View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Universiti Teknologi MARA Institutional Repository

On the Concept of "Comprehensiveness" in Information Services: The Case of the Online Translation Aid and Hosting Service Minna no Hon'yaku

KYO KAGEURA, kyo@p.u-tokyo.ac.jp
Graduate School of Education, University of Tokyo

TAKESHI ABEKAWA, abekawa@p.u-tokyo.ac.jp
Research and Development Center for Informatics of Association, National Institute of Informatics

ABSTRACT

The aim of this research is to clarify the concept of "comprehensiveness" and its relationship to the concept of "normativeness" in language reference tools and information for online translators, from the point of view of strategically providing useful reference information via a translation aid-service. The concept of "comprehensiveness" in reference information has not been explored fully to dateThe questions to be answered are: What are the factors that determine different levels of "comprehensiveness" and "normativeness"? How can "comprehensiveness" be classified in relation to different types of reference lookup, and what kind of strategies can we define and adopt in developing and providing useful reference resources automatically or semi-automatically? While it is widely held that careful user studies are important in the strategic design of information services, empirical studies of potential users are not sufficient in the conceptualisation and development of advanced information services and tools which incorporate innovative functions or features, because quite frequently users do not understand what they want from the new information technologies. This is all the more true for issues in which one or more of the key concepts are not understood clearly. The question we wished to address fall precisely within this category, as the concept of "comprehensiveness" has not yet been explored fully. We therefore took a deductive and analytical approach, firstly listing up the factors that affect the concept of "comprehensiveness" and related concepts, with special reference to the translation-aid site Minna no Hon'yaku (translation of/by/for all: http://trans-aid.jp/), and deriving the classification of and desiderata for language reference tools and information from the objective of helping online translators. Although we adopted an analytical and deductive approach, the whole argument is implicitly supported by our own experience with actual translators' behaviour on the site Minna no Hon'yaku. Results of the analysis revealed that, within a framework of providing language reference tools for translators in general and in the context of the online translation-aid environment in particular, three different types of combinations of "comprehensiveness" and "normativeness" are of prominence and importance, namely: (i) task-oriented normativeness/comprehensiveness; (ii) domain-oriented normativeness/comprehensiveness; and (iii) user-oriented normativeness/comprehensiveness.

Keywords: Translation aid; Reference resources; Comprehensiveness; Normativeness; Minna no Hon'yaku (MNH); Lexicon

INTRODUCTION

This paper examines the concept of "comprehensiveness" and related concepts such as "normativeness" from the point of view of strategically providing useful terminological reference resources via the translation-aid service Minna no Hon'yaku (MNH: translation of/for/by all: http://trans-aid.jp/), which we have been running since April 2009 (Utiyama, Abekawa, Sumita & Kageura, 2009).

While it is widely acknowledged that one of the essential traits that affect the quality of language reference tools as represented by dictionaries is the nature of entry words or headwords as a set (how many entries there are, which words are covered, etc.), there have not been many studies dealing with this aspect and no in-depth descriptions of this aspect are given in standard textbooks in lexicography (cf. Atkins & Rundell, 2008; Sterkenburg, 2003; Svensen, 2009). In the field of natural language processing (NLP), some recent studies emphasise the importance of this issue (Sato, 2010), but a full exploration is yet to be carried out, unfortunately. On the other hand, while librarians and library scientists are instinctively aware of the fact that "book collections themselves are intellectual instruments that transcend even the content that is within them" (Sandstrum, 2010), their scope is in general limited to textual collections and they

are typically unable to explicitly and concretely articulate this important concept within the context of establishing actual information systems.

This often makes engineers and technologically oriented researchers assume that they can define information services without input from librarians and library scientists – an unhappy state of affairs. To overcome this lose-lose situation, library scientists, computational linguists and translators have collaborated on the MNH project, in order to provide useful reference resources, part of which was/is automatically and/or dynamically constructed.

The key concepts that have become clear in this collaboration process, i.e. "comprehensiveness", and also "normativeness", are examined in terms of the concrete environment of the translation-aid service MNH. Our general strategic plan for the enhancement and augmentation of online reference tools was already reported in Kageura, et. al. (2006), but it set its starting point as existing reference tools, and did not delve into the requirements and desiderata for language reference tools for translators in general.

