Selecting Level-Specific Kyoto Tourism Vocabulary Using Statistical Measures

Kiyomi Chujo
Nihon University
chujo@cit.nihon-u.ac.jp

Masao Utiyama
NICT
mutiyama@nict.go.jp

Kathryn Oghigian
Tokyo International University
oghigian@gmail.com

The Japanese government's "Action Plan for Tourism Development" in 2003 has prompted colleges and universities to set up departments to specialize in tourism. In order to supply educators with keywords associated with tourism, this study selected beginner, intermediate and advanced level specialized vocabulary using statistical tools previously established to identify level-specific, domain-specific words (Chujo and Utiyama, 2005, 2006). In this study, a Kyoto tourism corpus was compiled from 'Kyoto-guide' texts that consists of four components: 'miru' (sight-seeing), 'kau' (shopping), 'taberu' (dining), and 'taikensuru' (hands-on activities). The corpus was then compared with the British National Corpus High Frequency Word List (Chujo, 2004) using statistical measures such as the log likelihood ratio and mutual information. An examination of the resulting vocabulary lists showed that each statistical measure extracted an appropriate level of domain-specific words by its vocabulary level, grade level, and school textbook vocabulary coverage.

BACKGROUND

According to the Japan National Tourist Organization, the total number of Japanese tourists abroad in 2005 reached 17.4 million, while the total number of international visitors to Japan was estimated to be 6.7 million[1]. This imbalance between outbound and inbound tourism was the impetus behind the Japanese government's 2003 "Action Plan for Tourism Development[2]." Measures such as the 'Visit Japan Campaign' have been implemented to focus on significantly increasing inbound tourism and have been giving a considerable boost to Japan's recent tourism development.

In response, many colleges and universities[3] have set up faculties and departments that specialize in tourism and its corresponding human resource development. One of the fundamental academic subjects taught is English for Tourism, an English for occupational purposes (EOP) course of study which is one of many types of English for specific purposes (ESP) (Robinson, 1991). One of the prominent characteristics of ESP is a heavy load of corresponding specialized vocabulary or "technical words that are recognizably specific to a particular topic, field, or discipline" (Nation, 2001:198). Since vocabulary expansion is essential for ESL and EFL learners to gain proficiency in English (Nation, 1994), it follows that tourism vocabulary would be essential to any academic tourism program.

REVIEW OF LITERATURE

Several subdivisions exist under the broad umbrella of "tourism English": language and communication for hotels, restaurants and catering, transportation, tours, ticketing and itineraries, resort facilities, and various support retail services as well as handling money, giving or dealing with complaints, health and safety issues, eco-tourism, business, marketing and accounting issues, etc. Even within these subdivisions there are further divisions, for example, a person in a hotel management position may have a different subset of vocabulary and phrases than a bell hop or a housekeeper; similarly the person handling ticketing at a travel agency may not necessarily also be doing marketing or accounting.

There are course books and resources available on tourism English, and some are more comprehensive than others. Wood's (2003) *Tourism and Catering* covers a wide range of aspects, as does *Check Your English Vocabulary for Leisure, Travel, and Tourism* (Wyatt, 2006). Resources that cast a net over a wider area tend not to be as comprehensive as those focused on a narrow subset, and those that are more comprehensive tend to focus only on a limited area. A good example of the latter is *Ready to Order* (Baude, Iglesias and Inesta, 2006), which provides in-depth language for chefs, bartenders and wait staff. So while tourism resources do exist, many seem to offer either a superficial view of many areas, or an in-depth look at one area. To the best of our knowledge, there is no definitive tourism resource that provides in-depth coverage for all aspects of tourism.

In addition, with regard to those resources that do provide more in-depth language, Walker (1995) reports that these have limited value because "a great deal of what is currently available (*English for Hotel Staff*, Nelson; *May I help you?* Cassell; etc.) is too job-specific for the requirements of those following courses in Travel and Tourism at Diploma or Degree levels, since many such students are often uncertain as to which of even the major divisions of tourism attracts them most." Given the inevitable nature of students whose target situations are still largely undefined, and the somewhat hit-or-miss resources currently available, it is apparent that a more comprehensive tourism vocabulary list applicable to wider divisions in tourism may be a useful resource.

