BEYTrans: A Free Online Collaborative Wiki-Based CAT Environment Designed for Online Translation Communities*

Youcef Bey[a,b], Kyo Kageura[b], and Christian Boitet[a]

[a]Laboratoire LIG-GETALP, L'Université Joseph Fourier, 385, rue de la Bibliothèque.
Grenoble, France.{youcef.bey, christian.boitet}@imag.fr [b]Graduate School of Education
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, 113-0033,Tokyo, Japan.kyo@p.u-tokyo.ac.jp

Abstract. This paper introduces BEYTrans (Better Environment for Your TRANSlation), the first experimental environment for free online collaborative computer-aided translation. The requirements and functionalities related to individual translators and communities of translators are distinguished and described. These functionalities have been integrated in a Wiki-based complete environment, equipped with all currently possible asynchronous linguistic resources and translation aids. Functions provided by BEYTrans are also compared with existing CAT systems and ongoing experiments are discussed.

Keywords: Language Resources, Computer-Aided Translation Tools, Translation Memory, Segmentation, Collaborative Translation, Volunteer Translation Communities, Wiki.

1. Introduction

Multilingual information exchange is rapidly increasing on the Internet. A substantial part of this multilingualization is carried out by volunteers (engaged in the translation of documents, articles, reports, etc. as well as the translation/localization of computer software). Many online translation communities are formed by translators, software engineers, and in general people sharing the same motivations and aims. In view of this, various useful projects for online translation communities have emerged. Yakushite.Net is one such example. In Yakushite.Net, the system aims to improve machine translation (MT) performance through the interaction with volunteer translators. Translators contributing in turn can take advantage of MT and other functions provided by Yakushite.Net, while they contribute to the improvement of the system by augmenting its language resources (YAKUSHITE, 2007).

Open environments for developing free online lexical resources also exist. Wiktionary – The linguistic companion of the huge free Wikipedia encyclopedia – is one example (WIKTIONARY, 2007). Papillon, another example, aims also to construct a large-scale, free multilingual lexical database with the help of volunteers (Mangeot, 2002). In addition, free

* Acknowledgements: Special thanks go to the DEMGOL organizers at Trieste (Italy), in particular Mr. Giovanni Zorzetti, who helped us to transfer hundreds of multilingual documents by developping special scripts, and Ms. Francesca Marzari, who has devoted much time and effort to evaluating and testing BEYTrans in a real-world environment.

* This research is partly supported by grant-in-aid (A) 17200018 "Construction of online multilingual reference tools for aiding translators" by the Japan Society for the Promotion of Sciences (JSPS).

standalone translation memory (TM) systems such as Omega-T are now available (OMEGAT, 2007).

While appreciating the importance of these online environments and systems for directly or indirectly promoting the multilingual exchange of information, we recognize the lack of an integrated environment to help online translator communities in a systematic way. Against this backdrop, a free online computer-aided translation (CAT) environment, BEYTrans (Better Environment for Your TRANSlation) has been developed, and the system is now in its experimental stage. This first experimental version has two levels of functionality. The first level corresponds to the translators, who act as separate entities and need specific functionalities (translation editor, linguistic help, etc.). The second level corresponds to the community, in which translators work as an integrated entity. Both functionalities have been integrated into our system using a collaborative Wiki-based technology which provides to volunteer translators with a user-friendly environment and helps them improve translation consistency (Schwartz et *al.*, 2004) (Augar et *al.*, 2004).

This paper explains the basic background, concepts and functionalities of BEYTrans. In section 2, we describe the flow of linguistic data and the interactions among translators, on the basis of which the system requirements have been identified. Section 3 describes the different functionalities that should ideally be made available to 1) individual translators and 2) translation communities. In section 4, a detailed explanation of BEYTrans and its position among existing CAT environments is given. The system is currently being used experimentally by the DEMGOL project and other communities, which is briefly examined in the final section.

