# A Translation Aid System with a Stratified Lookup Interface

**Takeshi Abekawa** and **Kyo Kageura**
Library and Information Science Course
Graduate School of Education,
University of Tokyo, Japan
{abekawa,kyo}@p.u-tokyo.ac.jp

## Abstract

We are currently developing a translation aid system specially designed for English-to-Japanese volunteer translators working mainly online. In this paper we introduce the stratified reference lookup interface that has been incorporated into the source text area of the system, which distinguishes three user awareness levels depending on the type and nature of the reference unit. The different awareness levels are assigned to reference units from a variety of reference sources, according to the criteria of "composition", "difficulty", "speciality" and "resource type".

## 1 Introduction

A number of translation aid systems have been developed so far (Bowker, 2002; Gow, 2003). Some systems such as TRADOS have proved useful for some translators and translation companies[1]. However, volunteer (and in some case freelance) translators do not tend to use these systems (Fulford and Zafra, 2004; Fulford, 2001; Kageura et al., 2006), for a variety of reasons: most of them are too expensive for volunteer translators[2]; the available functions do not match the translators' needs and work style; volunteer translators are under no pressure from clients to use the system, etc. This does not mean, however, that volunteer translators are satisfied with their working environment.

Against this backdrop, we are developing a translation aid system specially designed for English-to-Japanese volunteer translators working mainly online. This paper introduces the stratified reference lookup/notification interface that has been incorporated into the source text area of the system, which distinguishes three user awareness levels depending on the type and nature of the reference unit. We show how awareness scores are given to the reference units and how these scores are reflected in the way the reference units are displayed.

[1] http://www.trados.com/

[2] Omega-T, http://www.omegat.org/

## 2 Background

### 2.1 Characteristics of target translators

Volunteer translators involved in translating English online documents into Japanese have a variety of backgrounds. Some are professional translators, some are interested in the topic, some translate as a part of their NGO activities, etc[3]. They nevertheless share a few basic characteristics: (i) they are native speakers of Japanese (the target language: TL); (ii) most of them do not have a native-level command in English (the source language: SL); (iii) they do not use a translation aid system or MT; (iv) they want to reduce the burden involved in the process of translation; (v) they spend a huge amount of time looking up reference sources; (vi) the smallest basic unit of translation is the paragraph and "at a glance" readability of the SL text is very important. A translation aid system for these translators should provide enhanced and easy-to-use reference lookup functions with quality reference sources. An important point expressed by some translators is that they do not want a system that makes decisions on their behalf; they want the system to help them make decisions by making it easier for them to access references. Decision-making by translations in fact constitutes an essential part of the translation process (Munday, 2001; Venuti, 2004).

[3] We carried out a questionnaire survey of 15 volunteer translators and interviewed 5 translators.



*Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 5–8,
Prague, June 2007. 

Some of these characteristics contrast with those of professional translators, for instance, in Canada or in the EU. They have native command in both the source and target languages; they went through university-level training in translation; many of them have a speciality domain; they work on the principle that "time is money" [4]. For this type of translator, facilitating target text input can be important, as is shown in the TransType system (Foster et al., 2002; Macklovitch, 2006).

### 2.2 Reference units and lookup patterns

The major types of reference unit can be summarised as follows (Kageura et al., 2006).

**Ordinary words:** Translators are mostly satisfied with the information provided in existing dictionaries. Looking up these references is not a huge burden, though reducing it would be preferable.

**Idioms and phrases:** Translators are mostly satisfied with the information provided in dictionaries. However, the lookup process is onerous and many translators worry about failing to recognise idioms in SL texts (as they can often be interpreted literally), which may lead to mistranslations.

**Technical terms:** Translators are not satisfied with the available reference resources [5]; they tend to search the Internet directly. Translators tend to be concerned with failing to recognise technical terms.

**Proper names:** Translators are not satisfied with the available reference resources. They worry more about misidentifying the referent. For the identification of the referent, they rely on the Internet.

## 3 The translation aid system: QRedit

### 3.1 System overview

The system we are developing, QRedit, has been designed with the following policies: making it less onerous for translators to do what they are currently doing; providing information efficiently to facilitate decision-making by translators; providing functions in a manner that matches translators' behaviour.

