Computational Linguistics and Chinese Language Processing
Vol. 11, No. 4, December 2006, pp. 349-376

Data Management in QRLex, an Online Aid System for Volunteer Translators'

Youcef Bey*+, Kyo Kageura+, and Christian Boitet*

Abstract

This paper proposes a new framework for a system which will help online volunteers to perform translations on their PCs while sharing resources and tools and communicating via websites. The current status of such online volunteer translators and their translation practices and tools are examined, along with related work also being discussed. General requirements are derived from these considerations. The approach taken in this study for dealing with heterogeneous linguistic resources relies on an XML structure maximizing efficiency and enabling all of the desired functionalities. The QRLex environment is under development and implements this new framework.

Keywords: Computer-Aided Translation, Web Search for Translation, Memory Translation, Helping Volunteer Translators, Linguistique Ressources.

1. Introduction

There have been many misconceptions concerning Machine Translation. In the early days, some researchers promised to "replace translators", while others like Bar-Hillel warned against the impossibility of FAHQMT (Fully Automatic High Quality Machine Translation) *in general.* The famous ALPAC report negatively evaluated the performance of Machine Translation (MT) systems at the end of 1966[1]. It is also known as the “infamous” ALPAC

* Laboratoire CLIPS-GETA-IMAG, Université Joseph Fourier, 385, rue de la Bibliothèque, Grenoble, France
E-mail : {youcef.bey; christian.boitet}@imag.fr
+ Graduate School of Education, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
E-mail: kyo@p.u-tokyo.ac.jp

[1] As a matter of fact, the ALPAC committee worked on obsolete and incorrect data, and did not even investigate a significant project in the same city (Washington D.C.), the GAT (Georgetown Automatic Translation) project. Further, several members of the committee were themselves heads of labs that received finds to work on MT. These members preferred to work on theoretical linguistics and AI (as it later became known) instead of the necessary engineering.

[Received May 1, 2006; Revised October 20, 2006; Accepted November 7, 2006]

report because it was biased, as explained in detail in the no less famous counter-report by Zbigniew Pankowicz, the polyglot USAF official who oversaw MT funding at RADC (Rome Air Development Center) from the early days until 1985. The truth, as recognized by Bar-Hillel in 1972 at a seminar on the usability of MT organized at Austin, Texas[2], is that MT can be quite useful in practice even if the "translational quality" is medium or poor. Later, when MT became widely available (first at the European Community, then on the French Minitel, then on PCs, and finally on the Web), it became clear that commercial MT, *if properly used as a tool to help translators by offering them a kind of "pretranslation",* can be used efficiently as "MT for translators". According to [Allen 2001] and according to experiments presented on [MT POST-EDITING 2006], the productivity of translators is often multiplied by three, for a variety of tasks.

Nevertheless, it was, and still is, true that MT has three conflicting goals that cannot be achieved together: full automaticity, high quality, and general coverage. However, two of these goals can indeed be achieved together. For example, the METEO system[3] showed decisively that an MT system specialized to an "adequate" sublanguage, in that case the language of weather bulletins (as opposed to weather situations or warnings), can produce better translations than the humans previously employed . Indeed, before 1980—1985, it took about 5—10 minutes to post-edit a bulletin translated by a junior translator, while it has taken only one minute from 1985 on, when METEO reached its top quality. Here, MT is really "MT for revisers" in that MT output can be post-edited without reference to the source text.

By contrast, wide coverage fully automatic MT systems cannot be used in this way at all. Because of unsolved ambiguities, the number of possible valid translations with very different meanings (including nonsensical ones) is extremely high, and it is not feasible to show them all to the post-editor. Exactly the *same* systems that can be very useful to bilingual post-editors are useless to monolingual post-editors. Martin Kay wrote that:

"...this happens when the attempt is made to mechanize the non-mechanical or something whose mechanistic substructure science has not yet been revealed..." [Kay 1997]

Yes, but that is not really the point! The point is that the same problem arises with human

---

[2] One of the few research groups really working on MT that continued to be funded after 1966 in the US, such as that of Pr Wang (University of California, Berkeley), and the newly founded Systran and Logos.

[3] That is the name of the operational system, which was further developed, improved and deployed on PCs by John Chandioux and his team, starting from the TAUM-METEO prototype built by the TAUM group at Université de Montréal around 1975-76.

translators: if they are asked to translate texts far out of their domain, they also produce incomprehensible results, impossible or very difficult for monolingual domain specialists to post-edit. Their errors are different, and their translations are more grammatical, but the time required to obtain polished translations based on their draft (or to decide that they are too poor for that and the text must be retranslated by a specialist) is on the same order.