PURPOSE OF THE STUDY

The goal of this study is to provide a more comprehensive, broader-based tourism lexicon for Japanese educators and students. This was done by first determining what might be the most meaningful vocabulary based on research on popular Japanese destinations and activities, identifying an appropriate corpus, and then extracting various levels of tourism words by applying statistical measures to the corpus. Once identified, vocabulary level, grade level, and Japanese high school textbook coverage were investigated, resulting in the creation of beginner, intermediate and advanced level tourism vocabulary.

PROCEDURE

Corpus and Methodology

In order to determine how to target the most meaningful vocabulary, we researched statistics on inbound visitors' destinations and preferred activities in Japan. The most frequently visited prefectures by foreign visitors were Tokyo, Osaka, Kyoto, Kanagawa, and Chiba (Mukaiyama, 2003; METI Kansai, 2004; Kamio, 2005). Favored activities were experiencing the 'two-sides of Japan': modern Japan's culture and lifestyle (sightseeing in large cities, shopping and visiting fashionable areas) and its traditional culture (dining on traditional dishes and visiting places of scenic beauty and historic interest) (Kamio, 2005). We also studied the "Best 100" plans published by the Agency of Cultural Affairs (2005) and among these, the most preferred prefectures for Japanese travelers were Kyoto, Nara, and Tokyo. In addition, it was reported in a recent academic survey that the city that Japanese college students would most like to introduce to visitors from overseas was Kyoto, followed by Tokyo (Ichimura, 2004).

It was fortuitous that Kyoto was named as a highly ranked destination because one of the researchers in this study was previously involved in a project related to the above-mentioned 'Visit Japan Campaign' and developed a Kyoto-guide corpus in English. This Kyoto tourism data covers various aspects of modern and traditional Japan, including its history, culture, current events, and local tourist attractions. This corpus provides specialized vocabulary for both a highly ranked destination and a broad range of activities popular with tourists, and could be applicable as a broad-based database for tourism students as well as general English learners who want to be able to discuss Japan and Japanese culture in English (Dantsuji, 2001).

Lam (2004) reminds us that tourism English is very different from general English and that priority should be given to teaching the use of keywords. However, separating technical vocabulary (in this case tourism vocabulary) from general vocabulary has not been an easy task (Briggs and Lee, 2002) since this is time-consuming and heavily dependent on the selector's expertise in English education and specialist knowledge of the field (Utiyama et al., 2004). Chujo and Utiyama (2004) and Utiyama et al. (2004) have established an easy-to-use tool employing various statistical measures to identify level-specific, domain-specific words. Chujo and Utiyama (2005) created a list of written science vocabulary by applying those nine statistical measures to the 7.37-million-word written 'applied science' component of the British National Corpus (BNC). They found that each measure extracted a different level of domain-specific words by vocabulary level, grade level, and school textbook vocabulary coverage and that specific measures produced level-specific words, for example, the log likelihood ratio (*LLR*) identified intermediate-level technical words, and mutual information (*MI*) identified advanced level technical words. These measures were effective in separating technical vocabulary from general-purpose vocabulary, and provide a useful template as a means of identifying domain-specific vocabulary. Thus the Kyoto corpus was identified as our target database, and the statistical measures as our methodology.

Kyoto Tourism Word List

The Kyoto tourism corpus includes 885 Kyoto guide texts in four subcategories: (1) 160 'miru' (sight-seeing) texts, (2) 317 'kau' (shopping) texts, (3) 345 'taberu' (dining) texts, and (4) 63 'taikensuru' (hands-on activities) texts (see **Table 1**). Each text is about 47 words long on average and describes some aspect of tourism related to Kyoto,

for example: the history of a shrine, the best place to shop for a certain item, specialties of a restaurant, or a description of a hands-on pottery class. All the words in this corpus were first lemmatized to extract all the base forms using the CLAWS7 tag set[4]. (For example, *eat*, *eats*, *ate*, *eating*, and *eaten* are all forms of a single lemma and were listed under a base word *eat* with a frequency of five occurrences.) Secondly, all proper nouns and numerals were identified by their part of speech tags and deleted manually. This yielded a 2,786-word Kyoto tourism master list.

