Constructing and Evaluating Controlled Bilingual Terminologies

Rei Miyata
The University of Tokyo
rei@p.u-tokyo.ac.jp

Kyo Kageura
The University of Tokyo
kyo@p.u-tokyo.ac.jp

Abstract

This paper presents the construction and evaluation of Japanese and English controlled bilingual terminologies that are particularly intended for controlled authoring and machine translation with special reference to the Japanese municipal domain. Our terminologies are constructed by extracting terms from municipal website texts, and the term variations are controlled by defining preferred and proscribed terms for both the source Japanese and the target English. To assess the coverage of the terms/concepts in the municipal domain and validate the quality of the control, we employ a quantitative extrapolation method that estimates the potential vocabulary size. Using Large-Number-of-Rare-Event (LNRE) modelling, we compare two parameters: (1) uncontrolled and controlled and (2) Japanese and English. The results show that our terminologies currently cover about 45–65% of the terms and 50–65% of the concepts in the municipal domain, and are well controlled. The detailed analysis of growth patterns of terminologies also provides insight into the extent to which we can enlarge the terminologies within the realistic range.

1 Introduction

In this study, we construct controlled terminologies for the municipal domain and evaluate them in terms of the coverage and the quality of the variation control. Term variation management is essential in helping with the consistent use of terminology by not only authors but also translators and machine translation (MT) (Daille, 2005). On Japanese municipal websites, the case in point, we can find a number of variant forms of the same referent, such as '印鑑登録証明書' and '印鑑証明書' (seal registration certificate). As the former might be a preferred term in the municipal domain, we can define the latter as a proscribed term. In the target language texts, we also encounter various translations that correspond to the source terms, such as 'personal seals registration certificate' and 'seal proof certificate'. To maintain the consistency of the terminology use on the target side, we need to prescribe authorised translations.

Since there are no bilingual municipal terminologies that are well maintained and easily available, focusing on the municipal life information, we construct Japanese-English controlled terminologies from scratch by extracting terms from municipal texts and controlling the variant forms. To facilitate the manual extraction of terms, we developed a simple platform in which laypeople can collect terms efficiently.

While many attempts have been made to conduct extrinsic evaluation of terminological resources such as MT output evaluation (Langlais and Carl, 2004; Thicke, 2011), the intrinsic status of terminology such as coverage has not been examined much. The methodological difficulty in validating the coverage, i.e. how much of the potential terminology in a given domain is covered by the current terminology, is due to the fact that the population size of the terminology to compare is rarely available.[1] Sager (2001, p.763) pointed out, however, that statistical means 'can be used to decide when the addition of more text does not produce any new terms'. We can tackle this issue by employing a statistical method proposed for inspecting the current status of the corpus (Kageura and Kikui, 2006). It is also difficult to assess the quality of controlled terminology, i.e. how well the term variations are managed and standardised. In this paper, we present the idea of comparing the controlled terminologies of multiple languages to validate the quality of control.

[1] If it is available, we no longer need to *evaluate* such a gold standard.

Proceedings of the 5th International Workshop on Computational Terminology,
pages 83–93, Osaka, Japan, December 12 2016.

2 Controlled Bilingual Terminology

2.1 Compiling Parallel Corpus

To build bilingual terminologies, we first compiled a parallel corpus by (1) extracting Japanese and English sentences from the municipal websites including body texts, headings and texts in tables, and (2) aligning the sentences between the languages.

We chose three website texts as sources: **CLAIR**,[2] **Shinjuku**[3] and **Hamamatsu**.[4] Each website covers a full range of categories for municipal life information such as residential procedures, tax payment and child care. The **CLAIR** website provides general purpose life information independent of particular municipalities, while the **Shinjuku** and **Hamamatsu** websites provide life information pertaining to the particular municipalities. It should be noted that the source texts of **Hamamatsu** are written in *Easy Japanese*, the lexicon and grammar of which are simplified in order to make the texts easier to read for non-native speakers of the language. We can reasonably assume that the three websites cover a wide range of content and linguistic phenomena.

We first extracted all sentences from the three sources, obtaining 16741 Japanese sentences and 15503 English sentences. We then manually aligned the Japanese and English sentences and obtained 15391 aligned sentence pairs,[5] from which we extracted bilingual term pairs.

