A Self-Referring Quantitative Evaluation of the ATR Basic Travel Expression Corpus (BTEC)

Kyo Kageura† and Genichiro Kikui‡

† Graduate School of Education, University of Tokyo,
7–3–1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan.
kyo@p.u-tokyo.ac.jp

‡ ATR Spoken Language Translation Research Laboratories,
2–2–2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan.
genichiro.kikui@atr.jp

Abstract

In this paper we evaluate the Basic Travel Expression Corpus (BTEC), developed by ATR (Advanced Telecommunication Research Laboratory), Japan. BTEC was specifically developed as a wide-coverage, consistent corpus containing basic Japanese travel expressions with English counterparts, for the purpose of providing basic data for the development of high quality speech translation systems. To evaluate the corpus, we introduce a quantitative method for evaluating the sufficiency of qualitatively well-defined corpora, on the basis of LNRE methods that can estimate the potential growth patterns of various sparse data by fitting various skewed distributions such as the Zipfian group of distributions, lognormal distribution, and inverse Gauss-Poisson distribution to them. The analyses show the coverage of lexical items of BTEC vis-à-vis the possible targets implicitly defined by the corpus itself, and thus provides basic insights into strategies for enhancing BTEC in future.

1. Introduction

In this paper we present a quantitative method for evaluating the sufficiency of qualitatively well-defined corpora, and evaluate the Japanese part of the Basic Travel Expression Corpus (BTEC), developed by ATR (Advanced Telecommunication Research Laboratory), Japan. BTEC is a wide-coverage, qualitatively well-examined and consistent corpus that contains basic travel expressions with strong orientation toward real world conversation. It was constructed in order to provide basic data for developing high quality speech translation systems (Takezawa, 1999; Takezawa, et. al. 2002), and was used as a basic corpus for the development of speech translation systems at ATR.

Although BTEC has been evaluated as basic data for developing speech translation systems (for instance, Kikui, et. al. (2003) observe the cross-perplexity of BTEC and other corpora), the position of the corpus within the domain of travel conversations has not been examined so far and such an examination is long overdue. As a first step in this direction, we defined a self-referring method for evaluating the status of the corpus quantitatively, and observed some quantitative characteristics of lexical items in BTEC.

2. Basic Framework of Evaluation

Though many evaluations of various aspects of corpora have been reported (cf. papers in LREC conferences; Atwell, et. al., 2000; Han, et. al., 2002; Paik, et. al., 2005), there exists no standard procedure for evaluating corpora because different corpora constructed for different aims need to be evaluated from different points of view.

As the main aim of BTEC is to collect basic travel expressions for the development of high quality speech translation systems, the following points of evaluation are especially important:

(1) Sufficiency of the corpus vis-à-vis the domain of basic travel expressions;

(2) Sufficiency of the corpus for developing application systems (e.g. sufficiency of the corpus as training data for machine learning, etc.).

In the present study we focus on (1), because (a) it is necessary to obtain insight into the status of the corpus in terms of the varieties of real-world travel expressions in order to make strategic decisions about how to extend the corpus, and (b) (2) has been examined more than (1), in relation to the development of application systems.

Ideally, the current status of the corpus should be evaluated with respect to the overall range of actual and potential language expressions in the domain of travel conversations. This, however, is unfortunately not possible, because, in its bare state, any real-world language phenomenon is by definition potentially infinite and changes over time. However, there is a way of measuring the current status of a corpus, assuming that it is qualitatively well-designed vis-à-vis the idiosynchronic state of language. To the extent that a corpus qualitatively reflects what is assumed to be the proper characteristics of the whole range of relevant language phenomena in a given domain, the corpus itself can provide basic insights into the potential range of relevant language expressions in the domain, at least for the quantitative evaluation of the corpus.

The basic idea of self-referring quantitative evaluation of a corpus is simple:

(1) Extrapolate the corpus size to infinity and estimate the point of saturation of given points of observation; this can be used as the "model" of the observed language phenomenon within the domain that the corpus addresses;

(2) Evaluate the current status of the corpus vis-à-vis that saturated "model" point.