A BRIEF SKETCH OF THE NATURE OF THE PROBLEM

Let us first intuitively clarify the nature of the issue, with reference to partly corresponding concepts in information retrieval (IR) research. While the database is assumed in IR research as an a priori existence (cf. Tokunaga, 1999), the "comprehensiveness" of reference tools corresponds to the coverage of the database.

Reference to IR immediately provides us with some insights:

(1) Some databases, such as SCI, attain social importance precisely because they are selective. We may think of the "comprehensiveness" of reference tools in analogy with this, even though a rigidly identical situation may not exist;

(2) While such evaluation measures as the F-measure are held to make sense in IR, partly because IR presupposes that relevant documents are replaceable or missed documents can be compensated for by retrieved documents, this does not hold for language reference tools. One cannot make do by looking up the entry "brown" if one cannot find the entry "red". This incidentally suggests that it is not sufficient to evaluate the performance of automatic term recognition (ATR) by recall, precision and/or the F-measure, even though many ATR studies adopt these criteria for evaluation (cf. Bourigault, Jacquemin & L'Homme, 2001).

It should also be pointed out that there are no "comprehensive" reference tools in the factual or empirical sense. Take a simple general dictionary of one language. As we do not even know how many words exist in that language (putting aside related problems such as defining "one languageness" or the unit of "words" in the first place, etc.), we cannot expect that a dictionary will exhaustively contains all these words.

Thus the concept of "comprehensiveness" should be examined at the level of understanding of the players in the society, as in the concept of "reliability" etc. (cf. Yamagishi, 1998). The concept of "comprehensiveness" as seen from the social epistemological point of view can be postulated as a characteristic of a reference tool which enables users to give up the search for information if they cannot find that particular information in that reference tool. If a reference tool is socially understood to function as such, we can reasonably say that the reference tool is "comprehensive".

MINNA NO HON'YAKU (MNH)

Minna no Hon'yaku (MNH: translation of/for/by all) is a translation hosting site accessible at http://trans-aid.jp/, placing special emphasis on mechanisms that enable users to manage and make use of useful reference resources for translation. MNH was made public on April 2009. As of April 2011, more than 1,500 users have registered with MNH, over 6,000 documents have been translated using the translation-aid functions provided by MNH, and of these more than 2,500 have been published on the MNH site. Currently MNH accommodates the English-to-Japanese, Japanese-to-English, English-to-Chinese, Chinese-to-English, Japanese-to-Chinese, Chinese-to-Japanese and English-to-Catalan language pairs.

MNH provides functions to aid translators, including an flexible lookup of high-quality dictionaries and terminology resources, seamless access to Wikipedia and Google search, and access to translation memory (TM). Registered users can constitute groups and within which members can define translation projects and can share translation tasks and user-defined resources.

More concretely, the basic functions provided by MNH are as follows (Utiyama, et. al. 2009):

1. anybody can register with MNH anonymously, and is provided with her/his personal space;
2. users can publish their translations on the MNH site, if copyright permits;
3. a variety of social networking functions are provided, including social tagging, message exchange, question and answer, translation request, etc.;
4. users can define a group on MNH, in which they can co-edit translations, share registered terms, share translation memories;
5. register terms, upload and manage terminologies, register translation memory database;
6. search translation texts, translated sentence pairs (TM), translators, tags, and registered terms.

Figure 1: The Main Page of MNH

Translators who register with MNH can produce translations by using QRedit. QRedit is a two-pane translation-aid editor incorporated in MNH (Figure 1), which provides the following functions for online translators (Abekawa & Kageura, 2007; Takeuchi, et. al. 2007):

1. flexible (idiom variations can be matched to dictionary entries), stratified (important or difficult multi-word elements are emphasised) lookup in and copy-and-paste from a high-quality dictionary, some free dictionaries, and terminologies;
2. seamless connection to Wikipedia monolingual and bilingual entries;

3. seamless connection to Google search;
4. function to register terms in the process of translating and immediately enable their lookup;
5. An easy-to-use and effective interface which enables users to concentrate on translation.

