The Dynamics of Morphemes in Japanese Terminology

Kyo Kageura†

This paper quantitatively analyses the role of morphemes with respect to their types of origin. Static quantitative analysis of a given data set is not sufficient for this aim, as language data in general and terminological data in particular have the specific characteristic of being "incomplete" in the sense that many unseen elements are expected in the theoretical population. Thus, the quantitative structure of morphemes in terminology should be analysed dynamically, by observing the growth pattern of morphemes. In order to allow for that, we use binomial interpolation and extrapolation. Results of analyses of the terminologies of six different domains follow, revealing interesting characteristics of the role of morphemes of different types of origin that do not manifest themselves through static quantitative analysis.

KeyWords: *Terminology, Morphemes, Binomial Interpolation, Binomial Extrapolation, Types of Origin*

1 Introduction

Modelling the dynamic nature of vocabulary is important, not only theoretically but also practically. It is important theoretically because there has been little concrete work on the dynamics that underly the structure of vocabulary at an idiosynchronic slice of language although speculatively these dynamics are widely held to be important. It is important practically because it can give a basic perspective from which the problem of so-called "lexical bottleneck" in many NLP-related applications can be diagnosed. This task is especially important to the study of technical terminologies, due to their rapid growth in many different domains. This, however, is an area of research that has thus far gone unexplored.

Against this background, this paper analyses the role of morphemes in Japanese terminology — with respect to their type of origin — and clarifies the basic structural tendencies of dynamics of terminology, using a probabilistic method. The present study is both descriptive and theoretical — descriptive because it gives a concrete description of the growth patterns of morphemes; theoretical because it is concerned with examining the underlying structural dynamics of terminologies in such a way that they be properly visualised and their basic characteristics explained. Note here that the unique position of lexicology requires any theoretical

† Human and Social Informatics Division, National Institute of Informatics

research in the field to be concretely anchored to the existing vocabulary (Maeda 1989), which in turn requires the study to be descriptive (Kageura 2002). Thus, the descriptive content, together with the basic perspective from which the concrete descriptions are made, constitutes the theoretical contribution of the present study to the field of terminology.

We focus on the patterns of morphemes with regard to their type of origin because, in Japanese vocabulary in general and in terminology in particular, the roles of morphemes are said to differ according to origin type, i.e. whether they are borrowed mainly from Western languages (*gairaigo* morphemes), are of Chinese origin (*kango* morphemes) or are original Japanese morphemes (*wago* morphemes). Many studies have addressed the nature of these origin types (Saiga 1957; National Language Research Institute 1958; Miyaji 1982; Nomura 1984) and have argued qualitatively that there are differences in nature among morphemes of each type. Some have carried out quantitative analyses of Japanese terminologies with respect to the types of origin of the constituent elements or morphemes of terms (Ishii and Nomura 1984; Ishii 1987).

Note that the quantitative analyses carried out so far are mostly static; they describe the quantitative characteristics of a given set of data. Static analyses of a given data, however, are in general not sufficient when dealing with language, because there are events or items which may not appear in the sample but do exist in the theoretical population. In analysing morphemes in terminology, therefore, it is necessary to use a method by which the nature of morphemes — including those that do not appear in the data — are properly accounted for. This is not only technically important but also theoretically essential as it assigns the model a moment of dynamics and thus a basis for expectation. Only through revealing the structural characteristics not explicit in the data itself can the nature of vocabulary can be fully observed.

In the following, the nature of the terminological data used in this study is first briefly described. This is followed by the introduction of a theoretical model, that can properly treat a sample with unseen events, together with the method of binomial interpolation and extrapolation. This gives the basic framework for modelling the structural dynamics of morphemes in terminology. From there descriptions of the dynamic quantitative nature of morphemes of different types of origin in six sets of terminologies will ensue. This last section constitutes the central part of the present paper and its main contribution to the study of lexicology.

2 The Terminological Data

As the basic data for the analysis, we use lists of term types, as opposed to term tokens in texts, and analyse the quantitative nature of constituent elements or morphemes within the list of term types. There are two reasons for this. Firstly, as terms are basically created by lexical formation, the quantitative nature of morphemes in terminology is independent of the token frequency of terms (Sager 1990; Kageura 2002). Recent psycholinguistic studies also support this claim (Baayen, Lieber, and Schreuder 1997; Schreuder and Baayen 1997). Secondly, as the majority of terms are complex (Nomura and Ishii 1989) and new terms are constantly formed by compounding, the quantitative nature of morphemes in the construction of terminologies is a key element for the modelling of terminological structure.

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 (Zipf 1935; Yule 1944; Mandelbrot 1953; Simon 1955; Carrol 1967; Sichel 1975). Terms in the field of quantitative linguistics, such as "type", "token", etc., shall be used in this context.

Chose for the present study are the terminological data of the following six different domains: agriculture (AGR) (Japanese Ministry of Education 1986a), botany (BOT) (Japanese Ministry of Education 1990a), chemistry (CHM) (Japanese Ministry of Education 1986b), computer science (COM) (Aiso 1993), physics (PHY) (Japanese Ministry of Education 1990b) and psychology (PSY) (Japanese Ministry of Education 1986c). They were chosen, within the limited availability of terminological data from roughly the same period, to cover both "harder" and "softer" scientific and technological domains.