Table 1 Composition of the Kyoto-Guide Corpus

	Number of texts	Types	Tokens
Miru (Sight-seeing)	160	1,470	9,236
Kau (Shopping)	317	1,553	13,649
Taberu (Dining)	345	1,463	16,175
Taikensuru (Hands-on)	63	653	2,965
Total corpus	885	2,786	42,025

Three Control Lists

Three control lists were used for creating the extracted Kyoto tourism vocabulary and for investigating the vocabulary level, grade level, and school textbook vocabulary coverage of the statistically extracted vocabulary. These control lists were created using the same lemmatizing procedures described above.

(1) The British National Corpus High Frequency Word List (BNC HFWL) is a list of 13,994 lemmatized words representing 86 million BNC words that occur 100 times or more. (The compiling procedure is detailed in Chujo, 2004.) The British National Corpus (BNC) represents 100 million words of spoken and written British English. By comparing the tourism words in our master list to the BNC HFWL, we can statistically determine how they would appear differently from words in a general corpus.

(2) *The Living Word Vocabulary* (Dale and O'Rourke, 1981) includes more than 44,000 items, and each has a percentage score that rate whether the word is familiar to students in U.S. grade levels 4 through 16. For supplementing grade levels 1 through 3, reading grades from *Basic Elementary Reading Vocabularies* (Harris and Jacobson, 1972) were used. By comparing the tourism words in our master list to this list, we can determine the grade level at which the central meaning of a word can be readily understood.

(3) The junior and senior high school (JSH) textbook vocabulary list containing 3,245 different base words was compiled from the top selling series of Japanese high school textbooks (the *New Horizon 1, 2, 3* series and the *Unicorn I, II* and *Reading* series) in Japan. Japanese high school students generally use these or similar books to study English before entering a university. By comparing the tourism words in our master list to this list, we can determine which words have already been studied by most Japanese high school graduates.

Statistical Measures Used to Identify Outstanding Tourism Words

To extract level-specific vocabulary from the Kyoto tourism corpus, we used five

statistical measures: simple frequency (*Freq*), the complementary similarity measure (*CSM*), the log likelihood ratio (*LLR*), the chi-square test with Yates's correction (*Yates*), and mutual information (*MI*)[5]. The formula for each measure is available on the web[6]. A detailed description of each measure can be found in Utiyama et al. (2004) and Chujo and Utiyama (forthcoming 2006), and the notation for these kinds of statistics can be found in Scott (1997).

These statistical measures are widely used in computational linguistics. They automatically identify outstanding words in frequency of occurrence by making comparisons between one specified list (in this case, the tourism words in our master list) and another larger list (the BNC HFWL). Thus when each statistical measure is applied to the tourism master list, these statistics indicate whether a word is overused or underused compared with a list of general English. Applying each statistical score results in a list of extracted words. The statistical score for the extent of each word's "outstanding-ness" (Scott, 1999) in frequency of occurrence is computed, and the words are sorted from the most outstanding to the least outstanding. Thus the words near the top are ranked most outstanding in terms of each statistical measure's criteria. The goal in using these measures is to narrow down the number of candidates from the 2,786 different words in the tourism master list into more manageable sub-lists, but these sub-lists are not meant to be definitive. These statistical tools can help users to select technical vocabulary automatically without specialist knowledge, and by using these extracted lists, users can easily manually delete irrelevant words.

Verifying the Vocabulary Levels of the Extracted Lists

Each of the five measures produced a sub-list. All the extracted lists were initially examined for an overview comparison of the top 20 extracted words in order to get an indication of the general tendencies for each list. Next, the 300 most outstanding words of each list were examined for their potential for pedagogic applications by comparing them to the three control lists: (1) the BNC HFWL to understand how frequently these words are used compared to general English usage (frequency distribution); (2) the *Living Word Vocabulary* and *Basic Reading Vocabularie*s, to determine word familiarity based on grade level, and (3) the JSH vocabulary list to investigate the number of words covered by the JSH textbook vocabulary. Based on this information, each extracted list was designated as beginner, intermediate, or advanced. This is detailed in the next section.

RESULTS AND DISCUSSIONS

Top 20 Extracted Words Overview Comparison

The top 20 words from each of the five measures in descending order are shown in **Table 2**. The bottom two rows of each column show the average frequency score and average word length of the top 20 words extracted by each statistical measure. The lists in **Table 2** are very different from each other even though they were extracted from the same tourism master list. A glance at the top 20 words shows the general tendencies inherent in the extraction of these five measures.