In any case, this perception of MT has promoted research on *computer aided human translation*, which exploits the potential of computers to support translator skills and intelligence [Hutchins 1998]. Many industries have made large investments in developing useful translation-aid tools. These efforts have resulted in commercial Computer-Aided Translation (CAT) systems such as TM-2 (IBM), Trados, Déjà Vu, Transit, and Similis, which usually contain three components:

- *bilingual editors* (often embedded in text or document processors such as Word, WordPerfect, Ichitarou, Interleaf, etc.),
- *on-line terminology banks and dictionaries* (the latter being modifiable by translators and immediately updated), and
- *translation memory systems* (TM), which seek exact or fuzzy matches of the source segments to retrieve their translations as proposed translation [Bowker 2002].

There are two situations in professional translation. In the case of large and repetitive translation jobs such as successive versions of a product documentation, TM is quite useful, and MT is not used, even in the rare cases where its integration is foreseen, as in TM-2: the distance between users and developers is too great, so that MT dictionaries, especially for terminology, are not updated fast enough from the translators' dictionaries, whereas TM grows and becomes increasingly useful [Boitet 2005].

In the case of individual translators working on a variety of jobs, TM is not really useful, because the quantity of past translations of similar texts is too small. Accordingly, cheap commercial MT is used to obtain preliminary translation. (Even if no translations of complete segments are correct, many fragments are correct; hence MT functions as a kind of *dictionary in context).* Relatively few individual translators use commercial CAT tools anyway because of their high price.

What is the situation for online volunteer translators? Here again, there are two cases. In the first case, to which the QRLex system is addressed, translators are online in that they access a website to get documents to translate, retrieve resources such as dictionaries and TMs, deposit finished translations, and communicate with other translators. However, they don't translate online. Rather, they work on their PCs (or PDAs), just like many professional translators. However, they cannot afford to use commercial PC-oriented CAT tools, and, until now, they have not benefited from shared resources as do their professional counterparts.

In the second and more recently encountered case, volunteer translators do translate online, as on Translationwiki [TRANSLATIONWIKI 2006]. The documents to be translated are automatically segmented (paragraphs, sentences) and put up for translation. No CAT functions or resources are available.

In all cases, the CAT tools and resources available, if any, do not provide content and functions that fully satisfy all translators. There is thus a real need to aid online volunteer translators and their communities by providing them with a free environment with a rich set of linguistic resources and tools, and improved workflow and data management.

Recently, the number of volunteer translators has been growing sharply. Volunteers form or join communities, and they translate thousands of documents in different fields, thereby showing the true way to break the language barrier. These developments are mainly due to the Internet's crucial role in allowing translators to take part in such volunteer translation activities.

According to our study, volunteer translator communities are mainly of two types:

- *Mission-oriented translator communities:* strongly-coordinated groups of volunteers involved in translating clearly defined sets of documents (Linux-like communities). These communities translate what can be loosely called technical documentation, such as Linux documentation [TRADUC 2005], W3C specifications, and documentation as well as software (interface, messages, online help) of open source products. For example, in the W3C consortium, 301 volunteer translators are involved in translating thousands of specification documents into approximately 41 languages [W3C 2005]. Documentation in the Mozilla project exists in 70 languages, and is translated by hundreds of volunteer translators located in different countries [MOZILLA 2005].
- Network communities of s*ubject-oriented translators:* individual translators who translate online documents such as news, analysis, and reports and make translation available on personal or group Web pages [TEANOTWEAR 2005] [PAXHUMANA 2006].

  These translators are often involved in non-identified projects. They form translator groups with no *a priori* orientation, but they share similar opinions about events (anti-war humanitarian communities, translation of reports, news translation, humanitarian help, etc.).

In the following, the state of current online volunteer translation is first reviewed and related work intended to develop online computer aided translation tools according to the needs of online translator is presented. Then, several XML standards that are key components in the design of QRLex, and which solve the problem of managing heterogeneous data (such

as dictionaries, TMs, documents retrieved from the Web) are introduced. Finally, the results of the first two sections are used to justify the general architecture and the main features of the QRLex system, and the project's current status is presented.

2. Current Situation and Related Work

This paper will now review the existing translation environments and online tools designed to help online volunteer translators. In the first two subsections, the Lexical Knowledge Bases (LKB) by [Agirre *et al.* 2000] and an online Translationwiki system [Augar *et al.* 2004] [Schwartz 2004] [TRANSLATIONWIKI 2006] are presented. In the third subsection, a method for implementing a translation workflow demonstrating the usefulness of XML standards for managing the translation of documents and associated linguistic data is outlined.

2.1 Translator-Oriented Dictionary Systems

Back in 1994, X. Agirre and his team proposed developing Lexical Knowledge Bases (LKB) based on a model of "dictionary-use" by human translators [Agirre *et al.* 1994] [Agirre *et al.* 2000]. However, the structure of their LKB necessitates a complex transformation from existing dictionaries, implying, in turn, very heavy human labor. By contrast, this project wants to avoid intensive human work and simply give direct access in a uniform way to existing dictionaries, lexicons and term banks (Figure 1).