Within these sets of data, the terms are identified on the basis of their orthography and type of origin; polysemous morphemes are not semantically distinguished, nor are inflections stemmed (though there are not any in the data). The terms are segmented into morphemes according to the criteria given in (Nomura and Ishii 1989). Briefly, the method first defines a minimal element, the smallest unit that bears meaning in current Japanese. Then, according to the origin of linguistic elements (*wago*, *kango* and *gairaigo*), the morphemes are defined as follows: (i) for *wago* and *gairaigo*, a minimal element constitutes a morpheme, e.g. 手 ('te': hand) and コンピュータ (computer); (ii) for *kango*, a first-order combination of two minimal elements constitutes a morpheme, while a minimal *kango* element attached to a morpheme is also treated as a morpheme (e.g. 図書館員 has the structure [[図 書] 館] 員], so 図書, 館 and 員 are identified as morphemes); and (iii) for *kango* and *wago* mixture, a first-order combination of minimal elements is identified as a morpheme, e.g. 係員.

Table 1 Basic quantities of the terminology samples of the six domains

Domain	T	N (%)		V(N) (%)		N/T	N/V(N)	C_L
AGR All	15067	29142	(100.00 %)	9093	(100.00 %)	1.93	3.20	0.256
Borrowed		2610	(8.96 %)	1513	(16.64 %)		1.73	0.300
Native		26532	(91.04 %)	7580	(83.36 %)		3.50	0.247
BOT All	10956	22605	(100.00 %)	5348	(100.00 %)	2.06	4.23	0.224
Borrowed		3072	(13.59 %)	1678	(31.38 %)		1.83	0.283
Native		19533	(86.41 %)	3670	(68.62 %)		5.32	0.197
CHM All	12074	23577	(100.00 %)	6400	(100.00 %)	1.95	3.68	0.246
Borrowed		5998	(25.44 %)	2841	(38.77 %)		2.11	0.289
Native		17579	(74.56 %)	3559	(61.23 %)		4.94	0.212
COM All	14983	36640	(100.00 %)	5176	(100.00 %)	2.45	7.08	0.211
Borrowed		14696	(40.11 %)	2809	(54.27 %)		5.23	0.242
Native		21944	(59.89 %)	2367	(45.73 %)		9.27	0.174
PHY All	10635	25095	(100.00 %)	4745	(100.00 %)	2.36	5.29	0.228
Borrowed		5048	(20.12 %)	2081	(43.86 %)		2.43	0.269
Native		20047	(79.88 %)	2664	(56.14 %)		7.53	0.197
PSY All	6272	14314	(100.00 %)	3594	(100.00 %)	2.28	3.98	0.235
Borrowed		1541	(10.77 %)	995	(27.69 %)		1.55	0.309
Native		12773	(89.23 %)	2599	(72.31 %)		4.91	0.207

In the present analysis, types of origin are classified into two, i.e. *gairaigo* morphemes on the one hand and *kango* and *wago* morphemes on the other (henceforth, we will call the former "borrowed" morphemes and the latter "native" morphemes). *Kango* and *wago* morphemes are grouped together because: (i) the majority are *kango* and mixed morphemes (which behave roughly equivalent to *kango* and mostly written in Chinese characters), and the number of pure *wago* morphemes is very small, and (ii) we are here concerned more with the status of *gairaigo* morphemes in the recent development of terminologies (cf. (Ishii 1987)).

Table 1 gives the basic quantitative data of the six terminological data. T, N and $V(N)$ indicate the number of terms, the number of running morphemes (tokens), and the number of different morphemes (types), respectively. N/T indicates the average length of a term in terms of its constituent morphemes, and $N/V(N)$ represents the average frequency of a morpheme. The meaning of C_L will be explained shortly. Table 2 shows some examples of morphemes

Table 2 Some examples of morphemes

AGR	性 (497, n), 機 (306, n), 体 (195, n), 土壌 (192, n), 法 (183, n), エポキシ (1, b), インフルエンザ (1, b), こうじ (1, n), 雨害 (1, n), ＣＥＭ (1, n)
BOT	性 (467, n), 体 (431, n), 細胞 (337, n), 植物 (269, n), 酸 (240, n), アミラーゼ (1, b), アセト (1, b), つぼ (1, n), 果床 (1, n), 安全 (1, n)
CHM	酸 (424, n), 性 (308, n), 剤 (282, n), 化 (251, n), 油 (188, n), シャシ (1, b), 骨材 (1, b), りん (1, n), 剥離 (1, n), 行程 (1, n)
COM	システム (504, b), データ (499, b), 装置 (402, n), 制御 (368, n), の (339, n), ＶＬＳＩ (1, b), ＢＯＴ (1, b), スタディ (1, b), 思考 (1, n), 深度 (1, n)
PHY	の (594, n), 性 (246, n), 線 (236, n), 計 (216, n), 器 (210, n), 原色 (1, n), 誘体 (1, n), 標線 (1, n), アロイ (1, b), ストレージ (1, b)
PSY	的 (491, n), の (388, n), 性 (316, n), 法 (217, n), 学 (170, n), 付加 (1, n), 没 (1, n), 分節 (1, n), ベンダー (1, n), ホヴランド (1, n)

with their frequencies. It includes the top five morphemes and five randomly selected five hapax for each domain ("b" and "n" indicates "borrowed" and "native", respectively).

It is observed that, with the exception of the number of types in computer science (COM), both the type and token numbers of borrowed morphemes are smaller than those of native morphemes. The average frequency of the borrowed morphemes is smaller than that of the native morphemes in all of the data sets. From the terminological point of view, this tendency could be interpreted as follows: (i) The native morphemes are used to represent core conceptual elements which appear repeatedly in terminology (in terms of average frequency and token number), and (ii) although in terms of accommodating new concepts, the borrowed morphemes are used relatively more frequently, with the exception of computer science, the native morphemes still take a major role.

