QRcep: A Term Variation and Context Explorer Incorporated in a Translation Aid System on the Web

Takeshi Abekawa
University of Tokyo

Kyo Kageura
National Institute of Informatics

In this paper we describe the method of exploring term variations and the contexts in which terms occur using the Web, to help English-to-Japanese translators working online. Many English-Japanese terminological dictionaries are available in electronic form, but most of them do not provide rich examples of terminological use including variations. This is a problem for translators, who may not have sufficient knowledge on the use of terms in a specific subject they are translating. In order to augment this information gap, we have developed a system that explores actual use of terms using the information on the Web.

The system proceeds as follows:

- *when an electronic text in source language (English) is given, the system automatically looks up entries in terminological dictionaries including their variations, using the variation expansion rules;*
- *map the English entry to the Japanese translations;*
- *expand variations of Japanese terms on the basis of Japanese variation rules;*
- *search the Web and provide actual use of the term including variations within the actual context. For variation expansion, we are using Fastr Platform and defining corresponding rules for English and Japanese variations. The system is incorporated into the system that helps online volunteer translators and augments the terminology look-up functions.*

1. Introduction

This paper describes a method for exploring term variations and their context using the Web, within the framework of enhancing existing bilingual terminological dictionaries in order to help English-to-Japanese volunteer translators working online. In the case of English and Japanese pairs, many terminological dictionaries are available in both paper and electronic form. Most of them, however, do not provide users with detailed examples of terminological use and possible variations. This is a serious problem for volunteer translators, who may not have in-depth knowledge of the use of terms in the particular fields they are dealing with, as they are not (necessarily) subject specialists. In order to fill this gap, we have designed and developed a system, QRcep, which explores the actual use of terms using information on the Web.

The remainder of this paper is organized as follows. Section 2 describes the nature of terminological dictionaries in general. Section 3 elaborates on the mechanisms and functions of QRcep. Section 4 describes the way in which online translators look up terminological dictionaries as the framework within which the role of QRcep is defined.

2. The nature of terminological dictionaries

Unlike general dictionaries, most bilingual and multilingual terminological dictionaries only give source terms and their translations. Examples are listed in Table 1.

absolute temperature / zettai ondo / temperatura absoluta	(Kotani & Kori 1990)
A-B-X model / A-B-X moderu	(Japanese Ministry of Education 1986)
metaborato / metaborate	(Routledge 1997)

Table 1. Examples of entries in technical-term dictionaries

This is probably because terms have been thought to be stable in terms of their forms and clearly delimited in terms of the concepts they represent (Felber 1984), leading to the conclusion that it is sufficient to show (mostly one-to-one) translation relations among terms in their standard forms in different languages. Recent descriptive studies in terminology, however, have shown that terms are subject to rich variations when actually used and that there is a certain degree of subtlety in relation to the concepts they represent (Daille 2005; Pearson 1998; Temmerman 2000).

Another possible reason that most terminological dictionaries do not provide detailed information on term variations and context is because they assume users to be subject specialists, who can reasonably fill in any gaps with their own knowledge once the basic information has been provided by the dictionary. This assumption unfortunately does not hold for translators who need to deal with technical documents but are at the same time not subject specialists of that particular domain. For instance, translators may well have to check which, out of "koubun wo kaiseki suru" (to apply parsing), "koubun kaiseki wo suru" (to do parsing) or "koubun kaiseki suru" (to parse), is the most natural way of expressing the verbal form of "koubun kaiseki" (parsing), in order to produce good translations. Most technical term dictionaries do not provide the necessary information for this type of situation, but translators who are non-specialists will not be able to make an informed decision without it. So if we are to make full use of the range of existing technical-term dictionaries with the aim of helping translators, it is necessary to provide information on the usage and variations of terms.

