NLP MEETS LIBRARY SCIENCE: PROVIDING A SET OF ENHANCED LANGUAGE REFERENCE TOOLS FOR ONLINE TRANSLATORS

KYO KAGEURA
Graduate School of Education, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
kyo@p.u-tokyo.ac.jp

TAKESHI ABEKAWA
Graduate School of Education, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
abekawa@p.u-tokyo.ac.jp

Introduction. We are developing an online translation aid tool that provides enhanced language reference tools, specifically designed for translators working online. This paper introduces the concepts and framework for enhancing and organising the reference tools to maximally help translators.
Method. This paper is an essentially theoretical work, resorting to deduction from basic premises on language and translation. In order to design a maximally useful system, however, we also analysed translators' use of reference tools and examined the social arrangement of language reference tools.
Results. The main tenet of this paper is that existing natural language processing (NLP) technologies can be useful but for them to be actually used by language practitioners (translators in this context), it is necessary to take into account the social arrangement of the reference tools – the sort of knowledge that library science has accumulated. We have developed the Shiitake/QRedit translation aid system, in which a range of enhanced language reference sources are provided in a stratified and organised way that reflects the essential elements in the process of translation and the social arrangement of reference tools.
Conclusion. We have provided a sound basis for the design of the translation aid system and shown a way for NLP technologies to be fully utilised in real-world situations. The system we have developed is in experimental use (due to copyright restrictions on the core dictionary) by translators and has been enthusiastically received.

Introduction

In accordance with the rapid growth in the number of texts available online, translation activities targeting online texts have increased rapidly in recent years, covering a wide variety of languages, and the number of volunteer translators working on online texts is also growing (Salzberg, 2008). Many of these online volunteer translators either explicitly or vaguely feel that taking advantage of new language technology may help reduce the effort and time involved in translation, but few use existing translation aid systems or online dictionary lookup facilities (Kageura, et. al., 2006; see also Fulford and Zafra, 2004). Though a variety of factors contribute to this situation, one major reason is that the design of existing translation tools does not reflect the process of online translation, and consequently these tools do not meet the needs of online volunteer translators (Abekawa & Kageura, 2007; Kageura & Abekawa, 2007). While the availability of huge corpora has resulted in significant advancement in natural language processing (NLP) technologies, few substantial research has been carried out to address this mismatch, either theoretically or by means of developing actual systems (cf. Boitet, Bey and Kageura, 2005).

It was recognition of this lacuna that prompted us to develop a translation aid system specially designed to meet the needs of online volunteer translators. The system provides a set of enhanced language reference tools (content and lookup functions) to maximally aid online translators. This paper discusses the underlying philosophy of the system design, in the process clarifying the nature of translation, the reality of language strata as manifested in the translation process and how the proper arrangement of reference sources can help the translation process. The system, QRedit, is currently being used by several translators and has garnered extremely positive reactions[1]. We are currently only dealing with English to

[1] We only allow limited access to the system because the core dictionary incorporated in the system is copyrighted.

Japanese and Japanese to English translations, but the main part of the present work is language independent.

Translators' Needs and Relevant NLP Technologies

Kageura, et. al. (2006) reports the relation between volunteer translators' reference needs and existing language reference sources, based on consultations with around a score of translators via interviews and a questionnaire. Table 1 shows translators' degree of satisfaction with existing reference tools when looking up different language units. The degree of satisfaction is shown separately for content and lookup functions. A circle indicates that translators are generally satisfied with what is available, a triangle indicates that either translators are only partially satisfied or some translators are satisfied but others are not, and a cross indicates that they are not satisfied. Note that the reference tools listed are rough categories, which can be further grouped into two types: (a) those which take the form of dictionaries or lexica (dictionaries, proper name and technical term lexica, encyclopaedias); and (b) those which essentially provide texts (libraries and the Web).