However, as will be shown, this observation is too simplistic, if not incorrect. Technically, the problem is that, in language data in general, most statistical measures change systematically according to sample size (Tweedie and Baayen 1997). This makes it difficult to draw a reliable conclusion using summary statistics based on a particular sample or a snapshot of the target phenomena. This is related to the long-recognised fact that there are always events that do not appear in language data but do in fact exist (Yule 1944; Herdan 1960; Mizutani 1983). As we can in no way claim that our terminological data constitutes the population of terminology for each domain, even synchronically, we have to expect that there are morphemes that do not appear in our data[1]. In terms of terminology theory, this statistical peculiarity of

1 Theoretically, whether the interpretative framework of the present study is anchored to the synchronic state of

the data can be interpreted as a reflection of the dynamics of terminology, in the sense that the potentiality of terminology is manifested in the structure of a given terminological sample.

3 Theoretical Model and the Status of Data

We introduce here a dynamic statistical model which can treat the terminological data properly. In the process, we also confirm that the terminological data, like language data in general, anticipates unseen events, as we informally mentioned in the previous section.

3.1 Binomial/Poisson Model

The model we introduce here regards a terminology as a bag of morphemes, without any inter-morpheme dependencies. As a distributional model of morphemes in terminology, it offers a good and principled approximation to the behaviour of morphemes in terminology, for two reasons. Firstly, we can ignore the qualitative dependency of morphemes within individual complex terms in modelling the distribution of morphemes in terminology (Kageura 1998). Secondly, the order of terms in the data is basically arbitrary, unlike the order of words or sentences in texts (Baayen 1996b, 1996a). We can thus safely apply the binomial model — which assumes no inter-event dependency and sees the data as a bag of events — to the distribution of morphemes in terminology.

Assume that there are S different morphemes, i.e. w_i, $i = 1, 2, ...S$, in the population of a terminology, with a population probability p_i associated with each. Based on the binomial assumption, which in turn can be approximated by the Poisson model, the expected number of different morphemes, $E[V(N)]$, and the expected number of morphemes that appear 1, 2, 3, ... m times, $E[V(m, N)]$, in a given sample of size N, can be expressed as follows (Baayen 2001):

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

language or to the diachronic nature of language requires in-depth articulation. Here we simply assume that, as far as we are dealing with the internal structure of terminologies, this distinction is irrelevant. For further discussion, see (Kageura 2000, 2002).

This will be the starting point of the binomial interpolation and extrapolation which will be introduced shortly to trace the structural dynamics of morphemes.

Assuming this model, incidentally, we can check the statistical status of the terminological data. As discussed, it is widely held that there are always events that do not appear in a language sample. But do they really exist, for example, in the data in Table 1? Should we really take into account these morphemes that do not appear in the data? There is a convenient test to explore this, called the coefficient of loss (Chitashvili and Baayen 1993). The coefficient of loss calculates the ratio of the number of events that are lost by estimating the number of events in the original sample size, using the sample relative frequencies to estimate the population probabilities, based on the binomial model. Formally, the coefficient of loss (C_L) is defined as follows:

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

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 number of occurrence.

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

The column C_L in Table 1 indicates the values of the coefficient of loss for each data. The number of morpheme types is underestimated by around 20 per cent, which means that the sample relative frequency does not give a reliable estimate of the population probability. The data belong to what is called the LNRE (Large Number of Rare Events) zone of the sample range (Chitashvili and Baayen 1993; Khmaladze 1987), where the population events (morpheme types) are far from being exhausted in the sample. In this situation, not only the sample relative frequencies but also almost all of the statistical measures as well as the parameters of the distribution models change systematically according to the sample size (Baayen 2001; Tweedie and Baayen 1997). It is to overcome this problem that binomial interpolation and extrapolation is required.

3.2 Binomial Interpolation and Extrapolation

To overcome the problem of sample-size dependency among statistical measures, Good and Toulmin (Good and Toulmin 1956) propose a method of interpolating and extrapolat-

ing the sample and calculating the number of events as well as the spectrum elements for a (theoretically) arbitrary sample size.

The number of events and the number of the spectrum elements of a sample size N, conditional on the original sample of size N_0, can be expressed by the following formula:

$$E[V(\lambda N)] = V(N) - \sum_{k=0}^{\infty}(-1)^k(\lambda-1)^k E[V(k,N)] \tag{2}$$

$$E[V(m,\lambda N)] = \lambda^m \sum_{k=0}^{\infty}(-1)^k \binom{m+k}{m} (\lambda-1)^k E[V(m+k,N)] \tag{3}$$

Appendix A gives the derivation of (2) and (3) from (1), originally provided in (Good and Toulmin 1956).

Binomial interpolation and extrapolation provides the means of tracing the *developmental profile* of the growth of morphemes as well as the growth rate of morphemes (which will be introduced in 4.2). By employing this methods, we can make explicit the quantitative nature of morphemes implicit in a given data. In other words, through binomial interpolation and extrapolation, we can observe how the ratio between borrowed and native morphemes was, is, and will be when the data is changed in size, as opposed to simply how it is in a given set of data. Although the actual value diverges around $N = 2N_0$, the formula is sufficient for observing the basic dynamic characteristics of morphemes in a terminology within a realistic range of the terminological phenomena. Revealing the developmental profiles, therefore, explicates the structural characteristics of terminology.

Recall, incidentally, that, as discussed above, the randomness assumption behind the binomial/Poisson model holds for terminological data in general. This is confirmed to be valid in the terminological data used in the present study, as observed in Figure 1, which shows the developmental profiles of $E[V(N)]$ and $E[V(1,N)]$ obtained through binomial interpolation and extrapolation for up to twice the original sample size (lines), as well as the corresponding values obtained by 5,000 term-level (as opposed to morpheme-level) random permutations up to the original sample size for 20 equally-spaced intervals (dots). The results of binomial interpolation and extrapolation based on the randomness assumption of the distribution of morphemes are almost identical to the empirical results obtained through the random permutation of terms[2].