Several researchers in terminology have recognised this need, and developed mechanisms to explore term variations on the basis of existing terminological lexicons. Daille et al. (1996) defined the basic framework of describing term variations. Jacquemin (2001) developed software called Fastr to detect term variations and provided variation rules for English and French terms. Schmidt-Wigger (1999) and Yoshikane et al. (2003) developed variation rules for German and Japanese terms, respectively, using the Fastr platform. In bilingual settings, Carl et al. (2004) and Kageura et al. (2004) developed parallel variation rules for French and English and for Japanese and English, respectively. For now, however, these variation explorers have not been applied to exploring contexts to help volunteer translators working online. Our aim here is to develop a fully operating and practical system to help online translators look up term variations in the process of translation.

3. QRcep: A term variation explorer incorporated into a translation aid system

Taking advantage of the existing variation rules developed using the Fastr platform and analysing the information search behaviour of volunteer translators working online, we have developed a system called QRcep, which explores term variations in context using the Web as a corpus and which can be accessed from within a translation aid system specially designed for (currently only English to Japanese) online volunteer translators.

3.1. *Exploration of terminological information by translators and the framework of QRcep*

In the present setting, we target volunteer translators who are translating online English documents into Japanese. The translation aid system we assume allows automatic lookup of dictionary entries given the source document, and users (translators) activate the lookup function by clicking the relevant tokens in the source language area (the left-hand panel of Figure 1).

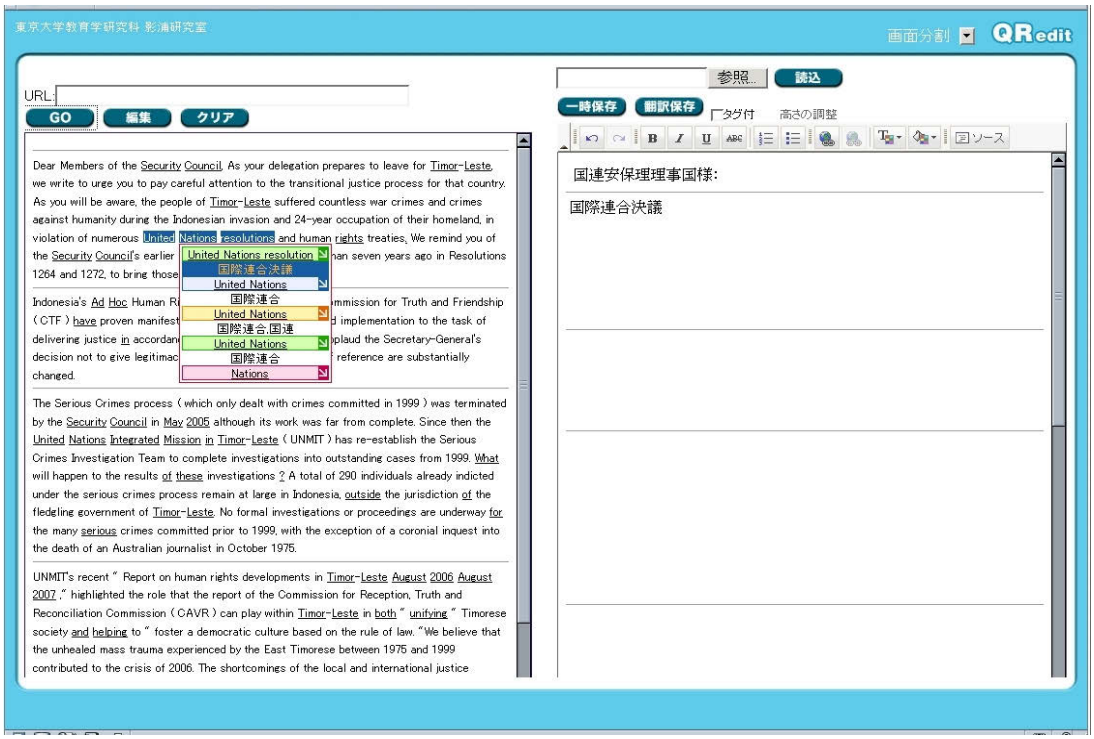

Figure 1. The translation aid system within which QRcep is called up

Within this environment, called QRedit (Abekawa 2007a; 2007b), which enables the lookup of terminological dictionaries, translators have indicated to us that they want to be able to refer to term variations in context when looking up basic translated terms. Thus the exploration of term variations in context is activated when translators look up terminological entries from the source language area of the translation aid system.