The top 20 words identified by *Freq* include general words such as *the*, *be*, and *of*,

which usually appear at the top of high frequency word lists. From *CSM* to *MI*, the top 20 extractions appear to include tourism words that gradually shift from simple and general to complex and specific. For example, we can recognize dining words such as *restaurant*, *dish*, *sweet*, and *tea* in the top 20 *CSM* words, while we can see *specialty*, *bakery*, *homemade*, *confectionery*, *paste*, *leek*, *mackerel*, *dish*, and *delicacy* in the top 20 *MI* list. This step verifies that the measures do in fact extract different vocabulary from the same master list.

As we see from the data in the bottom two rows of **Table 2**, the average frequency score of each list decreases from left to right or from *Freq* to *MI*. Inversely, the average word length increases from left to right, ranging from 3.3 to 7.7 letters. As Chujo, Utiyama, and Nishigaki (forthcoming 2006) have shown, difficulty levels increase with increasing word length. Although we are aware that word difficulty may be influenced by many more factors than frequency and word length, this might support the possibility that specific statistical measures can be used to target specific grade level vocabulary. This will be explored in the following sections.

Table 2 A Comparison of the Extracted Tourism Word Lists

Rank	***Freq***	***CSM***	***LLR***	***Yates***	***MI***
1	the	shop	restaurant	restaurant	specialty
2	be	this	shop	dish	long-established
3	of	restaurant	dish	bakery	bakery
4	and	dish	guest	shop	homemade
5	a	guest	sweet	specialty	precinct
6	in	serve	temple	temple	confectionery
7	to	in	serve	guest	paste
8	this	sell	visitor	sweet	well-established
9	shop	sweet	sell	paste	lacquer
10	it	enjoy	ingredient	ingredient	blossom
11	restaurant	use	enjoy	flavor	sundry
12	have	item	flavor	long-established	bamboo
13	for	visitor	item	homemade	ware
14	can	feature	tea	rice	leek
15	with	of	rice	seasonal	dye
16	as	tea	locate	well-established	ceramic
17	by	temple	atmosphere	precinct	mackerel
18	that	various	feature	ware	dish
19	dish	can	bakery	dye	delicacy
20	guest	produce	cake	locate	seasonal
Average Frequency	792.4	396.7	213.9	152.1	53.1
Average Word Length	3.3	4.9	5.8	7.2	7.7

Top 300 Word Frequency Distribution Comparisons

We next examined the frequency distribution of the top 300 extracted words by using the BNC HFWL, which was divided into fourteen 1000-word frequency bands[7]. ‘BNC frequency band 1000’ indicates words ranked from 1 to 1000, meaning the words appearing in these bands are the 1st to 1000th most frequently appearing words in English. ‘Frequency band 2000’ corresponds to a ranking of 1001 to 2000, etc. The percentages of the 300 words of each list that belong to each frequency band are shown in **Table 3**. Highlighted scores show the BNC frequency bands at which the top 60% of the extracted words belong[8].

We can see the clear graduations of frequency levels. Looking across the table from *Freq* to *MI*, the top 300 words belong to increasingly lower frequency bands. About 60% of the top 300 words from *Freq* belong to the top 1000 BNC band and about 60% of the *CSM* words belong to the top 2000 BNC words. About 60% of the *LLR* and *Yates* words belong to the top 5000 or the top 6000 BNC bands. Uniquely, *MI* extracts words evenly from all the frequency bands of BNC HFWL words and about 60% belong to the top 9000 BNC words. This table provides a graphic illustration of how frequently or infrequently these extracted words are used in English. We can also see that the five statistical measures clearly extract different outstanding levels of tourism words.