2 The z-score is available for up to half the original sample size, using the following formula (Baayen 2001):

$$\frac{|V(N) - E[V(N)]|}{\sqrt{V(2N) - V(N)}}$$

if we allow ourselves to estimate the variance of $V(N)$ by $V(2N) - V(N)$. The result showed no significant difference between the two.

4 The Growth of Morphemes and the Roles of Morphemes

As the binomial model of interpolation and extrapolation provides a valid estimation of the morphemes in a given terminology up to around twice the original sample size, we can now observe the developmental profiles of the behaviour of morphemes, in terms of their type of origin, in accordance with the changes in the size of the data within and beyond the original sample size. This allows us to form general expectations of how the morphemes of different types of origin should behave.

4.1 Patterns of the Growth of Morphemes

Figure 1 charts the developmental profiles of morphemes in each of the six terminological domains, according to their types of origin. We can observe that, in all the domains, the growth curves of the borrowed morphemes are more "straight" than the growth curves of the native morphemes. The developmental curves of the native morphemes tend to flatten out more quickly than the curves of the borrowed ones.

We can expect that, although the number of different borrowed morphemes is smaller in all but one domain (computer science) at the given as well as at twice the given sample size, the relation may well be reversed when the sample is further increased. In the domains of chemistry and physics, borrowed morpheme types are expected to outnumber native morphemes fairly soon. This general estimation is informally reinforced by the fact that, in computer science, where the number of different borrowed morphemes is greater at the original sample size, there is a greater number of native morpheme types at the beginning of the sampling range, i.e. $N < 14,000$.

To be rigorous, the actual ratio of borrowed to native morphemes should be observed. This is shown in Figure 2. A clear general pattern, irrespective of domain, is recognised in Figure 2, i.e. the more a terminology grows, the more it depends, in terms of type, on borrowed morphemes. In this sense, what is thought to be an exception in terms of static quantitative measures, i.e. the status of borrowed morphemes in computer science, follows a general pattern, the only difference being the degree of actual manifestation of the general pattern vis-à-vis the size of the terminology. This general pattern is also in accordance with the situation concerning the diachronic development of the Japanese vocabulary in general.

The actual ratio of borrowed to native morpheme types differs from domain to domain, revealing the characteristics of each domain within the general pattern of borrowed and native

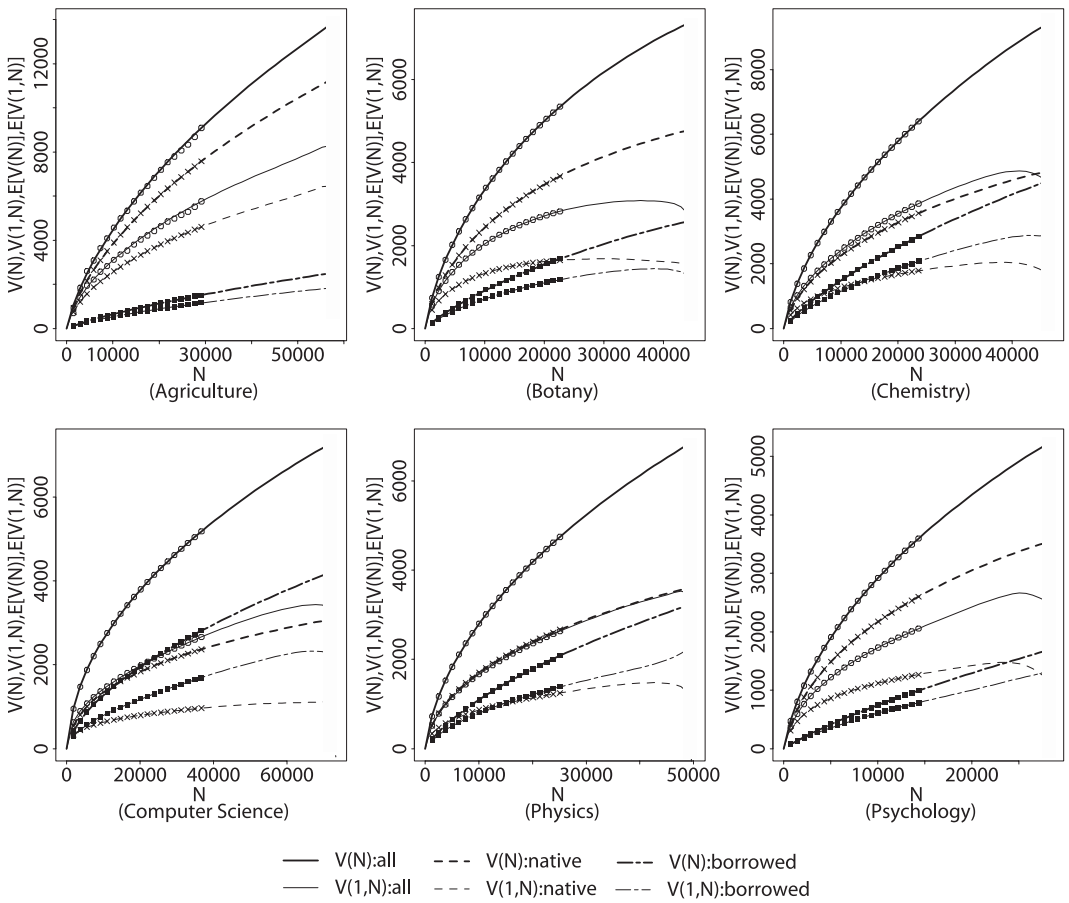

Fig. 1 Growth of morphemes based on binomial interpolation and extrapolation

morphemes. In computer science, as mentioned earlier, the number of different borrowed morphemes is already greater than that of native morphemes within the original sample size. In chemistry and physics, it is likely that the ratio will become greater than 1 within a realistic data size of the terminologies, say, $N = 3N_0$; while in botany and psychology, it is possible that the ratio will become bigger than 1 in due course, but at what data size this will occur is not clear. In agriculture, the opposite conclusion seems to be more reasonable. Thus it is the terminology of agriculture, not of computer science, that is exceptional in this respect. To confirm this informal and intuitive discussion more rigidly, it is useful to observe the growth rate of morphemes, to which we now turn.