2.2 Collecting Terms

2.2.1 Terms to be Collected

Our aim is to provide a practical terminology useful for authoring and (machine) translation. As Fischer (2010, p.30) pointed out, translators 'tend to consider terms in the broader sense, wishing to include everything which makes their work easier into a terminological database'. We thus decided to collect terms as widely as possible. The range of terms to be collected is defined as below.

1. Technical terms and proper nouns

   e.g. 外国人登録証明書/*gaikokujin-touroku-shomeisho* (alien registration card)
   e.g. JR 西日本/*JR-nishi-nihon* (JR West Japan)

2. More general words that refer to municipal services and activities

   e.g. 収入印紙/*shunyu-inshi* (stamp)
   e.g. 外交活動/*gaikou-katsudou* (diplomatic activity)

2.2.2 Extracted Terms

Ideally, the term extraction should be conducted by experts of the municipal domain. There is, however, a shortage of skilled municipal writers, and it is unrealistic to hire such experts. We thus employed four university students and asked them to manually extract bilingual term candidates from the parallel corpus. They are all native speakers of Japanese and have a sufficient command of English to correctly identify the translated terms.

In order to facilitate the extraction of terms, we developed a web-based platform to help with collaborative work. Figure 1 depicts the interface in which a pair of paralleled sentences is presented. This system enables users to capture the span of a term by clicking the starting word and the ending word.[6] At the bottom of the screen, terms that have been previously registered are also displayed. Registration of a pair of bilingual terms identical to an already identified pair is not allowed. These mechanisms support human decision-making and prevent duplicate registration, leading to improved efficiency of extraction.

[2]CLAIR (Council of Local Authorities for International Relations) Multilingual Living Information. `http://www.clair.or.jp/tagengo/`

[3]Shinjuku City, Living Information. `http://www.city.shinjuku.lg.jp/foreign/english/index.html`

[4]Hamamatsu City, Canal Hamamatsu. `https://www.city.hamamatsu.shizuoka.jp/hamaeng/`

[5]For some Japanese sentences, there were no corresponding English sentences, and vice versa.

[6]In this Figure, 'personal', the starting word of 'personal seal registration card', has been selected, and 'card', the ending word of the term, is about to be clicked.

Figure 1: Term registration platform

Another important feature of this platform is that it is designed to facilitate collaborative term extraction and validation. As soon as a user adds a comment to each pair of paralleled sentences and/or to each term, other users can refer to the comments and a task manager can promptly respond to the comment if necessary. The status of the work progress as well as the extracted terms can be checked online at any time, which helps conduct the task smoothly.

The identification of terms is difficult even for experts (Frantzi et al., 2000). To alleviate the individual differences of term identification and ensure comprehensiveness, we instructed the students to extract the terms as widely as possible. Finally, we validated all the terms they extracted to improve the accuracy of the terms.

A total of 3741 bilingual term pairs were collected from 15391 aligned sentence pairs. The number of distinct Japanese terms is 3012, while that of English terms is 3465, suggesting that in general the translated English terms are more varied than the Japanese source terms. This can be explained by the general tendency of greater inconsistencies in the translated terms, i.e. 'terminology inconsistencies often increase in frequency in the translated version compared to the original, due to the fact that there can be several ways to translate a given term or expression' (Warburton, 2015, p.649). She also pointed out an important factor leading to the terminology inconsistencies as follows:

> When a document or a collection of documents is divided into smaller parts which are translated by several translators, terminology in the target language will be more inconsistent than when only one translator is involved.

We can reasonably assume that several translators took charge of translating the municipal texts (terms) we deal with here, as the organisations in charge (CLAIR, Shinjuku City and Hamamatsu City) are different. Besides, the unavailability of bilingual municipal terminologies they can consult can aggravate the problem of terminology inconsistencies.

2.3 Controlling Term Variations

The range of the term variations to be addressed is dependent on foreseen applications (Daille, 2005). In this study, from the point of view of controlled authoring and MT, we cover a wide range of variations, including not only morphological and syntactic variations, but also synonyms and orthographic variations (Jacquemin, 2001; Yoshikane et al., 2003; Daille, 2003; Carl et al., 2004).