It should be emphasised again that, for the self-referring evaluation to be meaningful, we must assume that the corpus properly covers what it should cover within the practical limitations of its size. As BTEC is a well-designed

and well-planned corpus with the clear aim of providing basic travel expressions to be used in application development, we can reasonably assume that the self-referring method of quantitative evaluation of the corpus is applicable to BTEC. The self-referring method also requires that the point of evaluation saturates; this is a question to be checked empirically for each point of observation in each corpus.

3. Points of Observation

3.1. Elements to be Observed

As BTEC aims to cover basic expressions in travel conversation, it is important to pay attention not only to the formal aspects such as grammatical patterns or bigram POS patterns, but also to lexical and expressional coverage vis-à-vis potentially needed conversational variations. Among these, we focus on the evaluation of the lexical sufficiency of the content words of BTEC, because grammatical patterns have been addressed in the process of application development. More concretely, we observe the following classes of items:

- all lexical items
- nouns
- verbs
- adjectives

Nouns, verbs and adjectives are chosen as they are the core open class items.

3.2. Expected Number of Items

For these classes, we first estimate the population number of types using BTEC and then evaluate the quantitative status of BTEC itself with respect to the estimated population number of types for each class. The estimation of population item size is a traditional and much addressed problem in the field of quantitative linguistics (Efron & Thisted, 1976; Mizutani, 1983; Tuldava, 1995; Baayen, 2001) as well as in theoretical statistics (Shibuya, 2003). Here we adopt the LNRE methods described in detail in Baayen (2001), as they offer flexible models for estimation.

Let the population number of types be S, and for each type e_i ($i = 1, 2, ...S$) let the probability be p_i. Assuming the combination of binomial distribution and its Poisson approximation, E[$V(N)$], the expected number of types in a sample of size N, and E[$V(m, N)$], the expected number of types that occur m times in a sample of size N are given as follows:

$$E[V(N)] = S - \sum_{i=1}^{S}(1-p_i)^N = \sum_{i=1}^{S}(1-e^{-Np_i}). \quad (1)$$

$$\begin{aligned} E[V(m,N)] &= \sum_{i=1}^{S}\binom{N}{m}p_i^m(1-p_i)^{N-m} \\ &= \sum_{i=1}^{S}(Np_i)^m e^{-Np_i}/m!. \end{aligned} \quad (2)$$

Grouping the items by p_i, we can define the structural type distribution as:

$$G(p) = \sum_{i=1}^{S} I_{[p_i \geq p]},$$

where $I = 1$ when $p_i \geq p$ and 0 otherwise. $G(p)$ thus gives the cumulated number of types with probabilities equal to or greater than p.

Using $G(p)$, the equations (1) and (2) can be rewritten as:

$$E[V(N)] = \int_0^\infty (1-e^{-Np})\, dG(p). \quad (3)$$

$$E[V(m,N)] = \int_0^\infty (Np)^m e^{-Np}/m!\, dG(p). \quad (4)$$

We renumber the subscript of p in the ascending order of p for $p_j \neq 0$. As $G(p)$ is a step function, $dG(p) = G(p_j) - G(p_{j+1})$ where p_j and $dG(p) = 0$ otherwise.

Using some explicit distributions such as inverse Gauss-Poisson distribution, etc., we can estimate $V(N)$ and $V(m, N)$ for $N \to \infty$. It should be noted, however, that the parameters of the explicit distributions themselves depend on the sample size, so we should assume a sample size Z, where $G(p)$ can fit, as a parameter. We can now give the estimation of, for instance, the spectrum element as:

$$E[V(m,N)] = \int_0^\infty \frac{-(\frac{N}{Z}(Zp))^m)}{m!} e^{-\frac{N}{Z}(Zp)}\, dG(p)$$

3.3. Growth Rate of Items

The equation (1) above gives the expected number of items in the sample of size N. As $1-p_i < 1$, the number of item types that occur in the sample increases in accordance with N. The equation (1) thus expresses the growth curve of the number of items $V(N)$ in accordance with N. The first derivative of E[$V(N)$] therefore expresses the growth rate of the items, which gives the probability that unseen events are observed at that point.