The QRcep system detects the occurrence of terms with variations in English source texts by matching dictionary entries with term tokens in texts; extracts corresponding Japanese term(s) from the dictionary entries; looks for variations of both English and Japanese terms on the Web; and displays term variations in context to users. Figure 2 shows the basic framework of the QRcep system.

3.2. *Variation rules and their bilingual correspondences*

The crucial part of the QRcep system is the English and Japanese parallel term variation detection mechanism. This mechanism is used first for matching English dictionary entries to occurrences of terms in English source texts, and then for detecting term variations in context in English and Japanese Web documents.

As the English and Japanese Web documents are not parallel, to contrast and compare the possibly corresponding usage of terms can only be facilitated through the correspondence between English and Japanese variation patterns. Kageura et al. (2004) defined the corresponding variants at two different but interacting levels. The first is formal correspondence, or correspondence at the level of paraphrasing operations. For instance,

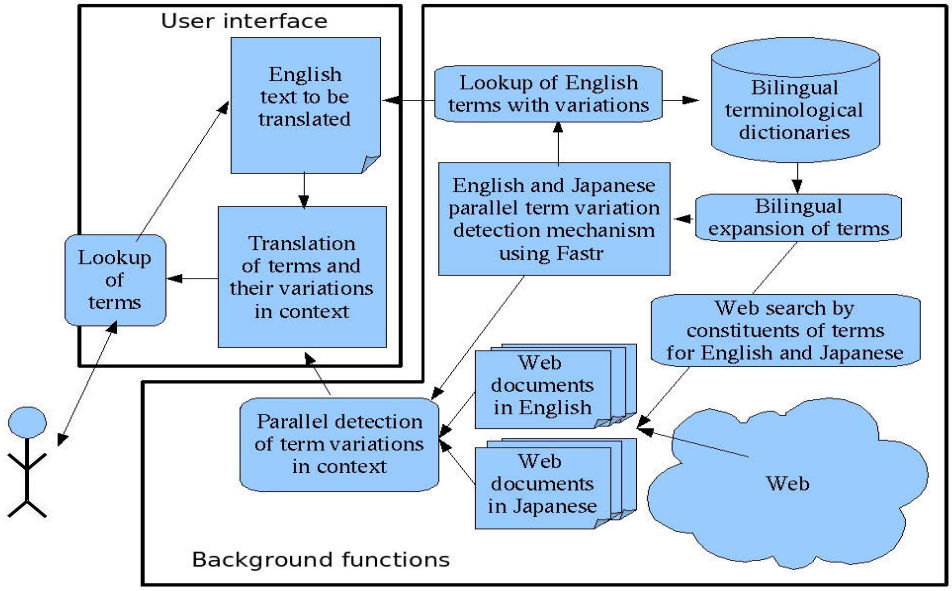

Figure 2. The basic framework of the system

changing a noun modifier to an adjectival modifier in English and Japanese is regarded as formal correspondence. The second is target correspondence, or correspondence at the level of produced variants. For instance, "retrieval of relevant information" as a variant of "information retrieval" is regarded as corresponding to "tekigou jouhou no kensaku" as a variant of "joho kensaku".

On the basis of this framework, we have extended English and Japanese variation rules in Fastr (Jacquemin 2001) and established the following classes of parallel variations:

1. Major category shifts: variations that change the grammatical category of original compounds. By adding the rules to the ones introduced in Kageura et al. (2004), we have established 12 Japanese rules and 5 English rules in total.

