A Statistical Analysis of Morphemes in Japanese Terminology

Kyo KAGEURA
National Center for Science Information Systems
3-29-1 Otsuka, Bunkyo-ku, Tokyo, 112-8640 Japan
E-Mail: kyo@rd.nacsis.ac.jp

Abstract

In this paper I will report the result of a quantitative analysis of the dynamics of the constituent elements of Japanese terminology. In Japanese technical terms, the linguistic contribution of morphemes greatly differ according to their types of origin. To analyse this aspect, a quantitative method is applied, which can properly characterise the dynamic nature of morphemes in terminology on the basis of a small sample.

1 Introduction

In computational linguistics, the interest in terminological applications such as automatic term extraction is growing, and many studies use the quantitative information (cf. Kageura & Umino, 1996). However, the basic quantitative nature of terminological structure, which is essential for terminological theory and applications, has not yet been exploited. The static quantitative descriptions are not sufficient, as there are terms which do not appear in the sample. So it is crucial to establish some models, by which the terminological structure beyond the sample size can be properly described.

In Japanese terminology, the roles of morphemes are different according to their types of origin, i.e. the morphemes borrowed mainly from Western languages (borrowed morphemes) and the native morphemes including Chinese-origined morphemes which are the majority. There are some quantitative studies (Ishii, 1987; Nomura & Ishii, 1989), but they only treat the static nature of the sample.

Located in the intersection of these two backgrounds, the aim of the present study is twofold, i.e. (1) to introduce a quantitative framework in which the dynamic nature of terminology can be described, and to examine its theoretical validity, and (2) to describe the quantitative dynamics of morphemes as a 'mass' in Japanese terminology, with reference to the types of origin.

2 Terminological Data

2.1 The Data

We use a list of different terms as a sample, and observe the quantitative nature of the constituent elements or morphemes. The quantitative regularities is expected to be observed at this level, because a large portion of terms is complex (Nomura & Ishii, 1989), whose formation is systematic (Sager, 1990), and the quantitative nature of morphemes in terminology is independent of the token frequency of terms, because the term formation is a lexical formation.

With the correspondences between text and terminology, sentences and terms, and words and morphemes, the present work can be regarded as parallel to the quantitative study of words in texts (Baayen, 1991; Baayen, 1993; Mandelbrot, 1962; Simon, 1955; Yule, 1944; Zipf, 1935). Such terms as 'type', 'token', 'vocabulary', etc. will be used in this context.

Two Japanese terminological data are used in this study: computer science (CS: Aiso, 1993) and psychology (PS: Japanese Ministry of Education, 1986). The basic quantitative data are given in Table 1, where T, N, and $V(N)$ indicate the number of terms, of running morphemes (tokens), and of different morphemes (types), respectively.

In computer science, the frequencies of the borrowed and the native morphemes are not very different. In psychology, the borrowed

Domain	T	N	V(N)	N/T	N/V(N)	C_L
CS all	14983	36640	5176	2.45	7.08	0.211
borrowed		14696	2809		5.23	0.242
native		21944	2367		9.27	0.174
PS all	6272	14314	3594	2.28	3.98	0.235
borrowed		1541	995		1.55	0.309
native		12773	2599		4.91	0.207

Table 1. Basic Figures of the Terminological Data

morphemes constitute only slightly more than 10% of the tokens. The mean frequency $N/V(N)$ of the borrowed morphemes is much lower than the native morphemes in both domains.

2.2 LNRE Nature of the Data

The LNRE (Large Number of Rare Events) zone (Chitashvili & Baayen, 1993) is defined as the range of sample size where the population events (different morphemes) are far from being exhausted. This is shown by the fact that the numbers of hapax legomena and of dislegomena are increasing (see Figure 1 for hapax).