To obtain the growth rate, we first rewrite the equation (1) as:

$$\begin{aligned} E[V(N)] &= S - \sum_{i=1}^{S} e^{-Np_i} \\ &= \sum_{i=1}^{S}(1-e^{-Np_i}) \\ &= \sum_{j}(1-e^{-Np_j})\Delta G(p_j) \\ &= \int_0^\infty (1-e^{-Np})dG(p), \end{aligned} \quad (5)$$

Taking the derivative of this, we obtain:

$$\begin{aligned} \frac{d}{dN}E[V(N)] &= \frac{d}{dN}\int_0^\infty (1-e^{-Np})\, dG(p) \\ &= \int_0^\infty -p \cdot -e^{-Np}\, dG(p) \end{aligned}$$

$$= \frac{1}{N}\int_0^\infty Npe^{-Np}\,dG(p)$$
$$= \frac{E[V(1,N)]}{N} \quad (6)$$

Incidentally, this equals the estimation of unseen events given by Good (1953).

4. Evaluation of BTEC

4.1. Basic Characteristics of BTEC

Table 1 shows the basic quantitative nature of the Japanese section of BTEC.

	♯sentence	♯all-words	♯nouns	♯verbs	♯adjs
Token	512,111	3,539,243	639,215	423,240	77,448
(%)	—	100%	18.1%	12.0%	2.2%
Type	312,757	36,031	19,139	8,854	1,198
(%)	—	10%0	53.1%	24.6%	3.3%

Table 1. Basic quantitative characteristics of BTEC.

For comparison, the ratio of nouns, verbs and adjectives given in National Language Research Institute (1958), which surveyed various cultural magazines, and in National Language Research Institute (1989), which surveyed geography textbooks for junior high school and high school, are given in Table 2 (HS = high school; JHS = junior high school). Although the data are either not new or not general enough, they are sufficient for our immediate aim of making a rough comparison in order to emphasise the unique characteristics of BTEC.

		nouns	verbs	adjectives
Token	Magazines	65.4%	22.6%	2.4%
	HS textbooks	57.2%	29.1%	3.0%
	JHS textbooks	43.8%	30.1%	5.0%
Type	Magazines	83.7%	11.6%	1.5%
	HS textbooks	78.5%	13.5%	1.3%
	JHS textbooks	71.6%	19.3%	2.3%

Table 2. Nouns, verbs and adjectives in other corpora.

Comparing Tables 1 and 2 immediately reveals that the ratio of nouns both tokenwise and typewise in BTEC is notably lower than that in the other data. The ratio of verbs in token in BTEC is also much lower than in the other data. Also notable is the type-token ratio. For BTEC, nouns that amount to more than 50% of the total number of types only cover 18% of the total number of tokens. Verbs also show the same tendency. These characteristics suggest that, in BTEC, expressions resort to repetitions of functional and related elements rather than content elements, and content words are not used as repeatedly as in the other corpora shown in Table 2. Although we do not intend to delve deeper into this discussion, these characteristics may well be a reflection of the particular way in which BTEC was constructed.

4.2. Population Types and Coverage Ratio

Table 3 gives the estimation of the population number of item types for each lexical class. It also shows the present coverage ratio (CR) of BTEC, as well as the best models used in the LNRE estimation and their X^2 values. IGP refers to the inverse Gauss-Poisson model, and GIGP refers to the generalised inverse Gauss-Poisson model. The multi-dimensional X^2 test values are not bad compared to, for instance, the data given in Baayen (2001) or Kageura (1998), so we can reasonably rely on the LNRE estimations.

	E[S]	$V(N)$	CR (%)	Model	X^2
All	56091	36031	64.24	IGP	176.78
Nouns	29483	19139	64.92	GIGP	149.40
Verbs	15964	8854	55.46	IGP	48.56
Adjectives	1945	1198	61.59	GIGP	48.54

Table 3. Population types E[S] and coverages CR.