Table 3 Frequency Distribution of the Top 300 Extractions

BNC Frequency	*Freq*	*CSM*	*LLR*	*Yates*	*MI*
1,000	60.7	39.0	22.3	16.0	4.3
2,000	14.0	20.3	17.0	13.3	4.3
3,000	7.3	10.3	10.7	10.0	6.0
4,000	4.7	8.0	9.3	8.7	7.3
5,000	2.7	4.7	7.7	8.0	7.7
6,000	2.0	3.0	5.7	6.0	7.7
7,000	1.3	2.3	5.7	6.7	7.3
8,000	1.7	2.7	3.3	5.0	8.0
9,000	2.3	3.3	6.0	7.0	10.7
10,000	1.0	2.7	4.7	6.3	9.3
11,000	1.0	1.7	2.0	4.0	7.7
12,000	0.7	0.7	2.3	3.7	5.0
13,000	0.7	0.7	1.3	2.0	6.3
13,994	0.0	0.7	2.0	3.3	8.3

Top 300 Word Grade Level Comparisons

Next we investigated at what US grade level the top 300 words would be understood by native English speaking children. *The Living Word Vocabulary* (Dale and O'Rourke, 1981:vii) is "an inventory of the written words known by children and young people in grades 4, 6, 8, 10, 12, 13, and 16." To supplement grades 1 through 3, we used reading grade word familiarity levels from Harris and Jacobson (1972). **Figure 1** shows at what grade level the majority of native speaking students (90%) would readily understand the central meaning of each word for the top 300 extractions produced by the statistical measures. Note that in **Figure 1**, 'N/A' denotes those words not appearing in either of the two resources.

In looking at the horizontal line and corresponding grade levels for each bar in the graph, we can see that 90% of the top 300 words from *Freq* are understood by 6th grade level students; those of *CSM* are understood by 8th grade students; those of *LLR* and *Yates* are known by 10th and 12th grade students respectively; those of *MI* are known by 13th grade students (college freshmen). It is interesting that similar results were obtained in a study using these measures to extract science vocabulary from a BNC science component (see Chujo and Utiyama, 2005), and these similar results support the effectiveness of using these measures to extract domain and level specific vocabulary.

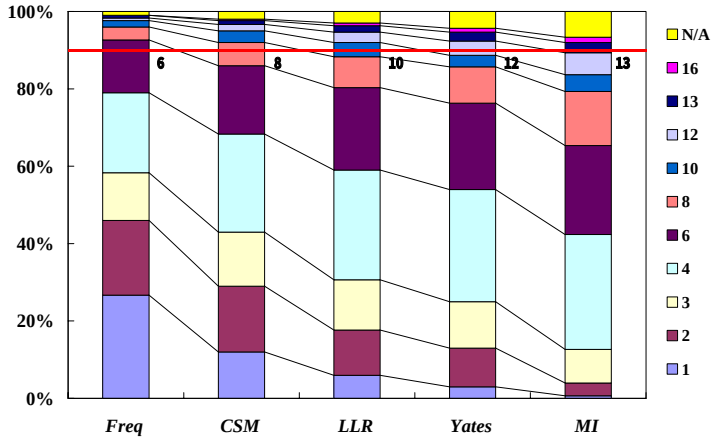

Figure 1 Word Familiarity Based on US Grade Level

Japanese High School Textbook Vocabulary Coverage

As educators in Japan, our primary interest is in how this extracted tourism vocabulary compares to what Japanese students may or may not already have studied and therefore how useful these lists might be. To do this, we compared the top 300 extractions to the vocabulary representing what most high school students have studied before entering university. This list, comprised of 3,245 different words, was compiled from the top selling series of junior and senior high school (JSH) textbooks in Japan from the 7th through 12th grades. **Figure 2** shows both what percentage of the top 300 extractions do appear in JSH textbooks in the lower section of the graph, and what percentage do not appear, in the upper section of the graph. We see from **Figure 2** that 82 percent of the *Freq* top 300 extractions are covered in the JSH school textbooks; 70 percent of the *CSM* extractions; 57 percent of the *LLR*, 50 percent of the *Yates*, and 32 percent of the *MI* extractions are covered in the JSH school textbooks. The data in **Figure 2** again verifies that the five different statistical measures extract quite different levels of words and provides us important information on which of the top 300 words might be appropriate for learners.