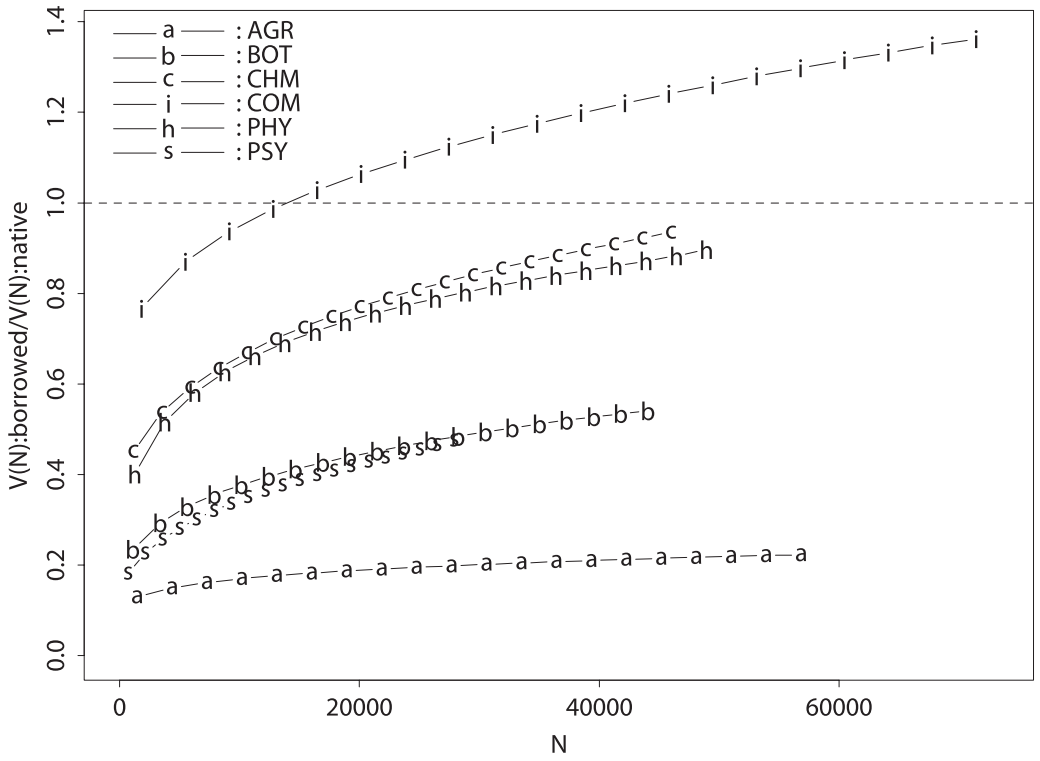

Fig. 2 Transitions in the ratio of borrowed to native morphemes

4.2 Patterns of the Growth Rate

The changing values of $E[V(N)]$ for changes in N provides the growth curve of the morpheme types, as illustrated in Figure 1. The next question to be asked is how we can obtain the *growth rate* of the morphemes at each point of observation. Interestingly, assuming the binomial/Poisson model, the growth rate of the morpheme types can be obtained by using the number of hapax legomena, or the morphemes that appear only once (Baayen 1991). Mathematically, the growth rate $\mathcal{P}(N)$ is defined as follows:

$$\mathcal{P}(N) = \frac{E[(V(1, N)]}{N}$$

A derivation of this formula from the binomial/Poisson model is explained in (Baayen 2001), and presented in Appendix B.

This index demonstrates the probability that new morpheme types will be encountered when the sample size is increased. Incidentally, this equals to the probability mass of unseen types obtained by well-known Good-Turing estimates (Good 1953).

