Chapter 1

TaLC in action: recent innovations in corpus-based English language teaching in Japan

Yukio Tono

This chapter discusses the effective use of corpora in English language teaching by introducing various areas of corpus-based applications in Japan, including the world's first TV English conversation programme based on corpora. The areas of application are divided into two, following Tudor's (1997) needs analysis framework: the creation of teaching resources and materials based on the analysis of native speaker corpora, viewed as the target for learners and the investigation of L2 learner and textbook corpora, viewed as their present situation. Various case studies show that a corpus-based approach has been very effective in reforming English language teaching in Japan, and can be applied in other countries.

1.1. Introduction

Corpus-based research has become increasingly popular in applied linguistics. Applications in English language teaching have three major areas: (a) indirect use of corpora, (b) direct use of corpora and (c) compilation of corpora for educational purposes (Leech 1997). Indirect use involves the development of corpus-based materials such as educational word lists, dictionaries, grammar books, conversation textbooks, etc. Direct use usually involves using corpora in the classroom, where the typical approach is Data Driven Learning (Johns and King 1991). The creation of educational corpora concerns specialised areas such as textbook corpora, classroom observation corpora, learner corpora and corpora of texts with controlled vocabulary. In this paper, I will share my experience of corpus-based approaches to various areas of English language teaching in Japan, in

order to illustrate the considerable potential of 'teaching and language corpora'.

To better situate the use of corpora in English language teaching, I shall adopt the framework of needs analysis proposed by Tudor (1997). Tudor distinguishes between two different kinds of needs analysis: Target Situation Analysis (TSA) and Present Situation Analysis (PSA). TSA deals with the analysis of learners' targets. Without a clear understanding of learners' goals, it is difficult to design a syllabus to achieve them. In the case of language teaching, TSA should be based on the analysis of the core components of the target language, and for this purpose native-speaker (NS) corpora can provide useful information about high-frequency lexis and grammar. PSA, on the other hand, analyses learners' present situation, thus showing the gap between this and the target. In ELT, this means assessing the language proficiency of prospective learners, and in a corpus-based approach, we can employ learner corpora to investigate interlanguage features of their speech and writing.

Within this needs-analysis framework, I have carried out several kinds of research, as illustrated in Figure 1.1. First, I have created a user-friendly interface for accessing mega-corpora which English teachers can use for teaching purposes. Second, I have produced corpus-based teaching materials in English, including TV conversation programmes, conversation books,

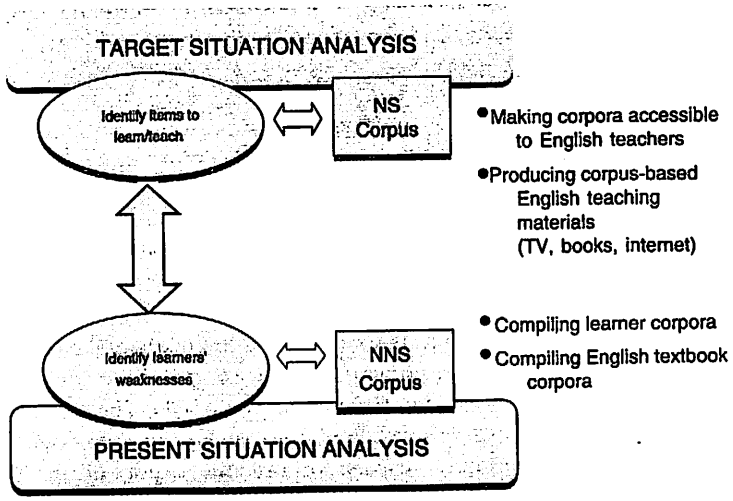

FIGURE 1.1 Needs analysis and corpus-based research in English language teaching

Internet e-learning contents and multimedia applications. Both these areas of research are based on the analysis of native speaker corpora, which is related to TSA, as explained above. Third, I have compiled two educational corpora: a learner corpus and an English textbook corpus. These corpora have been used to analyse the present situation of learners and how their output is affected by textbook input (PSA), and hence to diagnose the gap between the present and the target situations.

In this paper, I outline some of my research activities in these areas. First, the project for creating web-based corpus query services for English teachers will be illustrated. Secondly, I shall describe the development of the first corpus-based TV English conversation programme in Japan, and discuss its impact on English teaching there. Thirdly, I will mention some related educational materials based on corpora. The last part of my paper will be devoted to corpus-based analyses of learner language and textbooks. I will show how learner corpora shed light on the transitional characteristics of interlanguage development, and how textbook corpora reveal deficiencies in English textbooks published in Japan in comparison to those in other major Asian countries. In Japan, a corpus-based approach to ELT has proved very successful, and I will argue that the same approach can be taken in other countries.

1.2. Creating a user-friendly web-based corpus query system

Before 2000, little was known about corpus linguistics among ELT communities in Japan. The only exception was the COBUILD project and its products. I discussed the possibility of a corpus query system with Shogakukan Inc., one of the largest publishers in Japan, whose dictionary division went on to develop a set of Corpus Query Language (CQL) and web-based query tools called SAKURA (see Figure 1.2).

These tools were originally designed for in-house use by lexicographers, but the team was encouraged to build them in such a way that novice corpus users would be able to use them intuitively. At the end two versions of SAKURA were released, an advanced version and a simplified one. The latter was developed into a web-based interface for the Shogakukan Corpus Network (SCN), the first commercial web corpus query service in Japan (http://www.corpora.jp). This provided a unified interface to different corpora such as the British National Corpus (BNC), WordBanksOnline (WBO), the PERC Corpus and the JEFLL Corpus. While web query tools