Table 1: Translators' degree of satisfaction with existing reference tools

Language unit	Reference tools	Content	Lookup function
Ordinary words	Dictionaries	◯	◯
Idioms	Dictionaries	◯	×
Proper names (PN)	PN lexica, Encyclopaedias, Libraries/Web	△	△
Technical terms (TT)	TT lexica, Encyclopaedias, Libraries/Web	△	△
Quotations	Libraries/Web	×	×
Long collocations	Libraries/Web	×	×

If we try to address the weak part of the reference content using NLP technologies, the solution looks straightforward: to explore the huge corpora which have become available, possibly regarding the entire Web as an approximation of a universal corpus. Technologies for collecting large parallel or comparable corpora exist (Resnik & Smith, 2003; Fukushima, Taura & Chikayama, 2006), and the extraction of translation relations of various linguistic units from parallel or comparable corpora has been thoroughly investigated and reportedly achieved high performance (e.g. Fung, 1998; Huang, Zhang & Vogel, 2005; Morin, et. al., 2007; Nagata, Sato & Suzuki, 2001; Utsuro, et. al., 2006). Enhancing and augmenting lookup functions also looks like an easy problem to solve, or even a problem that has been solved already, as most translation memory tools currently in use employ flexible approximate matching methods. In short, recent NLP technologies based on large corpora are supposed to provide means to satisfy translators' need for the augmentation and enhancement of reference content and functions.

It is precisely at this point that we must consider a simple question is irresistibly upon us: why, then, do volunteer translators – even those who are willing to adopt new technologies into their work – not make use of existing systems based on these advanced NLP technologies? To understand this and to take real advantage of existing NLP technologies in translators' working environment, we need to consider the process of translation and the nature and arrangement of reference tools.

The Process of Translation Revisited

Basic Nature and Schema

Much work has been carried out on establishing a theory of translation and describing the process of translation (Baker, 1997; Eisteinsson, 2006; Munday, 2001; Venuti, 2004). On the basis of this existing work and discussion with translators, and with the aim of bridging the gap between NLP technologies and translators' needs, Kageura & Abekawa (2007) characterised the process of translation as follows:

- Translators deal with texts, not language as perceived by linguists;
- Translators make decisions, and do not simply follow the computation of language structure;
- Translators deal with texts as singular products, not as samples.

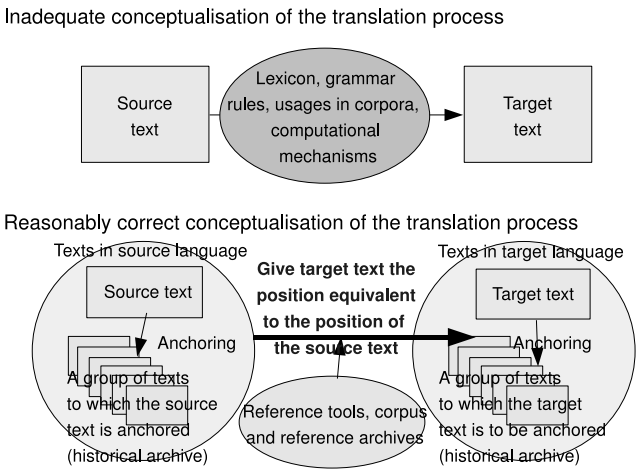

Figure 1. Inadequate and reasonably correct conceptualisations of the translation process

Incidentally, these correspond to the widely known fact that simply being bilingual does not automatically make a person a good translator.

These characteristics reflect the essence of the act of translation, i.e. translation is not merely concerned with transforming the source language expressions into target language expressions using lexica and grammar, which is analogous to the communication model à la Shannon, but is also essentially concerned with giving the target text a place in the set of existing relevant texts in the target language that corresponds to the place of the source text in the set of existing relevant texts in the source language. Figure 1 illustrates this point. It is also important to note that translators use reference sources not to obtain correct answers but to obtain relevant information which they examine in order to make their own decisions. This, incidentally, means that an IR-style evaluation by recall, precision, F-measure, etc. cannot guarantee usefulness of a system designed for translators.