Investigating all the term pairs extracted from the corpus, we identified 374 Japanese term variations (12.4% of 3012 Japanese term types) and 1258 English term variations (36.3% of 3465 English term types). What we need to do next is to define preferred terms and proscribed terms in both Japanese

	Term	Dic.	Freq.	Typology
1	**健康診査/*kenkou shinsa***	✓	30	
2	健康診断/*kenkou shindan*	✓	5	
3	検査/*kensa*	✓	51	(A-1) omission
4	健診/*ken-shin*	✓	12	(C-1) initialism
1	health medical examination		1	(A-1) insertion
2	health check-up		17	(B-3) hyphen
3	medical check-up		3	(B-3) hyphen
4	medical examination	✓	10	
5	**health checkup**	✓	37	
6	check-up		14	(A-1) omission, (B-3) hyphen
7	health check		1	
8	physical check-up		2	(B-3) hyphen

Table 1: Examination of term variations

	(a) Uncontrolled types	(b) Controlled types	b/a	Tokens
Japanese	3012	2802	93.0%	15313
English	3465	2740	79.1%	15708

Table 2: The basic statistics of the controlled terminologies

and English (Warburton, 2014). We take into account the following three criteria to examine the variant terms:

Dictionary evidence: If a term is registered as an entry form in general dictionaries,[7] we regard it as preferable.

Frequency evidence: Higher frequency in the corpus is preferable.

Typological preference: The following types of variations are not preferable:[8] (A-1) omitting necessary information/inserting unnecessary information, (A-2) possessive case/personal pronouns, (B-1) emphasis symbols, (B-2) Kana characters, (B-3) hyphens, (C-1) initialisms/acronyms, (C-2) clipping and (D-1) transliteration.

Table 1 shows some examples of how each term meets each of the criteria. From this, we can define, for instance, '健康診査' as a preferred term since it is registered in the dictionary and also observed frequently (30 times) in the corpus, while the other three can be defined as proscribed terms. On the other hand, for the English translated terms, we can choose 'health checkup' as a standard translation. Though 'health check-up' (with a hyphen) is also frequently used in the corpus, we prefer 'health checkup' (without a hyphen) based on the typological preference policy (B-3) we adopted above.

Table 2 gives the basic statistics of our terminologies, showing the reduced number of term types after the variations were controlled. It can be noted that the number of English term types was reduced by about 20%, and the number of controlled term types in Japanese and in English became closer. This is not surprising because it is reasonable to assume that Japanese terminology and English terminology should contain the same size of *concepts* (or *referents*) in the parallel corpus. Controlling the variant forms of terms can be regarded as assigning one (authorised) linguistic form to one concept. We can estimate that the number of municipal concepts in our corpus is around 2700–2800.

We are now in the position to address the question: How do we evaluate the terminology and the controlled terminology we constructed? In the following sections, we propose a way to quantitatively evaluate the coverage of terminology and the quality of variation control, and evaluate our terminologies.

[7] In this study, we consulted the Sanseido Grand Concise Japanese-English Dictionary and the Kenkyusha New Japanese-English Dictionary.

[8] A: Syntax/morphology, B: Orthography, C: Abbreviation, D: Translation

3 Method for Evaluating Uncontrolled and Controlled Terminologies

To present the basic idea and framework of the evaluation, henceforth we use the following symbols based on Baayen (2001):

$V(N)$: number of distinct terms (number of types).
N: number of term occurrences in the corpus (number of tokens).
m: index for frequency class (m is an integer value).
$V(m, N)$: number of types that occur m times in the corpus.

3.1 Self-Referring Coverage Estimation

To estimate the coverage of the terminologies without using external terminologies, we employ the self-referring quantitative evaluation method proposed by Kageura & Kikui (2006). The basic idea is (1) to extrapolate the size of N to infinity using the observed data and estimate the saturation point, and (2) to evaluate the current status of the $V(N)$ in comparison with the saturation point.

While Kageura & Kikui (2006) estimated the coverage of the lexical items of a Japanese travel expression corpus, specifically focusing on the content words (nouns, verbs and adjectives), we assume this method can be applied to our task of estimating the coverage of the terms (mostly noun compounds) that appeared in our municipal corpus. They also emphasised that this method presupposes that the corpus qualitatively represents the whole range of relevant language phenomena in the given domain. Though the size of our municipal corpus itself is not large, it is possible to apply the method to our case, as the corpus focuses on a narrow domain (municipal life information) and covers a wide and well-balanced range of linguistic phenomena.