The figures in Table 3 show that, assuming that BTEC is properly constructed according to its stated aim of collecting basic travel expressions, the necessary vocabulary for basic travel expressions in Japanese must be around 56,000 items, of which nouns constitute approximately 29,500 items, verbs approximately 16,000 items and adjectives approximately 2,000 items.

With respect to these saturation points, about 65% of nouns, 55% of verbs, and 60% of adjectives are covered by BTEC at present. The low coverage ratio of verbs and to some extent adjectives is notable here. If we think of a general domain of language expressions, we can reasonably expect that the coverage ratio of nouns might well be lower than that of verbs and adjectives, as the overall number of nouns are expected to be much larger than those of verbs and adjectives. In BTEC, however, the coverage ratio of verbs is lower than that of nouns, and the coverage ratios of nouns and adjectives are not very different.

4.3. Transition Patterns

So how is BTEC likely to be evaluated if the corpus were to be enlarged along the same lines it is currently constructed?

Figure 1 shows the growth curve of all the words (top-left), nouns (top-right), verbs (bottom-left) and adjectives (bottom-right), up to twice the original corpus size. The circles show the observed values and the lines show the estimated values. Figure 1 also shows the growth curve of hapax legomena, as they are the key to observing the growth rate. It appears that the number of verbs, and, to some extent, adjectives will increase rather constantly, while for all-words and nouns the growth curves become gentler beyond the current corpus size. Correspondingly, the curve of hapax legomena shows that the growth of hapaxes is flattened out for all words and nouns, while for verbs and adjectives the flattening out is not so obvious (the observed values for adjectives are not as flat as the estimated values).

To further the examination, it is necessary to examine the growth rate. Table 4 shows the growth rates of all-words, nouns, verbs and adjectives at BTEC's current size. Currently, 340 tokens should be added to obtain a new type, 122 noun tokens should be added to obtain a new noun