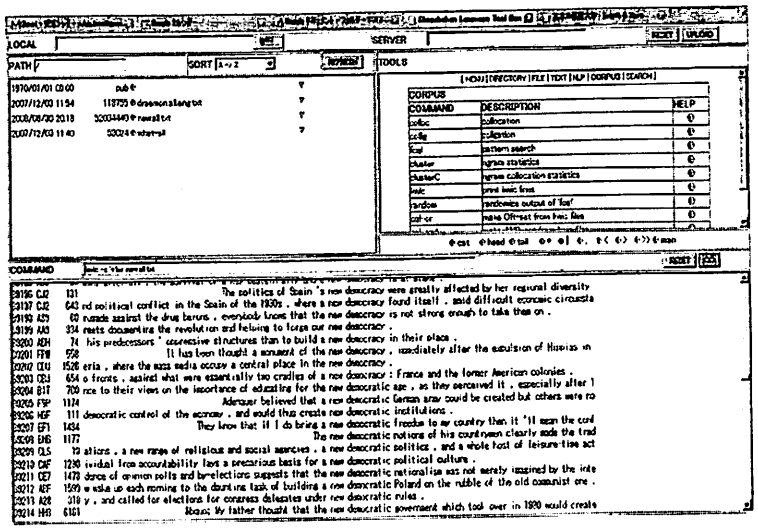

FIGURE 1.2 Shogakukan Corpus Query System, SAKURA BNC interface

for large corpora are often slow and fail to deal well with simultaneous access, SAKURA can handle a dozen mega-corpora simultaneously, and can be accessed by more than fifty people at the same time from the classroom.

1.3. Developing a corpus-based TV English programme in Japan

The second major area of corpus applications to ELT was the project to develop the world's first corpus-based English conversation TV programme. The public broadcasting center NHK (Nihon Hosou Kyokai, the Japan Broadcasting Center) is well known for excellent educational broadcasts, which include more than thirty programmes for the major foreign languages, including English, Chinese, German, French, Italian, Korean, Spanish, Russian and Arabic. In 2002, NHK reformed its foreign language broadcasting, and decided to produce a programme which would be screened every day. I suggested that this should consist of dozens of short episodes, each featuring one important English word and its use.

1.3.1. The '100-Go' programme

The basic features of the '100-go' programme were to be as follows:

(a) 10-minute English conversation TV programmes (4 units per week for 25 weeks, approximately 6 months);
(b) 100 units focusing on 100 key words;[1]
(c) Key words to be selected on the basis of frequency data from the BNC spoken component;
(d) Useful corpus data to be presented for each key word;
(e) Fun, exciting and educational;
(f) For all levels, including false beginners.

1.3.2. The making of '100-Go'

To design the syllabus, I extracted the frequency list of lemmas with part-of-speech information from BNC-spoken. This showed that core vocabulary does most of the work: out of 57,457 types in BNC-spoken (c. 10 million words), the most frequent 100 words cover 67% of the corpus tokens. Figure 1.3 shows the breakdown of the selected words, which are mainly

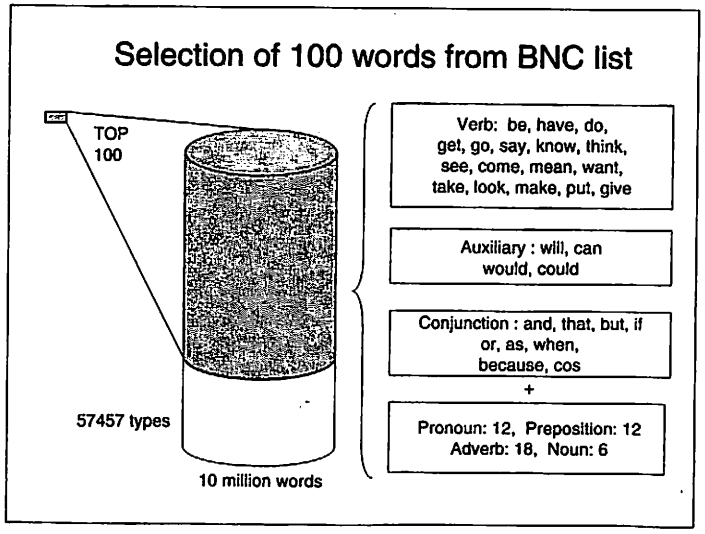

FIGURE 1.3 Selection of 100 words from the BNC-spoken list

major lexical verbs and function words (pronouns, prepositions, adverbs, conjunctions, auxiliaries).

The fact that only 100 words cover almost 70% of the entire spoken data shows that these basic words are used in a variety of ways. Most of the verbs have more than one meaning and take complex complementation patterns. The prepositions, conjunctions and auxiliaries are all polysemous. The adverbial particles are difficult to use when combined with verbs to form phrasal verbs. Thus, even though these words may look familiar, learners often lack sufficient knowledge of their usage.

The programme thus focused on these very core lexical items, featuring mainly verbs, auxiliary verbs, conjunctions and prepositions. For each key word, two kinds of corpus data were prepared: (a) collocation lists for given colligation patterns (e.g. 'verb + noun' or 'verb + prep/part'), and (b) n-grams using the key word. The verb 'give', for example, appeared three times as a key word, in three different constructions: 1. *give* + *sb* (=somebody) + regular noun, 2. *give* + preposition/particle, 3. *give* + deverbal noun. For each construction, lists of collocations were extracted from BNC-spoken to show the most frequent patterns of use. Thus, in the case of '*give* + deverbal noun', learners were taught the following collocation sets: (a) *give (sb) a ring*, (b) *give (sb) information*, (c) *give (sb) a kiss*, (d) *give (sb) some advice* and (e) *give (sb) an answer*. N-gram information was used when it was more suitable as a framework than colligation patterns.

Another original feature was the selection of topic vocabulary sets for each key word. For instance, the collocation pattern 'give (sb) a ring' can be used with temporal expressions as follows:

Give me a ring tomorrow / this evening / at seven o'clock

For each key word a set of ten additional words or phrases was prepared to go with the patterns in the collocation list. In this way, the key word multiplies, and learners will get to know how it is used in context. Figure 1.4 illustrates this design feature diagrammatically.

After selecting all 100 key words with their collocation rankings and accompanying topic vocabulary sets, native speaker writers were asked to write skits to show the use of those collocation patterns in real contexts.

