Cross-lingual Correspondences of Terms in Texts and Terminologies: Theoretical Issues and Practical Implications

Kyo Kageura
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
kyo@p.u-tokyo.ac.jp

Abstract

Terms are items in language that represent concepts. This relation of representation does not change through use. As such, terms have a unique status in language, second only to proper names. Due to this, clarifying the identity of concepts represented by terms becomes an important issue at the level of what is represented, and control of terms representing the same concept also becomes an important issue at the level of representation. These problems with which terminologists are concerned, though not clear at first glance, are in fact relevant to general words and vocabulary to a lesser extent. In this paper I first clarify theoretical issues of terms and terminologies and what they imply for terminology processing in particular and lexical and lexicological processing in general. I then pick up some terminological applications, examine their status and suggest a few issues that can be addressed in terminology processing.

Keywords: Terminology, Concept, Comparability

1. Background

1.1. Concepts, Knowledge and Terminology

Let me start this paper with a rather theoretical discussions.

Forgeries do not destroy science. Science is destroyed when people, including "scientists," start regarding claims and "arguments" based on forged or fake data as part of science. That we can safely assume that the concept and act of science, in its proper sense, exists and is shared enables us, not only practically but also *logically*, to identify what are to be identified as forgeries as forgeries.

An argument homomorphic to this holds for the changes in the meaning of words in general. When we say a word changes its meaning in accordance with its use, we *logically* presuppose the existence of the *identity* of meaning of the word. Otherwise we cannot talk about the meaning or a meaning or meanings of a word in the first place. This logical identity indeed restricts the *practical* range of changes in the meaning of a word: whenever I have responded "oh, yes, the meaning of a word is sweet and tasty, but it's too expensive" to a person who has asserted that "the meaning of a word changes in use," they have always been puzzled. In other words, the meaning of a word does not change beyond a certain limit, which reflects, at least within a certain range of duration, the identity of the meaning. One can say with confidence that the meaning of a word changes as long as – and precisely because – the underlying identity of the meaning of a word remains intact.

While this *identity* of the meaning tends to work implicitly in the background in the case of general words, it is one of the main and explicit concerns for technical terms. Crudely speaking, it is this *identity* represented by a term that is referred to as a *concept*. Though it is not easy to recognise the essential difference between the relationship between concept and term on the one hand and the relationship between meaning and word on the other (Kageura, 1995), especially when terms and words are handled in practical setups as in compiling dictionaries or terminologies, there is a logical necessity for terminologists to talk about concepts represented by terms rather than meanings of terms.

What is more, this concept-term relation as distinct from meaning-word relation constitutes a part of the essential language infrastructure that supports social construction and organisation, and issues related to this relation can cause practical – and sometimes serious – problems in our social life. A while ago, when US-based insurance companies started operating in Japan, the difference in the definition of "cancer" caused trouble in the application of insurance policies[1]. As this is a cross-lingual case, it is easily noticed that the issue is not to do with the change in the meaning in the process of use, but with the concept referred to by corresponding terms.

Now let us consider the following example:

> Responsibility accompanies freedom.

This clause is written in the draft revision to the Japanese constitution proposed by Liberal Democratic Party, which is the governing party of Japan as of this writing. How should we behave in the face of this statement? If one adopts the stronger version of the Firthian view of meaning, one must accept that freedom should be accompanied by responsibility, although to what degree one must accept that depends on how widespread this discourse is. From the point of view of terminology, this statement is just *false* from start to finish, simply in terms of the concept represented by the term "freedom". Freedom includes such passive forms of freedom as freedom from torture (Berlin, 1969). If we apply the LDP statement to the concept of freedom from torture, we end up with the following:

> If you do not take due responsibility, you may not
> be free from torture.

This reveals the following essential fact about the concept of "freedom":

[1] Personal communication with Professor Kazuhiko Ohe, Graduate School of Medicine, The University of Tokyo.

> That responsibility does not accompany freedom is the *sine qua non* trait of the very concept of "freedom," without which this word is nullified and we cannot talk about "freedom" at all.

So the statement "responsibility accompanies freedom" should not change the concept of "freedom." If such abuse of language spreads, however, it may become impossible to talk about freedom. In such a situation, we are not talking about the changes in the meaning of "freedom" as it becomes nothing to do with freedom if responsibility accompanies it. This is tantamount to killing the concept of freedom, and this is tantamount to killing the conditions which enable us to maintain the concept of freedom. Incidentally, *learning* for human being is not related to accepting the statement "responsibility accompanies freedom" as part of the determining feature of the concept of freedom, but to gain a system of judgement that enables one to properly identify this statement as false. The former is relevantly called *disciplinisation* or *indoctrination*, which is not – and indeed is the complete opposite of – learning.