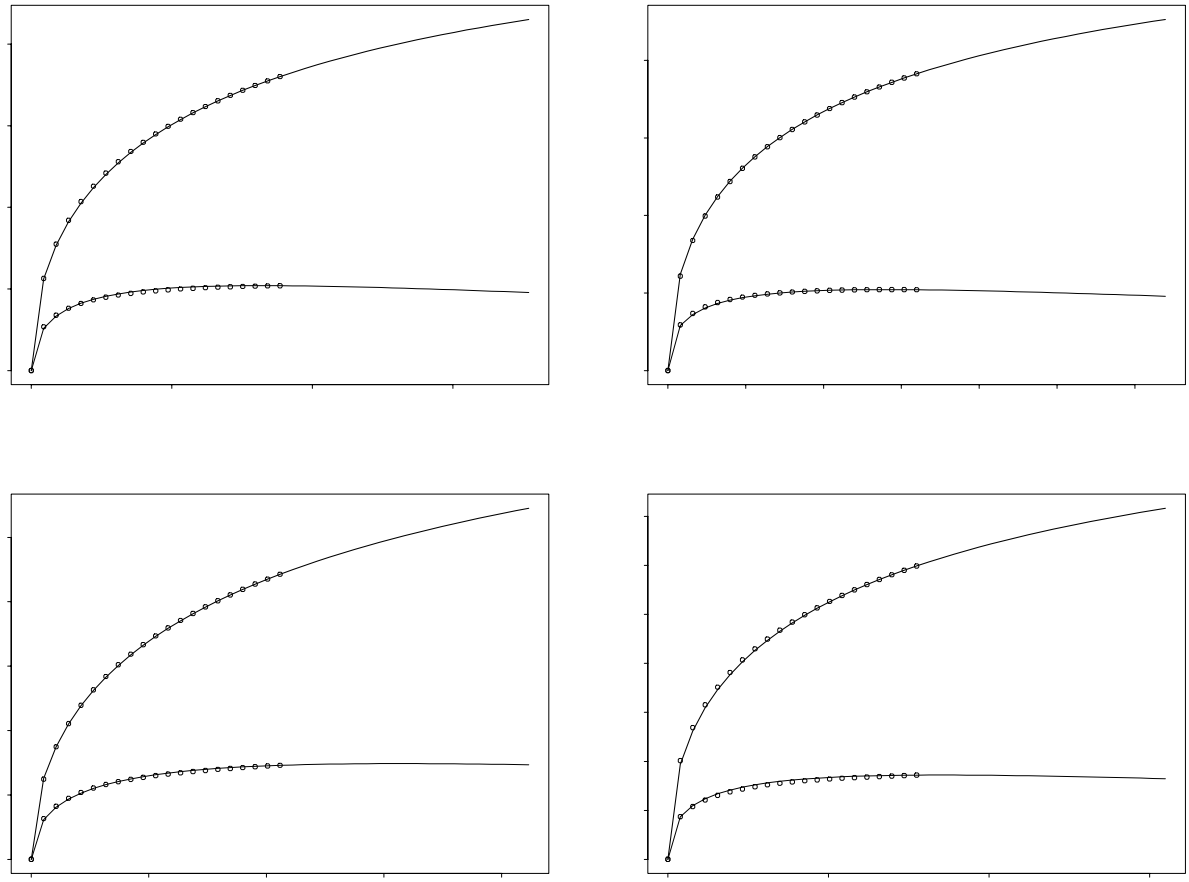

Figure 1. Growth curve of lexical items in BTEC corpus.

type, 144 verb tokens should be added to obtain a new verb type, and 227 adjective tokens should be added to obtain a new adjective type. Alternatively, the existing items should be repeated these numbers of times.

	all-words	nouns	verbs	adjs
Growth rate	0.00294	0.00815	0.00690	0.00444

Table 4. Growth ratio of BTEC.

Figure 2 traces the transitions in the growth rate up to twice the original BTEC size. N means the BTEC size, and alpha (0.0–2.0) indicates the relative size to the original N. The lines show, moving from top to bottom, nouns, verbs, adjectives and all-words. The left-hand panel of Figure 2 shows the overall pattern, and the right-hand panel limits the observation range to 0.0–0.03. Although the right-hand panel shows that we can still observe changes in the growth rate in accordance with the corpus size if we observe the transitions up close, the left-hand panel shows that the current growth rates are in fact very close to zero within the overall transition patterns. If we wish to enrich the vocabulary in the corpus, therefore, it may not be a good idea to try to enlarge BTEC in accordance with the same policy used to construct it.

Let us finally examine the changes in the coverage ratio, which is particularly important when planning the improvement and/or enlargement of a corpus. Table 5 and Figure 3 show the transitions in the coverage ratio of all-words, nouns, verbs and adjectives up to twice the original sample size.

token size	♯all-words	♯nouns	♯verbs	♯adjs
0.5N	51.55%	52.75%	43.35%	49.62%
N	64.24%	64.92%	55.46%	61.59%
1.5N	71.67%	71.99%	62.97%	68.75%
2N	76.72%	76.80%	68.31%	73.71%

Table 5. Changes in the coverage ratio.

Table 5 and Figure 3 show results that conform to the analyses so far, i.e. even if the corpus size is extended to twice its current size, the coverage of all-words, nouns, verbs and adjectives increases by only about 12 to 13 per cent, in other words, the coverage becomes only about 70 to 75 per cent. This confirms the case made above that it may not be a sensible idea to try to enlarge BTEC in accordance with the current policy if what we wish to achieve is high lexical coverage in the domain of travel expressions. For