Each unit in the monthly textbook accompanying the series has a six-page format. The first two pages show the key word title and model dialogues in English, with Japanese translations. Page 3 provides the collocation ranking and usage notes for the key word, while page 4 features key sentences in the skits. Finally, pages 5 and 6 provide exercises, including

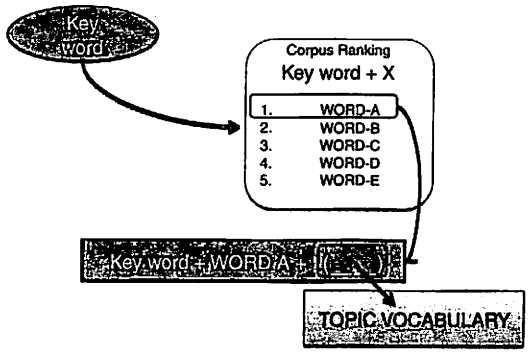

FIGURE 1.4 The concept of '100-Go': one key word will multiply

FIGURE 1.5 'Mr Corpus' introducing corpus ranking

mechanical drills using collocation sets, as well as situation-based, gap-filling oral compositions.

Given the success of the first series (Tono 2003), the TV programme ran for three years (2003–2005) with different varieties of English (US, Australian and UK). Each year there were a good-looking bilingual guy and a beautiful girl as MCs, and a special character called 'Mr Corpus', who introduced the corpus ranking (see Figure 1.5). I, as Dr Corpus, appeared for one minute to summarise the main points surrounding the key words. Tono (2004a) was essentially based on the same key word lists, but with improved collocation rankings and fashionable computer graphics

(a 'Ninja' version of Mr Corpus). In Tono (2005a), the key word list was expanded to 200 to incorporate basic nouns and adjectives.

1.3.3. Impact of '100-Go'

More than one million people watched the TV programme each year, and 'Mr Corpus' became popular among children. The first three monthly textbooks sold out, for the first time in the history of NHK's TV English programmes. The production team won the NHK President Award for the best programme and the best textbook of 2003. 'Corpus' became a buzz word, and many teachers became aware of the usefulness of corpus data as teaching and learning resources. Corpus-based English teaching materials gained increasing attention, and began to influence syllabus and materials development. The first TV series reran in 2008, and a brand new series was launched in 2009 (Tono 2009a).

1.4. Developing corpus-based English teaching materials in Japan

1.4.1. Conversation books

A series of books were published in relation to the NHK TV series. Tono (2004b) featured the programme's 100 key words and their collocation data with model sentences and exercises. Tono (2005b, 2006a) covered the second and third series respectively. DVD-book series were published for each edition of the programme (Tono 2004c, 2005c, 2006b). These were exported to other Asian countries and translated into Korean (Tono 2004d) and Chinese (Tono 2005d). Many became bestsellers, attracting attention from English learners as well as teachers.

I also published other types of conversation books. Tono (2005e) focused on 150 key words, where the information from corpora was centred on the phraseology surrounding each key word. Thus in the case of 'give', the phraseology was:

(a) I'm going to/I'll give you . . .
(b) give it to . . .
(c) give up . . .
(d) can you give me . . . ?
(e) let me give you . . .

As can be seen, this list is based on frequent phrase patterns rather than particular colligation patterns.

Tono (2005f, 2007a) focused on the use of nouns as productive vocabulary. Learners often only know a one-to-one relationship between English words and their translation equivalents, and find it difficult to use these words in actual contexts. For example, they know the word *hand* and its translation equivalent (*te* in Japanese), but they often have difficulties in expressing *te no hira* (*the palm of one's hand*), or *te wo ageru* (*put one's hand up*) in English. The books focused on fifty common nouns used in everyday conversation (Tono 2005f) and in business contexts (Tono 2007a). In each unit English collocation lists are presented in 5 x 5 tables (with the key word in the centre and 24 collocations), along with a corresponding table of Japanese translations (see Figure 1.6). Learners can use the Japanese table to test whether they can come up with the corresponding English phrases, and their lexical knowledge becomes more productive as they learn how to express different concepts using the key noun.

1.4.2. Reference materials

Another major area of corpus application in English teaching remains that of developing reference materials, such as dictionaries and vocabulary lists. Tono (2004e) is a semi-bilingual version of the *Cambridge Learner's Dictionary* (2nd edition). It provides translation equivalents for each sense of each word, so that learners can use it as a bridge between bilingual and monolingual dictionaries. It also provides special notes on word usage and common learner errors based on the Cambridge International Corpus and the Cambridge Learner Corpus.

Tono (2005g) is probably the world's first synonym dictionary to be based on corpora. It has 200 thematic entries in Japanese, each of which shows two to four English synonymous translation equivalents. For example, the Japanese thematic entry *sukuu* lists three translation equivalents, *save, rescue* and *relieve*. Each equivalent is illustrated with its five most salient collocation patterns in the BNC, so that these near synonyms can be compared.

save + N	*rescue* + N	*relieve* + N
(1) life	(1) hostage	(1) pressure
(2) planet	(2) economy	(2) pain
(3) queen	(3) country	(3) boredom
(4) child	(4) prisoner	(4) burden
(5) soul	(5) child	(5) poverty