3.2 Conditions for Evaluation

We compare two parameters: (i) controlled and uncontrolled and (ii) Japanese and English. Thus, four conditions of terminology were prepared: (1) uncontrolled Japanese terminology, (2) uncontrolled English terminology, (3) controlled Japanese terminology, and (4) controlled English terminology.

To estimate the coverage of *terms*, we investigate the uncontrolled conditions. Our previous observations showed that uncontrolled English terminology is more varied than uncontrolled Japanese terminology, which may affect the population size of the terminologies. On the other hand, investigating the controlled conditions is important to see the coverage of *concepts* in the domain.

From the point of view of validating how well our terminologies are controlled, we explore the controlled conditions of the terminologies. Our hypothesis is that if the terminologies are well controlled, the estimated population number of Japanese and English term types become closer, as both represent the same set of concepts.

3.3 Expected Number of Terms

A number of methods have been proposed to estimate the population item size (Efron and Thisted, 1976; Tuldava, 1995; Baayen, 2001). Here we adopt Large-Number-of-Rare-Event (LNRE) modelling, which has been used in the field of lexical statistics (Khmaladze, 1987; Baayen, 2001; Kageura, 2012). We outline the computational steps behind the method, following Baayen (2001).

Let the population number of types be S and let each type be denoted by w_i ($i = 1, 2, ..., S$). With each w_i population probability p_i ($i = 1, 2, ..., S$) is associated. Using the binomial theorem, we can express the expected number of types that occur m times in a sample of N as follows:

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

At the final step of (1), the Poisson approximation with parameter $\lambda = np$ is applied.

In order to express $E[V(N)]$, the expected number of types, we focus on the types that do not occur. Taking the complement of the probability that type w_i does not occur in the sample N tokens, we obtain

the probability that w_i occurs at least once in the sample N. Hence, the $E[V(N)]$ is given as follows:

$$E[V(N)] = \sum_{i=1}^{S}(1 - \binom{N}{0} p_i^0 (1-p_i)^{N-0}) = \sum_{i=1}^{S}(1 - e^{-Np_i}). \quad (2)$$

Note that the Poisson approximation is used again in the last step of (2).

For mathematical convenience, we rewrite the Poisson models in integral forms using the structural type distribution $G(p)$, the cumulative number of types with probabilities equal to or greater than p, which is defined as follows: $G(p) = \sum_{i=1}^{S} I_{[p_i \geq p]}$, where $I = 1$ when $p_i \geq p$, and 0 otherwise. We can renumber the subscript of p for $p_j > 0$, such that $p_j < p_{j+1}$ $(j = 1, 2, ..., \kappa)$. As $G(p)$ is a step function, jumps at the probabilities p_j, in other words, the number of types in the population with probabilities p_j, are given by $\Delta G(p_j) = G(p_j) - G(p_{j+1})$. We can now restate the equations (1) and (2):

$$E[V(m,N)] = \sum_{j=1}^{\kappa} \frac{(Np_j)^m}{m!} e^{-Np_j} \Delta G(p_j) = \int_0^{\infty} \frac{(Np)^m}{m!} e^{-Np} dG(p). \quad (3)$$

$$E[V(N)] = \sum_{j=1}^{\kappa} (1 - e^{-Np_j}) \Delta G(p_j) = \int_0^{\infty} (1 - e^{-Np}) dG(p). \quad (4)$$

Using some hypotheses about the form of distributions such as inverse Gauss-Poisson distribution, we can obtain models to extrapolate the $V(N)$ and $V(m,N)$ for $N \to \infty$.

3.4 Growth Rate of Terms

The constructed model also gives us insight into the growth rate, or how fast the number of types increases as we extract more terms from texts in the domain. The growth rate is obtained by taking the derivative of $E[V(N)]$ as follows:

$$\frac{d}{dN} E[V(N)] = \frac{d}{dN} \int_0^{\infty} (1 - e^{-Np}) dG(p) = \frac{1}{N} \int_0^{\infty} Npe^{-Np} dG(p) = \frac{E[V(1,N)]}{N}. \quad (5)$$

4 Results and Discussions

4.1 Population Types and Present Status of Terminologies

Table 3 gives the estimated population number of term types $E[S]$, together with the coverage ratio CR $(= V(N)/E[S])$.