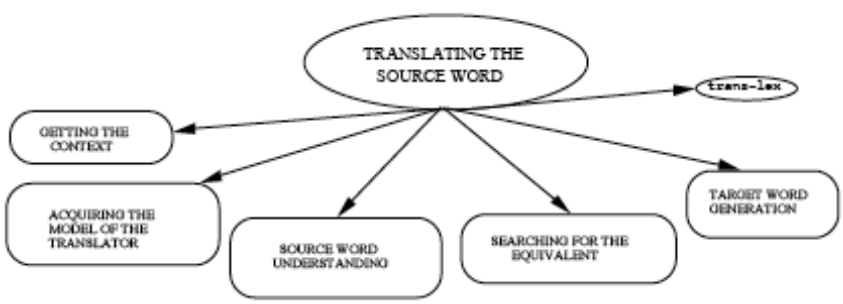


Figure 1. First level of the decomposition diagram of the tasks involved in the lexical translation process

First, they construct a monolingual model from monolingual French and Basque dictionaries. Then they develop a new French-Basque bilingual model. Two levels on top of the monolingual dictionaries have been proposed, which allow the establishment of links between the monolingual entities and facilitating translation from one language into the other.

The LKB is designed as an *"active tool",* which means that the dictionary tool works autonomously to present translators with potentially useful information and user functionalities during the translation process. The concept was inherited from [Martin 1990]

who stated:

"...the use of dictionary can be seen as a typical problem-solving activity, and user-orientation should involve both static and dynamic features of the intended user..."

[Agirre *et al*. 2000] add:

"Furthermore, along with the usual information about the meaning of the entries, dictionaries should show how to use words in context. In other words, we advocate that dictionaries should actively co-operate in finding the correct translation."

They emphasize that such LKBs are useful only if the translators are involved in the design and development of their functionalities. Speaking of their previous work (done in 1994), they say:

"The LKB provides various access possibilities to data. Even so, limitations are present when trying to exploit this knowledge in a lexical translation context. The cause of this limited usability is that the lexical organization was designed from a general perspective, without taking into consideration functional aspects. Incorporating this functionality means, in our case, transforming such LKB into a user-oriented dictionary system."

The study of translators' behavior during the translation process prompts one to take several important points into account when designing dictionary systems for translators:

- Expert and occasional translators need distinct and adapted sorts of help.
- Some translators (especially occasional ones) find bilingual dictionaries very useful.
- Multi-word terms are a source of failure when using normal dictionaries.
- Context is important when translating a text.
- Dictionaries for translation must give grammatical and usage information.
- The proximity between languages is helpful, but attention must be paid to "false friends"; dictionaries must prevent translation errors derived from them.

The approach followed for developing the LKB is very pertinent, especially because the model, behavior and needs of translators are taken in consideration. Integrating information related to translators' behavior in parallel with linguistic data is a new and promising direction for the design of future CAT tools. This idea should be extended to the integration in a CAT

environment of document and translation workflow management tools together with language-oriented resources and functionalities.

2.2 Online Collaborative Wiki-Based Translation Environment

Translationwiki.net is an online collaborative translation Web service [TRANSLATIONWIKI 2006]. It is based on a Wiki technology which allows translators/users to collaborate and share knowledge on the Web (Figure 2). There are several steps:

- choice of source documents, normalization of format and character encoding;
- automatic segmentation into translation units (TUs), which are paragraphs or if possible; sentences (see);
- translation proper;
- dissemination of translations in various formats.

In the Translationwiki environment, any user can upload a document for translation. The textual content is extracted and segmented automatically to TUs. No quality checking of translation is performed by the site manager (who only manage the environment and presumably has no time and does not know the target language(s)), but if translators or readers notice vandalism, modifications by suspect sources are erased and documents can be protected or semi-protected.

During translation, translators process only one TU at a time and can not see the whole document, as each TU is actually handled as a small Wiki document. Hence, translators need to navigate through the documents to check coherence and avoid translating the same expression differently in different places.

Figure 2. Main user interface of translation in translationwiki.net

Translationwiki is currently limited to five languages (Arabic, Chinese, French, German and Italian). Translators/users can sort and search documents in one language at a time. As for

the direction of translation, an uploaded document may be translated only into one of the supported languages. Translators cannot translate the same document into more than one target language in the same interface, and cannot manage the multilingual content of given document, if any (for instance, they cannot keep original fragments as citations in the translation).

2.2.1 Translation Methods and Interface for Editing

Documents are accessible directly from the main list. Volunteer translators are invited to select documents by clicking on their titles. A new screen appears which displays the source of the TU on the left side and an editing area for the target language on the right side. The editor is a simple text area without any formatting functionalities or linguistic aids (Figure 3).