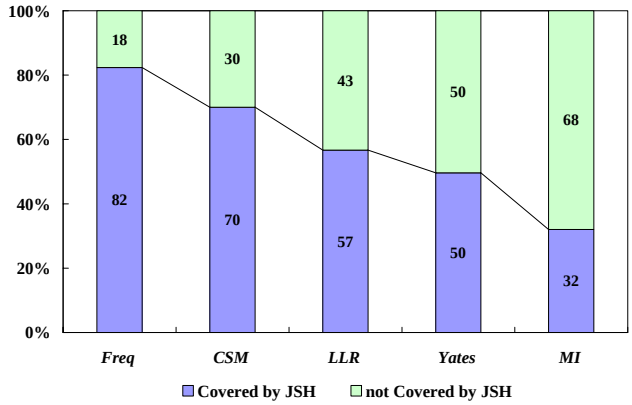

Figure 2 Percentage of Top 300 Words Covered/Not Covered by the JSH Textbook

Developing Level-Specific Tourism Lists

The results of these analyses support the finding that certain statistical measures tend to extract tourism vocabulary corresponding to a certain grade level. In terms of practical pedagogical application, we inferred from frequency distribution, grade level familiarity data, and JSH text coverage data in this current study, in addition to several similar previous studies (Chujo and Utiyama, 2004; 2005; 2006), that (1) the tourism words extracted by *Freq* and *CSM* might be most useful for beginner level Japanese EFL learners; (2) the *LLR* and *Yates* lists might be most useful for intermediate level Japanese EFL learners; and (3) the *MI* vocabulary might be most appropriate for advanced level Japanese EFL learners.

In order to aid retention, it would be more convenient for educators if the vocabulary were classified according to subtopics. The original Kyoto tourism corpus consists of four components relating to the target activities: 'miru' (sight-seeing), 'kau' (shopping), 'taberu' (dining), and 'taikensuru' (hands-on activities). We also developed similar level-specific tourism lists from these components using the same five statistical measures and created 20 sub-lists (5 measures x 4 components), whose vocabulary levels were similarly verified. The top 50 words from the advanced *MI* list are shown in **Table 4**.

In **Table 4**, we can easily visualize a wide range of tourism topics for the four sub-sections. For example, looking at the 'miru' list, we can enjoy seasonal changes of landscapes, particularly cherry blossoms in spring, lush greenery in early summer and crimson maple foliage in autumn while strolling temples and worshiping deities at shrines. In looking at the 'taberu' section, our appetite is stimulated by delicacies which embody the traditional culinary culture of Kyoto, and the authentic Kyoto-style dishes featuring tofu, seasonal ingredients, mackerel sushi, and homemade soybean paste served at long-established restaurants. Keep in mind that these are only 50 of the top 300 words for each list (beginner and intermediate are not shown), and that the list provides a manageable resource which can be further refined by educators and students.

CONCLUSION

In this study, tourism words were selected from a Kyoto tourism corpus using five statistical measures. An examination of the resulting vocabulary lists show that each statistical measure extracted an appropriate level of tourism words by its vocabulary level, grade level, and school textbook vocabulary coverage. They were classified into three proficiency level lists, which allow users to further refine the candidates based on their own appropriate contexts and level. Further research will include developing these lists into e-learning materials for vocabulary building.

Although the vocabulary lists produced by this study do not provide a definitive, comprehensive lexicon for all the sub-divisions in tourism (hotel, restaurant, tours, marketing, etc.), it has been possible to create level-specific, broad-based vocabulary lists not only useful for university teachers and students majoring tourism, but also useful for general students who are interested in developing skills to express Japan's culture in English. Since a Kyoto corpus was used, many of these key words are common to Japan's traditional culture and are therefore particularly applicable to tourism in Japan.