A convenient test to see if the sample is located in the LNRE zone is to see the ratio of loss of the number of morpheme types, calculated by the sample relative frequencies as the estimates of population probabilities. Assuming the binomial model, the ratio of loss is obtained by:

$$C_L = (V(N) - \hat{E}[V(N)])/V(N) = \frac{\sum_{m\geq 1} V(m,N)(1 - p(i_{[f(i,N)=m]}, N))^N}{V(N)}$$

where:

$f(i,N)$: frequency of a morpheme w_i in a sample of N.

$p(i,N) = f(i,N)/N$: sample relative frequency.

m : frequency class or a number of occurrence.

$V(m,N)$: the number of morpheme types occurring m times (spectrum elements) in a sample of N.

In the two data, we underestimate the number of morpheme types by more than 20% (C_L in Table 1), which indicates that they are clearly located in the LNRE zone.

3 The LNRE Framework

When a sample is located in the LNRE zone, values of statistical measures such as type-token ratio, the parameters of 'laws' (e.g. of Mandelbrot, 1962) of word frequency distributions, etc. change systematically according to the sample size, due to the unobserved events. To treat LNRE samples, therefore, the factor of sample size should be taken into consideration.

Good (1953) gives a method of re-estimating the population probabilities of the types in the sample as well as estimating the probability mass of unseen types. There is also work on the estimation of the theoretical vocabulary size (Efron & Thisted, 1976; National Language Research Institute, 1958; Tuldava, 1980). However, they do not give means to estimate such values as $V(N)$, $V(m,N)$ for arbitrary sample size, which are what we need. The LNRE framework (Chitashvili & Baayen, 1993) offers the means suitable for the present study.

3.1 Binomial/Poisson Assumption

Assume that there are S different morphemes w_i, $i = 1,2,...S$, in the terminological population, with a probability p_i associated with each of them. Assuming the binomial distribution and its Poisson approximation, we can express the expected numbers of morphemes and of spectrum elements in a given sample of size N as follows:

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

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

As our data is in the LNRE zone, we cannot estimate p_i. Good (1953) and Good & Toulmin (1956) introduced the method of interpolating and extrapolating the number of types for arbitrary sample size, but it cannot be used for extrapolating to a very large size.

3.2 The LNRE Models

Assume that the distribution of grouped probability p follows a distribution 'law', which can be expressed by some structural type distribution $G(p) = \sum_{i=1}^{S} I_{[p_i \geq p]}$, where $I = 1$ when $p_i \geq p$ and 0 otherwise. Using $G(p)$, the expressions (1) and (2) can be re-expressed as follows:

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

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

where $dG(p) = G(p_j) - G(p_{j+1})$ around p_j, and 0 otherwise, in which p is now grouped for the same value and indexed by the subscript j that indicates in ascending order the values of p.

In using some explicit expressions such as lognormal 'law' (Carrol, 1967) for $G(p)$, we again face the problem of sample size dependency of the parameters of these 'laws'. To overcome the problem, a certain distribution model for the population is assumed, which manifests itself as one of the 'laws' at a pivotal sample size Z. By explicitly incorporating Z as a parameter, the models can be completed, and it becomes possible (i) to represent the distribution of population probabilities by means of $G(p)$ with Z and to estimate the theoretical vocabulary size, and (ii) to interpolate and extrapolate $V(N)$ and $V(m,N)$ to the arbitrary sample size N, by such an expression:

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

The parameters of the model, i.e. the original parameters of the 'laws' of word frequency distributions and the pivotal sample size Z, are estimated by looking for the values that most properly describe the distributions of spectrum elements and the vocabulary size at the given sample size. In this study, four LNRE models were tried, which incorporate the lognormal 'law' (Carrol, 1967), the inverse Gauss-Poisson 'law' (Sichel, 1986), Zipf's 'law' (Zipf, 1935) and Yule-Simon 'law' (Simon, 1955).