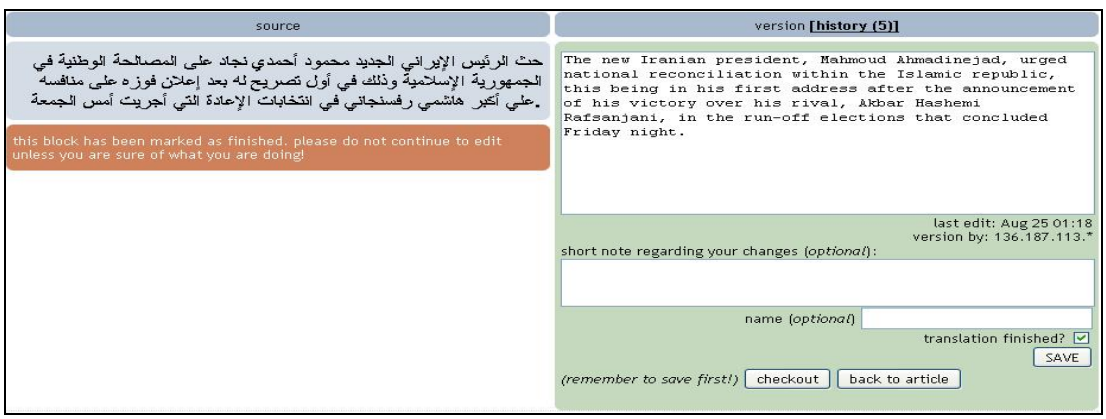


Figure 3. Translation editor

2.2.2 Translation Units (TU) and Versioning

In this environment, the versioning module keeps the history of the modifications. This allows translators to check the evolution of a translation and avoid losing content. Translators/users can easily restore old translations deleted erroneously or by vandals. When the translation of a TU is finished, the system keeps puts its translation in its repository and allows the translators to check the differences between different translations in a user-friendly interface (In Figure 4, the red terms are the result of comparing two versions of a translation into English from Arabic. All versions are listed and can be compared pairwise.)

Metadata is attached to each modification, so that, for all versions, it is easy to determine the date of the last modification and to identify the users by their profiles. Hence, users can follow the introduction and modification of content, and can distinguish which other translators produce high quality translations and which ones don’t.

Figure 4. Wiki-based version management of document.

2.3 Translation of Open Software and Associated Documents by Volunteer Translators

There are many projects aiming at the translation and localization of open software and associated documents. Two quite interesting projects are [MOZILLA 2005] and [TRADUC 2005]. Mozilla is a set of open software tools that includes a web navigator, an HTML page composer, and an e-mail manager. It is available in 70 languages. Translation in this project is a continuous process because each new version has new documentation and a new interface that must be translated. Two main categories of documents have to be translated by volunteer translators:

- *Interface translation:* messages are text stored in various files (Table 1). Volunteer translators download them for translation using CVS (Concurrent Versioning System).
- *Online help documents:* documents are HTML pages. They are translated at the end of the new release because they often contain screenshots which may change up to the last minute.

It is interesting that most online translators show similar behavior. In all the TRADUCT and MOZILLA localization sub-projects (one per language), volunteer translators are invited to translate a list of documents which have been put up on a website (one for each sub-project) in different formats (XML, SGML, HTML, HLP, plain text, etc.). First, they check whether the relevant document has been translated; if not, they make a reservation and announce to other translators, via a discussion list or via e-mail, that they have begun to translate it. To obtain the source document, they download it directly from the CVS (Concurrent Version System) or ask the coordinator to send it via e-mail (Figure 5).

Table 1. Document types in the Mozilla project

File type	Extension	Description
Module description files	RDF	XML files containing metadata (version number, language, etc.)
Interface files	DTD	XML files containing a textual part (Text for user interfaces).
Property files	Properties	Files containing messages to be displayed in dialog boxes.
HTML files	HTML	These files contain the online help.