QRedit operates on the client server model. It is implemented by Java and run on Tomcat. Users access the system through Web browsers. The integrated editor interface is divided into two main areas: the SL text area and the TL editing area. These scroll synchronically. To enable translators to maintain their work rhythm, the keyboard cursor is always bound to the TL editing area (Abekawa and Kageura, 2007).

### 3.2 Reference lookup functions

Reference lookup functions are activated when an SL text is loaded. Relevant information (translation candidates and related information) is displayed in response to the user's mouse action. In addition to simple dictionary lookup, the system also provides flexible multi-word unit lookup mechanisms. For instance, it can automatically look up the dictionary entry "with one's tongue in one's cheek" for the expression "He said that *with his* big fat *tongue in his* big fat *cheek*" or "head screwed on *right*" for "head screwed on *wrong*" (Kanehira et al., 2006).

The reference information can be displayed in two ways: a simplified display in a small popup window that shows only the translation candidates, and a full display in a large window that shows the full reference information. The former is for quick reference and the latter for in-depth examination.

Currently, *Sanseido's Grand Concise English-Japanese Dictionary*, *Eijiro*[6], List of technical terms in 23 domains, and Wikipedia are provided as reference sources.

## 4 Stratified reference lookup interface

In relation to reference lookup functions, the following points are of utmost importance:

1. In the process of translation, translators often check multiple reference resources and examine several meanings in SL and expressions in TL. We define the provision of "good information" for the translator by the system as information that the translator can use to make his or her own decisions.
2. The system should show the range of available information in a manner that corresponds to the translator's reference lookup needs and behaviour.

[4]Personal communication with Professor Elliott Macklovitch at the University of Montreal, Canada.

[5]With the advent of Wikipedia, this problem is gradually becoming less important.

[6]http://www.eijiro.jp/

The reference lookup functions can be divided into two kinds: (i) those that notify the user of the existence of the reference unit, and (ii) those that provide reference information. Even if a linguistic unit is registered in reference sources, if the translator is unaware of its existence, (s)he will not look up the reference, which may result in mistranslation. It is therefore preferable for the system to notify the user of the possible reference units. On the other hand, the richer the reference sources become, the greater the number of candidates for notification, which would reduce the readability of SL texts dramatically. It was necessary to resolve this conflict by striking an appropriate balance between the notification function and user needs in both reference lookup and the readability of the SL text.

### 4.1 Awareness levels

To resolve this conflict, we introduced three translator "awareness levels":

- Awareness level -2: Linguistic units that the translator may not notice, which will lead to mistranslation. The system always actively notifies translators of the existence of this type of unit, by underlining it. Idioms and complex technical terms are natural candidates for this awareness level.
- Awareness level -1: Linguistic units that translators may be vaguely aware of or may suspect exist and would like to check. To enable the user to check their existence easily, the relevant units are displayed in bold when the user moves the cursor over the relevant unit or its constituent parts with the mouse. Compounds, easy idioms and fixed expressions are candidates for this level.
- Awareness level 0: Linguistic units that the user can always identify. Single words and easy compounds are candidates for this level.

In all these cases, the system displays reference information when the user clicks on the relevant unit with the mouse.

### 4.2 Assignment of awareness levels

The awareness levels defined above are assigned to the reference units on the basis of the following four characteristics:

***C(unit)*:** The compositional nature of the unit. Single words can always be identified in texts, so the score 0 is assigned to them. The score -1 is assigned to compound units. The score -2 is assigned to idioms and compound units with gaps.

***D(unit)*:** The difficulty of the linguistic unit for a standard volunteer translator. For units in the list of elementary expressions[7], the score 1 is given. The score 0 is assigned to words, phrases and idioms listed in general dictionaries. The score -1 is assigned to units registered only in technical term lists.

***S(unit)*:** The degree of domain dependency of the unit. The score -1 is assigned to units that belong to the domain which is specified by the user. The score 0 is assigned to all the other units. The domain information is extracted from the domain tags in ordinary dictionaries and technical term lists. For Wikipedia entries the category information is used.