It is often the case that the concepts represented by terms are not constitutively accessible and can only be *presumed* as a regulatory ideal (Kant, 1781). In other words, the identity of the concept represented by a term may not be described fully. But this does not mean that the identity of the concept does not exist and everything depends on usage. Reflecting this theoretical status of concepts and terms, practical study of terminology is also concerned with the identity of concepts.

1.2. Machine Learning/Disciplinisation

One of the standard ways of handling the "meaning" of words is word embedding or distributed representation of words. That representations obtained by `word2vec` enabled such operations as follows showed the power of distributed representation of words (Mikolov et al., 2013):

$$\text{Madrid} - \text{Spain} + \text{France} = \text{Paris}.$$

In the same manner, it is pointed out that the following also becomes possible:

$$\text{Doctor} - \text{Male} + \text{Female} = \text{Nurse}$$[2]

We can immediately see the qualitative difference between these two cases, i.e. the former reflects the relationships among the meanings of these words, while the latter has nothing to do with the meanings of "doctor," "nurse," "female," or "male." and just reflects gender biases that exist in society and in social discourse. We can also recall what happened to Microsoft Tay, soon started tweeting about its admiration for Hitler and using racist slurs against Jewish and black people. Using the term we introduced above, we have to say that machines did not learn, but rather were disciplinised or indoctrinated[3].

Can corpus-based or data-oriented terminology processing get around these or similar issues? We have been (mostly unconsciously) assuming yes, for the following reasons:

- Specialised knowledge is created and expressed in the proper manner, and biases are filtered out through peer review in each specialised domain of knowledge;
- Popularisation and wider dissemination of specialised knowledge is also carried out in a due manner, reducing the granularity of discourse but essentially keeping the wholesomeness of the specialised knowledge.

Assuming these hold, we can safely use domain corpora for a narrower or wider range for different domains in different languages, even if machines can only be disciplinised and cannot learn in the proper sense of this word.

Unfortunately, however, a range of recent events indicate that relying on these assumptions is becoming more and more dangerous:

- Forgeries have repeatedly come to light and a number of papers have been retracted;
- Some authors have tried to cheat journal editors by supplying fake e-mail addresses for real scientists as potential reviewers;
- Unfounded historical revisionism and views based on such revisionism has appeared in descriptions of history in some school textbooks in Japan (and perhaps in other countries as well);
- Funding bodies require more and more short-term social "impact";
- Mass media pick up more and more sensational aspects of research with improper use of terms.

Together, these blur the distinction between scientific activities which are carried out in accordance with established norms of science and those activities that are not. Recall that science is destroyed when people, including "scientists," start regarding claims and "arguments" based on forged or fake data as part of science.

In such a situation, automatic terminology processing may contribute to the destruction of science through unconsciously extracting the abuse of concepts as normal and spreading them. Daille once argued for the necessity of detailed text profiling (Daille, 2008). If we start from corpora or textual data, text profiling becomes more and more important. Theoretically, however, the relation between concepts (and terms) and texts is the other way round. Texts are constructed in such a way that they make proper sense and concepts and terms are assumed beforehand. Text profiling is concerned with providing machines with appropriate information while assuming that machines are disciplinised rather than that they learn. Can we add the ingredient of learning rather than only avoid inappropriate disciplinisation? What does this mean?

This is the situation which terminology processing currently is facing. Having this in mind, I introduce some practical terminological tasks and some trials. In fact, since the mid-1990s, at the background of terminology processing, I have kept thinking of these issues. Words are grandiose, deeds are miserably tiny. Worse still, the practical tasks introduced below are only remotely related to what we have discussed so far. But let us move on anyway.

[2] An example cited in the Q&A session for Steedman, M., "On distributional semantics," invited talk at the Australian Language Technology Association 2016 Workshop.

[3] I owe this recognition to Dr. Hideto Kazawa of Google.

2. Issues in Terminology Processing

2.1. Terminology and Textual Corpora

Research in and the practice of terminology as an independent area of activity was first consolidated in Wüster's seminal work (Wüster, 1959), in which he put emphasis on the rigidity of concepts and terms. Felber states that terminology starts with concepts rather than terms, is concerned with the system of concepts in its synchronic state, and is not concerned with the linguistic features of terms that are unrelated to concepts (Felber, 1984).

In terminology, terms and concepts are defined as follows (de Besse et al., 1997):

term: A lexical unit consisting of one or more than one word which represents a concept inside a domain.

concept: An abstract unit which consists of the characteristics of a number of concrete or abstract objects which are selected according to specific scientific or conventional criteria appropriate for a domain.