Levels of Language in the Process of Translation

Reflecting the nature of the act of translation, translators deal, consciously or unconsciously, with three levels of language in its broader sense in the process of translation (Kageura, 2007):

1. Core language level (grammar and lexicon): Translators have sufficient command of the grammar of the source and target languages, but need high quality lexica as a standard reference for the basic meaning and the potential range of use of words in the source/target languages.
2. Text-archive level: Translators, knowing the group or the basic range of actual, historically and socially accumulated texts to which the target text is anchored, need to refer to these texts to find concrete expressions for possible use in translation.
3. Text-corpus level: Translators sometimes need to refer to the existing usage patterns in general open corpora to check the permissible or standard range of expressions.

Three points should be emphasised in relation to this stratification. Firstly, these three levels are not mutually interchangeable. Although translators sometimes use a source from one level as a surrogate for a source from another level, this is because of the practical limitations of reference sources and not because of the theoretical interchangeability of different resources.

Second, these three levels correspond to different realities of language and discourse as perceived in linguistics, philosophy, social studies of discourse, literature and natural language processing. Saussure (1910/11) regarded lexica as the core of *la langue*, and grammar constitutes the core of Chomskian program of linguistics (Chomsky, 1968). The text-archive level is referred to by Foucault (1968):

> La question que pose l'analyse de la langue, à propos d'un fait de discours quelconque, est toujours: selon quelles règles tel énoncé a-t-il été construit, et par conséquent selon

> quelles régles d'autres énoncés semblables pourrairent-ils être construits? La description du discours pose une tout autre question: comment se fait-il que tel énoncé soit apparu et nul autre à sa place?[2]

The text-corpus level is the one currently being dealt with by many NLP studies and such applications as automatic term extraction, translation extraction, information retrieval (IR), and open-domain information extraction (IE), etc [3].

Third, as these three levels operate in the process of translation and translators recognise information needs at each level, it holds that the reference sources should be organised in accordance with this three-level classification. Looking back from this point, we can see the issues to be dealt with for enhancing reference sources not only from the point of view of augmenting the content and providing advanced lookup functions, but also from the point of view of how reference sources have been provided and how they should be provided. Traditionally, high-quality dictionaries and lexica have covered reference needs for the core language level (translators should have command of grammar, so no grammatical references are needed). Libraries provided sources potentially useful for both the text-archive and text-corpus levels, but due to the limitations of language-level lookup functions, translators mostly used them for referring to information at the text-archive level. Note that this issue of the social and functional arrangement of reference sources has been well addressed in the study of reference services and reference sources in library and information science (cf. Nagasawa, 1993; Nagasawa, 1995). The use of NLP technologies, on the other hand, focuses on corpus exploration and consequently collapses the core language level and text-archive level, regarding them as functional subsets of the text-corpus level (Kageura, 2007). The configuration of a system straightforwardly relying on current NLP technologies can thus be illustrated as in Figure 2. Though this is a simplified illustration and existing systems provide rich and advanced functions, from the point of view of the arrangement of reference sources this illustration holds. Although the performance of potentially relevant NLP technologies such as translation pair extraction and term extraction has been improved, some essential problems can be observed from the point of view of aiding translators:

- The three levels of language are not clearly distinguished;
- As a corollary, the mechanism to distinguish general corpora and relevant archive has not been studied much;
- Also as a corollary, the gap between lexical information belonging to the dictionary or *la langue* and the information of use in corpora assigned to lexical items is not sufficiently distinguished (cf. Kageura, 2004).

These are the important reasons – related to NLP technologies – why many translators do not use available translation aid systems.