Though there are several models of LNRE, we chose the following two models, which were shown to be effective in this estimation task: Generalised Inverse Gauss-Poisson (GIGP) model (Sichel, 1975) and finite Zipf-Mandelbrot (fZM) model (Evert, 2004; Evert and Baroni, 2005).[9]

The lower χ^2-value and higher p-value indicate a better fit of the LNRE model, and Baayen (2008, p.233) remarks that a p-value above 0.05 is preferable. Though all of the p-values are below 0.05, the χ^2-values are not bad compared to the related work by, for example, Kageura (2012) or Baayen (2001), so we can reasonably assume that the estimation results are meaningful.

The estimated population size $E[S]$ ranges from 4299 to 7616, and the coverage ratio CR ranges from 42.7% to 64.0%. Though the values of $E[S]$ and CR depend on the models used,[10] we can observe several important points of the result.

Firstly, focusing on the uncontrolled terminologies, we recognise very different results between Japanese and English: the population number of types of Japanese, 5505 (GIGP) and 4626 (fZM), is much smaller than that of English, 7616 (GIGP) and 6083 (fZM). Consequently, the coverage ratio of Japanese is generally higher than that of English. This may reflect the higher diversity of the uncontrolled English terminology. As we have seen in Section 2.3, the ratio of variations in the English uncontrolled

[9] Though we tried two other LNRE models, the lognormal model (Carroll, 1969) and the Yule-Simon model (Simon, 1960), the fit of the models to our data was not good compared to the GIGP and fZM models, so we did not adopt these models.

[10] For all conditions, the fZM model produced higher values of $E[S]$ than the GIGP model.

		Model	$E[S]$	$V(N)$	$CR(\%)$	χ^2	p
Uncontrolled	Ja	GIGP	5505.3	2953	53.6	35.260	0.0008
		fZM	4626.2	2953	63.8	33.930	0.0012
	En	GIGP	7616.4	3255	42.7	23.857	0.0325
		fZM	6083.0	3255	53.5	28.197	0.0085
Controlled	Ja	GIGP	5111.9	2753	53.9	34.620	0.0010
		fZM	4299.0	2753	64.0	27.905	0.0093
	En	GIGP	5380.2	2611	48.5	35.354	0.0007
		fZM	4444.5	2611	58.7	36.525	0.0005

Table 3: Population types $E[S]$ and coverage CR

terminology is much higher than that in the Japanese one, which suggests the potential diversity of translated English terminology in the population.

Secondly, the controlled terminologies tend to exhibit a lower $E[S]$ and higher CR than the uncontrolled terminologies. For example, the CR of controlled terminology when fZM is adopted is 64.0% for Japanese and 58.7% for English, which means that around two thirds of the concepts in the domain are included in our terminologies. It is worth noting that the coverage of the controlled terminologies exceeds that of the uncontrolled ones. This result is fairly good as a starting point and encourages the practical use of the terminologies.

Finally, related to the second point, the differences of $E[S]$ and CR values between Japanese and English in the controlled conditions are much smaller than those in the uncontrolled conditions. In principle, the (population) size of the concepts in the parallel data of a given domain should be the same across the languages. The closer values of $E[S]$ between Japanese and English demonstrate that our constructed terminologies have a desirable nature. We should, however, remain aware that there are still differences between the Japanese and English controlled terminologies. We believe this is mainly because (1) the English translated sentences in the parallel corpus are sometimes not word-for-word translations of the original Japanese sentences, which may affect the distribution of terms in the corpus, and (2) the term variation control was performed solely by the authors of the paper, i.e. native speakers of Japanese, and there is still room for improvement in English term variation control.

4.2 Growth Patterns of Terminology

From the practical point of view, it is impossible to observe an infinite size of N within the limited textual data that is available. Our next question is to what extent we can enlarge the size of the terminologies and extend their coverage within the realistic range. To address this question, we take a closer look at the dynamic trends of the terminology growth.

We first observe how the expected number of term types $V(N)$ shifts as the number of term tokens N increases. Figure 2 draws for each LNRE model the growth curves of $V(N)$, as N grows to 100000, which is approximately 6.5 times as large as the present N.[11] The vertical dotted line indicates $N = 15000$, which is close to the present N.

Comparing the growth curves of the four conditions, we can easily recognize the general tendencies that conform to what we pointed out in Section 4.1. We summarise them as follows:

1. The English uncontrolled terminology grows more rapidly than the Japanese one.
2. The controlled terminologies shows more moderate growth than the uncontrolled ones.
3. The growth curves of the controlled Japanese and English align very closely.