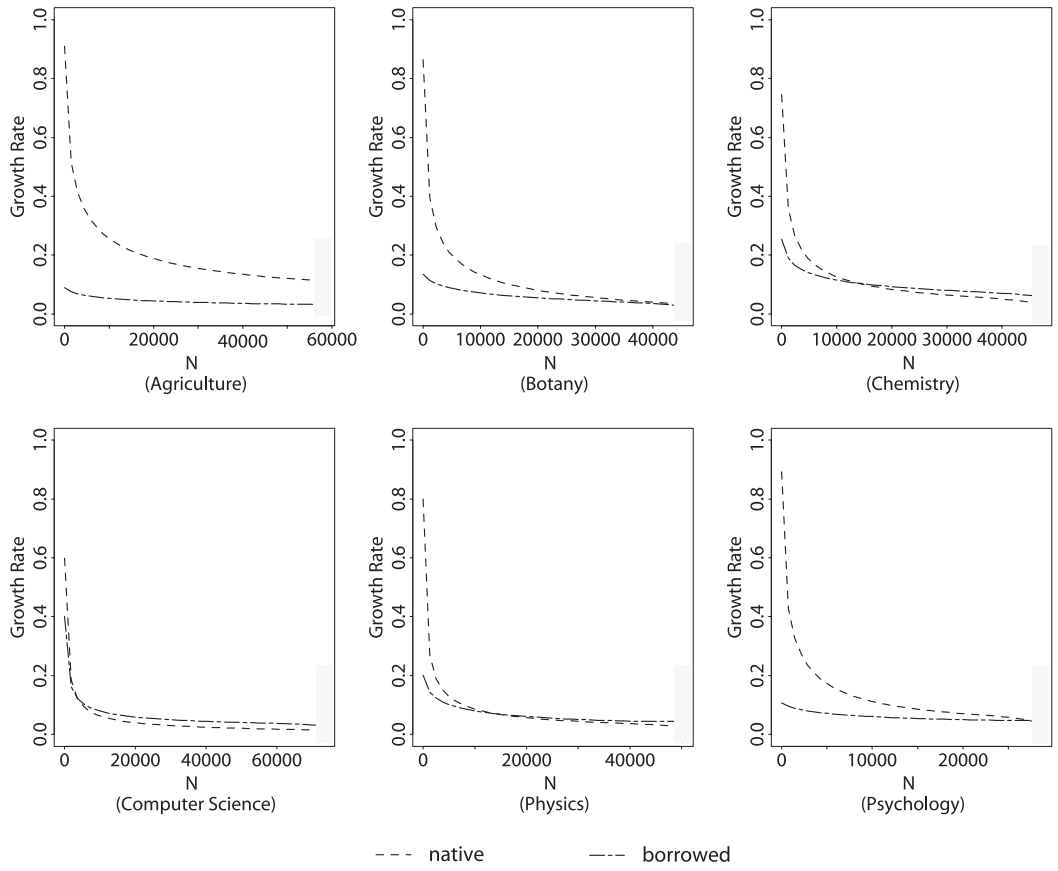

Fig. 3 Transitions of the growth rate of borrowed and native morphemes

Figure 3 shows the transition profile of the growth rate of the borrowed and native morphemes in the terminologies of the six domains, in accordance with increases in the sample size to up to twice the original size. The transition profiles of borrowed and native morphemes take different forms, with the same basic pattern observed in all six domains, i.e. at the beginning of the sample range, the growth rate of native morphemes is much higher than that of borrowed morphemes[3], but the former quickly decreases while the latter decreases very slowly as the sample is increased.

From the terminological point of view, this difference can be interpreted as follows: Native morphemes are first used to constitute the *core* set of morphemes in a terminology, but as the terminology grows, it begins to depend more and more on borrowed morphemes, in order to

3 At the outset, i.e. $N \simeq 0$, the growth rate of morphemes of each origin type is equal to the ratio of N of each origin type to the total number of running morphemes.

accommodate new concepts.

Within this general tendency, the actual values of the growth rates in the six domains show the concrete nature of the terminology of each domain[4]. In computer science, the growth rate is already reversed around $N = 3500$. In chemistry and physics, the growth rate is reversed around $N = 10,000$ to $15,000$, well within the original sample size. Botany and psychology show a similar pattern, and the growth rate is reversed or expected to be reversed around just $N = 2N_0$.

Focusing on the earlier stage of the sample range, the terminology of computer science is exceptional in its high dependency on borrowed morphemes. If we interpret the beginning stage of the sample size to be the stage at which core morphemes are introduced and consolidated, then computer science can be characterised by its heavy reliance on borrowed morphemes in the role of core morphemes.

On the other hand, when the size of a terminology becomes bigger, newly introduced morphemes are expected to be used to add new concepts to the existing structure. As N approaches ∞, the ratio of borrowed to native morpheme types converges to the ratio of their growth rates. From Figure 3, we can expect that, as $N \to \infty$, there will be a greater number of different borrowed morphemes than the number of native morphemes in all of the domains but agriculture. In that sense, the informal observation given in 4.1 based on Figure 2 has been rigidly confirmed.

This leaves us with one domain, i.e. agriculture. In agriculture, it is not clear whether the growth rate of the borrowed and native morphemes will be reversed at all. In this sense, among the six different domains we observe here, it is agriculture that is exceptional in the use of morphemes of different types of origin in the construction of terminology.

4.3 Summary of the Observations

Summarising the observations above, we can conclude, from the developmental profiles of the morphemes and the transitions in their growth rates, that native morphemes tend to be used to constitute the *core* of a terminology. Because the first and main role of the native morphemes is to contribute to expressing the core conceptual elements, it is natural that the native morphemes are used more frequently than borrowed morphemes. As the terminology grows, on the other hand, the use of borrowed morphemes grows, in order to incorporate new concepts. As the new concepts are incorporated into the existing terminological structure, the core of which has already become stable (Sager 1990), the average use of a borrowed mor-

4 In the discussion here, we use both the absolute size of the data and the sample scale relative to the original sample size of each domain, though the emphasis is on the latter.

pheme remains relatively low, as is manifested by the low average frequency of the borrowed morphemes. This general tendency can be observed irrespective of domain.

Turning our eyes to the differences among the domains, we can observe the following:

(1) From the point of view of the tendency of borrowed morphemes to be used to incorporate new concepts, the terminology of agriculture is an exception, in that the native morphemes will continue to be used more often for incorporating new morphemes than borrowed morphemes, even if the size of terminology becomes very large. This tendency is expected to continue possibly for $N \to \infty$. All the other five domains come to use, or will come to use, more borrowed morphemes than native morphemes for incorporating new morphemes. Among these domains, chemistry and physics show similar tendencies. Botany and psychology are also similar.

(2) From the point of view of the basic tendency of the native morphemes to be used to constitute the core set of morphemes in a terminology, it is computer science that is exceptional, in light of the high presence of borrowed morphemes from the beginning of the sample range, i.e. in the core morpheme set.