Kageura showed that, theoretically, terminology as a coherent set of terms conceptually precedes individual terms; terms are items within a terminology which in its totality reflects the conceptual system of a domain (Kageura, 2015). Two features of concepts and terms can be pointed out here:

1. A concept represented by a term may be updated, but does *not change* through casual use. This update of the concept is understood as a step towards the ideal state of that concept, which exists as a regulatory ideal.
2. Terms and the concepts they represent are attributed to the system of knowledge of the domain.

Since the 1990s, more descriptive approaches have appeared (Budin and Oeser, 1995; Temmerman, 2000). While these approaches have advanced *how* concepts and terms can be described, understanding of *what* concepts are seems to have remained intact behind the scenes[4]. The Wüstarian view of terms has always been there as the regulatory ideal for terminology. This also holds for corpus-based automatic terminology processing. After all, without this regulatory ideal, we do not need to and we cannot talk about terms and terminologies as something different from ordinary words, compounds or collocations anymore.

In corpus-based terminology processing such as monolingual and bilingual automatic term extraction, this regulatory ideal that links the work to terminology is implicitly taken into account when domain corpora are defined. Domain corpora are the discoursal part of the linguistic representation of the system of knowledge of the domain. This discoursal part, to be relevant, makes use of the terminological part, which is the other part of the linguistic representation of the system of concepts and knowledge of the domain. Though every now and then concepts are updated through discourse, specialised discourse at the same time critically depends on the system of concepts and the corresponding terminology.

Thus term extraction thus should *not* be the task of extracting linguistic elements that are relevant to a given set of texts or domain corpora; it is the task of extracting terminology that represent a system of concepts and thus the system of knowledge of the domain *through* domain corpora. This contrasts with keyword extraction, which is defined as the task of extracting linguistic elements that are relevant to texts. One can extract keywords from a document which consists only of fake information and the extracted keywords can be valid, but one cannot extract terms from such a document.

This is the theoretical reason why text profiling becomes critical in corpus-based terminology processing (Daille, 2008). The practical result that text profiling can improve the performance of such tasks as bilingual term extraction (Morin et al., 2010) can be a reflection of this theoretical point. For text profiling, we can resort to external information at a variety of levels, such as the reliability of authors, of institutions authors are affiliated with, of journals, thus of publishers, or of the format of documents, etc. Unfortunately, it is not sufficient. We can see this from the example we observed above, i.e. the planned insertion of the statement "responsibility accompanies freedom" into the Japanese constitution. The agent trying to do this is the governing party and once inserted the statement will constitute a part of the Japanese constitution. In view of the external criteria, this statement is to be regarded as "reliable," even if it is nonsense. Ultimately, therefore, we need to resort to knowledge *itself* to avoid this sort of misjudgement. But how? Note that here the problem has gone beyond text profiling.

2.2. Conceptual Systems and Normativity

Two clues exist that guide us when dealing with this issue, though neither of them provides us with direct solutions to the problem we have discussed so far.

First, at a certain stage in the process of learning, human beings start judging information or a chunk of knowledge that is given to them and start refusing to accept it. This is because they have nurtured their *system* of belief, which is supported by the *system* of knowledge. One of the core parts of this system of knowledge is a vocabulary, which is not just a set of words but "a coherent, integrated system of concepts" (Miller, 1986). In the arena of sciences, the most basic part of this system of knowledge is reflected in terminology, which represents a coherent, integrated system of the concepts of the domain. A system of concepts is not just a set of concepts randomly collected. It embodies normativity, to the extent that we can talk about degree of systematicity and whether something is relevant to the system or not. Explicitly dealing with the terminology as a reflection of the system of concepts rather than dealing with individual terms or a set of given terms, therefore, can be a step towards properly handling terms, terminology and concepts, i.e. dealing with terms consistently and systematically in such a way that they collectively reflects the meaningful part of the system of concepts of the domain.

Let me cite an example here, though it is not terminological. Suppose we are interested in extracting words from textual corpora to construct a dictionary. Suppose that we

[4] This perception may bring us back to Frege but we do not elaborate on this further here.

extracted a set of words from a corpus of 10,000 word tokens, and obtained two words that indicate types of fruit, i.e. "orange" and "apple". From the point of view of constructing a dictionary, given the range of words referring to fruit that are used in daily life in many English-speaking areas in the world, it is most natural that a dictionary which includes "apple" and "orange" as entries would also have "banana" as an entry.