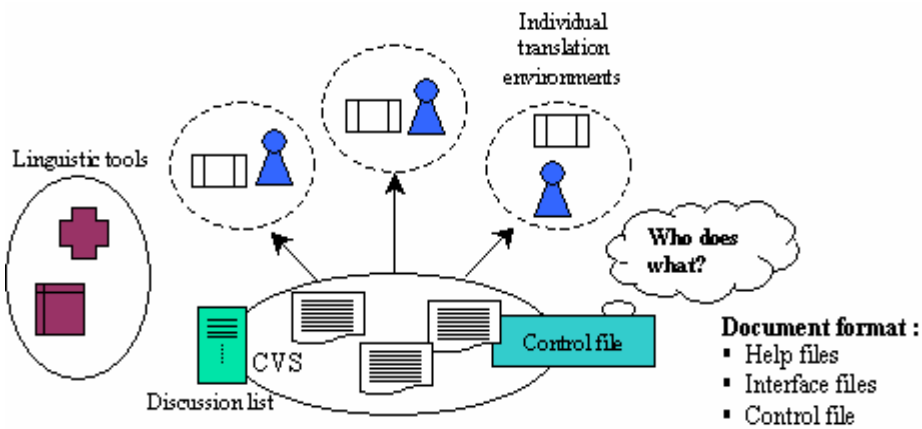


Figure 5. Translation method in Linux communities

Once the document is obtained, each translator uses his or her individual translation environment, which often varies from person to person. A typical personal environment consists of a set of tools: textual editor, dictionaries (electronic, paper version, or online), glossaries, terminology, and sometimes Translation Memory (TM). In addition, the importance of the Internet should be noted, as it has become a precious linguistic resource for translators, who use it to recover existing translation segments (such as quotations, collocations, technical terms, etc.).

Documents are translated into their original format. For example, in the TRADUCT project, documents are structured in XML DocBook[4]; translators translate only the text which lies between XML markers. After the translation is finished, they send the whole target document in the same structure to the coordinators who transform it to readable document and disseminate it on the Web.

[4] A rich XML format used to produce readable HTML with OpenJade; for further information refer to http://www.docbook.org.

Translation-aided tools offered on web servers for aiding translator communities may contain quite poor linguistic resources, but may also contain some useful management-oriented facilities:

- *a set of local free dictionaries*, glossaries, and links to other linguistic web sites.
- *a discussion list* used for exchanging skills and resolving most issues faced during the translation process.
- *control files* for checking "who does what" [MOZILLA 2005]. This is useful for the collaborative translation of a given document by several volunteer translators: before starting to add to a translation, a translator finds the most recent endpoint, and starts translating from there.
- *a server for managing the document versions like the CVS server* (Concurrent Version System).

The linguistic resources are almost never maintained and updated, because of the lack of automatic tools for synchronizing modifications made off-line, and because a process and a team to validate and consolidate updates are lacking.

3. Elements Reusable from Current Professional Practice

3.1 Translation Workflow and XML Standards: the IBM Localization Model

The localization process generally consists of several steps, from document creation to the final translation [IBMLOCALIZATION 2005]:

- document creation
- preparation of translation
  - normalization of format and character encoding
  - automatic or semi-automatic segmentation into TUs
- translation proper, followed by quality checking
- dissemination of translated document, possibly in various formats (PDF, HTML, etc.).

After segmentation, a document is uploaded to the server and added to the list of documents to be translated, and then is assigned to a translator. (This study's procedure is different mainly in that volunteer translators decide which documents they will translate. Otherwise, the processes are similar.)

The translator downloads the document in a textual format suitable to his or her CAT editor (TM-2 in the case of IBM), together with a "kit" containing a document-specific translation memory and a dictionary, extracted from the resources available on the server.

Each step presents some problems. For example, a good localization (*e.g.* a user's guide for electrical appliances, websites, slide shows, scripts for advertisements, etc.) should keep the document format intact and produce a high-quality translation. If the localization is a software element, the localization should fit nicely into the interface without causing any trouble.

In the following paragraphs, a translation workflow method exploited by IBM™ called *reverse conversion* is presented, which shows the usefulness of XML standards for managing crucial data during translation. Accordingly, these standards for QRLex data management will be adopted here.

3.2 Localization Methods: the Reverse Conversion Workflow

Document creation is a step performed by an individual or by an independent group. The resulting document may be in various formats, such as those of graphical user interfaces or of help and manual files.

The translation of documents containing heterogeneous data in various formats is a heavy task for translators, who must attend closely to these documents both during and after translation (during post-translation). Other difficulties are related to the duplication of content, which increases translation time, and related to the production of the final version.

Another problem is that documents need some adjustment in the post-translation stage, for example, because text length varies from language to language.

To overcome such problems at each step of the translation, the IBM teams have adopted the *reverse conversion* translation workflow.

It consists of extracting only the relevant part of the materials to be translated, and putting it in an XML format for transfer to translation services. Figure 6 illustrates the translation workflow from the creation of a document to the production of the translated documents.

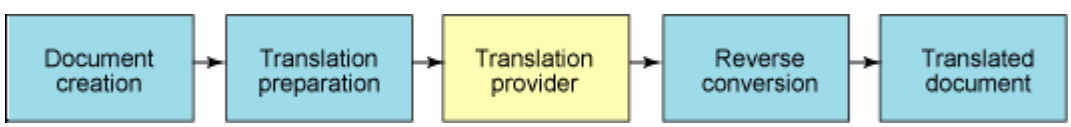


Figure 6. The reverse conversion localization workflow[5]

As some critical data, such as confidential data, may have to be kept secret, the producer of the document should not put it in any part of the source code which goes out for translation.