5 Conclusions

We have analysed the role of native and borrowed morphemes in the construction of the terminologies of six different domains, tracing the developmental patterns of the growth and the growth rates of morphemes. A few general as well as domain-dependent patterns in the use of morphemes were clarified. In the process, we introduced a theoretical framework based on the binomial/Poisson assumption, which was proved to provide a very useful and powerful method of analysing the dynamic patterns of morphological growth in terminology.

The work reported here should be extended further, at least in three aspects. Firstly, we should extend the observation to the terminologies of other domains. This is not only in itself crucial as a descriptive quantitative study of terminology but also important for uncovering the general tendencies of terminological structure across different domains, which in turn would lead to the characterisation of technical terminology as a whole.

Secondly, in order to fully explore the morphological structure in terminology, it is important to obtain reliable extrapolated values beyond $N < 2N_0$. Chitashvili and Baayen (Chitashvili and Baayen 1993; Baayen 2001) formulate the method of incorporating parametric models of word frequency distributions (Zipf 1935; Yule 1944; Simon 1955; Carrol 1967; Sichel 1975) to the framework of binomial interpolation/extrapolation. This opens the possibility of describing what is left open here, e.g. the possibility of the reversal of the growth

rate of the morphemes in agriculture.

The last point is related to the theoretical modelling of terminology. We have focused on the general difference between borrowed and native morphemes *en masse* in Japanese terminology, effectively ignoring the differences in such factors as term length distributions among different domains. To fully exploit the quantitative modelling of terminology, however, it will be necessary to take into account the wider characteristics terms, including the intra-term dependency patterns of morphemes, in addition to the nature of morphemes in terminology *en masse*.

Acknowledgement

This work is supported by the Grant-in-Aid C(2) 14580465 of the Japan Society of the Promotion of Sciences. I am indebted to Dr. H. Baayen of the Max Plank Institute for Psycholinguistics for guiding me to the dynamic analysis of language data.

Reference

Aiso, H. (1993). *Joho Syori Yogo Daijiten*. Ohm, Tokyo.

Baayen, R. H. (1991). "Quantitative Aspects of Morphological Productivity." In Booij, G., and van Marle, J. (Eds.), *Yearbook of Morphology 1991*, pp. 109–149. Kluwer, Dordrecht.

Baayen, R. H. (1996a). "The Effects of Lexical Specialization on the Growth Curve of the Vocabulary." *Computational Linguistics*, *22*(4), pp. 455–480.

Baayen, R. H. (1996b). "The Randomness Assumption in Word Frequency Statistics." In Perissinotto, G. (Ed.), *Research in Humanities Computing 5*, pp. 17–31. Clarendon, Oxford.

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

Baayen, R. H., Lieber, R., and Schreuder, R. (1997). "The Morphological Complexity of Simplex Nouns." *Linguistics*, *35*, pp. 861–877.

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*, pp. 406–424. Brown University Press, Province.

Chitashvili, R. J., and Baayen, R. H. (1993). "Word Frequency Distributions." In Hrebicek, L., and Altmann, G. (Eds.), *Quantitative Text Analysis*, pp. 54–135. Wissenschaftlicher Verlag, Trier.

Good, I. J. (1953). "The Population Frequencies of Species and the Estimation of Population Parameters." *Biometrika*, *40*(3/4), pp. 237–264.

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), pp. 45–63.

Herdan, G. (1960). *Type-Token Mathematics*. Mouton, The Hague.

Ishii, M. (1987). "Economy in Japanese Scientific Terminology." In Czap, H., and Galinski, C. (Eds.), *Proc. Terminology and Knowledge Engineering 1987*, pp. 123–136 Trier. Indeks Verlag.

Ishii, M., and Nomura, M. (1984). "Word-Formation of Compounds in 'Japanese Scientific Terms: Mechanical Engineering', on the Basis of Classification of Stems." *Mathematical Linguistics*, *14*(4), pp. 163–175.

Japanese Ministry of Education (1986a). *Japanese Scientific Terms: Agriculture*. Gakujutu-Sinkokai, Tokyo.

Japanese Ministry of Education (1986b). *Japanese Scientific Terms: Chemistry* (2nd edition). The Chemical Society of Japan, Tokyo.

Japanese Ministry of Education (1986c). *Japanese Scientific Terms: Psychology*. Gakujutu-Sinkokai, Tokyo.

Japanese Ministry of Education (1990a). *Japanese Scientific Terms: Botany*. Maruzen, Tokyo.

Japanese Ministry of Education (1990b). *Japanese Scientific Terms: Physics* (2nd edition). Baifukan, Tokyo.

Kageura, K. (1998). "The Effect of Intra-Term Morphological Coherence on the Growth Curve of Morphemes in Terminology." *Mathematical Linguistics*, *21*(7), pp. 311–323.

Kageura, K. (2000). "The Dynamics of Phenomena and the Dynamics of Data: On the Relationship between Events and Structures in Terminology (in Japanese)." *Mathematical Linguistics*, *22*(7), pp. 281–302.

Kageura, K. (2002). *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth*. John Benjamins, Amsterdam.

Khmaladze, E. V. (1987). "The Statistical Analysis of a Large Number of Rare Events." Technical report MS-R 8804, Department of Mathematical Statistics, Center for Mathematics and Computer Science.

Maeda, T. (1989). "Goi Souron." In Tamamura, F. (Ed.), *Nihongo no Goi to Imi*, pp. 1–22. Meiji Syoin.

Mandelbrot, B. (1953). "An Informational Theory of the Statistical Structure of Language." In Jackson, W. E. (Ed.), *Communication Theory*, pp. 486–502. Academic Press.