A System to Aid Online Translators

Basic Framework

We can now postulate a framework for the use of NLP technologies in helping translators by providing reference tools. Firstly, it is essential to base the system on the way translators actually work (one of the criticisms of existing translation aid systems is that they are detached from reference sources that translators are using). It is thus necessary for the system to provide one or more of the existing high-quality dictionaries used by translators; as mentioned, corpus exploration cannot substitute dictionaries, both theoretically and as a matter of fact. In the case of online volunteer translators, it is also necessary to take into account the terminating condition of the corpora exploration: the Web search. Online volunteer

[2] The question that the analysis of "la langue" asks, in the face of some fact of discourse, is always: from what kind of rules was this "énoncé" constructed, and in consequence from what kind of rules can other "énoncés" that resemble this one be constructed? The description of discourse asks a completely different question: how is it that this "énoncé," and nothing else in its place, appeared?

[3] In relation to translation aid systems, translation memory (TM) sometimes provides text-archive level information, and sometimes text-corpus level information, depending on how data are constructed and used (Bowker, 2002). TM that is successfully used tends to focus on the text-archive level, as in the case of patent translation.

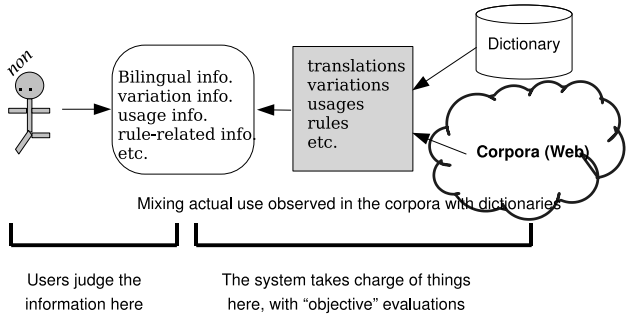

Figure 2. Configuration of the standard use of NLP technologies

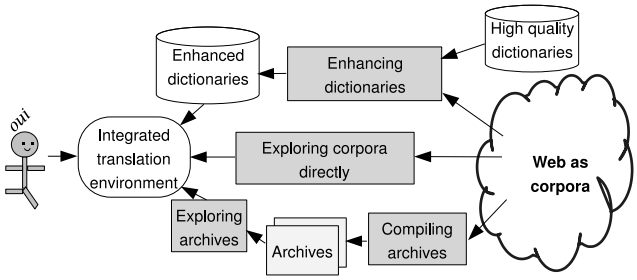

Figure 3. Framework of a desirable translation aid system

translators' Web searches using standard search engines currently functions, like libraries, as a place in which they can give up looking for further information (rather than the source of correct information they can rely on). Thus the system should be able to deal with an equivalent range of Web documents as can be checked through direct Web searches. Secondly, it is of utmost importance to maintain the three levels of language and corresponding arrangement of reference sources, i.e. dictionaries, archive reference, and corpus lookup. Enhancing the content and lookup functions of reference tools should be realised within this overall framework. Also, though we do not discuss this issue here, the system's functions should be provided via an integrated environment with simple and easy-to-use interfaces (another criticism translators have about existing translation aid systems is that the interfaces are too complex).

Figure 3 illustrates the general framework of a translation aid system that reflects these requirements. From the point of view of standard NLP approaches to corpus exploration, this configuration adds two new technical issues: (a) the lexicographical issue of dictionary compilation, in which the pursuit of improved lexical extraction or translation extraction should be integrated with the issue of editing or augmenting high-quality dictionaries, the computational formalisation of which is still at a very embryonic stage; and (b) the issue of archive construction from corpora, in which present techniques for text classification and textual evaluation should be fine-tuned to reflect a variety of features that contribute to consolidating a set of texts as archive relevant to the text that a translator is translating.

Shiitake: Reference Enhancement Systems

Taking into account the current state of reference tools given in Table 1 and the basic configuration of the system given in Figure 3, we defined the most important issues to be addressed, and developed the reference enhancement systems/modules, as follows.

Core Language Level (Dictionary Reference)

As the basic high-quality dictionary, we adopted Sanseido's Grand Concise English-Japanese Dictionary (Sanseido, 2004). In addition, the following three issues were dealt with:

(a) Improvement of idiom lookup functions, by realising flexible matching of idiom entries in dictio-

naries to idiom occurrences in texts with variations. The system, QRidiom, matches, for instance, the occurrence "He said that with his big fat tongue in his big fat cheek" with the dictionary entry "with one's tongue in one's cheek" (Takeuchi, et. al., 2007).