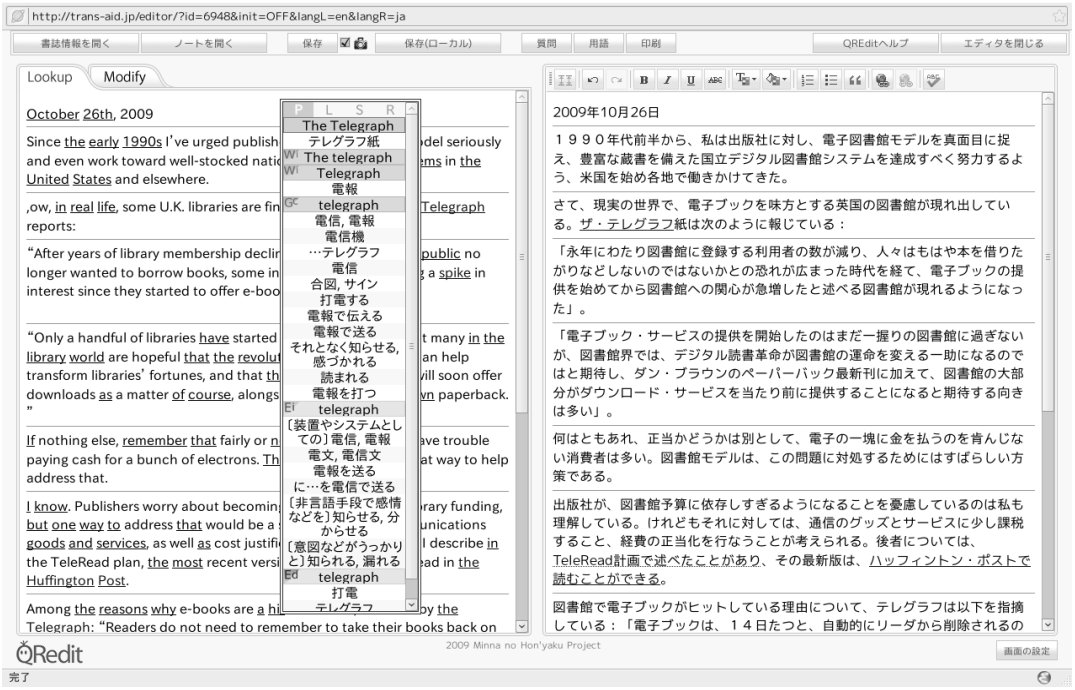

Figure 2: An Image of Translation-aid Editor QRedit

One of the main characteristics of an open system such as MNH is that it is used by a wide variety of translators to translate documents of many different types and registers. Correspondingly, requirements for reference tools vary. As a result, such questions as “what sort of reference tools should be preferentially supplied” and “what kind of characteristics should these reference tools have” are of central importance in designing the strategic development of MNH translation aid functions; thus the nature of the problem briefly described in the previous section.

SOCIAL SPACES RELEVANT TO "COMPREHENSIVENESS"

Types of Social Spaces

Through informal interactions with several online translators using MNH, we identified three main spaces to which translators refer when making decisions in translation:

1) Audience space: This space consists of, among others, potential readers, editors, fellow translators;
2) Task space: This space consists of the source document that the translator is translating, the target document, and the set of documents which are closely related to the document under translation in the source and target languages;
3) Language and information space: Lexicons (we use the plural form "lexicons" instead of "lexica", following the convention in lexicography), textual data or corpora, factual information are some of the elements that constitute this space.

These three spaces are shown in Figure 3.

Note that these are not necessarily mutually exclusive. A text may be regarded as residing in the task space from the point of view of one reference lookup, while also being regarded as part of the language and information space from the point of view of another type of reference lookup.