[5] Blue boxes in the graphics represent a process that takes place in house; yellow boxes show tasks done by the translation agency.

Such data is not extracted, but kept in the "skeleton" of the document.

The *reverse conversion* consists in extracting the translated TUs from the XML documents returned by the translators and merging the translation with the remaining data from the original document in order to produce the final translated document [IBMLOCALIZATION 2005].

3.3 XML Standards for Data Management: TMX (Translation Memory eXchange)

Using XML standard formats for structuring data to be translated makes the management of documents and translation easier. In the first place, such standardization facilitates management of linguistic resources and documents in several CAT (Computer Aided Translation) tools. In addition, using standard formats reduces costs and increases the productivity of translation.

Before terminology standards appeared, terminology specialists and lexicographers exchanged data in various formats. In order to facilitate reuse of linguistic resources and increase communication exchange inside or outside a given organization, a unified format for structuring data seems important. The main aims of international standards are these:

- To facilitate the reuse of existing lexical databases, terminology term bases, translation memories, dictionaries, etc.;
- To increase the data flow between people;
- To facilitate the exchange of data between CAT tools;
- To decrease future programming efforts and avoid the need to define new structures.

TMX is developed by LISA (the Localization Industry Standards Association) for managing multilingual translation memories [LISA 2006]. It is an XML format, where the *seg* element includes translation units in several languages. Figure 7 gives an example of a source/target text within an HTML document. The same content is presented after conversion into TMX in Figure 8.

```
<p>The big<b>black</b>cat</p>
<p>Le gros chat<b>noir</p>
```

Figure 7. Data in tagged HTML documents

Most CAT tools already offer a text-based input/output format that can be used to transfer data to and from other instances of the same application [SIMILIS 2005] [TRADOS 2005]. Adjusting to the use of the TMX format, as an alternative to managing translated segments, should be relatively straightforward as it is not a complex format and there are

plenty of freely available XML parsers. Here are some advantages of this format:

- *Exchange of memories:* the most immediately obvious benefit of TMX is that it allows translation memory information to be exchanged between existing CAT tools, which permits increased communication between linguists.
- *Choice:* once a standard has been provided and its use has been encouraged, experts are free to change tools, which ensures that they don't become locked in to a particular product.
- *Openness:* given a clearly defined standard, developers of other tools have the opportunity to complement existing translation functionalities with new or proprietary features that can benefit the translation process (Figure 9).

```
<body>
 <tu tuid="0001">
   <tuv xml:lang="EN-US">
     <seg>
      The big
      <bpt id="1"x="1"><b></bpt>
      black
      <ept id="1"></b></ept>
      cat
    </seg>
      </tuv>
```

```
<tuv xml:lang="FR">
  <seg>
    Le gros chat
    <bpt id="1" x="1"><b></bpt>
    noir
    <ept id="1"></b></ept>
  </seg>
  </tuv>
   </tu>
  </body>
```

Figure 8. TMX example for translation memory management

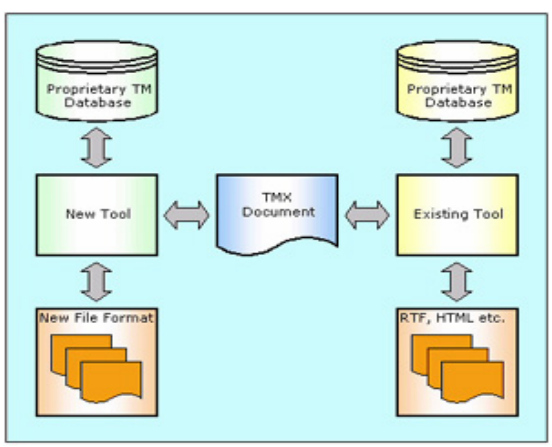


Figure 9. Openness feature of the TMX Standard

4. Data Management in QRLex

In designing the QRLex environment, the designers tried to take into consideration all relevant strengths of the existing environments described above and combine them with the advantages of the use of XML for managing heterogeneous data. Following LISA, the designers use TMX for handling multilingual content and aligned translatable units. To manage heterogeneous linguistic data, however, the designers have developed and used the new XLD (XML Linguistic Data) format [Bey *et al.* 2005].

4.1 Specification

4.1.1 Managing Heterogeneous Linguistic Data: the XLD Format

The following are among existing reference data for which the management structure is defined (Table 2):

- "Eijiro" and "Grand Concise" are two high-quality English-Japanese unidirectional dictionaries widely used by many translators;
- "Nichigai" is specifically used for proper names;
- "Medical Scientific Terms" is included to check the structure of terminological dictionaries;
- "Edict" is a free Japanese-English dictionary, included for checking the directionality of the bilingual dictionaries.