   Japanese example rules in simplified Fastr format are:

   X1 NS1 → X1 "wo" NS1 VS ("gainen gakusyu" → "gainen wo gakusyu suru")
   NS1 X1 → X1 "wo" NS1 VS ("jissou sisusemu" → "sisutemu wo jissou suru")
   NA1 NS2 → NA1 S NS2 VS ("aimai bunrui" → "aimai ni bunrui suru")
   in which X stands for any part of speech, NS stands for verbal noun stem, VS stands for verbal suffix, and NA stands for adjectival noun stem. The number indicates the correspondence between the terms on the left-hand side and those on the right-hand side of the rules.

   English example rules in simplified Fastr format are:

   N1 N2 → V2 ART? N1 ("word category" → "categorise words")
   X1 N2 → V1 N2 ("implemented system" → "implement system")
   A1 N2 → ADV1 V2 ("ambiguous classification" → "ambiguously classify")
   in which N stands for noun, V stands for verb, ART stands for articles, and A stands for adjectives. In both Japanese and English example rules, the first two are argument-verb constructions and the last is a modification-verb construction.

2. Head shifts: variations that change the content head. No additional rules were introduced to the 14 Japanese rules and 9 English rules originally established in Kageura et al. (2004).

   Japanese example rules in simplified Fastr format are:

   NS1 NX2 → NX2 "no" NS1 ("tsuika siryou" → "siryou no tsuika")
   NX1 TPNS1 NX2 → NX2 NX1 ("kyouyuuka memori" → "memori kyouyuu")

where NX stands for any sort of noun and TPNS stands for the suffix that turns nouns into verbal nouns.

Roughly corresponding English rules are:

V1 N2 → N2 N1 (“added material” → “material addition”)
V1 N2 → N2 V1 (“shared memory” → “memory sharing”)

3. Internal variants: variations that retain the major content/root of the head element and the overall category of compounds. These are further classified into three types:

(a) addition or deletion of functional elements (8 Japanese rules and 5 English rules).

Japanese example rules are:

NX1 NX2 → NX1 “no” NX2 (“kansuu keisan” → “kansuu no keisan”)
NX1 “no” NX2 → NX1 NX2 (“kansuu no keisan” → “kansuu keisan”)

English example rules are:

N1 N2 → N2 PREP N1 (“job amount” → “amount of job”)
N1 PREP N2 → N2 N1 (“amount of job” → “job amount”)

where PREP stands for prepositions.

(b) addition or deletion of content elements (19 Japanese rules and 23 English rules). Rules are divided further into modifications and coordinations.

A Japanese example rule for modifications is:

NX1 NX2 → NX1 {NX TPX?} NX2 (“kaihatu kankyou” → “kaihatu shien kankyou”)

where TPX indicates a suffix that can be attached to NX.

An English example rule for modifications is:

X1 N2 → X1 {A|N|V} N2 (“word type” → “word class type”)

where {A|N|V} stands for either adjective, noun or verb.

A Japanese example rule for coordinations is:

NX1 NX2 → NX1 C NX S NX2 (“gakusyu seigyo” → “gakusyu to taiwa no seigyo”)

where C stands for coordinating element and S stands for postpositions.

An English example rule for coordinations is:

X1 N2 → N1 C N N2 (“word class” → “word and concept class”)

(c) morphological operations in which one or more constituent elements change their POS category (20 Japanese rules and 13 English rules).

Japanese example rules are:

NX1 TPNS NX2 → NX1 NX2 (N to N) (“kyoyuka memori” → “kyoyu memori”)
NS1 N2 → NS1 VS N2 (N to V) (“bousou syaryo” → “bousou suru syaryo”)
NX1 VS NX2 → NX1 NX2 (V to N) (“bunrui suru kikai” → “bunrui kikai”)
NA1 NX2 → NA1 MD NX2 (N to A) (“aimai jouhou” → “aimai na jouhou”)
NX1 TPNA NX2 → NX1 NX2 (A to N) (“kikateki moderu” → “kika moderu”)
NX1 TPNA2 MD NX3 → NX1 TPNA2 NX3 (A to A) (“kika teki na hyougen” → “kika teki hyougen”

where VS stands for sa-inflexional verb, MD stands for adjectival inflexional suffix, and TPNA stands for adjectival suffix.