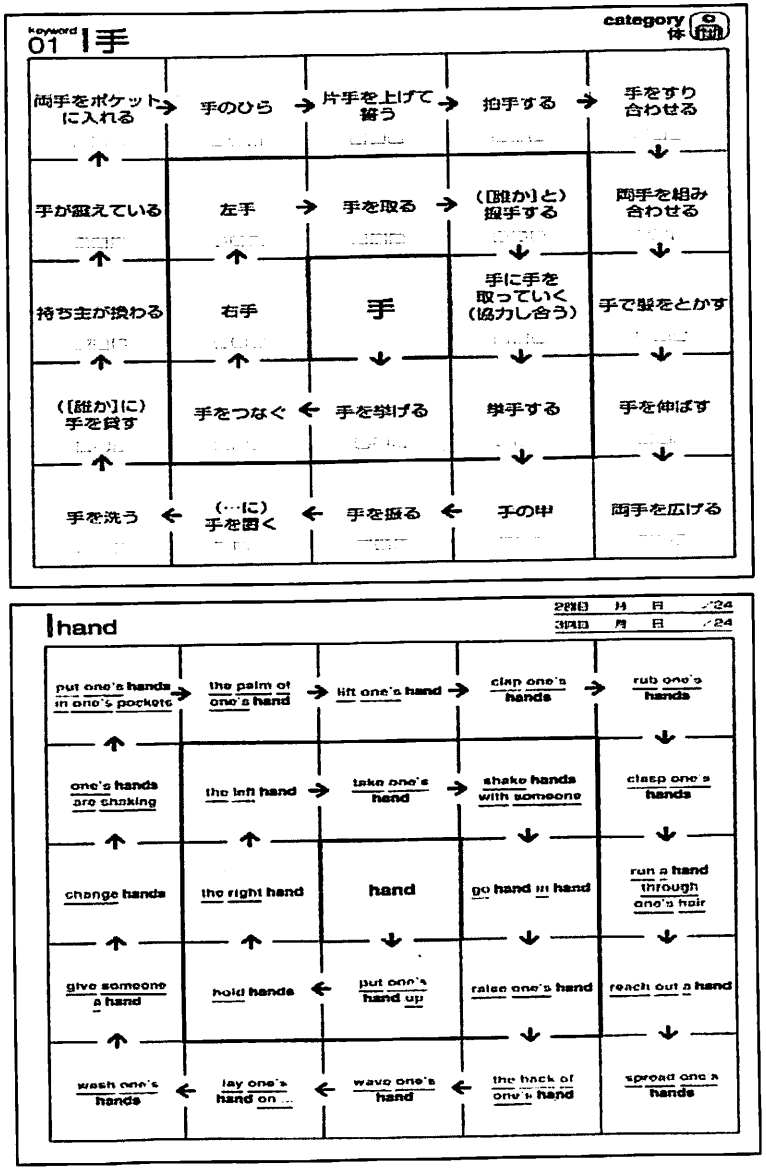

FIGURE 1.6 Japanese vs. English noun collocation table (Tono 2005f)

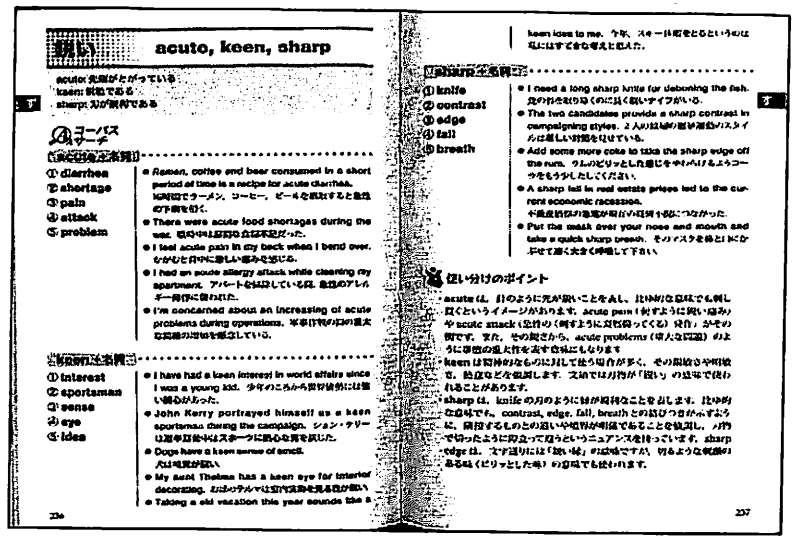

FIGURE 1.7 Sample page for the entry '*surudoi*' – *acute, keen, sharp* (Tono 2005g)

This approach is a direct application of Firth's (1957) notion of 'knowing a word by the company it keeps'. 'Save' is the most general word, covering physical life, concrete persons or objects, as well as a spiritual one ('soul'). 'Rescue' implies saving someone or something in immediate danger, while 'relieve' involves reducing pain or making a problem less difficult. Figure 1.7 shows a sample page of the entry for *surudoi* (*acute, keen, sharp*).

Tono (2008) is a bilingual learner's dictionary targeted at elementary to lower-intermediate level learners of English at secondary school. It has special 'Focus Pages', which are basically corpus-based summary pages of core vocabulary. Figure 1.8 shows the focus page for the word *have*.

Each focus page has various sections:

(a) *Core image*: an image showing the core meaning of the word
(b) *Semantic map*: a list of senses in a dictionary entry
(c) *Structure ranking*: a frequency list of major verb patterns
(d) *Corpus phrases*: major colligation patterns with collocation lists
(e) *Spoken phrases*: phrases taken from spoken corpus data
(f) *Textbook phrases*: phrases taken from a textbook corpus

There are approximately a hundred of these focus pages, each providing corpus-based information on the usage of the word and the most useful

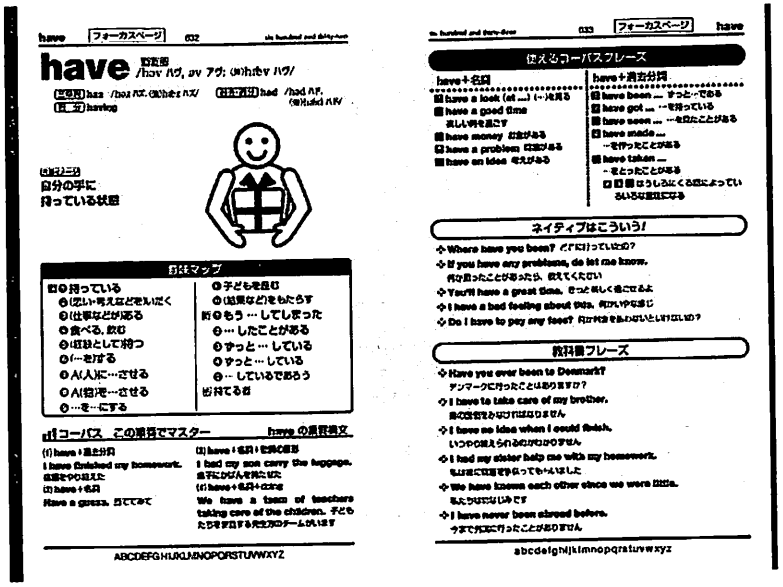

FIGURE 1.8 Focus Page for *have* (Tono 2008)

structures and phrases in which it occurs. Again, this builds on the idea that it is essential to acquire a working knowledge of the core vocabulary that covers 70% of speech. The dictionary proposes a clear boundary between essential vocabulary and the rest: 100 core words are featured in the focus pages, an additional 2,000 words are treated with an emphasis on their use as active vocabulary, and the next 3,000 words are highlighted typographically but treated less extensively. Words below the 5,000 word level are given minimal space, with only translation equivalents and a few illustrative examples. In this way, Japanese learners of English at the secondary school level should be able to see which words they should learn to use, which words they should know the meaning of and which words they can just look up if necessary. Use of the dictionary is thus closely linked to vocabulary learning strategies.