The growth curves also enable us to visually grasp the shift of the growth rate. We can observe that all of the curves grow rapidly in the beginning and become gentler when N reaches around 30000, about twice the size of the present N. Although within the size of 100000, all the growth curves do

[11] Note that the actual present N is 15313 for Japanese and 15708 for English as shown in Table 2.

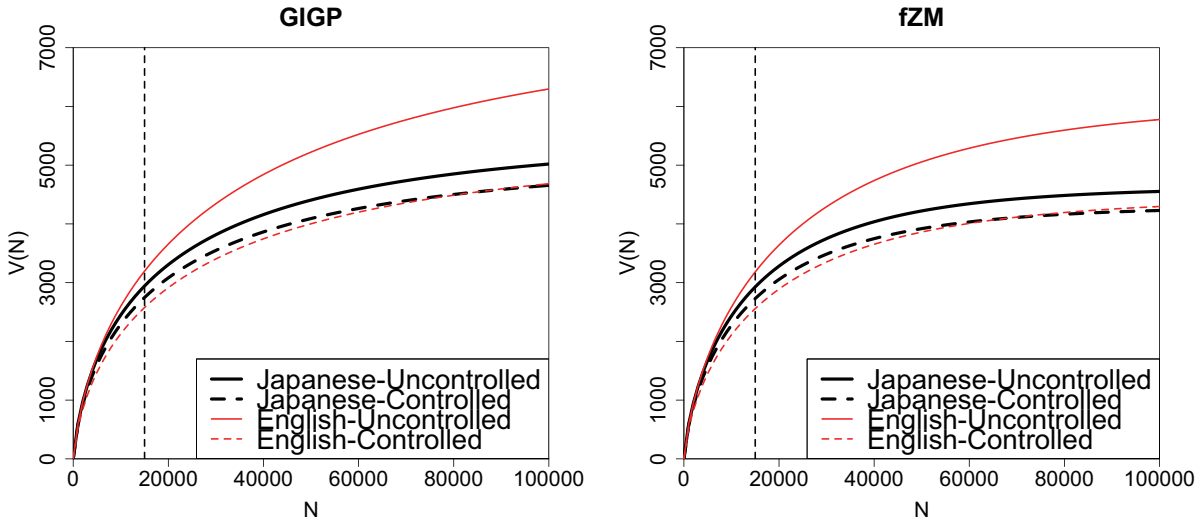

Figure 2: Growth curves of the terminologies

			0.5N		N		1.5N		2N		2.5N		3N	
			CR	GR	CR	GR	CR	GR	CR	GR	CR	GR	CR	GR
Uncont.	Ja	GIGP	39.0	0.144	53.9	0.082	63.2	0.055	69.8	0.040	74.6	0.030	78.4	0.024
		fZM	45.9	0.146	63.8	0.081	74.5	0.051	81.5	0.035	86.4	0.024	89.8	0.018
	En	GIGP	29.9	0.162	42.9	0.100	51.6	0.072	58.1	0.055	63.1	0.044	67.2	0.036
		fZM	37.2	0.163	53.5	0.099	64.2	0.069	71.8	0.050	77.4	0.038	81.8	0.030
Cont.	Ja	GIGP	39.4	0.133	54.2	0.075	63.4	0.051	69.9	0.037	74.7	0.028	78.4	0.022
		fZM	46.3	0.134	64.0	0.074	74.6	0.047	81.5	0.032	86.3	0.023	89.8	0.016
	En	GIGP	35.0	0.125	48.9	0.074	57.8	0.051	64.3	0.038	69.2	0.030	73.1	0.024
		fZM	41.9	0.126	58.7	0.073	69.3	0.049	76.5	0.035	81.8	0.025	85.7	0.019

Table 4: Shift in the coverage ratio (CR: %) and the growth rate (GR)

not seem flattened out, we can gain insight into how to effectively extend the size of the terminologies. Considering the difficulty in compiling bilingual (or multilingual) parallel municipal corpora on a large scale, we further restrict ourselves to a realistic size of N. Table 4 shows the shift in the estimated coverage ratio CR and the growth rate GR at $0.5N$ intervals up to $3N$ (about 450000 tokens). These two measures give us different perspectives for the terminology extension.