2. Virtual translation network

Online volunteer translators are organized in communities in which they perform translation together or separately, and disseminate multilingual content on the Web. In so doing, they use their private translation environments and communicate frequently with their counterparts (other translators). To build a support system that deals adequately with this situation, for this situation, we need to understand the requirements and the needs of communities, considered as integrated entities, and those of translators, considered as separate entities (Figure 1). In other words, it is important to help translators at two levels: the translator level and the community level. Distinguishing between these two levels allows us to more effectively identify and fulfill existing needs in the translation process.

For example, the translation progress in the context of the ArabicMozilla project (ARABICMOZILLA, 2007) needs to be checked constantly and translators should be kept aware of changes in the content. At the same time, each translator, as individual entity, performs translation separately and sends it to the repository (web location dedicated to the relevant community) where the new content is updated.

The translation process led us to differentiate two categories of practices:

(i) Individual translators working as separated entities;
(ii) A translators working as an integrated entities.

The former is related to translator behavior where she/he uses private environment (editors, dictionaries, etc.). The later needs separate functionalities for increasing collaboration and consistency.

The communication and data exchange in ARABICMOZILLA can be schematized as shown in Figure 1. The community itself is divided into three sub-communities and modeled as a network where nodes represent translators and edges represent change and data manipulation – it reflects also the action that data is subjected to – Furthermore, nodes reflect the translator's position and requirements inside her/his community and edges represent the requirements and needs of the community.

For example, translator *A* adds an entry in the community dictionary (requirement category (i): dictionary addition function). The same translation could be used by *B* for the translation of a

source word (requirement category: (i) search function and (ii) control change function). After that, it could be updated by *C* (requirement category (i): dictionary update function). Translators *D* and *E* may exchange comments about the possible validation of the new entry (requirement category (i): communication). In fact, in the process of translation, these tasks are done manually and still need to be controlled and supported by an integrated environment.

To further clarify our method and show how it applies to the real situation of translation communities in a concrete way, according to these categories of functionalities, we take as an example the ARABICMOZILLA community, in which the translation process is as follows:

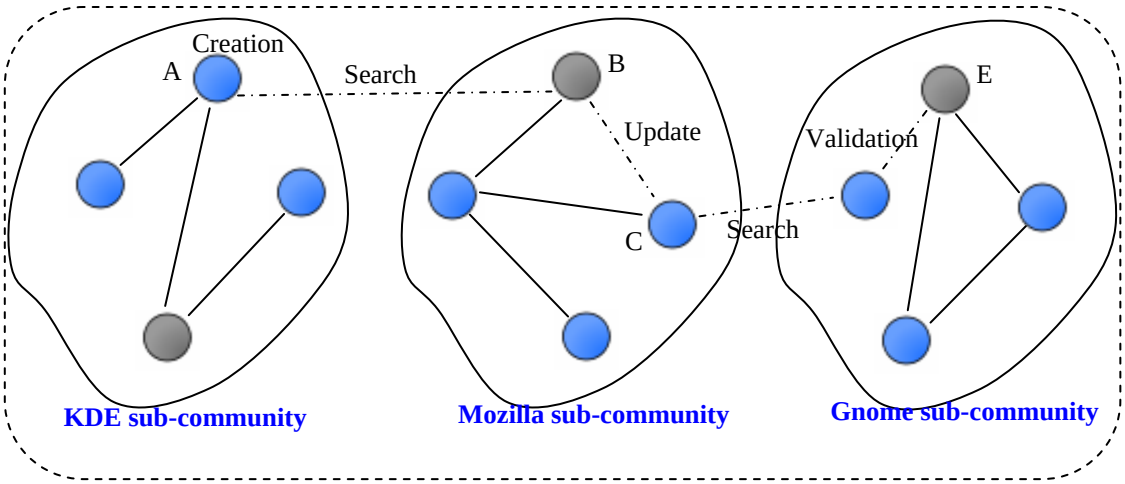

Figure 1: Translator and sub-community interactions in the ARABICMOZILLA project.