Table 2. Reference data in the QRLex framework

Reference Data	Description	Entries	Format
Eijiro 86	General English/ Japanese dictionary (EDP 2005)	1576138	Textual
Edict	Free Japanese/English Dictionary	112898	Textual
Nichigai	Guide for spelling foreign proper names in Katakana	112679	Textual
Medical Scientific Terms	Medical terms (terminology)	211165	Textual
Grand Concise	Japanese/English Dictionary	360000	XML

For each reference work, there are a few requirements: (a) various levels of recyclable units should be dealt with in a unified framework; (b) existing high-quality content should be properly accommodated; and (c) unnecessary information contained in existing content should be properly excluded, while necessary information appropriated for reference data and useful for translators should be incorporated. To satisfy these requirements, we need an internal XML structure for storing and exchanging content within different QRLex modules.

Existing high-quality reference data in electronic form takes a variety of formats. After examination of existing XML standard formats for terminologies such as TBX (TermBase eXchange) and MARTIF (Machine-Readable Terminology Interchange Format), the authors found that these formats, unfortunately, did not satisfy the above requirements [LISA 2006]. The authors considered using the XML CDM [Mangeot 2002] format for the unified presentation of many monolingual, bilingual and multilingual usage dictionaries in the Papillon database, but it is somewhat too complex for the needs of this structure.

The authors have thus defined the basic XML structure of the linguistic data by reference to the data elements of various dictionaries and terminological lexicons. Figure 10 illustrates the XLD (XML linguistic data) format that has been developed for managing heterogeneous reference data.

The XLD format consists of three main parts:

- *Source reference data description:* contains the description of the original linguistic resources and their content. This XLD header information includes the creation date, the author profiles, the encoding, the number of entries, the source language, etc.
- *Source element:* contains the source entry and its description, *e.g.* the language (xml:lang).
- *Target element:* this element is multilingual. Its sub-elements contain the translations of the segment into several target languages.

```
<!-- XLD (XML Linguistic Data) structure definition -->
<!-- Part 1: General description of original version and content -->
<!ELEMENT resource    (res-info, content)>
<!ATTLIST res-info    name CDATA #REQUIRED>
<!ATTLIST res-info    author CDATA #IMPLIED>
<!ATTLIST res-info    version CDATA #IMPLIED>
<!ATTLIST res-info    date-creation CDATA #IMPLIED>
<!ATTLIST res-info    last-modification CDATA #IMPLIED>
<!ATTLIST res-info    original-codage CDATA #IMPLIED>
<!ATTLIST res-info    entries-number CDATA #IMPLIED>
<!ATTLIST res-info    description CDATA #IMPLIED>
<!ELEMENT content      (entry*)>
<!ELEMENT entry        (source, target)>
<!ATTLIST entry        id CDATA #IMPLIED>
<!-- Part 2: Source element definition -->
<!ELEMENT source       (#PCDATA)>
<!ATTLIST source       xml:lang CDATA #REQUIRED>
<!ATTLIST source       additional-info CDATA #IMPLIED>
<!-- Part 3: Target element definition -->
<!ELEMENT target       (expression+)>
<!ATTLIST target       xml:lang CDATA #REQUIRED>
<!ELEMENT expression      (#PCDATA)>
<!ATTLIST expression add      kata-pronunciation CDATA #IMPLIED>
```

Figure 10. DTD (Document Type Definition) of XLD format

Source and target elements contain additional information expressed using a set of attributes:

- *Additional-info:* description of a source element. This will be useful if one transforms the relevant resource direction from, for example, Japanese-English into English-Japanese.
- *Kata-pronunciation:* In the case of Japanese linguistic data this attribute contains the pronunciation in katakana of the foreign words.

Figure 11 shows an entry after the compilation of the "Nichigai" in XLD format. The source element contains the English transliteration of an Arabic proper name and its transliteration in katakana.

```
- <entry id="Nichigai100001280">
    <source xml:lang="en" additional-info="">Abdeslam</source>
  - <target xml:lang="jp" kata-pronunciation="">
      <expression id="Nichigai100001280-1">アブデスラム</expression>
    </target>
  </entry>
- <entry id="Nichigai100001290">
    <source xml:lang="en" additional-info="">Abdessadki</source>
  - <target xml:lang="jp" kata-pronunciation="">
      <expression id="Nichigai100001290-1">アブデサドキ</expression>
    </target>
  </entry>
```

Figure 11. Japanese "Nichigai" entries in XLD format

The developers compiled all of the linguistic resources cited in Table 2 in the XLD structure, in other words, the existing resources have been preprocessed, filtered, and, after that, passed to the structure manager for transforming them in XLD XML documents.