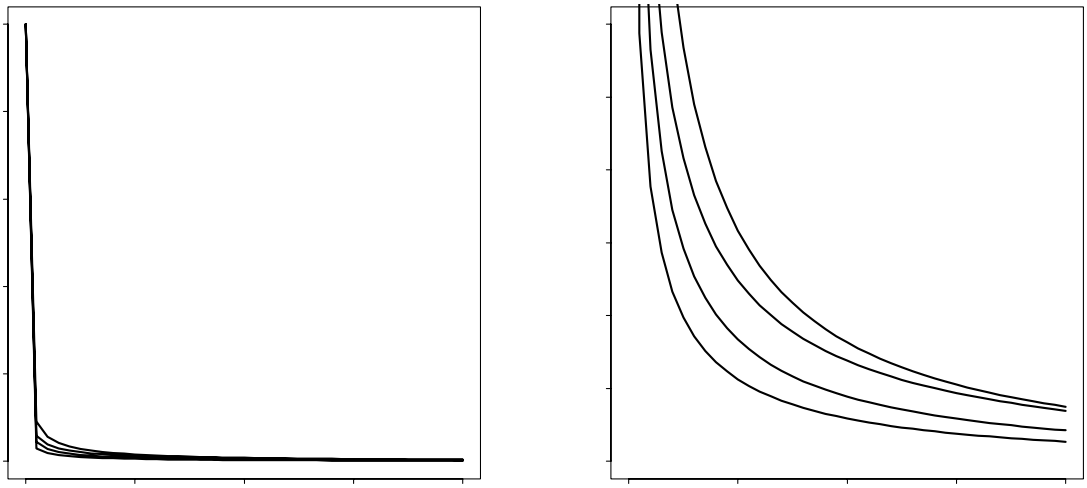

Figure 2. Growth rate of lexical items in BTEC.

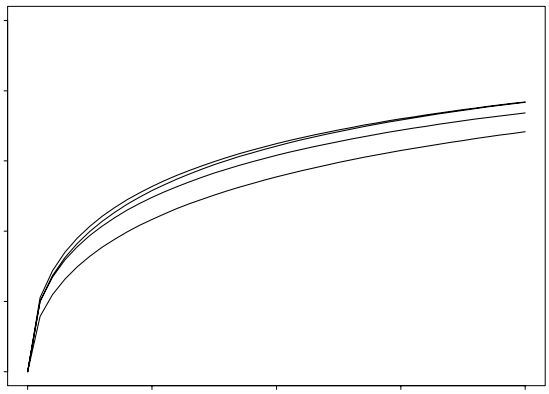

Figure 3. Coverage ratio of lexical items.

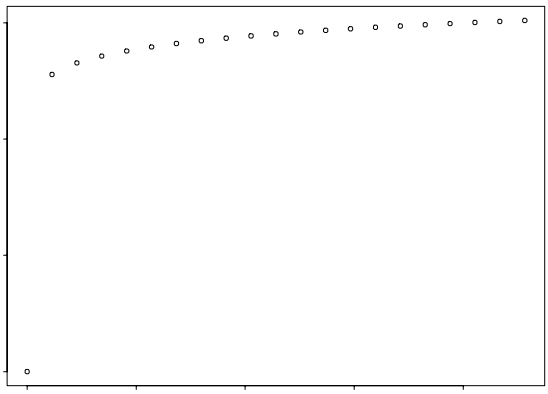

Figure 4. Transitions in POS bigrams

instance, assuming for simplicity that the growth of coverage does not flatten out further (which is far from the reality), it is still necessary to more than triple the corpus size in order to achieve coverage of 90% for overall lexical items.

Whether this is worthwhile attempting or whether other methods for enriching the lexical coverage should be adopted depends on additional factors concerning the corpus. One such factor is the coverage of grammatical or formal patterns. To obtain a rough sense of this coverage, we analysed the growth patterns of POS bigrams (151 types in the corpus) as a rough approximation to the status of grammatical patterns in the corpus, by adopting the second level POS categories of ChaSen. Figure 4 shows the transition in the number of bigram patterns up to BTEC's current size.

As can be expected from the small number of patterns (151), the growth almost flattens out at BTEC's current size. In fact, only 5 new bigram patterns occur when the corpus size is extended from $0.5N$ to N. This suggests that covering a greater variety of grammatical patterns would not be a strong driving force for enlarging BTEC in accordance with the current construction policy, either.