(b) Augmentation of the proper name lexicon by exploring the Web. The Maitake system currently collected about 150,000 personal names from the Web and newspaper corpora, providing a coherent reference source for personal names (Sakakibara & Sato, 2008).

(c) Augmentation of the technical term lexicon by exploring the Web. The Eryngii system explores technical corpora and extracts possible translations of complex terms on the basis of the compositional translation of constituent elements (Utsuro, et. al. 2006).

QRidiom and Maitake have achieved a performance level sufficient for actual use by translators, while Eryngii needs further improvement to reach this level.

Text-Archive Level (Archive Reference)

We developed a translation archive construction mechanism, QRselect, which reflects individual translators' requirements. A translator registers keywords or a set of Web sites that publish translated documents, and the system automatically collects existing relevant translation document pairs on the basis of information provided by the translator (Kaguera, Abekawa & Sekine, 2007). The collected translation document pairs or archive can be automatically looked up through the interface as translation memory. The lookup function is not yet implemented.

Text-Corpus Level (Corpus Lookup)

The following two issues were dealt with:

(a) Augmentation of information given in dictionaries by providing usage information observed in the Web. We focused on enhancing technical term dictionaries by providing usage and variation information extracted from the Web. The system, QRcep, dynamically provides usage and variation information in the process of terminological dictionary lookup (Abekawa & Kageura, 2008).

(b) General Web exploration functions, such as lookup of Wikipedia and mechanisms that simulate the Web searches performed by translators. This is provided as built-in functions in the QRedit interface.

Both of them are in full operation through the integrated translation aid environment QRedit.

QRedit: An Integrated Translation Aid Environment

The overall Shiitake resources and functions are provided as a Web service through the integrated translation aid environment QRedit (Abekawa & Kageura, 2007). The main philosophy of the QRedit environment and interface design is: to provide a seamless reference lookup environment while at the same time maintaining a clear conceptual distinction concerning what level of information (core language, text-archive, text-corpus) is referred to. Users can register the QRedit bookmarklet in the bookmark bar of the Web browser. This enables the user to activate the QRedit environment with one click, when she or he is reading a Web text and feels like translating it.

Figure 4 shows the basic interface of the QRedit environment. When the user activates QRedit through the QRedit bookmarklet, or specifies an URL or copies and pastes text in the source text area and clicks on "Go" button, the reference lookup function is fully activated and the user can access reference information by moving the mouse over relevant words or phrases. As details are given in Abekawa & Kageura (2007), we focus here on the aspect related to the different reference source levels. Two levels of reference display are defined in QRedit:

- The first step, small popup window: Basic translations from existing dictionaries and lexica are given in this window, as well as reference sources that provide information on the unit, as shown in Figure 4.
- The second step, large browser window: Full information from a dictionary entry, candidates from Web-based enhanced dictionaries (Maitake and Eryngii), usage and variation information on technical terms taken from the Web (QRcep), and relevant translations found in the archive (QRselect) are provided here, depending on which reference source is chosen in the first popup window. For