CR is a goal-oriented measure, which tells us how much addition of term tokens (or texts) is needed to attain a certain coverage of the potential terminology in the domain. If we double the token size N, we achieve nearly 80% coverage of the Japanese terms, 70% coverage of the English terms and 80% coverage of the concepts in the domain (when estimating by fZM), showing an increase of more than 15% compared to the original size N. If we treble N, we achieve an additional increase of at most 10% in the coverage ratio, with some of the values reaching nearly 90%. Setting goals for terminological (lexical) development is crucial in practical applications such as MT dictionary development (Dillinger, 2001; Kim et al., 2005). Using this measure, we can set the goal of terminology construction in terms of coverage.

GR is an ROI (return on investment)-oriented measure, which tells us how much addition of term tokens (or texts) is needed to obtain a new term or concept. At the current size of the terminologies, to obtain a new term type, 12 ($\approx 1/0.08$) term tokens should be added to the Japanese terminology, and 10 ($= 1/0.10$) to the English terminology. To obtain a new concept, 14 ($\approx 1/0.07$) term tokens should be added in the Japanese or English terminology. When we reach the token size of $2N$, to obtain a new term, 25 Japanese term tokens and 20 English term tokens should be added, showing the reduced efficiency in enlarging the terminologies as we examine more term tokens. This estimation enables us to decide when to stop collecting term tokens/texts in terms of cost effectiveness.

5 Conclusion and Future Work

In this study, we constructed controlled bilingual municipal terminologies and evaluated their status. The outcomes and contributions of this study are summarised as follows:

1. Using the term collection tool we developed, we efficiently extracted 3741 Japanese-English term pairs from a municipal text corpus. We then controlled the term variations by defining the preferred and proscribed terms to construct controlled bilingual terminologies.
2. The evaluation results showed that our terminology currently covers (1) about 55–65% (Japanese) and 45–55% (English) of the terms and (2) about 55–65% (Japanese) and 50–60% (English) of the concepts in the municipal domain. Also, the closer values of the population number of the term types and the similar shapes of the terminology growth curves for Japanese and English demonstrated that our terminologies are well controlled.
3. We proposed a method to evaluate the coverage of terminology. Though the self-referring method employed in this paper has difficulty in obtaining a good fit of the model for the observed data, we consider our method to respond to the practical need for estimating the potential size of terminology.

As future work, we plan to utilise the terminologies in our controlled authoring and MT environment, and evaluate their effectiveness and utility. We are now developing a real-time interactive terminology checker that detects term variations in the source text and suggests a preferred term (Miyata et al., 2016). The list of synsets of preferred terms and proscribed terms constructed in this study will be implemented in the checker. Furthermore, controlled authored source texts can be consistently translated by MT systems if their user dictionaries register pairs of preferred source and target terms.

We will also expand the size of our terminologies. Based on the estimation presented above, to achieve about 80–90% coverage of municipal terms and concepts, we need to check 15000–30000 more term tokens. At this stage, automatic term extraction (ATE) would be a viable option to efficiently collect term candidates (Itagaki et al., 2007; Macken et al., 2013; Aker et al., 2013; Kilgarriff et al., 2014). We also intend to adopt a 'generate and validate' method (Sato et al., 2013), which makes use of constituents of terms to obtain new term candidates. The terminologies constructed in this study enable us to employ this method.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 16J11185, the Research Grant of Tokyo Institute of Technology, and the Research Grant Program of KDDI Foundation, Japan.

References

Ahmet Aker, Monica Paramita, and Robert Gaizauskas. 2013. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 402–411, Sofia, Bulgaria.

Harald Baayen. 2001. *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

Harald Baayen. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press, Cambridge.

Michael Carl, Ecaterina Rascu, Johann Haller, and Philippe Langlais. 2004. Abducing term variant translations in aligned texts. *Terminology*, 10(1):101–130.

John B. Carroll. 1969. A rationale for an asymptotic lognormal form of word-frequency distributions. In *Research Bulletin*. Educational Testing Service, Princeton, New Jersey.

Béatrice Daille. 2003. Conceptual structuring through term variations. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment (MWE 2003)*, pages 9–16, Sapporo, Japan.

Béatrice Daille. 2005. Variations and application-oriented terminology engineering. *Terminology*, 11(1):181–197.

Mike Dillinger. 2001. Dictionary development workflow for MT: Design and management. In *Proceedings of the Machine Translation Summit VIII*, pages 83–88, Galicia, Spain.