1.4.3. Computer game applications

Fun and excitement can be incorporated into language learning through computer game applications. Tono (2009b) is a newly developed iPhone English vocabulary learning programme (see Figure 1.9), which has become

FIGURE 1.9 English Vocabulary iPhone application

one of the most popular iPhone applications for language learning. It has ten different vocabulary levels, each of which has special training menus with amusing exercises involving reading, listening and writing. People can win belts for certain grades by obtaining high scores in the tournament.

The NHK programme '100-go' was so popular that the plan to develop a Wii application is now underway (Tono, forthcoming), and this may well be the world's first corpus-based language learning game application.

To sum up, corpus-based English teaching materials have been hugely successful in Japan, and this corpus-based approach to developing teaching materials should work equally well in other countries. There are several good reasons. First, people are eager to find the most economical way of learning a foreign language. I do not mean that there is any magical, instant way, but simply that some are more cost-effective. Corpus rankings based on frequencies provide users with very clear guidelines for learning lexical combinations. By focusing on frequent collocation or phrase patterns, learners will be able to acquire the core component of L2 knowledge faster and more effectively than they can with materials without information from corpora. Second, people like to get an objective measure of the importance of items to learn, for which corpora provide evidence. Viewers of my TV programmes responded very positively to the corpus rankings as an efficient way of making their goals concrete and objective. Knowing the importance of the items to learn leads

to higher motivation. Finally, corpus data give learning materials greater credibility. In the past, the design of syllabuses and teaching materials depended largely on the materials developers' experience and intuitions. The advent of computer technology and corpus linguistics has made it possible to access large amounts of naturally occurring language data, showing how the language is actually used by native speakers. People are now aware that the use of corpus data makes learning materials more reliable.

In the following section, I move on to describe my second major area of research, based on L2 learner corpora. This will illustrate the present situation of L2 learners in Japan, the second kind of input required for an effective needs analysis (see Figure 1.1 above).

1.5. L2 learner corpus research

1.5.1. Compilation of the NICT JLE Corpus and the JEFLL Corpus

Compiling a learner corpus is a major project. I have been involved in two such projects for Japanese-speaking learners of English. The first, the NICT JLE Corpus, is a collection of transcripts of more than 1,200 subjects' oral proficiency test interviews. The test is the Standard Speaking Test (SST) developed by ALC Press, which is a customised version of the ACTFL Oral Proficiency Interview. Each 15-minute interview has five parts: warm-up, picture description, storytelling, role play and wind-down. Each interview script has an individual proficiency score on nine levels: beginner (level 1) to near-native (level 9). The corpus is available as a book with a CD-ROM (Izumi et al. 2004) and also in electronic format under license.

The other L2 learner corpus, the JEFLL Corpus (Tono 2007b), is a collection of more than 10,000 Japanese secondary school students' English compositions. It contains timed, in-class, free compositions in English on six different topics (argumentative or narrative). Each task was given as part of regular classroom activities, not as homework. Subjects were not allowed to use dictionaries, but if there were any words they could not come up with in English, they were allowed to write them in Japanese. The average length of each essay is rather short (about sixty to seventy words), but with more than 1,000 compositions in each school year category, we could approximate the patterns of use and possibly paths of learning. The JEFLL Corpus is available via the Shogakukan Corpus Network (http://scn02.corpora.jp/~jefll04dev/), where it can be accessed using a web-based query tool in English.

1.5.2. Some findings

The main objectives of these projects are (a) the description of inter-language development in terms of overuse vs. underuse as well as correct use vs. misuse of certain linguistic features, (b) the identification of criterial features which distinguish one proficiency level from another and (c) the development of a list of language features which are learnt particularly slowly, and perhaps call for revision or modification of teaching syllabi or methodologies. Theoretically, we also hope to distinguish those patterns of overuse/underuse/misuse which are specific to the learners' L1, and those which appear common to learners from whatever background. In this way, we could possibly redesign syllabuses, adjusting them to the L2 learning path, and ask people working on action research in the classroom to test the effects of the modifications. This will all lead us to a better understanding of the gap between the target situation and the present situation illustrated in Figure 1.1 above, and how best to fill that gap.

1.5.3. Identification of criterial features

Tono (2000a) investigated the relationship between the subjects' school year and frequencies of part-of-speech (POS) tag sequences in the JEFLL Corpus (sequences of three tags = trigrams). By looking at these sequences, we can observe frequent patterns of use, which helps us understand the process of acquiring syntactic patterns in the target language. This was done by POS tagging the learner data and extracting tag sequences, and then performing a data reduction statistical procedure called Correspondence Analysis over the frequencies of tag sequences across different school-year groups. The results are shown in Figure 1.10.

The analysis shows that the beginning level has a tendency to be more closely associated with verb-related patterns, while noun- and preposition-related trigrams are more closely associated with lower-intermediate and advanced learners respectively. This clearly shows that more advanced students have a tendency to use more complex noun phrases and prepositional phrases. We also observed constant underuse of auxiliaries and articles. In the early stages of acquisition, learners tend to use short sentence units, consisting mainly of verbs and arguments (e.g. subjects or objects, etc.) with minimum modifiers. At more advanced levels, on the other hand, learners start to expand arguments by adding adjectival or adverbial modifiers. This is one of the first studies in learner corpus research to investigate the syntactic features characterising different stages of acquisition.