To obtain the word "banana" from the corpus, we may have to extend the corpus to 100,000 word tokens. We would then obtain "banana", but would also obtain "mango" and "kiwi fruit". We would most probably think that a dictionary that contains "mango" and "kiwi fruit" as entry words should also have "papaya" as an entry. Otherwise, the set of entries lacks systematicity and coherency. To obtain "papaya", we may have to extend the size of the corpus to, say, 1,000,000 word tokens. In addition to "papaya," then, we would obtain "kiwano" and "star fruit," in which case we would need "dragon fruit" to make the set of entries in the dictionary coherent and systematic. This is the so-called "orange, apple, banana problem" [5].

Although this description is imaginary, a situation equivalent to it can happen in real-world dictionary-making situations. Kilgarriff et al. (2014) found that, in a project that aimed at developing monolingual and bilingual word lists for language learning using corpora, for nine languages and thirty-six language pairs, it was preferable to define a set of common key domains and populate the domains with words independently for each language. As domains they defined calendar, i.e. days of week, months, time, celebrations, colours, clothes, numbers, etc (Kilgarriff et al., 2014). This is partly because there is no guarantee that all the names for the days of week exist in a given corpus. This also indicates that the system of vocabulary or terminology is not a secondary, artificial derivation while discourse and texts are the first-hand manifestation of languages (Wilks et al., 1996).

The concept of normativity is also relevant to the textual or discoursal sphere. For instance, we do not refer to the New York Times, let alone AmericanNews.com, to make a legally *learned* argument about the Indonesian invasion of East Timor in 1975[6]. We refer to binding international law and authentic political records[7]. Indeed, researchers, irrespective of their research area, should be fully aware of what is called normativity here; they refer mostly to peer-reviewed and other academically reliable papers. They do not regard these papers and anonymous blog posts as equivalent. This observation is closely related to the proposal of text profiling mentioned above.

Documents thus have a degree of normativity. This concept is also frequently valid within cross-lingual setups. We often observe that a bilateral contract made between institutions in different countries in two languages adds such a statement as "in case of discrepancies between the versions in two languages, a preference is given for the interpretation according to a version in which the contract was originally drawn up." In bilingual or multilingual situations, it is usually the case that the document is written in one language, which is the source language. The corresponding documents in other languages are created through translation. This implies that, not infrequently, when context vectors for corresponding terms in different languages have some gaps, they are not relative to each other. We sometimes need talk about *deviations* of the usage of a term or the concept represented by a term, as in the case of the definition of cancer in insurance policies.

Linguists may say that normativity of terms and documents is not inherent in languages. May be true, but terms and terminologies are the functional class of languages and the determining factor is social and/or conceptual, which are not linguistic in the first place. What we see is that linguistics in its narrower sense falls short of addressing the issue we have observed so far. I see no merit to sticking to the purity of linguistics or whatever that cannot counter the destruction of the very conditions which enable us to sensibly communicate with each other, without resorting to physical violence. Freedom, in its essence, is never accompanied by responsibility, even if 99 percent of people claim that this is so. The rights of individuals, in their essence, are never accompanied by duties, even if the governing party of a nation declares this to be so. The concepts of freedom and rights should be properly maintained, logically, even when oppressive and discriminative discourse becomes prevalent.

3. Directions in Terminological Research

We can define a range of terminological studies and terminology-processing tasks that take into account the concepts of normativity and/or systematicity, both in monolingual and in bilingual or multilingual situations. To do so, we can conveniently distinguish two phases of terms and terminology: individual terms and concepts they represent as they are and in their use in texts, and the system of terminology and conceptual system.

3.1. Terms, Concepts and Use of Terms

If we focus on individual terms, their relation with concepts is the point of central importance, as has been pointed out by theoretical terminologists. Terminologists pay great attention to how to define concepts properly. Note that terminological lexicons with proper entries and reliable definitions are used as a resource that people commonly refer to and attain normative status. Normativity inevitablly accompanies tasks dealing with concepts represented by terms. Referring to terminologists and other specialists activities, we can define, for instance, the following tasks as taking into account the issues we raised in the previous section:

[5] Personal e-mail communication with the late Dr. Adam Kilgarriff on April 1, 2014. Though the conversation took place on April Fool's Day, the content was academic.

[6] The Indonesian invasion of East Timor began on 7 December 1975, one day after then U.S. Secretary of State Henry Kissinger left Jakarta.

[7] The National Security Archive of George Washington University revealed the conversation between then U.S. President Gerard R. Ford and Kissinger and then Indonesian president Suharto, responsible for the invasion and the massacre that followed. The record showed that U.S. had given "greenlight" to Suharuto's planned invasion. See https://nsarchive2.gwu.edu/NSAEBB/NSAEBB62/.