Social Spaces in the Translation Process

We can elaborate further on these three spaces and the elements of which they constitute in terms of the translation process:

a) Reference tools and reference sources are located in one or more of these spaces and gain certain social characteristics in relation to these spaces. For instance, from the point of view of neutral observers, bilingual dictionaries constitute the "lexicons" space in the language and information space; encyclopaediae fell into the "factual information" space; well-maintained translation memory belongs to the task space. Which reference tools fall into which space may differ from application to application, and from task to task.

Audience Space	**Task Space**	**Language and Information Space**
- Readers - Editors - Translators, etc	- Text to be translated - Related texts in source language - Related texts in target language etc.	- Lexicons - Corpora - Factual Information

Figure 3: Three Reference Spaces in Translation Process and Their Constituents

b) The act of translation itself belongs to the task space. Task space is defined in relation to the document to be translated, and, in accordance with the definition of the task space, some of the information resources or reference tools are moved from the language and information space to the task space and relocated within the task space;
c) The act of translation is ultimately a decision-making process which aims to maximise the acceptance of the translations by potential readers. In order to support this decision-making process, the position, status, and/or characteristics of reference tools which reside in the task space and in the language and information space are adjusted and fixed;
d) The audience space is more like an expected community or a projected image of certain concrete communities perceived by translators, rather than a communication space consisting of "the other". Accordingly, translators can assume a more or less concrete image of the audience community and the corresponding arrangements of and requisites for the task space and the language and information space.

It is within these arrangements of the spaces and elements that translators take due procedures for decision making in translation so that their translations can be accepted by readers in accordance with the way they expected. The concept of "comprehensiveness" of reference tools plays an important role in the process, supporting the due procedure defined in the society in decision making by translators. From this point of view, the three spaces, i.e. the audience space, the task space, and the language and information space (in which most reference tools reside) mutually affect or control each other, within the relationships of which the concept of

"comprehensiveness" is consolidated. In the discussions that ensue, we will clarify the strata of the concept of "comprehensiveness" of reference tools in more concrete terms, with these factors in mind.

CLASSES OF "COMPREHENSIVENESS" AND "NORMATIVENESS"

From our experience in MNH, translators' reference lookup can be classified mainly into two types, i.e. (i) situations in which translators must look up a particular reference (unless they are sure that they have accurate knowledge concerning the issue) and (ii) situations in which lookup by translators may be convenient or produce good results. The concept of "comprehensiveness" is determined within these situations. It is within the first type of reference lookup that "comprehensiveness" becomes an essential issue. At this point, we can understand that the concept of "comprehensiveness", which is socially defined, is inherently related to the "normativeness" prescribed by the particular society. For instance, a canonical terminological lexicon to which everybody in the domain refers, even if it contains a relatively small number of entries compared to other terminological lexicons, can be understood to attain its own "comprehensiveness" in terms of its target range of headwords and should be referred to by translators (note that the situation is somewhat similar to the SCI database). From the opposite angle, we can say that what is canonically normative does not need to be "comprehensive".

Basic Classification

In the process of observing a small number of translators' reference lookup, we identified the following three classes of "comprehensiveness"/"normativeness" for reference tools (we will refer to these three classes as classes of reference tools/information, or classes of reference lookup, depending on the context):

a) Task-oriented normativeness/comprehensiveness: Certain types of reference lookup in the process of translation are obligatory for maintaining consistency of translation or to satisfy the quality criteria defined in the task. For instance, referring to an already-translated section of text to keep translations of key terms or cited phrases consistent falls into this category, as does referring to past documents translated by the same group of translators to maintain group consistency (for example, Amnesty International and many other NGOs place great importance on maintaining consistency not only within individual documents but within the overall group of translated documents). In such cases, the target range against which the "comprehensiveness" of the reference information is determined is defined objectively, or, alternatively, the normativeness defines comprehensiveness. The mission of translation-aid systems for this class of reference information is to provide translators with the full range of relevant reference information, because the range of "comprehensiveness" can be defined objectively and empirically. Note that normativeness is derived from the requirements for deciding target language expressions.
b) Domain-oriented normativeness/comprehensiveness: As is typically the case with the translation of technical terms, translations must refer to and follow the conventions of the domain in relation to linguistic expressions. Thus comprehensiveness of terminological lexicons, for instance, should be examined at this level. This is also a requirement derived from deciding the due target language expressions. Unlike task-oriented comprehensiveness, domain-oriented comprehensiveness cannot be accomplished objectively or empirically. It is therefore necessary to define relative comprehensiveness, referring to the mutual understanding among relevant players. In this class, there can often be cases in which it is obligatory to refer to highly canonical sources which may be comparatively less comprehensive – normativeness thus precedes comprehensiveness. Another example that belongs to this class is international treaties. This case is close to class (A) in the sense that the range of expressions against which "comprehensiveness" is defined can be determined empirically (the