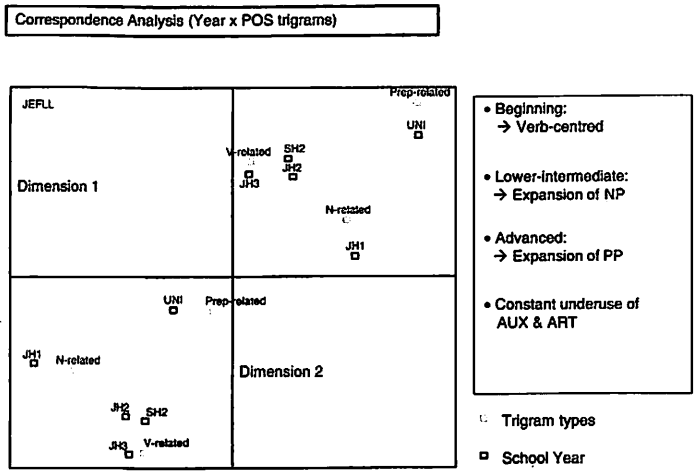

FIGURE 1.10 Analysis of POS tag sequences across school years (Tono 2000a)

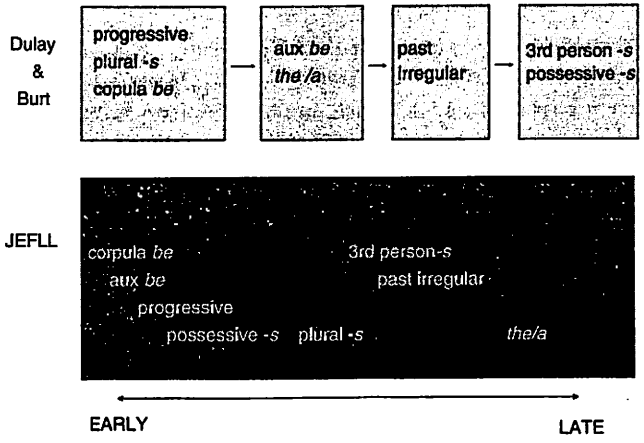

FIGURE 1.11 Dulay and Burt's (1972, 1974) morpheme-order study replicated using learner corpora (Tono 2000b)

1.5.4. Error frequencies across proficiency

A second area of studies has been the analysis of error frequencies across proficiency levels. Tono (2000b) was one of the first learner-corpus-based studies of acquisition order for grammatical items (see Figure 1.11).

This replicated the English grammatical morpheme studies by Dulay and Burt (1972, 1974), and showed that while there were many similarities in the order of acquisition, Japanese learners also showed distinctive tendencies. In the so-called universal order of acquisition, the article system is supposed to be acquired in the middle of the acquisition order, while the possessive *–s* is acquired very late. But the article system was the last to be acquired by the Japanese learners, while the possessive marker was acquired relatively early. The article system is difficult for the Japanese, because there is no article system in our language. On the other hand, the possessive *–s* seems relatively easy because the Japanese genitive marker *-no* behaves in a very similar way. These findings are consistent with previous empirical studies on Japanese EFL learners (cf. Shirahata 1988).

1.5.5. Automatic error identification

In order to analyse learner errors, the entire JEFLL Corpus was examined by a native speaker and corrected for errors. Tono and Mochizuki (2009) used Dynamic Programming (DP), a sequence alignment method of finding corresponding patterns in text, in order to identify the similarities and differences between the original essays and the corrected ones. Automatic extraction of omission, addition and misformation errors (see James 1998) was performed and the output further processed by Correspondence Analysis. Figures 1.12 and 1.13 show the results for omission/addition errors in relation to school year.[2]

Omission errors are the type that learners make when they omit words or morphemes in a position where they are obligatory. As Figure 1.12 shows, lower-level learners (J1, J2) tend to omit verbs, conjunctions and personal pronouns, whereas learners at higher levels tend to omit determiners, prepositions and adverbs. There is a strong tendency for elementary level learners (J-level) to omit core sentence elements, like verbs or pronouns used as subjects or objects. Conjunctions work as phrase/clause connectors, and elementary level learners have difficulties handling these as well. Intermediate learners (H-level), on the other hand, have overcome these fundamental omission errors, but make new types of errors involving more complex noun/verb phrase structures. Thus they tend to omit prepositions when modifying noun phrases, and adverbs when modifying verbs and adjectives.

Figure 1.13 shows patterns of addition errors. Noun addition errors are common at the very beginning stage of learning (see Circle C, closely

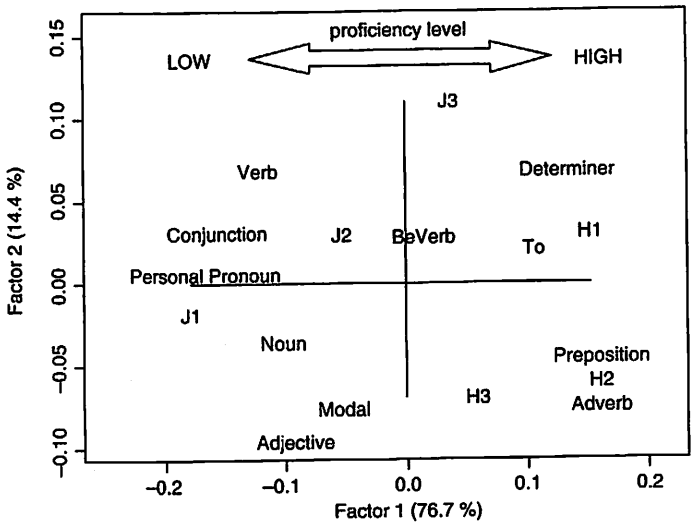

FIGURE 1.12 Correspondence Analysis – omission errors by part of speech vs. school year