4.4. Summary of Observations

We have so far depicted the lexical status of BTEC using the following methods:

(1) Estimating the population number of all-words, nouns, verbs and adjectives. From this we obtained the concrete target of the number of lexical items, i.e. 56,000 words, with 29,500 nouns, 16,000 verbs and 2,000 adjectives under the current corpus policy.

(2) Calculating the growth curve, growth rate and coverage ratio as well as their transitions for all-words, nouns, verbs and adjectives in BTEC vis-à-vis the population characteristics estimated in (1). We ob-

served that at present BTEC covers 64% of the estimated population for all words, 65% for nouns, 55% for verbs and 62% for adjectives. We also demonstrated that enlarging the corpus to twice its current size contributes to only about a 12% to 13% coverage increase, if we assume that the enlargement is done in accordance with the current BTEC construction policy.

For secondary reference, we also checked the status of grammatical patterns by means of POS bigrams, which suggests the rough saturation of grammatical patterns.

5. Conclusions

By introducing a self-referring method of quantitatively evaluating qualitatively well-designed corpora, we have characterised the current status of the Japanese section of BTEC. With respect to the methodology, the self-referring method crucially depends on the qualitative well-foundedness of the corpus. If we introduce some sound viewpoints for comparing different corpora using this method, this condition of qualitative well-foundedness on the side of the corpora could be loosened. We would like to examine to what extent this can be done in the next stage, by using other corpora for comparison.

With respect to BTEC, we have only dealt with the Japanese section; examination of the other language sections of BTEC by the same methodology reported here is a task for the immediate future.

In terms of actual examination of the corpus, the observations made in this paper provide some concrete quantitative information that can be used when planning the strategic enhancement and/or enlargement of BTEC. For this information to be properly used in corpus extension, however, the basic strategic framework should be defined in advance, e.g. how much coverage of lexical items is needed; for what purpose are they to be used; whether the preferable level of coverage could be achieved in the current corpus or could be dealt with by some other means, etc. The same information — an estimated 12% increase in coverage by doubling the corpus, for instance — might well lead to different conclusions if these external conditions vary. At this point, we need to go back to the qualitative consideration of corpus construction.

Acknowledgements

This research was supported in part by the National Institute of Information and Communication Technology (NiCT).

References

Atwell, E. et. al. (2000). A comparative evaluation of modern English corpus grammatical annotation schemes. ICAME Journal, 24, 7–24.

Baayen, R. H. (2001). Word Frequency Distributions. Dordrecht: Kluwer.

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? Biometrika, 63(4), 435–447.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. Biometrika, 40(3), 237–264.

Han, C-H. et. al. (2002). Development and evaluation of a Korean treebank and its application to NLP. Language and Information, 6(1), 123–138.

Kageura, K. (1998). A statistical analysis of morphemes in Japanese terminology. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (pp. 638–645).

Kikui, G., Sumita, E. Takezawa, T. and Yamamoto, S. (2003). Creating corpora for speech-to-speech translation. In Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH) (pp. 381–382).

Mizutani, S. (1983). Goi. Tokyo: Asakura.

National Language Research Institute (1958). Research on Vocabulary in Cultural Reviews. Tokyo: National Language Research Institute.

National Language Research Institute (1989). Research on Vocabulary in School Textbooks. Tokyo: National Language Research Institute.

Paik, K. et. al. (2005). Quantitative analysis of corpora with different source languages. Journal of Natural Language Processing, 12(4), 117–136.

Shibuya, M. (2003). Number of categories with a singleton in sample and population. Proceedings of the Institute of Statistical Mathematics, 51(2), 261–295.

Takezawa, T. (1999). Building a bilingual travel conversation database for speech translation research. In Proceedings of the 2nd International Workshop on East-Asian Language Resources and Evaluation — Oriental COCOSDA Workshop '99 (pp. 17–20).

Takezawa, T. et. al. (2002). Toward a broad-coverage bilingual corpus for speech translation of travel conversation in the real world. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (pp. 147–152).

Tuldava, J. (1995). Methods in Quantitative Linguistics. Trier: WVT Wissenschaftlicher Verlag.