Creating/extracting definitions: We can define a task of creating or extracting a normative definition(s) for a given term using corpora or other resources. In its ordinary sense, definition extraction is a well-established NLP task (Sierra et al., 2009). We can also regard word embeddings as the task of defining word meanings.

Detecting deviations: In the context of what we have discussed so far, what matters about definitions is their normativity. So one possible application – or evaluation scheme of definition extraction through application – can be the task of detecting deviated use of terms in terms of their definitions. Automatically judging that the statement "responsibility accompanies freedom" is misusing the concept of *freedom* gives a concrete image of the objective of this task. It is somewhat similar to word sense disambiguation and also outlier detections used for evaluating word embeddings (Camacho-Collados and Navigli, 2016), though these tasks regard meanings as relative. At a different level, this task is related to detecting logically inappropriate statement. We started a research for detecting deviations of usage of technical terms in Japanese mass media, currently focussing on the domain of law and politics. A very embryonic observation was reported in (Tang and Kageura, 2017). When term variations exist, i.e. different representational forms are regarded as representing the same concept (Daille, 2017), controlling the surface form of terms also becomes an issue accompanying deviation detection.

Detecting cross-lingual gaps: As in the case of "cancer" and its Japanese "equivalent," terms in different languages that are regarded as representing the same concept can be different in some critical details. Not only judging the degree of correspondence but also evaluating the critical difference will be an important task as an extension of bilingual term extraction from parallel or comparable corpora. If we take into account the fact that very frequently corresponding documents in two or more languages do *not* have the same status (the goodness of TL texts should be evaluated by using original SL texts as the norm), the task is defined as directed, using the normative concepts represented by SL terms to judge the concepts (also normative in a monolingual setup) represented by TL terms[8]. At a different level, dealing with representational variations also becomes an issue.

Text profiling: Social profiling of texts can provide corpus-based processing with external criteria of normativity. To what extent texts themselves can be used to evaluate their normativity is also an important issue. This is technically related to text clustering or classification.

As technical problems involved in these tasks are similar to related tasks that have been well established, methods proposed for these related tasks may be adopted to tackle these problems. The difference resides in definitions of problems. Note that issues related to above topics are recently being dealt with in NLP. For instance, fact checking and analysing and detecting biased language are listed as topics relevant to the Workshop "Natural Language Processing meets Journalism." In the field of terminology, most of these have been dealt with manually.

3.2. Terminologies and System of Concepts

While we have witnessed a great advance in methods of both monolingual and bilingual automatic term extraction (ATE), the systematicity or coherency of extracted terms have not been taken into account when these methods were evaluated. Indeed, it is not stated as one of the aims of ATE in most cases. It is understandable, as we do not really know how to measure systematicity or coherence of terminologies. In lexicography, selecting a coherent set of headwords for a dictionary is left to the expertise of experienced lexicographers and remains one of the last frontiers of lexicography yet to be systematised[9]. Nevertheless, as we discussed above, systematicity is one of the essential features in knowledge and thus to address this issue is critical to the study of terminology.

Taking into account that terminologies represent conceptual systems, we can define, among others, the following tasks:

Evaluating systematicity of terminologies in terms of conceptual systems they represent: Terms are relatively motivated. Complex terms, which constitute 70 to 85 percent of all the terms in most domains in many languages, represent concepts by showing their main conceptual characteristics and their relations through constituent elements (Sager, 1990). A terminological representation thus reflects conceptual system to a certain extent. How systematic a terminology represents the corresponding conceptual system depends on domains and languages. Here we can define the issue of systematicity of terminological representations vis-à-vis the conceptual system. Once we can establish a method that can evaluate the systematicity of terminological-conceptual system, we may be able to judge to what extent a newly obtained term is relevant to the conceptual system and thus to the domain. The dynamic modelling of terminological and conceptual growth can be considered as an extension of this task. Ontology building shares a similar concern, though it focuses on conceptual system rather than representations.

Evaluating cross-lingual differences in the systematicity of terminologies: Terminologies of different languages represent the same conceptual system[10] differently. Evaluating the difference in the systematicity of terminological representations in different languages not only is important for the theoretical terminology but also contributes to cross-lingual terminological applications.

These studies are important not only from the theoretical point of view but also for real-world applications. Recall

[8] One may argue that MT deals with this issue indirectly when TL expressions are selected. For professional translators, being able to explain the difference among possible choices of TL expressions and the reason why a particular expression was chosen is not only part of their competence but also part of the *end*-product; the definitions of end-to-end in MT and in human translation are different.