4 Analysis of Terminology

4.1 Random Permutation

Unlike texts, the order of terms in a given terminological sample is basically arbitrary. Thus term-level random permutation can be used to obtain the better descriptions of sub-samples. In the following, we use the results of 1000 term-level random permutations for the empirical descriptions of sub-samples.

In fact, the results of the term-level and morpheme-level permutations almost coincide, with no statistically significant difference. From this we can conclude that the binomial/Poisson assumption of the LNRE models in the previous section holds for the terminological data.

4.2 Quantitative Measures

Two measures are used for observing the dynamics of morphemes in terminology. The first is the mean frequency of morphemes:

$$X(V(N)) = \frac{N}{V(N)} \quad (5)$$

The repeated occurrence of a morpheme indicates that it is used as a constituent element of terms, as the samples consist of term types. As it is not likely that the same morpheme occurs twice in a term, the mean frequency indicates the average number of terms which is connected by a common morpheme.

A more important measure is the growth rate, $P(N)$. If we observe $E[V(N)]$ for changing N, we obtain the growth curve of the morpheme types. The slope of the growth curve gives the growth rate. By taking the first derivate of $E[V(N)]$ given by equation (3), therefore, we obtain the growth rate of the morpheme types:

$$P(N) = \frac{d}{dN} E[V(N)] = \frac{E[(V(1,N)]}{N} \quad (6)$$

This "expresses in a very real sense the probability that new types will be encountered when the ... sample is increased" (Baayen, 1991).

For convenience, we introduce the notation for the complement of $P(N)$, the reuse ratio:

$$R(N) = 1 - P(N) \quad (7)$$

which expresses the probability that the existing types will be encountered.

For each type of morpheme, there are two ways of calculating $P(N)$. The first is on the basis of the total number of the running morphemes (frame sample). For the borrowed morphemes, for instance, it is defined as:

$$P_{fb}(N) = E[V_{borrowed}(1,N)]/N$$

The second is on the basis of the number of running morphemes of each type (item sample). For instance, for the borrowed morphemes:

$$P_{ib}(N) = E[V_{borrowed}(1,N)]/N_{borrowed}$$

Correspondingly, the reuse ratio $R(N)$ is also defined in two ways.

P_i reflects the growth rate of the morphemes of each type observed separately. Each of them expresses the probability of encountering a new morpheme for the separate sample consisting of the morphemes of the same type, and does not in itself indicate any characteristics in the frame sample.

On the other hand, P_f and R_f express the quantitative status of the morphemes of each type as a mass in terminology. So the transitions of P_f and R_f, with changing N, express the changes of the status of the morphemes of each type in the terminology. In terminology, P_f can be interpreted as the probability of incorporating new conceptual elements.

4.3 Application of LNRE Models

Table 2 shows the results of the application of the LNRE models, for the models whose mean square errors of $V(N)$ and $V(1,N)$ are minimal for 40 equally-spaced intervals of the sample. Figure 1 shows the growth curve of the morpheme types up to the original sample size (LNRE estimations by lines and the empirical values by dots). According to Baayen (1993), a good lognormal fit indicates high productivity, and the large Z of Yule-Simon model also means richness of the vocabulary. Figure 1 and the chosen models in Table 2 confirm these interpretations.

Domain	Model	Z	S	$V(N)$	$E[V(N)]$
CS all	Gauss-Poisson	236	56085	5176	5176.0
borrowed	Lognormal	419	75296	2809	2809.0
native	Gauss-Poisson	104	6095	2367	2362.6
PS all	Lognormal	1283	30691	3594	3594.0
borrowed	Yule-Simon	38051	∞	995	995.0
native	Gauss-Poisson	231	10191	2599	2599.0