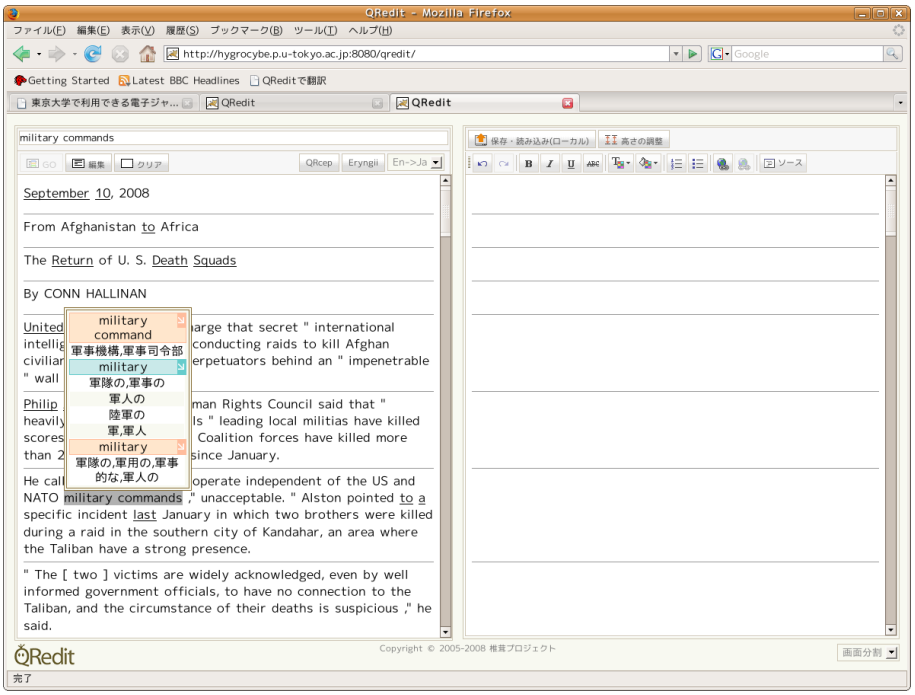

Figure 4. QRedit integrated translation aid environment

instance, Figure 5 shows the second-step display of QRcep results for the term "information retrieval". One can explore variations and usages from this window.

The different reference sources are indicated first by the colour of the entries in the small popup window, through which we materialise a seamless lookup mechanism while clearly indicating the different strata to which different reference sources belong.

QRedit has been experimentally available for volunteer translators since September 2007. To date, the system has been used to translate two full-length books, which are due out shortly, and over 100 online articles. The Global Voices Japanese translation team is using the system, and it received an enthusiastic reception among participants (including volunteer translators) in the translation session of the 2008 iSummit conference (iSummit, 2008). Though for now the system is only available on an experimental basis due to copyright restrictions on the core dictionaries, we are negotiating with dictionary publishers in order to make the system publicly available at the beginning of 2009.

Conclusions

In this paper we have discussed how reference sources should be organised in a translation aid system designed for online volunteer translators, and introduced a system that realises these basic requirements. It was recognition of the fact that online volunteer translators do not use translation aid tools incorporating advanced NLP techniques that initially prompted us to launch this work. After examining translators' recognition and evaluation of existing reference sources (as summarised in Table 1), we were all the more puzzled because the deficiencies in existing reference sources indicated by translators are precisely those problems that the exploration of corpora by NLP techniques would be expected to address. Yes, the fact is that translators are not willing to use systems that provide these functions.

Upon closer examination of the process of translation and the translators' use of reference sources, we found – though it would not be correct to attribute translators' not using translation aid systems to a single reason – that the process of translation is related to three different levels of language in its broader sense, and the act of reference lookup also comprises three aspects corresponding to these three levels, which are not mutually interchangeable. Studies of reference tools and reference services in library and information science corroborated this observation. On the basis of this understanding and taking advantage of exist-