> The community is comprised of 10–20 volunteer translators (ARABICMOZILLA, 2007). Translators constantly check new English releases; if available, they proceed to the translation/localization into Arabic. In the process, many problems arise. For example, the Arabic script is RTL (Right To Left), hence, translation needs to be tuned with the software compilation. An estimated of 100 strings can be translated in one hour. For example, for the last Gnome release, version 2.18 had about 6,000 extra strings which were translated in 60 hours. The translation was performed by 6 translators and took one month, including checking the translation of the whole text (which contained about 35,000 strings) in order to ensure consistency. As for linguistic resources, translators are very active and have created under Wiki a technical terms dictionary, which contains around 18,000 entries (ARABEYES, 2007). Translators refer sometimes also to commercial standalone dictionaries like the English-Arabic 'Al-warid', which is largely regarded as authoritative in Arabic-English translation.

As is usual in any community, some group-motivated behaviors appear in this translation community. In this case, they are related to the organization of translators and to data manipulation, and reflect separated and group functions. It is very difficult to delimit such behaviors either at translator or group level. But some lacks and needed functionalities are common to this and other communities (FRENCHMOZILLA, W3C, etc.). These functionalities, which correspond to the practices (i) and (ii) cited above, can be categorized as follows:

a. Functionalities for translator who operates on translation in autonomous way and disposing of private environments (Bey et *al.*, 2006 (a)) (Abekawa et *al.*, 2007).
b. Functionalities for collaboration and work between translators (Bey et *al.*, 2006 (b)).

In the next section, details of these two categories of practice are described and the specification for functionalities implementation is given.

3. Description of integrated functionalities: the first step toward a solution

In the process of translation, a volunteer translator will want individually exploit the potential of various automatic linguistic functionalities (e.g., dictionaries and MT suggestions) and at the same time interact with her/his counterparts by using community functionalities (e.g. asking for translation community aid when linguistic aids aren't sufficient). From this point of view, translator and community functionalities should be implemented as follows.

3.1. Individual translator functionalities

The environment should be open to all volunteers without any restriction. Individual translators should be able to take advantage at the following different functionalities:

- Text extraction and tokenization: documents to be translated undergo a process of text extraction and tokenization for the identification of Translation Units (TU) (LINGPIPE, 2007) (Walker et *al.*, 2001).
- Linguistic aids: linguistic aids are integrated and activated automatically in an online asynchronous manner (see 4.3).
- Online translator-oriented editor: translators read the source TUs synchronized with their corresponding target TUs and input the translation in the source TU in order, segment by segment, or jumping to the segment they wish to work on. The target TUs are replaced in fact automatically or manually by the "best" pre-translations of MT or TM –

3.2. Translator community functionalities

Extended functionalities for the translation community have been implemented using the Wiki technology. As explained above, Wikis are user-friendly environments that recently have provided to be a great success. Using this technology, any user/translator can upload documents and share them with his/her community; thereby making it possible to implement the following functionalities (Schwartz et *al.*, 2004) (Augar et *al.*, 2004):

- Collaboration: documents of the same community can be grouped in a specific space where they are freely accessible by all translators of the same community. Each translator has an information access that allows her/him to interact with other translators.
- Progression control: collaborative translation needs access control and historic content revision. The progress of the translation can be checked comparing different versions of the same document. Translators can check the content and make comparisons between versions at the sentence or word levels.

The translator and community functionalities outlined above have been combined to produce the integrated collaborative environment BEYTrans, which is described in the next section.

4. BEYTrans: first experimental version

In this section, the position of BEYTrans compared with existing CATs is discussed, and then its concrete application is illustrated.

4.1. Environment features

A first experimental version of BEYTrans has been completed and deployed on the Web (BEYTRANS, 2007). However, it is necessary at this stage to clarify the position of our environment in comparison with other CAT functionalities. The comparison here is limited to Trados™ and Déjà vu™ (TRADOS, 2007) (DÉJÂ VU, 2007), but it could be extended to other CAT environments (CAT-COMPARISON, 2007).