Table 4 Advanced Level Domain-Specific Tourism Words

miru	kau	taberu	taikensuru
precinct	bakery	homemade	dye
blossom	specialty	specialty	potter
shrine	long-established	long-established	ware
temple	lacquer	leek	ceramic
maple	sundry	mackerel	cedar
greenery	confectionery	dish	lacquer
deity	bamboo	restaurant	meditation
foliage	well-established	paste	sash
stroll	ware	delicacy	indigo
enshrine	paste	stew	engraving
worshipper	ceramic	seafood	precinct
scenic	incense	cafe	temple
repose	footwear	well-established	incense
relocate	secondhand	fluffy	sermon
crimson	homemade	culinary	bamboo
sundry	accessory	seasonal	kiln
panoramic	fragrance	authentic	artisan
popularly	sweetness	ingredient	unparalleled
confectionery	handmade	confectionery	blossom
bamboo	high-quality	entertainer	aromatic
spacious	adjoining	curry	napkin
thrilling	cherish	rice	visitor
erect	sash	grill	confectionery
inscribe	flavor	guest	memento
open-air	indigo	chef	brighten
cherry	manually	blossom	well-established
visitor	soy	flavor	handkerchief
moss	entertainer	sweetness	affordable
bloom	fluffy	fragrance	doll
ceramic	dye	gluten	wholesaler
entertainer	additive	teatime	alcove
extant	sweet	dessert	burner
fame	souvenir	manually	handmade
birthplace	seasonal	eel	bracelet
promenade	teatime	pastime	goblin
seclude	gluten	affordable	adjoin
pavilion	boutique	edible	souvenir
crystallization	grill	sauce	sect
tranquil	showcase	curd	vase
curd	ingredient	gourmet	foil
gourmet	comb	broth	purification
backdrop	bread	vegetarian	reservation
potter	yeast	meticulous	seedling
observatory	artisan	soy	beginner
boutique	found	cherish	paraffin
pastime	affordable	precinct	regrettable
seasonal	shop	octopus	enshrine
masterpiece	rice	abundant	culinary
rustle	bracken	locate	manually
annex	cosmetic	crystallization	accessory

Notes

1 Tourism statistics: http://www.tourism.jp/english/

2 Action Plan for Tourism Development:
http://www.kantei.go.jp/jp/singi/kanko2/kettei/030731/keikaku.pdf

3 As of 2005, there are 25 universities that have departments of tourism in Japan.

4 Claws7: http://www.comp.lancs.ac.uk/computing/users/eiamjw/claws/claws7.html.

5 References for each measure are as follows: *CSM* (Wakaki and Hagita, 1996), *LLR* (Dunning, 1993), *Yates* (Hisamitsu and Niwa, 2001), and *MI* (Church and Hanks, 1989).

6 The formulas are available at http://www2.nict.go.jp/jt/a132/members/mutiyama.

7 From our previous studies, we can assume that the top 300 words (about 10% of the master word list) provide a clear illustration of each measure's unique extractions. For a more detailed discussion, see Utiyama, et al. (2004).

8 The '60%' level was chosen arbitrarily in order to easily observe the graduations of frequency levels. An '80%' level will provide a similar graphic illustration as shown in **Table 3**.

REFERENCES

Agency for Cultural Affairs. (2005). My journey: Best 100. Retrieved April 5, 2006 from http://www.bunka.go.jp/1tabi/pdf/tabi2005pamphlet.pdf

Baude, A., Iglesias, M., & Inesta, A. (2006). *Ready to order.* NY: Pearson Publishers.

Briggs, S. & Lee, D. (2002, November). *Developing a lexical database of academic spoken English (LDASE) for language testing: Problems and prospects.* Paper Presented at the 4th North American Symposium on Corpus Linguistics and Language Teaching, Indianapolis, Indiana.

Chujo, K. (2004). Measuring vocabulary levels of English textbooks and tests using a BNC lemmatized high frequency word list. In J. Nakamura, N. Inoue, & T. Tabata (Eds.), *English Corpora under Japanese Eyes* (pp. 231-249). Amsterdam: Rodopi.

Chujo, K. & Utiyama, M. (2004). Toukeiteki shihyou wo shiyoushita tokuchougo chuushutsu ni kannsuru kenkyuu [Using statistical measures to extract specialized vocabulary from a corpus]. *KATE Bulletin*, *18*, 99-108.

——— (2005). Selecting level-specific BNC applied science vocabulary using statistical measures. *Selected Papers from the Fourteenth International Symposium on English Teaching*, 195-202.

——— (forthcoming 2006). Selecting level-specific specialized vocabulary using statistical measures. *SYSTEM, 34*, 2.

Chujo, K., Utiyama, M., & Nishigaki, C. (forthcoming 2006). Towards building a usable corpus collection for the ELT classroom. In E. Hidalgo, L. Quereda, & J. Santana (Eds.), *Corpora in the Foreign Language Classroom*. Amsterdam: Rodopi.