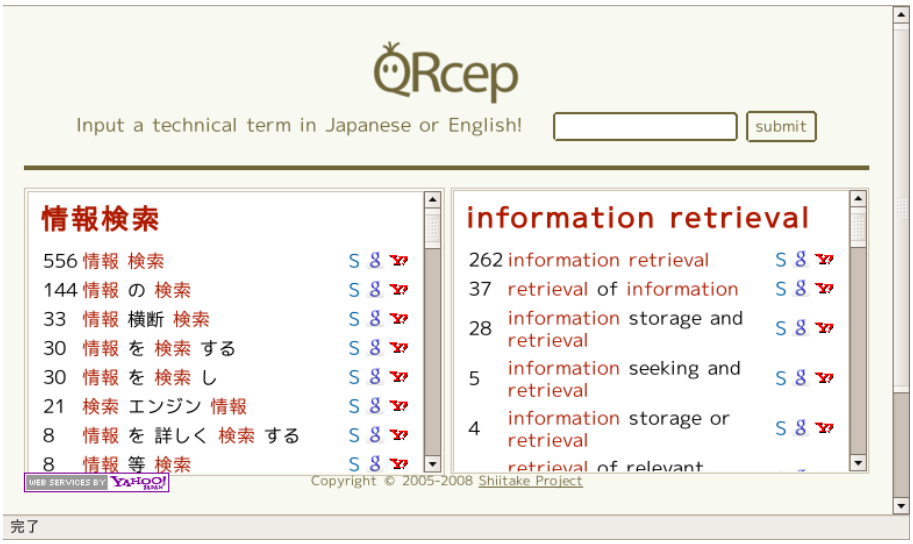

Figure 5. Second-step reference lookup in QRedit: When QRcep has been selected

ing NLP techniques, we developed the system Shiitake/QRlex that helps online volunteer translators by providing a set of enhanced language reference tools in a manner that reflects the three strata of language information at work in the process of translation. Initial experimental use of the system by translators has been very positive. This shows that, for advanced NLP techniques to be used by language practitioners in real world applications, it is useful for NLP to study the sort of knowledge accumulated in the study of library and information services.

From the point of view of NLP research, we have recognised two rather underrepresented research issues: (a) automatic editing of or aids for editing dictionaries and lexica as distinguished from automatic extraction of lexical items or translation expressions; and (b) automatic compilation of translation text archives relevant to texts being translated. Both issues require specialist knowledge to fully explore, the former the knowledge of lexicographers and the latter the knowledge of translators. We are currently developing NLP techniques for these two tasks, taking advantage of the feedback from translators and users of the Shiitake/QRedit system.

Acknowledgements

This work is partly supported by the Japan Society for the Promotion of Sciences (JSPS) grant-in-aid (A) 17200018 “Construction of online multilingual reference tools for aiding translators” and by the 2008 NII research cooperation project “Research on the modelling and reconstruction of lexicographical space from textual corpora”.

References

Abekawa, T. and Kageura, K. (2007). A translation aid system with a stratified lookup interface. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Demos and Poster Sessions* (pp. 5–8).

Abekawa, T. and Kageura, K. (2008). QRcep: A term variation and context explorer incorporated in a translation aid system on the Web. In *Proceedings of the 13th Euralex International Congress* (pp. 915–922). Barcelona: IULA.

Baker, M. (Ed.). (1997). *Routledge Encyclopedia of Translation Studies*. London: Routledge.

Boitet, C., Bey, Y. and Kageura, K. (2005). Main research issues in building web services for mutualized, non-commercial translation. In *Proceedings of the 6th Symposium on Natural Language Processing* (pp. 451–454).

Bowker, L. (2002). *Computer-Aided Translation Technology: A Practical Introduction*. Ottawa: Univer-

sity of Ottawa Press.

Chomsky, N. (1968). *Language and Mind*. New York: Harcourt Brace Jovanovich.

Eisteinsson, A. (Ed.). (2006). *Translation – Theory and Practice: A Historical Reader*. Oxford: Oxford University Press.

Foucault, M. (1968). Sur l'archéologie des sciences : Réponse au cercle d'épistémologie," *Cahiers peur l'Analyse* 9, 9–40.

Fukushima, K., Taura, K. and Chikayama, T. (2006). Fast and accurate method for detecting English-Japanese parallel texts. In *Proceedings of the COLING/ACL Workshop on Multilingual Language Resources and Interoperability* (pp. 60–67).

Fulford, H. and Zafra, J.G. (2004). The uptake of online tools and web-based language resources by freelance translators. In *Proceedings of the Second International Workshop on Language Resources for Translation Work, Research and Training* (pp. 37–44).

Fung, P. (1998). A statistical view on bilingual lexicon extraction. In *Proceedings of AMTA'98* (pp. 1–16).