Miyaji, Y. (1982). "Gendaigo no Gokousei." In *Gendai no Goi*, Vol. 7 of *Kouza Nihongo no Goi*, pp. 67–90. Meiji Syoin, Tokyo.

Mizutani, S. (1983). *Goi*. Asakura Syoten, Tokyo.

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

Nomura, M. (1984). "Gosyu to Zougoryoku." *Nihongogaku*, *3*(9), pp. 40–54.

Nomura, M., and Ishii, M. (1989). *Gakujutu Yogo Goki-Hyo*. National Language Research Institute, Tokyo.

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

Saiga, H. (1957). "Gokousei no Tokushitsu." In Iwabuchi, E., Hayashi, O., Ohishi, H., and Shibata, T. (Eds.), *Kouza Gendai Kokugogaku 2*, pp. 217–248. Chikuma, Tokyo.

Schreuder, R., and Baayen, R. H. (1997). "How Complex Simplex Words can be." *Journal of Memory and Language*, *37*, pp. 118–139.

Sichel, H. S. (1975). "On a Distribution Law for Word Frequencies." *J.Am.Stat.Assoc.*, *70*(351), pp. 542–547.

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

Tweedie, F. J., and Baayen, R. H. (1997). "How Variable May a Constant be? Measures of Lexical Richness in Perspective." *Computers and the Humanities*, *31*(3), pp. 153–167.

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

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

Appendix

A Derivation of the formula for binomial interpolation/extrapolation

On the basis of the binomial model and under the assumption that the population events are known, the number of morphemes that occur exactly m times in the sample of size λN can be given by the equation (1):

$$E[V(m, \lambda N)] = \sum_{j=0}^{S} \binom{\lambda N}{m} p_j^m (1 - p_j)^{\lambda N - m}$$

This can be transformed as follows:

$$E[V(m, \lambda N)] \quad = \quad \sum_{j=0}^{S} \binom{\lambda N}{m} p_j^m (1 - p_j)^{\lambda N - m}$$

$$
\begin{aligned}
&= \sum_{j=0}^{S} \binom{\lambda N}{m} p_j^m (1-p_j)^{N-m} \left(1+\frac{p_j}{1-p_j}\right)^{-(\lambda-1)N} \\
&= \sum_{j=0}^{S} \binom{\lambda N}{m} p_j^m (1-p_j)^{N-m} \cdot \sum_{k=0}^{\infty} \binom{-(\lambda-1)N}{k} p_j^k (1-p_j)^{-k} \\
&= \sum_{k=0}^{\infty} \binom{\lambda N}{m} \binom{-(\lambda-1)N}{k} \sum_{j=0}^{S} p_j^{m+k} (1-p_j)^{N-(m+k)} \\
&= \sum_{k=0}^{\infty} \frac{\binom{\lambda N}{m}\binom{-(\lambda-1)N}{k}}{\binom{N}{m+k}} E[V(m+k,N)].
\end{aligned}
$$

If we only use here the range $m+k \leq N$ and $k \leq (\lambda-1)N$ for actual calculation, the term for combinatorics in the last line can be rewritten as:

$$
\begin{aligned}
\frac{\binom{\lambda N}{m}\binom{-(\lambda-1)N}{k}}{\binom{N}{m+k}} &\simeq \frac{(\lambda N)^m (-(\lambda-1)N)^k (m+k)!}{m!\, k!\, N^{m+k}} \\
&= (-1)^k \lambda^m (\lambda-1)^k \binom{m+1}{m},
\end{aligned}
$$

This leads to equation (3). Equation (2) immediately follows.

B Derivation of $\mathcal{P}(N)$

Firstly, let us introduce the structural distribution of the morpheme types, which can be expressed as follows:

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

where $I = 1$ when $p_i \geq p$ and 0 otherwise. The value of $G(p)$ represents the number of morpheme types whose occurrence probability is greater or equal to p. Let us then focus on the value of p, and introduce the new subscript j, such that p_j indicates, in ascending order, the values of p which at least one morpheme type takes, i.e. $p_j < p_{j+1}$ if $j < j+1$, and there is at least one w_i such that $p_i = p_j$ if $p_j > 0$.

Using $G(p)$ with the re-indexed subscript j of p, the equation that estimates the number of morpheme types can be rewritten in the integral form as follows:

$$
\begin{aligned}
E[V(N)] &= S - \sum_{i=1}^{S} (1-p_i)^N \\
&= \sum_{i=1}^{S} (1-e^{-Np_i}) \\
&= \int_0^{\infty} (1-e^{-Np})\, dG(p)
\end{aligned}
$$

where $dG(p) = G(p_j) - G(p_{j+1})$ around p_j, and 0 otherwise.

As this indicates the growth curve of the vocabulary, the first derivative of this formula expresses, in mathematical sense, the growth rate of the vocabulary, which can be expressed as follows:

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

Kyo Kageura: Kyo Kageura was born in 1964. He received his Ph.D. from the University of Manchester in 1993. He was working at the Department of Research and Development, the National Center for Science Information Systems (NACSIS), Japan, since 1988. Since 2000, he has been working as an associate professor of the Human and Social Informatics Division, the National Institute of Informatics, Japan. His main research interest is in media studies. He is also carrying out research in qualitative and quantitative modelling of term formation and terminological growth as well as in the logical foundation of the theory of terminology. He is a member of the Japan Society of Library and Information Science, the Association for Natural Language Processing, the Mathematical Linguistic Society of Japan, and the International Quantitative Linguistic Society.

(Received November 1, 2002)
(Reviced February 10, 2003)
(Accepted April 10, 2003)