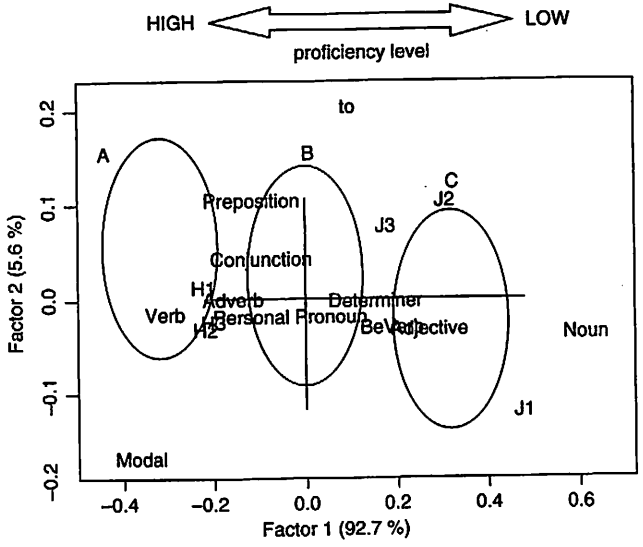

FIGURE 1.13 Correspondence Analysis – addition errors by part of speech vs. school year

associated with J1 level), but soon the number decreases dramatically. This is partly due to the fact that sentences by novice learners are often a series of nouns without verbs and function words. Once this stage is over, they acquire the basic SV+X pattern of English. Another significant tendency is shown in Circle A. Intermediate learners tend to make addition errors for prepositions, conjunctions, verbs and adverbs. This may seem to contradict the findings for omission errors, but actually both types are frequent. Intermediate-level learners risk using more complex sentence structures, making extensive use of modifiers and sentence connectors, leading to more addition as well as omission errors.

Learner corpus analysis has not yet been integrated into the mainstream of syllabus construction and materials design, despite some fragmentary references in usage notes in learner dictionaries (e.g. *Cambridge Advanced Learner's Dictionary*, *Longman Dictionary of Contemporary English*). Identifying criterial features of L2 developmental stages will surely help towards a more learner-centred, acquisition-conscious materials design.

1.6. English textbook corpora across Asian countries

My last major area of research has involved compiling textbook corpora in order to assess the quality and quantity of L2 input for Japanese learners of English. This too is a part of what Tudor (1997) calls Present Situation Analysis. I have compiled corpora of English textbooks in Japan, Korea, Taiwan and China in order to make comparisons across different Asian countries where English is taught as a foreign language.

Figure 1.14 shows the overall text size of the junior high school English textbooks in the four countries. Compared to Korea and Taiwan, English textbooks published in Japan are 3 to 4.5 times smaller in terms of total amount of text. Textbooks in China are four to six times larger. This means that the amount of exposure to English provided by the textbooks used in Japan is relatively limited.

Table 1.1 shows the coverage of vocabulary in senior high school English textbooks in the four countries for the top 10,000 words in the entire BNC.

The first column shows the vocabulary level (1,000 to 10,000 based on the BNC frequencies). The second column indicates the number of these types found in the textbooks across the four countries. Out of the top 1,000 words in the BNC, for instance, 972 types were found across the textbooks in the four countries, 89.81% of those 972 words were found in Korean textbooks, and so on. The coverage of the 5,000–10,000 level words is

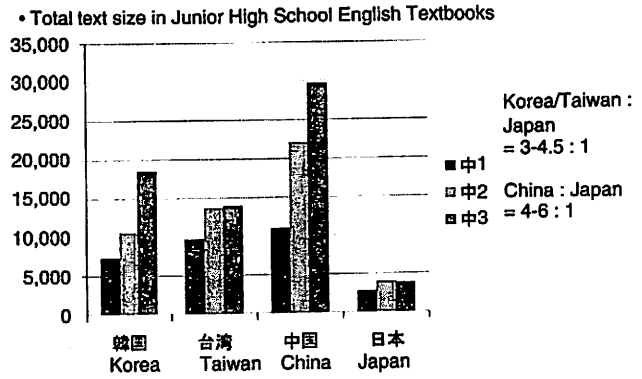

FIGURE 1.14 Text size in Junior High English textbooks in four Asian countries

Table 1.1 Textbook coverage of the BNC top 10,000 words

BNC frequency ranks	Words found	Korea	Japan	China	Taiwan
1–1000	972	89.81%	77.98%	84.16%	96.71%
1001–2000	863	63.73%	45.54%	52.72%	87.02%
2001–3000	656	43.14%	24.70%	32.01%	76.98%
3001–4000	454	37.67%	16.96%	23.57%	68.94%
4001–5000	342	35.96%	19.30%	14.91%	69.30%
5001–6000	250	30.40%	14.40%	18.40%	66.80%
6001–7000	150	33.33%	8.00%	16.00%	63.33%
7001–8000	129	28.68%	10.08%	9.30%	69.77%
8001–9000	93	30.11%	10.75%	10.75%	70.97%
9001–10000	70	27.14%	12.86%	15.71%	62.86%

around 30% in Korean textbooks, and 15% in Chinese textbooks respectively. Textbooks in Taiwan, on the other hand, tend to use very high-level vocabulary in the texts, indicated in the higher coverage rates of the

vocabulary levels 5,000–10,000. This suggests that English textbooks in Taiwan use native-like authentic texts as they are, while Korea and China provide more vocabulary control. China especially seems to have a sensible policy of providing a large amount of easy-to-read text in order to promote better reading skills.

Compared to these three countries, high school English textbooks in Japan have serious problems. First, coverage of the 1,000 word level vocabulary is less than 80%, which is low compared with the other three countries. This is probably due to the fact that the size of the textbooks is so much smaller. The same can be said about the coverage of the 2,000 word level. With little exposure to these basic lexical items, Japanese learners at high school are forced to read more difficult English texts without having had the chance to read simpler ones. This is the exact opposite of the way Chinese textbooks expose learners to the language.

In May 2008 I made a report on my comparative study of textbooks to the Education Reform Round Table set up under the auspices of ex-Prime Minister Yasuo Fukuda. The committee subsequently recommended a drastic reform of English education in Japan. This was a case where corpus data caused a change in national language policy.

1.7. Conclusions

In this paper, I have shared some of my research activities and practice in Japan. Corpus-based approaches to language teaching have made a significant difference in redesigning language policy, language syllabuses and modes and methods of learning. Japan is one of the most technology-aware nations in the world, which could be one of the reasons why people are attracted by the idea of using computational analysis to improve language education. This same approach, however, may be applicable to other countries, and some of the methodologies and materials developed can be shared among researchers and practitioners. I hope that my research and experience – based on data of attested language use – may also contribute to a better practice and pedagogy in foreign language teaching and learning elsewhere.