Bradley Efron and Ronald Thisted. 1976. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447.

Stefan Evert and Marco Baroni. 2005. Testing the extrapolation quality of word frequency models. In *Proceedings of the Corpus Linguistics 2005*, Birmingham, UK.

Stefan Evert. 2004. A simple LNRE model for random character sequences. In *Proceedings of the 7es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2004)*, pages 411–422, Louvain-la-Neuve, France.

Márta Fischer. 2010. Language (policy), translation and terminology in the European Union. In Marcel Thelen and Frieda Steurs, editors, *Terminology and Lexicography Research and Practice: Terminology in Everyday Life*, volume 13, pages 21–34. John Benjamins, Amsterdam.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Masaki Itagaki, Takako Aikawa, and Xiaodong He. 2007. Automatic validation of terminology translation consistency with statistical method. In *Proceedings of the Machine Translation Summit XI*, pages 269–274, Copenhagen, Denmark.

Christian Jacquemin. 2001. *Spotting and Discovering Terms through Natural Language Processing*. The MIT Press, Cambridge.

Kyo Kageura and Genichiro Kikui. 2006. A self-referring quantitative evaluation of the ATR Basic Travel Expression Corpus (BTEC). In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1945–1950, Genoa, Italy.

Kyo Kageura. 2012. *The Quantitative Analysis of the Dynamics and Structure of Terminologies*. John Benjamins, Amsterdam.

Estate V. Khmaladze. 1987. *The Statistical Analysis of Large Numbers of Rare Events*. Technical Report MS-R8804, Department of Mathematical Sciences, CWI, Amsterdam.

Adam Kilgarriff, Miloš Jakubíček, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2014. Finding terms in corpora for many languages with the Sketch Engine. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 53–56, Gothenburg, Sweden.

Young-Gil Kim, Seong-Il Yang, Munpyo Hong, Chang-Hyun Kim, Young-Ae Seo, Cheol Ryu, Sang-Kyu Park, and Se-Young Park. 2005. Terminology construction workflow for Korean-English patent MT. In *Proceedings of the Machine Translation Summit X*, pages 55–59, Phuket, Thailand.

Philippe Langlais and Michael Carl. 2004. General-purpose statistical translation engine and domain specific texts: Would it work? *Terminology*, 10(1):131–153.

Lieve Macken, Els Lefever, and Véronique Hoste. 2013. TExSIS: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology*, 19(1):1–30.

Rei Miyata, Anthony Hartley, Kyo Kageura, Cécile Paris, Masao Utiyama, and Eiichiro Sumita. 2016. MuTUAL: A controlled authoring support system enabling contextual machine translation. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016), System Demonstrations*, Osaka, Japan. (forthcoming).

Juan C. Sager. 2001. Terminology compilation: Consequences and aspects of automation. In Sue Ellen Wright and Gerhard Budin, editors, *Handbook of Terminology Management*, volume 2: Application-Oriented Terminology Management, pages 761–771. John Benjamins, Amsterdam.

Koichi Sato, Koichi Takeuchi, and Kyo Kageura. 2013. Terminology-driven augmentation of bilingual terminologies. In *Proceedings of the Machine Translation Summit XIV*, pages 3–10, Nice, France.

Herbert S. Sichel. 1975. On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351a):542–547.

Herbert Simon. 1960. Some further notes on a class of skew distribution functions. *Information and Control*, 3(1):80–88.

Lori Thicke. 2011. Improving MT results: A study. *Multilingual*, January/February:37–40.

Juhan Tuldava. 1995. *Methods in Quantitative Linguistics*. Wissenschaftlicher Verlag Trier, Trier.

Kara Warburton. 2014. Developing lexical resources for controlled authoring purposes. In *Proceedings of LREC 2014 Workshop: Controlled Natural Language Simplifying Language Use*, pages 90–103, Reykjavik, Iceland.

Kara Warburton. 2015. Terminology management. In Sin-Wai Chan, editor, *Routledge Encyclopedia of Translation Technology*, pages 644–661. Routledge, New York.

Fuyuki Yoshikane, Tsuji Keita, Kyo Kageura, and Christian Jacquemin. 2003. Morpho-syntactic rules for detecting Japanese term variation: Establishment and evaluation. *Journal of Natural Language Processing*, 10(4):3–32.