***R(unit)*:** The type of reference source to which the unit belongs. We distinguish between dictionaries and encyclopaedia, corresponding to the user's information search behaviour. The score -1 is assigned to units which are registered in the encyclopaedia (currently Wikipedia[8] ), because the fact that factual information is registered in existing reference sources implies that there is additional information relating to these units which the translator might benefit from knowing. The score 0 is assigned to units in dictionaries and technical term lists.

The overall score $A(unit)$ for the awareness level of a linguistic unit is calculated by:

$$A(unit) = C(unit) + D(unit) + S(unit) + R(unit).$$

Table 1 shows the summary of awareness levels and the scores of each characteristic. For instance, in an the SL sentence "The airplane *took* right *off*.", the $C(\text{take off}) = -2$, $D(\text{take off}) = 1$, $S(\text{take off}) = 0$ and $R(\text{take off}) = 0$; hence $A(\text{take off}) = -1$.

A score lower than -2 is normalised to -2, and a score higher than 0 is normalised to 0, because we assume three awareness levels are convenient for realising the corresponding notification interface and

[7] This list consists of 1,654 idioms and phrases taken from multiple sources for junior high school and high school level English reference sources published in Japan.

[8] As the English Wikipedia has entries for a majority of ordinary words, we only assign the score -1 to proper names.

| $A(unit)$ : awareness level | <= -2 | -1 | >= 0 | |
|---|---|---|---|---|
| Mode of alert | always emphasis | by mouse-over | none | |
| Score | -2 | -1 | 0 | 1 |
| $C(unit)$ : composition | compound unit with gap | compound unit | single word | |
| $D(unit)$ : difficulty | | technical term | general term | elementary term |
| $S(unit)$ : speciality | | specified domain | general domain | |
| $R(unit)$ : resource type | | encyclopaedia | dictionary | |

Table 1: Awareness levels and the scores of each characteristic

are optimal from the point of view of the user's search behaviour. We are currently examining user customisation functions.

## 5 Conclusion

In this paper, we introduced a stratified reference lookup interface within a translation aid environment specially designed for English-to-Japanese online volunteer translators. We described the incorporation into the system of different "awareness levels" for linguistic units registered in multiple reference sources in order to optimise the reference lookup interface. The incorporation of these levels stemmed from the basic understanding we arrived at after consulting with actual translators that functions should fit translators' actual behaviour. Although the effectiveness of this interface is yet to be fully examined in real-world situations, the basic concept should be useful as the idea of awareness level comes from feedback by monitors who used the first version of the system.

Although in this paper we focused on the use of established reference resources, we are currently developing (i) a mechanism for recycling relevant existing documents, (ii) dynamic lookup of proper name transliteration on the Internet, and (iii) dynamic detection of translation candidates for complex technical terms. How to fully integrate these functions into the system is our next challenge.

## References


Takeshi Abekawa and Kyo Kageura. 2007. Qredit: An integrated editor system to support online volunteer translators. In *Proceedings of Digital Humanities 2007 Poster/Demos*.

Lynne Bowker. 2002. *Computer-aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press.

George Foster, Philippe Langlais, and Guy Lapalme. 2002. User-friendly text prediction for translators. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 148–155.

Heather Fulford and Joaquín Granell Zafra. 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*, pages 37–44.

Heather Fulford. 2001. Translation tools: An exploratory study of their adoption by UK freelance translators. *Machine Translation*, 16(3):219–232.

Francie Gow. 2003. *Metrics for Evaluating Translation Memory Software*. PhD thesis, Ottawa: University of Ottawa.

Kyo Kageura, Satoshi Sato, Koichi Takeuchi, Takehito Utsuro, Keita Tsuji, and Teruo Koyama. 2006. Improving the usability of language reference tools for translators. In *Proceedings of the 10th of Annual Meeting of Japanese Natural Language Processing*, pages 707–710.

Kou Kanehira, Kazuki Hirao, Koichi Takeuchi, and Kyo Kageura. 2006. Development of a flexible idiom lookup system with variation rules. In *Proceedings of the 10th Annual Meeting of Japanese Natural Language Processing*, pages 711–714.

Elliott Macklovitch. 2006. Transtype2: the last word. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*, pages 167–172.

Jeremy Munday. 2001. *Introducing Translation Studies: Theories and Applications*. London: Routledge.

Lawrence Venuti. 2004. *The Translation Studies Reader*. London: Routledge, second edition.