Notes

[1] This notion of 'key word' is different from what Scott (2004) calls a key word. Here it simply means a core lexical item or word feature in a unit.

[2] The accuracy rate for misformation errors was relatively low in this experiment, and clearly in need of further refinement. Consequently data for this class of errors is not given here.

References

Dulay, H.C. and Burt, K.M. (1972), 'Goofing: an indicator of children's second language learning strategies', *Language Learning*, 22, (2), 235–252.

Dulay, H.C. and Burt, K.M. (1974), 'Natural sequences in child second language acquisition', *Language Learning*, 24, (1), 37–53.

Firth, J.R. (1957), 'A synopsis of linguistic theory 1930–1955', in *Studies in Linguistic Analysis* (Special Volume of the Philological Society). Oxford: Blackwell, pp. 1–32.

Izumi, E., Uchiyama, K. and Isahara, H. (eds.) (2004), *Nihonjin 1200-nin no Eigo Speaking Corpus.* (An Oral Corpus of 1200 Japanese Learners of English). Tokyo: ALC Press.

James, C. (1998), *Errors in Language Learning and Use: Exploring Error Analysis.* London: Longman.

Johns, T. and King, P. (eds.) (1991), *Classroom Concordancing. English Language Research Journal*, 4. Birmingham: University of Birmingham.

Leech, G. (1997), 'Teaching and language corpora: a convergence', in A. Wichmann, S. Fliegelstone, T. McEnery and G. Knowles (eds.), *Teaching and Language Corpora.* London: Longman, pp. 1–22.

Scott, M. (2004), *WordSmith Tools 4.* Oxford: Oxford University Press.

Shirahata, T. (1988), 'The learning order of English grammatical morphemes by Japanese high school students', *The JACET Bulletin*, 19, 83–102.

Tono, Y. (2000a), 'A corpus-based analysis of interlanguage development: analysing part-of-speech tag sequences of EFL learner corpora', in B. Lewandowska-Tomaszczyk and J. Melia (eds.), *PALC '99: Practical Applications in Language Corpora.* Frankfurt am Main: Peter Lang, pp. 323–340.

Tono, Y. (2000b), 'A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes', in L. Burnard and T. McEnery (eds), *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the Third International Conference on Teaching and Language Corpora.* Frankfurt am Main: Peter Lang, pp. 123–132.

Tono, Y. (2003), *100-go de Start! Eikaiwa.* (Let's start English with 100 Words! USA Edition), 6 vols. Tokyo: NHK Publishing.

Tono, Y. (2004a), *100-go de Start! Eikaiwa.* (Let's start English with 100 Words! Australia Edition), 6 vols. Tokyo: NHK Publishing.

Tono, Y. (2004b), *Corpus Renshucho.* (*Corpus Drill Books*). Tokyo: NHK Publishing.

Tono, Y. (2004c), *100-go de Start! Eikaiwa: DVD-book.* (Let's start English with 100 Words! USA Edition: DVD-book). Tokyo: NHK Publishing.

Tono, Y. (2004d), 우선순위 100 단어로 티뜨리는 영어회화. Seoul: Darakwon.

Tono, Y. (2004e), (ed.) *Cambridge Learner's Dictionary: Semi-bilingual Version.* Tokyo: Cambridge University Press and Shogakukan Inc.

Tono, Y. (2005a), *100-go de Start! Eikaiwa.* (Let's start English with 100 Words! UK Edition), 6 vols. Tokyo: NHK Publishing.

Tono, Y. (2005b), *Super Corpus Renshucho.* (*Super Corpus Drill Books*). Tokyo: NHK Publishing.

Tono, Y. (2005c), *100-go de Start! Eikaiwa: DVD-book.* (Let's start English with 100 Words! Australia Edition: DVD-book). Tokyo: NHK Publishing.

Tono, Y. (2005d), 關鍵100生活英會話. Taipei: Kai Hsin Publishing.

Tono, Y. (2005e), *Mimi kara Oboeru Corpus Gensen 150-go.* (Learn 150 Words by Listening: a Corpus-based Approach). Tokyo: Takarajimasha.

Tono, Y. (2005f), *Eikaiwa Corpus Drill: Nichijo Kaiwa Hen.* (English Conversation Corpus Drills: Daily Conversation). Tokyo: ALC Press.

Tono (2005g), (ed.) *Shogakukan Corpus-Based Dictionary of English Synonyms.* Tokyo: Shogakukan Inc.

Tono, Y. (2006a), *Corpus Renshucho Plus.* (Corpus Drill Books Plus). Tokyo: NHK Publishing.

Tono, Y. (2006b), *100-go de Start! Eikaiwa: DVD-book.* (Let's start English with 100 Words! UK Edition: DVD-book). Tokyo: NHK Publishing.

Tono, Y. (2007a), *Eikaiwa Corpus Drill: Business Hen.* (English Conversation Corpus Drills: Business English). Tokyo: ALC Press.

Tono, Y. (2007b), *Nipponjin Chukousei 10,000-nin no Eigo Corpus.* (A Corpus of English Compositions by 10,000 Japanese Secondary School Students). Tokyo: Shogakukan Inc.

Tono, Y. (2008), (ed.) *Sanseido's ACE CROWN English-Japanese Dictionary.* Tokyo: Sanseido.

Tono, Y. (2009a), *Corpus 100! De Eikaiwa.* (English Conversation with Corpus 100), 6 vols. Tokyo: NHK Publishing.

Tono, Y. (2009b), *Tono Yukio no Eitango Dojo.* (The English Vocabulary Boot Camp by Yukio Tono). Tokyo: Tokyo Shoseki.

Tono, Y. (forthcoming), *100-go de Start! Eikaiwa.* NINTENDO Wii version. (Let's start English with 100 Words!). Kyoto: NINTENDO.

Tono, Y. and Mochizuki, H. (2009), 'Toward automatic error identification in learner corpora: a DP matching approach', Paper presented at Corpus Linguistics conference 2009, University of Liverpool, UK, 22 July 2009.

Tudor, I. (1997), *Learner-Centredness as Language Education.* Cambridge: Cambridge University Press.