Church, K. & Hanks, P. (1989). Word association norms, mutual information, and lexicography. *Proceedings of ACL-89*, 76-83.

Dale, E. & O'Rourke, J. (1981). *The living word vocabulary*. Chicago: World Book-Childcraft International, Inc.

Dantsuji, M. (2001). Multimedia Eigo CALL kyouzai: Introduction to the beauties of Kyoto [Using multimedia English CALL: Introduction to the beauties of Kyoto]. In S. Sakamoto (Ed.), *Koutou kyouiku kaikaku ni shisuru multimedia no koudo riyou ni kansuru kenkyu [A Study on high-grade multimedia use conducive to the reform of higher education*], 119-124.

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

Harris, A. J. & Jacobson, M. D. (1972). *Basic elementary reading vocabularies*. New York: The Macmilan Company.

Hisamitsu, T. & Niwa, Y. (2001). Topic-word selection based on combinatorial probability. *NLPRS-2001,* 289-296.

Ichikawa, Y., Yasuyoshi, I., & Hestand, J.R., et al. (2002). *Unicorn English Course I & II*. Tokyo: Bun'eido.

Ichikawa, Y., Yasuyoshi, I., & Hestand, J.R., et al. (2003). *Unicorn English Reading*. Tokyo: Bun'eido.

Ichimura, Y. (2004). Edogawa daigakkusei ga erabu gaikokujin ni shoukaishitai Nihon no toshi, kankou meisho [Cities and tourist attractions appropriate for introducing to overseas visitors, chosen by Edogawa University students]. Retrieved February 19, 2006 from http://www.edogawa-u.ac.jp/~ichi/zemi2004.htm

Kamio, S. (2005, July). *Inbound tourism ni okeru marketing senryaku [A marketing strategy for inbound tourism]*. The 14th Workshop Presented at the Graduate School for Cities, Asian Business Studies, Osaka City University. Retrieved April 4, 2006 from http://www.gscc-asianbusiness.jp/workshop.html

Kasajima, J., Asano, H., & Shimomura, Y., et al. (2002). *New Horizon English Course 1, 2, & 3*. Tokyo: Tokyo Shoseki.

Lam, P. (2004, July). *A corpus-driven lexical analysis of English tourism industry texts and the study of its pedagogic implications in English for specific purposes.* Paper Presented at the Sixth Teaching and Language Corpora Conference, Granada, Spain.

METI Kansai (Kansai Bureau of Economy, Trade and Industry). (2004). Kansai Internationalization Data File 2004. Retrieved April 4, 2006 from http://www.kansai.meti.go.jp/3-1kokusai/index2004.htm

Mukaiyama, H. (2003). Nihon kakuchi no gaikokujin ryokousha no tokuchou ya ryokou no jittai ga akirakani [Review of overseas visitors to Japan] JNTO Houdou Shiryou, Retrieved January 13, 2003 from www.jnto.go.jp/info/pdfs/0131chousa.pdf

Nation, I. S. P., (1994). *New ways in teaching vocabulary*. Virginia: TESOL, Inc.

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

Robinson, P. (1991). *ESP today: A practitioner's guide*. New York: Prentice Hall.

Scott, M. (1997). PC analysis of key words and key key words. *System, 25,* 2, 233-245.

——— (1999). WordSmith tools manual [Computer software]. Oxford: Oxford University Press.

Utiyama, M., Chujo, K., Yamamoto, E. & Isahara, H. (2004). Eigokyouiku no tameno bunya tokuchou tango no sentei shakudo no hikaku [A comparison of measures for extracting domain-specific lexicons for English education]. *Journal of Natural Language Processing, 11*, 3, 165-197.

Wakaki, M. & Hagita, N. (1996). Recognition of degraded machine-printed characters using a complementary similarity measure and error-correction learning. *IEICE Trans. Inf. & Syst*. E79-D, 5.

Walker, R. (1995). Teaching the English of tourism. *IATEFL ESP SIG Newsletter,* No.4. Retrieved March 27, 2006 from http:// www.unav.es/espSig/walker4.htm

Wood, N. (2003). *Tourism and catering*. London: Oxford Publishers.

Wyatt, R. (2006). *Check your English vocabulary for leisure, travel, and tourism*. London: A&C Black Publishers.