Almost all the individual functionalities are present in these environments. The networked version is absent in Déjà Vu but is available in a separate version of Trados. Our environment integrates the basic translator functionalities for translators and the online collaborative functionalities for communities. Adding to that, all linguistic translation aids are suggested automatically and simultaneously in "proactive" behavior. Furthermore, BEYTrans is platform-independent – translators need only a PC with an Internet connection and browser –

4.2. Translator-oriented edition in a collaborative Wiki environment

The online translation editor has an Excel-like interface where all source and target TUs are displayed in parallel (Figure 2, area I). The editor allows translators to exploit dictionaries and machine translation asynchronously (Figure 2, area II).

Fuzzy matching detection is proposed at the same time the TU is selected in the main grid (Figure 2, area III). In fact, during the translation process, the editor proposes suggestions by computing similarity scores between the current source TU and the TUs stored in multiple translation memories (each translation community has its own TM, which stores previous translations). Accordingly, edition functionalities are important: translators are able to add, delete and split TU cells, which is useful for the manual enhancement of the tokenization (Figure 2, area IV). Translators can however choose the rate of similarity (Figure 2, area V) and fine-tune the environment parameters.

After a translation is complete, BEYTrans creates a new version of the TC (translation companion is an XML structure in which are managed the TUs) and sends it to its repository. As the content is saved in Wiki mode, it is easy to track the modifications by comparison that has been made to a translation, which enhances the efficiency and increases the consistency of the translation. Finally, translators are able to generate a target document that can be directly disseminated, and that also becomes available to readers on the BEYTrans Web site (BEYTRANS, 2007).

Table 1: Comparison of BEYTrans with April Déja Vu™ and Trados™.

	Atril Déja Vu™	**Trados™**	**BEYTrans**
Editor			
Looks like	Excel	Word processor	Excel
Bold/Italic/Underline formatting	No	Yes	Yes
Comments on sentence level	No	Yes	Yes
Backup/Restore facilities	No	No	Yes
Connection to Machine Translation	No	Yes	Yes
Translation memory			
Fuzzy matching quality	6/10	9/10	Needs a huge MT but works efficiently
Search results on the same screen	No	Yes	Yes
Difference marking	Yes	Yes	Yes

Automatic search	No	Yes	Yes
Multiple Translation Memory	No	Yes	Yes
Percentages shown	Yes	Yes	Yes
Terminology search			
Search results on the same screen	No	Yes	Yes
Multiple words	No	Yes	Yes
Automatic search	No	Yes	Yes
Multiple Terminology databases	No	Yes	Yes
Translation memory management			
Global search/replace	Yes	Yes	Yes
Translation unit edition	Yes	Yes	Yes
Networking			
Collaborative translation	No	Separate networked version	Yes
Networked terminology	No	Separate networked version	Yes

4.3. Linguistic resource functionalities

BEYTrans has been equipped with modules that allow for dictionary management. Functionalities related to dictionary management are as follows:

- Entire dictionary importation: a dictionary in its original format has to be preprocessed and transformed into the XLD format used by BEYTrans importation module, which extracts headwords and their corresponding translations. The newly imported dictionary is also activated automatically, and becomes immediately available to all translators (Bey et *al.*, 2006(a)).
- Progressive dictionary construction: translators are able to create temporary or permanent dictionaries. This process has been made quite simple: translators can specify the community and dictionary name, and can then automatically create a new dictionary for their own use.
- Look-up functions: the search function can be called up manually or automatically during translation. The selection of the TU in the editor area (Figure 2, area I) activates the sentence tokenization and dictionary suggestions.

Direct dictionary creation enhances translators' ability to work on their own data rather than on the data of the whole community. Related look-up functionalities are integrated with the translation aids, which makes it possible to select and update entries "on the fly" while translating.

