hung, J., & Petch-Tyson, S. (2002). Computer Learner Corpora, *Second Language Acquisition and Foreign Language Teaching.* Amsterdam: John Benjamins.

Kucera, H., & Francis, W. (1967). *Computational Analysis of Present-Day American English.* Providence, RI: Brown University Press.

Ozasa, T., & Erikawa, H. (2004). *Eigo Kyokasho no Rekishiteki Kennkyu [A Historical Study of English Textbooks].* Tokyo: Jiyusha.

Scott, M. (1998). *Wordsmith Tools Version 3,* Oxford, UK: Oxford University Press.

Silberztein, M. (2005). NooJ: A Linguistic Annotation System for Corpus Processing. In *Proceedings of HLT/EMNLP on Interactive Demonstrations,* Vancouver, 10–11.

Weir, G. R. S. (2007). The Posit Text Profiling Toolset. *Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics.* Pan-Pacific Association of Applied Linguistics, Bangkok, Thailand, December 2007.

Weir, G. R. S., & Anagnostou, N. K. (2007). Exploring Newspapers: A Case Study in Corpus Analysis. *Proceedings of ICTATLL 2007,* Hiroshima, Japan, 12–19.

Weir, G. R. S., & Ozasa, T. (2007). Estimating naturalness in Japanese English textbooks. *Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics.* Pan-Pacific Association of Applied Linguistics, Bangkok, Thailand, December 2007.

Weir, G. R. S., & Ozasa, T. (2008). Multiword Vocabulary in Japanese ESL Texts. *Proceedings of the 13th Conference of Pan-Pacific Association of Applied Linguistics.* Pan-Pacific Association of Applied Linguistics, Honolulu, Hawai'i, August 2008.

ENDNOTES

[1] For detailed analysis, the system outputs data files in spreadsheet-compatible format.

[2] Individual words (or word sequences) that appear only once in a collection of texts.

[3] The frequency counts treat each word as lower case.