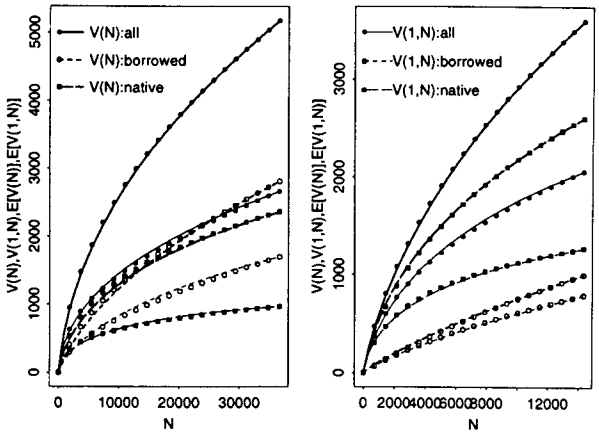
* Z : pivotal sample size ; S : population number of types

Table 2. The Applications of LNRE Models

From Figure 1, it is observed that the number of the borrowed morpheme types in computer science becomes bigger than that of the native morphemes around $N = 15000$, while in psychology the number of the borrowed morphemes is much smaller within the given sample range. All the elements are still growing, which implies that the quantitative measures keep changing.

Figure 2 shows the empirical and LNRE estimation of the spectrum elements, for $m = 1$ to 10. In both domains, the differences between $V(1,N)$ and $V(2,N)$ of the borrowed morphemes are bigger than those of the native morphemes.

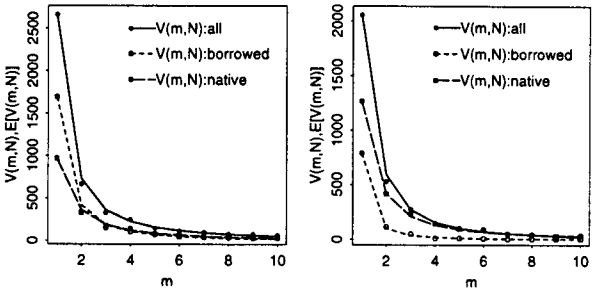
Both the growth curves in Figure 1 and the distributions of the spectrum elements in Figure 2 show, at least to the eye, the reasonable fits of the LNRE models. In the discussions below, we assume that the LNRE based estimations are



lines : LNRE estimations ; dots : empirical values

(a) Computer Science (b) Psychology

Fig. 1. Empirical and LNRE Growth Curve



lines : LNRE estimations ; dots : empirical values

(a) Computer Science (b) Psychology

Fig. 2. Empirical and LNRE Spectrum Elements

valid, within the reasonable range of N. The statistical validity will be examined later.

4.3.1 Mean Frequency

As the population numbers of morphemes are estimated to be finite with the exception of the borrowed morphemes in psychology, $\lim_{N\to\infty} X(V(N)) = \infty$, which is not of much interest. The more important and interesting is the actual transition of the mean frequencies within a realistic range of N, because the size of a terminology in practice is expected to be limited.

Figure 3 shows the transitions of $X(V(N))$, based on the LNRE models, up to $2N$ in computer science and $5N$ in psychology, plotted according to the size of the frame sample. The mean frequencies are consistently higher in computer science than in psychology. Around $N =$

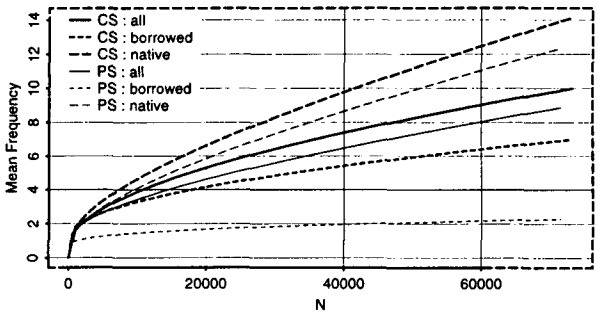


Fig. 3. Mean Frequencies

70000, $X(V(N))$ in computer science is expected to be 10, while in psychology it is 9. The particularly low value of $X(V(N_{borrowed}))$ in psychology is also notable.