Huang, F., Zhang, Y. and Vogel, S. (2005). Mining key phrase translations from web corpora. In *Proceedings of HLT/EMNLP'05* (pp. 483–490).

iSummit. (2008). Local context global commons track [Online]. Accessed 11 September 2008 at http://icommonssummit.org/programme/labs/local-context-global-commons.html

Kageura, K. (2004). Quantitative portraits of lexical elements. In *Proceedings of the 3rd International Workshop on Computational Terminology* (pp. 75–78).

Kageura, K. (2007). Terminological lexicons and terms in context: The translator's perspective. In Dieng-Kuntz, R. & Enguehard, C. (Eds.), *7^{e} Conférence Terminologie et Intelligence Artificielle* (pp. 1–10). Grenoble: Presses universitaires de Grenoble.

Kageura, K. and Abekawa, T. (2007). Types of translators and aspects of translation. In *Proceedings of the 13th Annual Meeting of the Japanese Natural Language Processing Society* (pp. 392–395).

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

Kageura, K. et. al. (2006). Improving the usability of language reference tools for translation. In *Proceedings of the 12th Annual Meeting of the Japanese Natural Language Processing Society* (pp. 707–710).

Morin, E., et. al. (2007) Bilingual terminology mining – using brain, not brawn comparable corpora. In *Proceedings of ACL'07* (pp. 664–671).

Munday, J. (2001). *Introducing Translation Studies*. London: Routledge.

Nagasawa, M. (1993). *Searching Information and Literature*. 3rd. ed. Tokyo: Maruzen.

Nagasawa, M. (1995). *Reference Service*. Tokyo: Maruzen.

Nagata, M., Saito, T. and Suzuki, K. (2001). Using the web as a bilingual dictionary. In *Proceedings of the Workshop on Data-driven Methods in Machine Translation* (pp. 95–102).

Resnik, P. and Smith, N.A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(4), 349–380.

Sakakibara, Y. and Sato, S. (2008). Large-scale extension of the dictionary of foreign personal names: Automatic editing of 150,000 entries. In *Proceedings of the 14th Annual Meeting of the Japanese Natural Language Processing Society* (pp. 833–836).

Salzberg, C. (2008). Translation and participatory media: Experiences from Global Voices. Translation Journal, 12(3) [Online], Retrieved 10 September 2008 from http://accurapid. com/journal/45global.htm

Sanseido. (2004). *Grand Concise English-Japanese Dictionary*. Tokyo: Sanseido.

Saussure, F. de. (1910–11). *3ème Cours de Linguistique Générale*. Geneve: Bibliothèque Publique et Universitaire.

Takeuchi, K. et. al. (2007). Flexible automatic look-up of English idiom entries in dictionaries. In *Machine Translation Summit XI Proceedings* (pp. 451–458).

Utsuro, T., et. al. (2006). Collecting novel technical terms from the Web by estimating the domain specificity of a term. In Matsumoto, Y. et. al. (Eds.), *Computer Processing of Oriental Languages: Beyond the Orient: The Research Challenges Ahead* (pp. 173-180). Berlin: Springer.

Venuti, L. (Ed.). (2004). *The Translation Studies Reader* (2nd ed.). London: Routledge.

About the Authors

Takeshi Abekawa holds a PhD degree from the Tokyo Institute of Technology. He is currently working as a postdoctoral research fellow at the University of Tokyo, specialising in natural language processing. He was awarded a best presentation award at the 2001 Annual Meeting of Japanese Natural Language Processing Society.

Kyo Kageura obtained his PhD from the University of Manchester. He is working as an associate professor of library and information science at the University of Tokyo. His main research interests are: the social philosophy of information and media, quantitative modelling of information, media and language, and their use in real-world applications. He is an editor of the journal *Terminology* and the book series *Terminology and Lexicography: Research and Practice* (together with Professor Marie-Claude L'Homme of the University of Montreal).