[9] Personal communication with Dr. Judy Pearsall of the Oxford University Press on 8 July 2010 at the occasion of Euralex 2010 held in Leeuwarden, The Netherlands.

[10] Though the case of "cancer" indicates that it is not necessary safe to assume the identity of conceptual systems represented in corresponding terminologies in different languages, we can assume that they are approximately the same.

that in the keynote presentation in BUCC 2017, Professor Philippe Langlais stated "Despite numerous studies devoted to mining parallel material from bilingual data, we have yet to see the resulting technologies wholeheartedly adopted by professional translators and terminologists alike (Langlais, 2017). One of the reasons that not many advanced term extraction methods are not used in the real-world terminology management tasks or in translation is that it is difficult to know what are extracted and what are missed. If one could judge the status of extracted terms in relation to the existing set of terms, terminologists would be able to take advantage of the results of advanced methods more comfortably. The problem of dealing with the systematicity of terminologies within data-oriented or corpus-based language processing framework is that the size of terminologies are small. This may be one of the reasons why not much work has been carried out that deals with terminologies *per se*[11].

4. Two Concrete Studies

We introduce here two concrete studies we have been and are carrying out, which (remotely) take into account the issues of systematicity and normativity of terminologies and terms. The first is augmentation of bilingual terminologies, and the other is controlling term translations.

4.1. Augmentation of Bilingual Terminologies

In some languages pairs, as in the case of Japanese and English, manually constructed high-quality bilingual terminologies exist, and there is a strong demand for updating these terminologies. Standard corpus-based bilingual term extraction, unfortunately, cannot satisfy this demand, because new terms mostly occur with low frequency in the corpus and often hard to extract, and the relationship between extracted terms and entries in the existing terminologies is not transparent. Against this backdrop and taking into account issues we have discussed so far, we are developing a terminology-driven method for augmenting existing bilingual terminologies (Iwai et al., 2016a; Iwai et al., 2016b). The framework is simple:

1. Assuming that terminologies systematically reflects conceptual systems, we define terminological network which represents termino-conceptual structure of the domain with terms as vertices and edges as common constituent elements among terms. Figure 1 shows a terminology network of a small putative terminology.
2. Apply partitive clustering to the terminological network to obtain subclusters of terms which represent conceptual subsystem (Figure 2). Corresponding terminologies in different languages show similar tendencies, though differences are not small.
3. Complex terms are formed in accordance with the dynamics of these subclusters. Head-modifier bipartite graphs are created for terms in these subclusters, and new term candidates are generated by interpolating the missing links.
4. Bilingual candidates are generated by compositional matching, assuming that terms are motivated roughly in the same manner.
5. Candidates are validated by Web search.

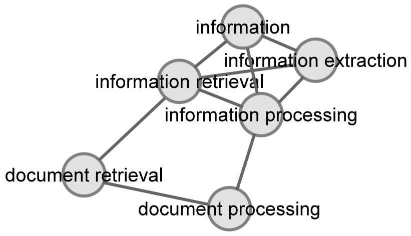

Figure 1: An exemplar terminological network.

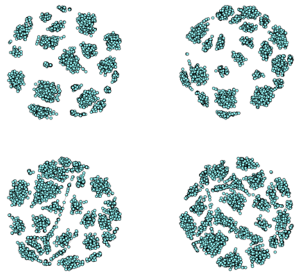

Figure 2: Subclusters in computer science (top) and economy (bottom) in Englih (left) and Japanese (right).

As of now, the method has several shortcomings:

1. The degrees of systematicity at the representational level vis-à-vis the conceptual systems have not been taken into account. The method just assumed that terms are reasonablly well motivated and thus terminologies systematically reflect the underlying conceptual systems to a degree that we can simply use terminological representations as a key to approximate conceptual systems. This assumption holds in monolingual situation, but as shown in Figures 1 and 2, different languages represent different parts of conceptual systems.
2. The gap between the systematicity of English and Japanese terminologies as reflected in the terminological networks can be explored to further capture the terminological structures that reflect conceptual systems. We have not elaborated on this.
3. Terminology networks are defined in a very rudimentary way. As edges are made when two terms (vertices) have common constituents, the hierarchical relations encoded in the forms of terms are not reflected in the network. Also, the dependency relations between constituent elements within terms are not encoded in the networks. This is the other side of the fact that the method currently does not take into account the

[11] Some other reasons are: terminologies and dictionaries are generally regarded as secondary creations compared to documents, which is based on the misunderstanding of languages; and terminologies are not too large and stable so it has been held that manual handling suits better than automatic processing.

conceptual system (see 1 above) and carries out candidate term generation purely at the level of terminological representations. It would be more theoretically proper to define the conceptual system separately from terminological networks, make correspondences between these two layers, and resort to the information at these two levels. To define the conceptual system that corresponds to a given set of terms, we are currently examining the use of distributional representation of constituent elements of terms in terminologies.