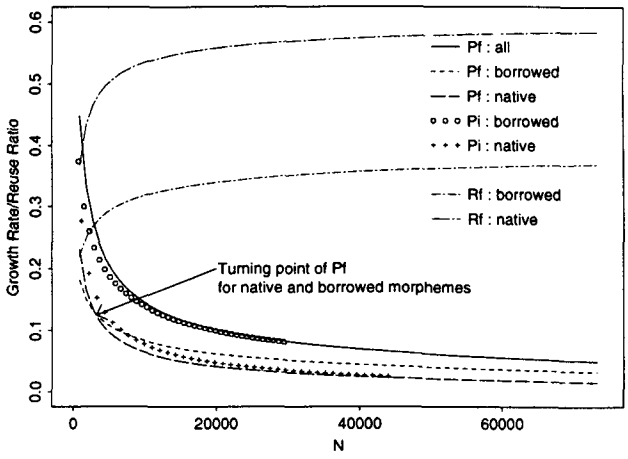
4.3.2 Growth Rate/Reuse Ratio

Figure 4 shows the values of P_f, P_i and R_f, for the same range of N as in Figure 3. The values of $P_{ib}(N)$ and $P_{in}(N)$ in both domains show that, in general, the borrowed morphemes are more 'productive' than the native morphemes, though the actual value depends on the domain.

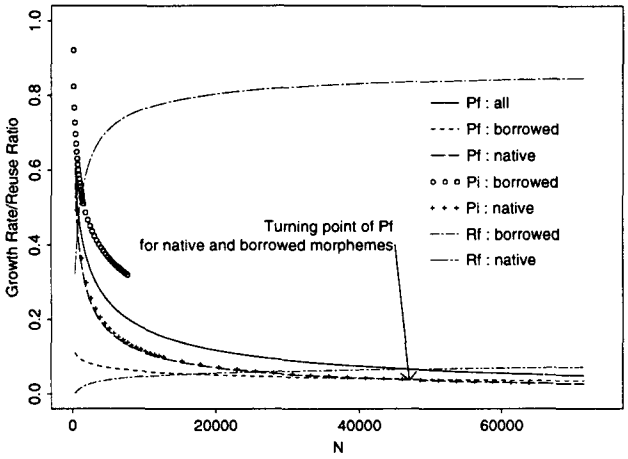
Comparing the two domains by $P_{fall}(N)$, we can observe that at the beginning the terminology of psychology relies more on the new morphemes than in computer science, but the values are expected to become about the same around $N = 70000$.

P_fs for the borrowed and native morphemes show interesting characteristics in each domain. Firstly, in computer science, at the relatively early stage of terminological growth (i.e. $N \simeq 3500$), the borrowed morphemes begin to take the bigger role in incorporating new conceptual elements. $P_{fb}(N)$ in psychology is expected to become bigger than $P_{fn}(N)$ around $N = 47000$. As the model estimates the population number of the borrowed morphemes to be infinite in psychology, that the $P_{fb}(N)$ becomes bigger than $P_{fn}(N)$ at some stage is logically expected. What is important here is that, even in psychology, where the overall role of the borrowed morphemes is marginal, $P_{fn}(N)$ is expected to become bigger around $N \simeq 47000$, i.e. $T \simeq 21000$, which is well within the realistic value for a possible terminological size.

Unlike P_f, the values of R_f show stable transition beyond $N = 20000$ in both domains,



(a) Computer Science



(b) Psychology

Fig. 4. Changes of the Growth Rates

gradually approaching the relative token frequencies.

5 Theoretical Validity

5.1 Linguistic Validity

We have seen that the LNRE models offer a useful means to observe the dynamics of morphemes, beyond the sample size. As mentioned, what is important in terminological analyses is to obtain the patterns of transitions of some characteristic quantities beyond the sample size but still within the realistic range, e.g. $2N$, $3N$, etc. Because we have been concerned with the morphemes as a mass, we could safely use N instead of T to discuss the status of morphemes,

implicitly assuming that the average number of constituent morphemes in a term is stable.