Roughly corresponding English rules are:

X1 N2 → X1 N2 (N to N) (“word classification” → “word class”)
N1 N2 → V1 N2 (N to V) (“index grammar” → “indexed grammar”)
V1 N2 → N1 N2 (V to N) (“indexed grammar” → “index grammar”)
N1 N2 → A1 N2 (N to A) (“category grammar” → “categorial grammar”)

A1 N2 → N1 N2 (A to N) (“categorial grammar” → “category grammar”)
A1 N2 → A1 N2 (A to A) (“syntactic information” → “syntactical information”)

Given the fact that conceptually equivalent English and Japanese terms may not have formal correspondence, as well as the fact that different variations may be used in corresponding contexts, the validity of these correspondences is limited. We nevertheless use them to display the English and Japanese term variations detected in Web documents, together with the number of hits, as these provide translators with useful information even if they then check relevant corresponding variations by themselves.

3.3. *The interface of the QRcep prototype*

Once the user triggers the term lookup function through the translation interface shown in Figure 1, the QRcep system is activated and the results are returned in a new window (Figure 3). In Figure 3, variations are ordered according to the number of hits obtained from the Web. Users can look up the Web snippets as well as the original pages to check how these variants are used in English and Japanese.

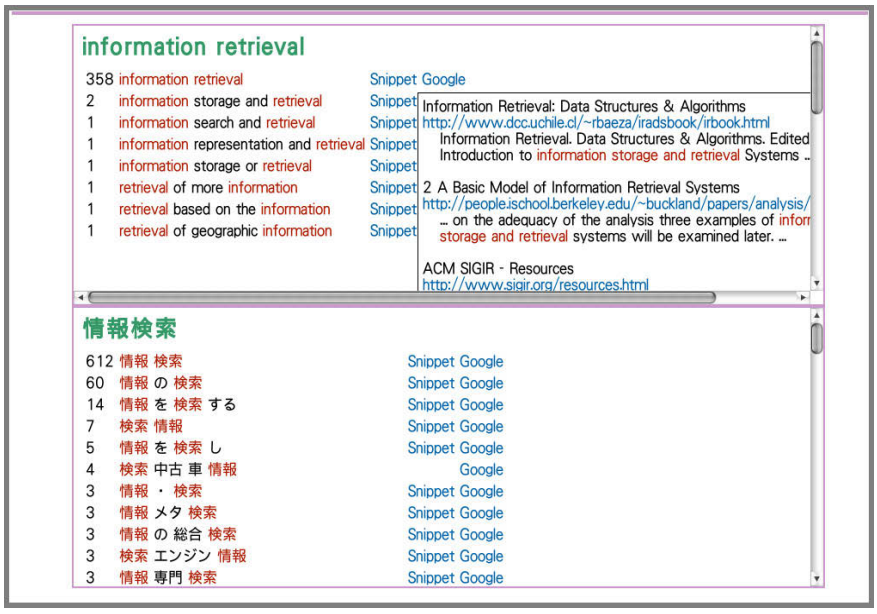

Figure 3. QRcep interface

4. QRcep as part of the augmentation of terminological dictionaries

QRcep augments the lack of information related to term usage with possible variations in context, using the Web as an open-ended corpus. Compared to existing high-quality ordinary dictionaries which give usage information, QRcep has both an advantage and a deficiency, which are in fact two sides of the same coin.