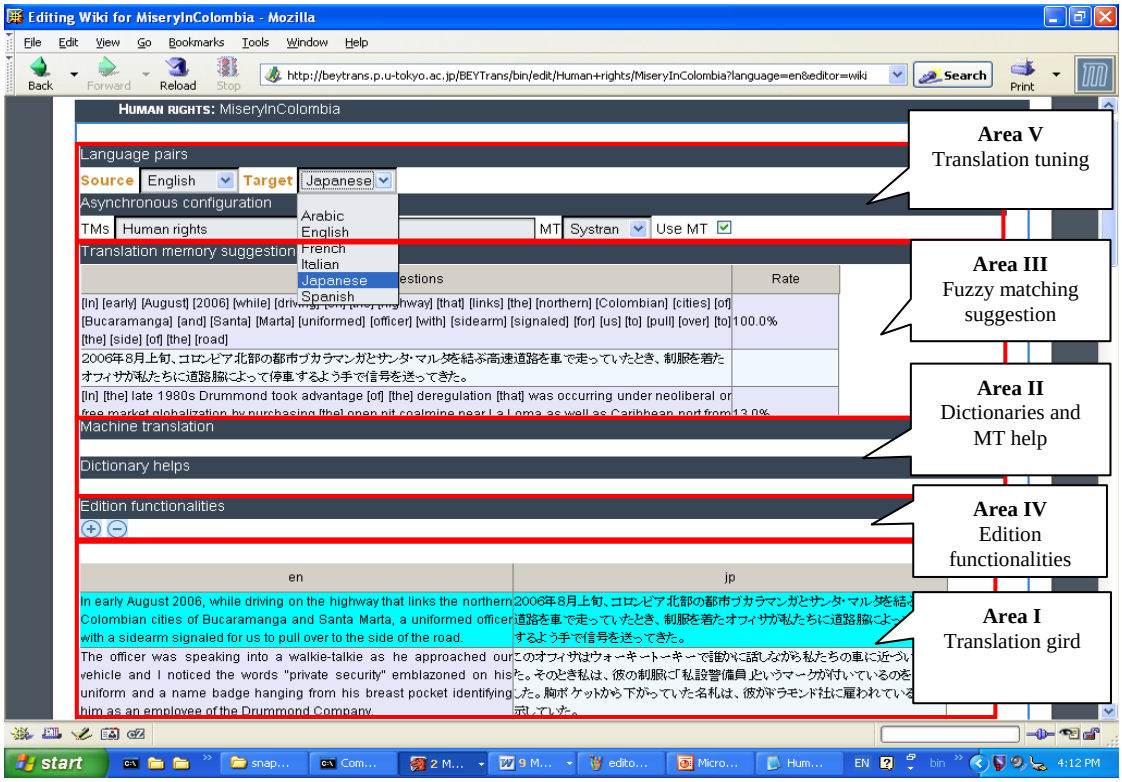

Figure 2: Multilingual online translator-oriented editor.

5. Experimentation

The environment is deployed on the Web for free translation (BEYTRANS, 2007) and is being used experimentally by the DEMGOL translation community (DEMGOL, 2007) and other translation communities. DEMGOL is an Italian research project which aims at the construction of an etymological dictionary of the Greek mythology and Homeric references. It contains about 1200 Italian documents which volunteers are translating into French, and which will later be translated into Spanish.

Beside DEMGOL, there is another ongoing experiment testing of multilingual functionalities in the translation of human rights documents from English into Japanese (Figure 2). Furthermore, as many volunteer communities aren't aware of our environment, we have recently studied the nature of individual and community translation work in ARABICMOZILLA (ARABICMOZILLA, 2007) and introduced our environment to these translators. We have imported an English-Arabic technical dictionary with around 18,000 entries for them to facilitate their work.