number of treaties that a particular country has ratified is finite and limited). However, as treaties do not in themselves exist in the task space and a realistic task for developing reference tools for treaties is to construct relatively superior reference resources in terms of comprehensiveness, we can regard them as belonging to this class.

c) Audience-oriented normativeness/comprehensiveness: From the point of view of translators, peer translators or editors represent the "audience". This class of reference lookup is mainly carried out to resolve misunderstanding or careless mistakes or to make "good" translations. As such, this class is conceptually closer to the case of reference lookup in which reference lookup may be convenient or may produce good results. In this case, the normativeness requirement from target language expressions is weak, and relatively more comprehensive reference tools (among those which satisfy certain quality criteria) are preferred by translators. For instance, MNH provides English-to-Japanese translators with Sanseido's Grand Concise English-Japanese Dictionary (Sanseido, 2001), precisely because the coverage of this dictionary is among the best among existing English-Japanese dictionaries. In relation to textual data lookup, we need to use Google as a preferential choice because it is socially agreed (if only implicitly) that if you cannot find certain information using Google it is reasonable to give up searching for it on the web.

As for the second case, i.e. the case in which lookup by translators may be convenient or produce good results, there is no concept of normativeness involved, and the range of phenomena against which the concept of comprehensiveness can be defined cannot be identified globally. Reference lookup in this case thus depends mostly on translators' competence. Thus we can essentially restrict our discussion to the first case in defining a strategic plan to develop reference tools which are "comprehensive".

Relationships among the Classes

Among the three classes introduced above, a good reference tool corresponding to (A) can and should satisfy both normativeness and comprehensiveness simultaneously; one corresponding to (B) should give preference to normativeness if both normativeness and comprehensiveness cannot be achieved at the same time, and one corresponding to (C) should give importance to comprehensiveness because normativeness is not that binding. As mentioned informally, deciding on due target expressions in translation seems to be the main driving force for requiring normativeness in reference tools. Though the unit and information type to be looked up differ in these three classes, they are not necessarily mutually exclusive, either. Apart from the stage at which the lookup of these types is carried out (draft translation, revision, review, etc.), the overall ordering of the lookups follows the order (A) -> (B) -> (C). For instance, in the reference lookup of the source language expression "crimes againt humanity", if the translator can confirm the information in the lookup class (A), then s/he does not need to look up reference resources belonging to (B) or (C). On the other hand, even if the translator could find the information related to "crimes against humanity" in reference tools belonging to (C), it would still be obligatory for her/him to look up the phrase in reference tools belonging to (A) and/or (B). This indicates the ideal situation for reference tools in translation, i.e. if possible, all the reference tools and information sources should be provided as reference information belonging to (A). Let us also note that, as long as normativeness is related to comprehensiveness, the comprehensiveness which is achieved by deviating from due normativeness is not appreciated at the level of the required normativeness.

Application Design

From these observations, we can put forward strategic directions for improving language reference tools for translators:

a) Materialisation of relative superiority in comprehensiveness: for reference tools belonging to (B) and (C) (and in some cases to (A) as well), what is required for "good" reference tools in relation to comprehensiveness is to extend the coverage so that relative superiority in comprehensiveness can be achieved while at the same time maintaining the requirements of normativeness within each reference lookup class;
b) Raising the level of normativeness and shifting the class: To increase the level of normativeness and thus change the class to which the reference tool belongs from (C) to (B) and from (B) to (A).