The advantage is derived from the fact that the QRcep system explores the Web. Because of this, QRcep is able to provide a much greater amount of usage information compared to ordinary dictionaries. On the other hand, QRcep cannot, for exactly the same reason, indicate to users which are standard, authentic or well-established usages as distinct from peripheral, minor or erroneous usages. As discussed in Kageura (2007), translators differentiate between three levels of information in seeking information sources, including established reference tools such as terminological dictionaries: (1) standard information about conceptually equivalent terms; (2) information about the standard and/or well-established usages and other information about the terms in actual use; and (3) wide, non-filtered information about the usage of terms. QRcep provides (3) in addition to (1) which is provided by existing bilingual terminological dictionaries.

Although (2) constitutes a subset of (3), simply providing translators with information at the level of (3) cannot compensate for the lack of information at the level of (2). What is required

for the further augmentation of terminological dictionaries with the mechanisms we have described so far is to filter or select an authentic set of English and Japanese documents that are relevant to the term from among all the Web documents. Selection at the level of documents has been shown to improve performance in identifying bilingual terms (Morin et al. 2007), so adding a document filtering routine to provide users with information at the level of (2) will be a natural extension of QRcep to fully address the current deficiencies in existing terminological dictionaries. As an approximation to the level of (2), we have developed QRselect, which collects translation document pairs that are relevant to the documents translators are dealing with (Kageura et al. 2007). We are also developing a system called Eryngii that augments existing bilingual terminological dictionaries (Utsuro et al. 2006a; Utsuro et al. 2006b). Together with QRselect and Eryngii, QRcep constitutes an overall environment for terminological reference assistance to online volunteer translators. Figure 4 illustrates the structure of the overall terminological reference assistance system.

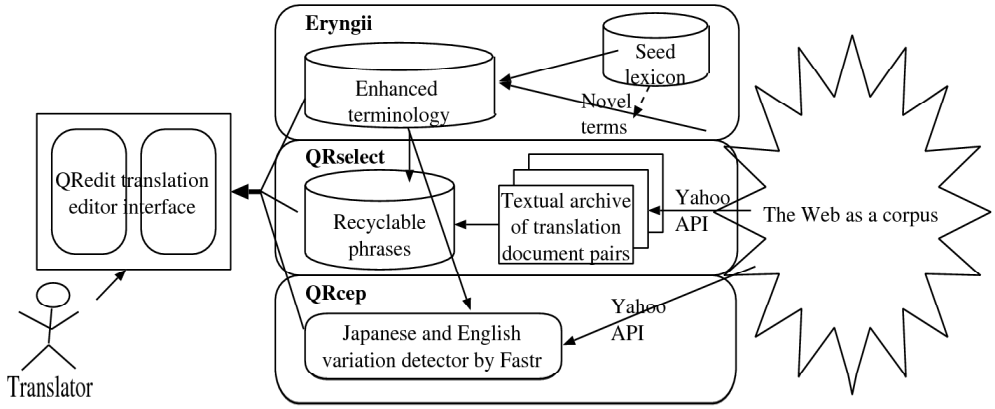

Figure 4. The overall structure of the terminological reference assistance system for online volunteer translators

5. Conclusions

In this paper we have described the system QRcep, which augments existing terminological dictionaries by providing users with usage information including variations in context. The system is specifically aimed at online volunteer translators, who are working online and thus have familiarity with Web language resources but may not have in-depth knowledge of particular subject areas, including the usage of technical terms.

Because we have this practical application in mind and because we are using the Web, which is regarded as an open-ended, growing corpus, IR-style evaluations on the basis of precision and recall are neither relevant nor useful. What is more important is how useful the system is for translators. The baseline for evaluating the system is that it should provide at least the same amount of information from the Web as translators can already obtain on their own without using QRcep, but in a way that reduces their burden. The overall Shiitake system, which contains QRcep as one of its parts, is currently being tested by a few English-to-Japanese volunteer translators.