Among the measures we used in the analysis of morphemes, the most important is the growth rate. The growth rate as the measure of the productivity of affixes (Baayen, 1991) was critically examined by van Marle (1991). One of his essential points was the relation between the performance-based measure and the competence-based concept of productivity. As the growth rate is by definition a performance-based measure, it is not unnatural that the competence-based interpretation of the performance-based productivity measure is requested, when the object of the analysis is directly related to such competence-oriented notion as derivation. In terminology, however, this is not the case, because the notion of terminology is essentially performance-oriented (Kageura, 1995). The growth rate, which concerns with the linguistic performance, directly reflects the inherent nature of terminological structure[1].

One thing which may also have to be accounted for is the influence of the starting sample size. Although we assumed that the order of terms in a given terminology is arbitrary, it may not be the case, because usually a smaller sample may well include more 'central' terms. We may need further study concerning the status of the available terminological corpora.

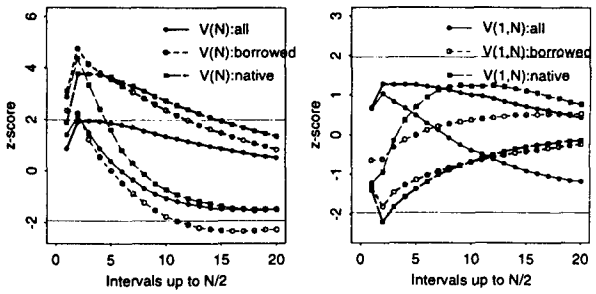
5.2 Statistical Validity

Figure 5 plots the values of the z-score for $E[V]$ and $E[V(1)]$, for the models used in the analyses, at 20 equally-spaced intervals for the first half of the sample[2]. In psychology, all but one values are within the 95% confidence interval. In computer science, however, the fit is not so good as in psychology.

Table 3 shows the χ^2 values calculated on the basis of the first 15 spectrum elements at the original sample size. Unfortunately, the χ^2 values show that the models have obtained the fits which are not ideal, and the null hypothesis

[1]Note however that the level of what is meant by the word 'performance' is different, as Baayen (1991) is text-oriented, while here it is vocabulary-oriented.

[2]To calculate the variance we need $V(2N)$, so the test can be applied only for the first half of the sample



(a) Computer Science (b) Psychology

Fig. 5. Z-Scores for $E[V]$ and $E[V(1)]$

is rejected at 95% level, for all the models we used.

Data	Model	χ^2	DF
CS all	Gauss-Poisson	129.70	14
borrowed	Lognormal	259.08	14
native	Gauss-Poisson	60.30	13
PS all	Lognormal	72.21	14
borrowed	Yule-Simon	179.36	14
native	Gauss-Poisson	135.30	13

Table 3. χ^2 Values for the Models

Unlike texts (Baayen, 1996a;1996b), the ill-fits of the growth curve of the models are not caused by the randomness assumption of the model, because the results of the term-level permutations, used for calculating z-scores, are statistically identical to the results of morpheme-level permutations. This implies that we need better models if we pursue the better curve-fitting. On the other hand, if we emphasise the theoretical assumption of the models of frequency distributions used in the LNRE analyses, it is necessary to introduce the finer distinctions of morphemes.

6 Conclusions

Using the LNRE models, we have succesfully analysed the dynamic nature of the morphemes in Japanese terminology. As the majority of the terminological data is located in the LNRE zone, it is important to use the statistical framework which allows for the LNRE characteristics. The LNRE models give the suitable means.

We are currently extending our research to integrating the quantitative nature of morphological distributions to the qualitative model of term formation, by taking into account the po-

sitional and combinatorial nature of morphemes and the distributions of term length.