6. Conclusion

This paper has described the two levels of functionality in the experimental computer-aided translation environment BEYTrans. The first level corresponds to translators who act as separate entities and need specific functionalities (editors, linguistic help, etc.). The second level corresponds to the community level at which several translators work as an integrated entity (communication, progression control, etc.). Both functionalities have been integrated using a collaborative Wiki-based technology which provides volunteer translators with a user-friendly environment. However, the editor was the most important component for translators. When translators are working on a translation, the system suggests a variety of linguistic aids

(dictionaries, glossaries, etc.) and translation suggestions (MT, TM, etc.) to them in an asynchronous and "proactive" manner.

As Wiki-based environments are designed basically to support small documents, in the near future, BEYTrans will be extended for managing and translating high-scale data in multilingual format and will also be enhanced by tackling translator and community problems that are identified from the feedback from different translator communities.

References

Abekawa, T. and K. Kageura. 2007. QRedit: An Integrated Editor System to Support Online Volunteer Translators. *Digital Humanities*. 3-5.

ARABICMOZILLA. 2007. http://www.arabeyes.org/project.php?proj=Mozilla.

ARABEYES. 2007. http://www.arabeyes.org.

DÉJÀ VU. 2007. http://www.atril.com.

Augar, N., R. Raiitman and Z. Wanlei. 2004. Teaching and Learning Online with Wikis. *Proceedings of the 21st Australasian Society for Computers in Learning in Tertiary Education Conference*, Perth, pp. 95-104.

Bey, Y., C. Boitet and K. Kageura. 2006(a). The TRANSBey Prototype: An Online Collaborative Wiki-Based CAT Environment for Volunteer Translators. Yuste, E., Ed., *Proceedings of the 3rd International Workshop on Language Resources for Translation Work, Research & Training (LR4Trans-III). LREC 2006 - Fifth International Conference on Language Resources and Evaluation*. Paris: ELRA / ELDA (European Language Resources Association, European Language Resources Distribution Association), Genoa, Italy. 49-54.

Bey, Y., K. Kageura and C. Boitet. 2006(b). Data Management in QRLex, an Online Aid System for Volunteer Translators. *International Journal of Computational Linguistics and Chinese Language Processing*, 11(4), 349-376.

BEYTRANS. 2007. http://beytrans.p.u-tokyo.ac.jp/BEYTrans/.

CAT-COMPARISON. 2007. http://www.geocities.com/fmourisso/CAT.htm.

DEMGOL. 2007. http://demgol.units.it/show.do?action=base.

DHTMLGRID. 2007. http://sourceforge.net/projects/dhtmlgrid/.

FRENCHMOZILLA. 2006. http://frenchmozilla.online.fr.

LINGPIPE. 2005. http://alias-i.com/lingpipe/demo.html.

Mangeot, M. 2002. An XML Markup Language Framework for Lexical Databases Environments: the Dictionary Markup Language. *Proceedings of International Standards of Terminology and Language Resources Management Workshop*, Spain, pp. 37-44.

OMEGAT. 2007. http://www.trados.com.

PAXHUMANA. 2006. Translation of Various Humanitarian Reports in French, English, German, Spanish. http://paxhumana.info.

Schwartz, L., S. M. Clark Cossarin and J. Rudolph. 2004. Educational Wikis: Features and Selection Criteria. *International Review of Research in Open and Distance Learning*, 1(5), Australia, 95-104.

TRADUCT. 2007. http://wiki.traduc.org.

TRANSLATIONWIKI. 2007. http://www.translationwiki.net.

WIKTIONARY. 2007. http://www.wiktionary.org.

Walker, D. J., D. E. Clements, M. Darwin and J. W. Amtrup. 2001. Sentence Boundary Detection: A Comparison of Paradigms for Improving MT Quality. *Proceedings of the 8th Machine Translation Summit*, Santiago de Compostela, Spain.

YAKUSHITE. 2007. http://www.yakushite.net/cgi-bin/WebObjects/YakushiteNet.woa/wa/main.