4. Currently all the candidates are treated equally. Using the information contained in termino-conceptual structure, we can give weight to candidates in terms of their status within the termino-conceptual system.

We are currently working to overcome these issues.

4.2. Controlling Term Translations

In translating terms, one TL term for an SL term is the basic principle for properly controlled documents (Sharoff and Hartley, 2012). In practice, it is frequently the case that several different TL terms are used for a single SL term. Terminology control should be made at the early stage of translation projects, i.e. controlled bilingual terminologies should be provided with translators involved in the project before they start translating documents. While language service providers generally adopt this procedure, it is still difficult to control terms properly. For instance, across Japanese municipalities, Japanese terms for administrative procedures are the same, but their translations vary because each municipalities translate their documents independently to each other. In these cases, "posterior" terminology control is essential; it is posterior in relation to already translated documents, but prior in relation to future documents to be written and translated.

One of the theoretically essential and practically important issues is to estimate the coverage of collected terms[12]. This issue is related to several other questions, i.e. whether or not the size of the corpus should be extended to collect sufficient number of terms, how many more texts should be checked, and how controlling terms affect these tasks.

We carried out TL term control for Japanese municipality documents manually (Miyata and Kageura, 2018). We collected three Japanese-English parallel documents that describes municipal procedures and extracted bilingual terms from them. Table 1 shows the number of terms ($V(N)$ indicates the number of term types, N the number of term tokens). Although we collected corresponding terms from parallel documents, the number of terms both in types and in tokens differ between two languages. We identified 374 Japanese term variations (12.4%) and 1258 English term variations (36.3%); TL terms have more variations than SL terms (Warburton, 2015).

Variations were groped and a preferred term for each group was assigned, based on three types of evidence, i.e. frequency evidence, topological evidence (expressions of terms) and dictionary evidence. After the terminology control, the numbers of term types in Japanese and English became closer.

	$V(N)$	N	$N/V(N)$
Japanese	3012	15313	5.08
English	3465	15708	4.53

Table 1: The number of extracted terms

	$V_c(N)$	$V(N)_c/V(N)$	$N/V_c(N)$
Japanese	2802	93.0%	5.47
English	2740	79.1%	5.73

Table 2: The number of extracted terms

What do they mean for the status of terms we collected? First, to evaluate the status of terms we collected vis-à-vis potential terminology we are dealing with in the domain, we adopted self-referring evaluation of collected terms. The idea is simple: (a) estimate the population number of terms using the distributional information of the terms we collected, and (b) evaluate the coverage of the collected terms against the estimated size of terminology. For the estimation of population number of terms, we used LNRE models (Baayen, 2001; Evert and Baroni, 2007).

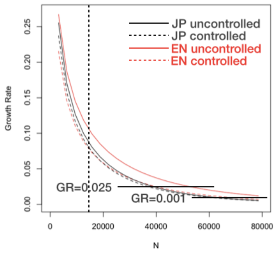

Figure 3: Growth rate of terms to the corpus size.

Figure 3 shows the growth rate of terms, before and after terminology controll was applied. We can observe several points: coverage became higher when terms were controlled and if we extend the corpus size to 40,000 word tokens, only one out of 40 terms is expected to be new. These enable us to evaluate the status of terms and terminology controll and ROI-based evaluation of the usefullness of extending the corpus.

5. Conclusions

We examined theoretical and social issues related to terminology, and clarified the position of terms and terminologies in relation to textual corpora together with issues in corpus-based terminology processing. We argued that the identity of concepts represented by terms is supported by the regulatory ideal, which provides the conditions upon which we can rationally communicate with each other in the first place. The concepts of systematicity and normativity were then introduced as on-the-ground concepts that reflect the regulatory ideal of the identity of concepts. We defined a range of tasks that take into account these issues and introduced two concrete studies as examples.

[12] That extracted terms be evaluated in terms of coverage is a prerequisite for evaluating systematicity of terminologies.

Much of what we discussed here is yet to be fully pursued, although relevant technologies exist. Indeed, the same technologies that can be used to pursue the tasks defined here can easily be used to promote pseudo-communication, including "fake news" and other forms of communication that promote hatred and discrimination. Unfortunately, current data-driven ML technologies do not internalise the regulatory ideal that human beings have tried to pursue painstakingly, so it is still upon us to decide how these advanced technologies are used. Cross-lingual comparable corpora contain interesting and important gaps, which we can explore to promote mutual understanding, as understanding starts from the recognition and identification of differences.