Acknowledgement

I would like to express my thanks to Dr. Harald Baayen of the Max Plank Institute for Psycholinguistics, for introducing me to the LNRE models and giving me advice. Without him, this work coudn't have been carried out. I also thank to Ms. Clare McCauley of the NLP group, Department of Computer Science, the University of Sheffield, for checking the draft.

References

[1] Aiso, H. (ed.) (1993) *Joho Syori Yogo Daijiten.* Tokyo: Ohm.

[2] Baayen, R. H. (1991) "Quantitative aspects of morphological productivity." *Yearbook of Morphology 1991.* p. 109–149.

[3] Baayen, R. H. (1993) "Statistical models for word frequency distributions: A linguistic evaluation." *Computers and the Humanities.* 26(5–6), p. 347–363.

[4] Baayen, R. H. (1996a) "The randomness assumption in word frequency statistics." *Research in Humanities Computing 5.* p. 17–31.

[5] Baayen, R. H. (1996b) "The effects of lexical specialization on the growth curve of the vocabulary." *Computational Linguistics.* 22(4), p. 455–480.

[6] Carrol, J. B. (1967) "On sampling from a lognormal model of word frequency distribution." In: Kucera, H. and Francis, W. N. (eds.) *Computational Analysis of Present-Day American English.* Province: Brown University Press. p. 406–424.

[7] Chitashvili, R. J. and Baayen, R. H. (1993) "Word frequency distributions." In: Hrebicek, L. and Altmann, G. (eds.) *Quantitative Text Analysis.* Trier: Wissenschaftlicher Verlag. p. 54–135.

[8] Efron, B. and Thisted, R. (1976) "Estimating the number of unseen species: How many words did Shakespeare know?" *Biometrika.* 63(3), p. 435–447.

[9] Good, I. J. (1953) "The population frequencies of species and the estimation of population parameters." *Biometrika.* 40(3–4), p. 237–264.

[10] Good, I. J. and Toulmin, G. H. (1956) "The number of new species, and the increase in population coverage, when a sample is increased." *Biometrika.* 43(1), p. 45–63.

[11] Ishii, M. (1987) "Economy in Japanese scientific terminology." *Terminology and Knowledge Engineering '87.* p. 123–136.

[12] Japanese Ministry of Education (1986) *Japanese Scientific Terms: Psychology.* Tokyo: Gakujutu-Sinkokai.

[13] Kageura, K. (1995) "Toward the theoretical study of terms." *Terminology.* 2(2), 239–257.

[14] Kageura, K. and Umino, B. (1996) "Methods of automatic term recognition: A review." *Terminology.* 3(2), 259–289.

[15] Mandelbrot, B. (1962). "On the theory of word frequencies and on related Markovian models of discourse." In: Jakobson, R. (ed.) *Structure of Language and its Mathematical Aspects.* Rhode Island: American Mathematical Society. p. 190–219.

[16] Marle, J. van. (1991). "The relationship between morphological productivity and frequency." *Yearbook of Morphology 1991.* p. 151–163.

[17] National Language Research Institute (1958) *Research on Vocabulary in Cultural Reviews.* Tokyo: NLRI.

[18] Nomura, M. and Ishii, M. (1989) *Gakujutu Yogo Goki-Hyo.* Tokyo: NLRI.

[19] Sager, J. C. (1990) *A Practical Course in Terminology Processing.* Amsterdam: John Benjamins.

[20] Sichel, H. S. (1986) "Word frequency distributions and type-token characteristics." *Mathematical Scientist.* 11(1), p. 45–72.

[21] Simon, H. A. (1955) "On a class of skew distribution functions." *Biometrika.* 42(4), p. 435–440.

[22] Tuldava, J. (1980) "A mathematical model of the vocabulary-text relation." *COLING'80.* p. 600–604.

[23] Yule, G. U. (1944) *The Statistical Study of Literary Vocabulary.* Cambridge: Cambridge University Press.

[24] Zipf, G. K. (1935). *The Psycho-Biology of Language.* Boston: Houghton Mifflin.