Acknowledgements and information

This work is supported by grant-in-aid (A) 17200018 “Construction of online multilingual reference tools for aiding translators” from the Japan Society for the Promotion of Sciences (JSPS). We would like to thank anonymous reviewers for their useful comments, not all of which we could incorporate in the final article. For those who are interested in testing the Shiitake system and QRcep, please contact the first author by e-mail at: abekawa@p.u-tokyo.ac.jp.

References

Abekawa, T.; Kageura, K. (2007a). "QRedit: An integrated editor system to support online volunteer translators". *Digital Humanities 2007*. 3-5.

Abekawa, T.; Kageura, K. (2007b). "A translation aid system with a stratified lookup interface". In *Proceedings of the 45th ACL Annual Meeting Demos and Poster Sessions*. 5-8.

Carl, M.; Rascu, E.; Haller, J.; Langlais, P. (2004). "Abducing term variant translations in aligned texts". *Terminology* 10 (1). 101-130.

Daille, B. (2005). "Variations and application-oriented terminology processing". *Terminology* 11 (1). 181-197.

Daille, B.; Habert, B.; Jacquemin, C.; Royaute, J. (1996). "Empirical observation of term variations and principles for their description". *Terminology* 3 (2). 197-257.

Felber, H. (1984). *Terminology Manual*. Paris: Unesco and Infoterm.

Jacquemin, C. (2001). *Spotting and Discovering Terms through Natural Language Processing*. Cambridge: MIT Press.

Japanese Ministry of Education (ed.) (1986). *Japanese Scientific Terms: Psychology*. Tokyo: Gakujutu Shinkokai (this series covers some 30 scientific domains in separate volumes).

Kageura, K. (2007). "Terminological lexicons and terms in context: The translator's perspective." In Dieng-Kuntz, R.; Enguehard, C. (eds.). *7e Conference: Terminologie et Intelligence Artificielle*. Grenoble: Presses Universitaires de Grenoble. 1-10.

Kageura, K.; Abekawa, T.; Sekine, S. (2007). "QRselect: A user-driven system for collecting translation document pairs from the web". *Proceedings of the 10th International Conference on Asian Digital Libraries (ICADL 2007)*. 131-140.

Kageura, K.; Yoshikane, F.; Nozawa, T. (2004). "Parallel bilingual paraphrase rules for noun compounds". In *Proceedings of the Fourth Workshop on Asian Language Resources*. 54-61.

Kotani, T.; Kori, A. (eds.) (1990). *Dictionary of Technical Terms*. Tokyo: Kenkyusya.

Morin, E.; Daille, B.; Takeuchi, K.; Kageura, K. (2007). "Bilingual terminology mining—using brain, not brawn comparable corpora". In *Proceedings of the 45th ACL Annual Meeting*. 664-671.

Pearson, J. (1998). *Terms in Context*. Amsterdam: John Benjamins.

Routledge, ed. (1997). *Routledge Spanish Technical Dictionary*. 2 vols. London: Routledge.

Schmidt-Wigger, A. (1999). "Term checking through term variation". In *TKE'99*. 570-581.

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

Utsuro, T.; Kida, M.; Tonoike, M.; Sato, S. (2006). "Collecting novel technical terms from the Web by estimating domain specificity of a term". In Matsumoto, Y.; Sproat, R.; Wong, K-F.; Zhang, M. (eds.). *Computer Processing of Oriental Languages: Beyond the Orient: The Research Challenges Ahead*. Springer. 173-180.

Utsuro, T.; Kida, M.; Tonoike, M.; Sato, S. (2006). "Towards automatic domain classification of technical terms: Estimating domain specificity of a term using the Web". In Ng, H. T.; Leong, M.-K.; Kan, M.-Y.; Ji, D. H. (eds.). *Information Retrieval Technology: Third Asia Information Retrieval Symposium.* Springer. 633-641.

Yoshikane, F.; Tsuji, K.; Kageura, K.; Jacquemin, C. (2003). "Morpho-syntactic rules for detecting Japanese term variation". *Journal of Natural Language Processing* 10 (4). 3-32.