The basic configuration of these points are depicted in Figure 4. Additionally, although not directly related to the concept of comprehensiveness, improvement of the environment in which reference tools are referred to should accompany these enhancements of the reference tools themselves (returning to the analogy with IR, this can be interpreted as improving IR methods at the same time as enhancing the databases).

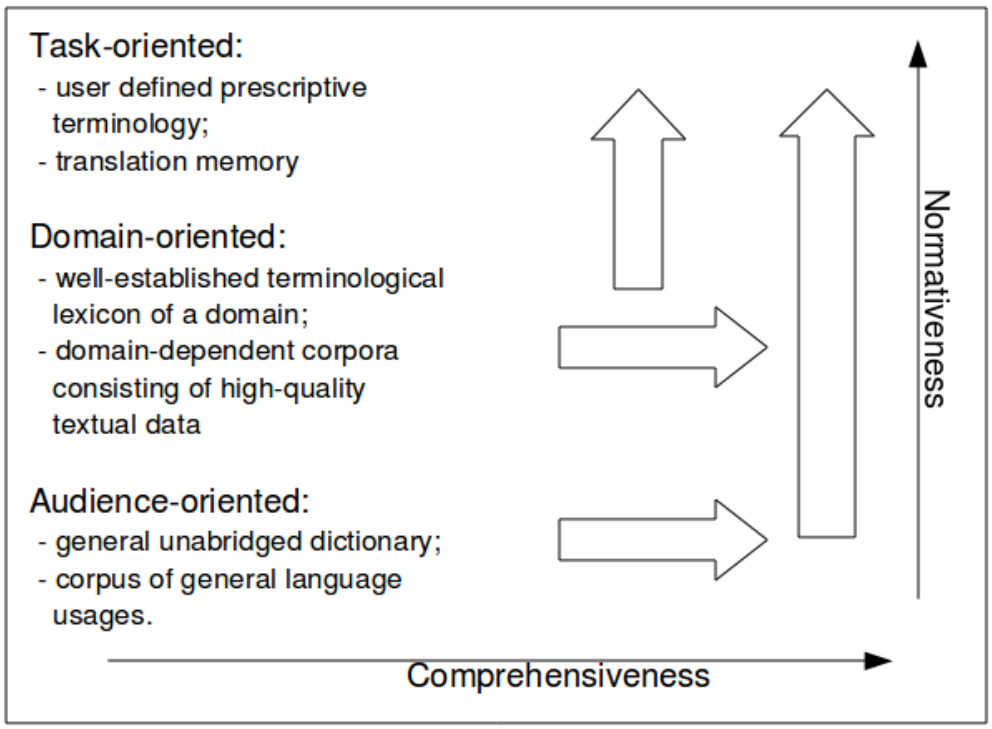

Figure 4: Axes and Configurations of the Application Design

COMPREHENSIVENESS AND NORMATIVENESS REVISITED

In section 4, we classified the desiderata for reference tools and strategic directions for their improvement in terms of comprehensiveness. A remaining issue from the practical point of view is within what frame of reference the strategic directions can be made concrete. The issue can be defined as a choice between two alternatives, i.e. to define the improvement of reference tools as a dynamic process in reference to the task space, or to define the improvement of reference tools in more task-independent manner and within the language and information space. As mentioned, the ideal for reference tools for translation would be to locate all the reference lookups in class (A). In fact, the preliminary research process often carried out by book translators can be interpreted as a procedure to reorganise and relocate relevant reference information belonging to (A), (B) and (C) into class (A). As class (A) is defined vis-a-vis individual tasks, we can perhaps say that aiming at developing a method of dynamically

organising reference information in the task space and constructing reference resources that belong to class (A) would be the path to take.

However, in actual system design, we need to take into account not only the technical feasibility of organising all the reference information in class (A) but also the theoretical issue of the division of labour between human decision-making and computational aids for translations. Taking these factors into account, designing a proper interaction between human and machine will be an important issue for the proper design and development of mechanisms which contribute to enhancing reference tools in terms of comprehensiveness. A natural choice would be for the automatic module to collect possible reference information as comprehensively as possible and for human translators or domain specialists to filter the information in the process of translation in order to raise the normativeness level. Note that this issue is not specific to MNH, nor to the problem of translation aid in general, but is a general issue related to the "usefulness" of information systems as seen from the point of view of "comprehensiveness".