6. Acknowledgements

I would like to thank the organisers of BUCC Workshop, Professor Reinhard Rapp, Professor Pierre Zweigenbaum and Professor Serge Sharoff for giving me an opportunity to give this talk. Terminology-driven augmentation is a joint work with Dr. Koichi Takeuchi and Mr. Kazuya Ishibashi of Okayama University and Ms. Miki Iwai and Mr. Long-Huei Chen of the University of Tokyo, and was partly supported by JSPS Research Grants 24650122 (2012–2014). Evaluation of controlled terminology is a joint work with Dr. Rei Miyata of Nagoya University, and was partly supported by JSPS Research Grants 25240051 (2013–2017) and KDDI Foundation Research Grant Program (2014–2016).

7. Bibliographical References

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

Berlin, I. (1969). *Four Essays on Liberty*. Oxford University Press, Oxford.

Budin, G. and Oeser, E. (1995). Controlled conceptual dynamics: From 'ordinary language' to scientific terminology — and back. *Terminology Science and Research*, 6(2):3–17.

Camacho-Collados, J. and Navigli, R. (2016). Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. In *ACL Workshop on Evaluating Vector Space Representations for NLP*, pages 43–50.

Daille, B. (2008). Terminologie et traitement automatique des langues. In *TAMA 2008*, Ottawa.

Daille, B. (2017). *Term Variation in Spericalised Corpora*. John Benjamins, Amsterdam.

de Besse, B., Nkwenti-Azeh, B., and Sager, J. C. (1997). Glossary of terms used in terminology. *Terminology*, 4(1):117–156.

Evert, S. and Baroni, M. (2007). zipfr: Word frequency distributions in r. In *45th ACL Poster and Demo Session*, pages 29–32.

Felber, H. (1984). *Terminology Manual*. UNESCO, Paris.

Iwai, M., Takeuchi, K., Ishibashi, K., and Kageura, K. (2016a). A method of augmenting bilingual terminology by taking advantage of the conceptual systematicity of terminologies. In *Computerm 2016*, pages 30–40.

Iwai, M., Takeuchi, K., and Kageura, K. (2016b). Cross-lingual structural correspondence between terminologies: The case of english and japanese. In *TKE 2016*, pages 14–23.

Kageura, K. (1995). Toward the theoretical study of terms: A sketch from the linguistic viewpoint. *Terminology*, 2(2):239–257.

Kageura, K. (2015). Terminology and lexicography. In Hendrik J. Kockaert et al., editors, *Handbook of Terminology*, pages 45–59, Amsterdam. John Benjamins.

Kant, I. (1781). *Critique of Pure Reason*. Cambridge University Press (1999), Cambridge.

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., janne Bondi Johannessen, Khalil, S., Kokkinakis, S. J., Lew, R., Sharoff, S., Vadlapudi, R., and Volodina, E. (2014). Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1):121–163.

Langlais, P. (2017). Users and data: The two neglected children of bilingual natural language processing research. In *BUCC 2017*, pages 1–5.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates.

Miller, G. (1986). Dictionaries in mind. *Language and Cognitive Process*, 1:171–185.

Miyata, R. and Kageura, K. (2018). Building controlled bilingual terminologies for the municipal domain and evaluating them using a coverage estimation approach. *Terminology*, 24(2) (to appear).

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2010). Brains, not brawn: The use of 'smart' comparable corpora in bilingual terminology mining. *ACM Transactions on Speech and Language Processing*, 7(1).

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

Sharoff, S. and Hartley, A. (2012). Lexicography, terminology and ontologies. In Alexander Mehler et al., editors, *Handbook of Technical Communication*, pages 317–346, Boston. Mouton De Gruyter.

Sierra, G., Pozzi, M., and Torres, J.-M. (2009). *Proceedings of the 1st Workshop on Definition Extraction*. ACL, Borovets.

Tang, L. and Kageura, K. (2017). 'Fighting' or 'conflict'? An approach to revealing concepts of terms in political discourse. In *EMNLP Workshop on Natural Language Processing meets Journalism*, pages 90–94.

Temmerman, R. (2000). *Towards New Ways of Terminology Description: The Sociocognitive Approach*. John Benjamins, Amsterdam.

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

Wilks, Y., Slator, B. M., and Guthrie, L. M. (1996). *Electric Words*. MIT Press, Cambridge, Mass.

Wüster, E. (1959). Das Worten der Weld, schaubildlich und terminologisch Dargestellt. *Sprachforum*, 3(3):183–204.