CONCLUSIONS

We have discussed the concept of "comprehensiveness" in relation to language reference tools for translators. We started from postulating the nature of the problem, then gives a brief description of MNH, which we are running and on which the reference resources are to be provided to translators. On the basis of observing the behaviours of translators and talking with several translators, we consolidated the three spaces to which translators refer to in the translation process.

In the process of discussion, we have clarified three spaces which guide the clarification of the concept of "comprehensiveness", and introduced another important concept of "normativeness". These concepts shed light on some of the less understood features of reference tools in general and dictionaries and lexicons in particular, i.e. what are good characteristics that entries as a set should have in good and truly useful reference tools? As most information services are based on information units (in the case of dictionaries, each entry constitutes an information unit), the insights should be useful not only for reference tools but also for any kind of information services, including library services.

So far, our program for enhancing reference tools has emphasised the extension of comprehensiveness to achieve relative superiority at the same normativeness level (e.g. Abelawa & Kageura, 2009). Currently, we are designing next-generation mechanisms to enhance reference tools that explicitly aim at achieving comprehensiveness and raising levels of normativeness.

ACKNOWLEDGEMENTS

This work is partly supported by the Japan Society for the Promotion of Sciences (JSPS) grant-in-aid (A) 21240021 "Developing an integrated translation-aid site which provides comprehensive reference sources for translators" and by the 2009 National Institute of Informatics (NII) research cooperation project "Construction and use of practical terminological resources from a variety of information sources."

REFERENCES

Abekawa, T. & Kageura, K. (2007). *A translation aid system with a stratified lookup interface*. Proceedings of the 45th Association for Computational Linguistics Demos and Poster Session, 5-8.

Abekawa, T. & Kageura, K. (2009). QRpotato: A system that exhaustively collects bilingual technical term pairs from the web. *Proceedings of the 3rd International Universal Communication Symposium*, 115-119.

Atkins, S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford UP.

Bourigault, D., Jacquemin, C. & L'Homme, M-C. (eds.) (2001). *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins.

Kageura, K., Sato, S., Takeuchi, K., Utsuro, T., Tsuji, K. & Koyama, T. (2006). Hon'yakusha shien no tame no gengo refarensu tu-ru koudoka houshin [Enhancing language reference tools to aid translators]. *Proceedings of the 12th Conference on Natural Language Processing*, 707-710.

Sandstrum, J. (2010). *Saving the Warburg library*. Retrieved December 10, 2010 from http://centeredlibrarian.blogspot.com/ 2010/09/saving-warburg-library.html.

Sanseido. (2001). *Grand Concise English-Japanese Dictionary*. Tokyo: Sanseido.

Sato, S. (2010). Jisho jido hensan no tame no tekunoroji [Technologies for automatic construction of dictionaries]. *Symposium on the Development of NLP Technologies and New Prospects for Theory and Practice*.

Sterkenburg, P. (2003). *A practical guide to lexicography*. Amsterdam: John Benjamins.

Svensen, B. (2009). *A Handbook of Lexicography*. Cambridge: Cambridge University Press.

Takeuchi, K., Kanehila, T., Hilao, K., Abekawa, T. & Kageura, K. (2007). Flexible automatic look-up of English idiom entries in dictionaries. Proceedings of the MT Summit XI, 451-458.

Tokunaga, T. (1999). *Joho Kensaku to Gengo Shori* [Information retrieval and language processing]. Tokyo: University of Tokyo Press. [In Japanese]

Utiyama, M., Abekawa, T., Sumita, E. and Kageura, K. (2009). Hosting volunteer translators. *Proceedings of the XIIth Machine Translation Summit*.

Yamagishi, T. (1998). *Shinrai no Kozo - Kokoro to Shakai no Shinka Gemu* [Structure of reliability - evolutional games of mind and society]. Tokyo: University of Tokyo Press.