4.1.2 Managing Textual Data: the TMX Format

The document data structure should satisfy two requirements: (a) maximal facilitation of the provision of recyclable units and (b) unified management of translated documents. The first requirement comes from translators, who avidly seek existing translations of linguistic units (especially collocations and quotations) in related translations. The second requirement comes from the mission-oriented community in which translators take part. Although no readily usable reference data format was found, an existing standard framework TMX suitable for the developers' aims was found.

This standard simplifies the storage of textual data extracted from documents that contain formatting information such as HTML tags. It allows one to represent and manage translation memories as well as "multilingual" documents, that is, documents containing source and target translation units in the same file [LISA 2006]. Figure 12 illustrates TMX as used for an

English-French-Italian example translated by volunteer translators of the PAXHUMANA community [PAXHUMANA 2005].

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <tmx version="1.4">
  - <body>
    - <tu tuid="0001">
      - <tuv xml:lang="en">
          <seg>I recently caught a glimpse of the effects of torture in action at an
            event honoring Maher Arar. The Syrian-born Canadian is the world's most
            famous victim of "rendition," the process by which US officials outsource
            torture to foreign countries...</seg>
        </tuv>
      - <tuv xml:lang="fr">
          <seg>J'ai récemment eu un aperçu en action des effets de la torture lors
            d'un événement en l'honneur de Maher Arar. Ce Canadien d'origine
            syrienne est la plus célèbre victime d'un genre d'extradition spécial
            appelé « restitution » [rendition], qui est un procédé par lequel les
            fonctionnaires des États-Unis sous-traitent la torture dans d'autres
            pays...</seg>
        </tuv>
      - <tuv xml:lang="it">
          <seg>Ho recentemente avuto un compendio in azione degli effetti della
            tortura durante un'avvenimento in onore di Maher Arar. Questo Canadese
            di origine siriana è la vittima più famosa di un genere di estradizione
            speciale chiamato "restituzione" [rendition] un procedimento con il quale
            i funzionari degli Stati Uniti subappaltano la tortura in altri paesi...</seg>
        </tuv>
      </tu>
    </body>
</tmx>
```

Figure 12. Source and translated document in TMX format

Structuring linguistic data in an XML format is appropriate according to the needs of online volunteer translators (as already explained) and accords with the overall design of QRLex for managing linguistic data in heterogeneous formats. Such structuring also makes it easy to construct a parser and to develop improved functionalities. At the management level, all imported data (reference data or textual data) and the information flowing between modules will be stored in XLD and TMX format.

4.2 Modular Architecture of QRLex Environment

By talking with volunteer translators and coordinators and by examining existing translation aid systems [SIMILIS 2005] [TRADOS 2005], the developers clarified a few essential general requirements:

- Content of language reference tools cannot be separated from system functionality.
- Translators look for information on (i) ordinary words, (ii) idioms and set phrases, (iii) technical terms, (iv) proper names, (v) easy collocations, and (vi) quotations. In general they conceptually distinguish these six classes but want to look them up with unified functionality and interfaces.

With respect to reference contents, therefore, the developers' needs are as follows:

- Use and update good reference material whenever it is available.
- Enhance the material when it is not sufficient.
- Make reusable translation units available from existing relevant translated documents.

Taking these general desiderata into consideration, the developers have defined a system that implements the QRLex framework by means of six functional modules (Figure 14), each of which covers specific tasks and deals with different types of data: a structure manager, a document manager, a database manager, a data control manager, a functionalities manager, and the Akin system.

- **Structure manager:** a module that transforms reference data and textual data into a structured XML format. The authors have thus compiled linguistic data including dictionaries, a Japanese pronunciation guide, technical terms, and proper name resources in XLD format. This module preprocesses and filters original linguistic data in various formats before transforming them to XLD and storing them in the centralized database. In the same manner, source documents and corresponding translated documents are processed in the documents manager module and converted into structured LISA TMX standard (Translation Memory eXchange) format [LISA 2006].
- **Document manager:** This module is based on three functionalities for the detection/extraction of textual content from document. The following paragraphs explain these functions:
  (i) *Direct document detection:* QRLex gives online users the ability to upload source documents and their translations for internal storage. This functionality needs no access to Web searches; instead, it involves extracting the content from both source and target documents and aligning the uploaded documents.
  (ii) *External document detection:* volunteers can search the Web to detect translated documents which can help them translate current documents. The search is carried out by crawling the Web in search of documents with bilingual content. For this purpose, the Akin-I[6] system has been developed. Currently, it detects only English-Japanese (in both directions), but enhancement is under way to generalize the system for the detection of other

[6] Developed at the Graduate School of Education (the University of Tokyo) by the team of Professor Kyo Kageura.

language pairs. Other functionalities such as the detection of bi-segments would also seem very helpful since translators need to know how segments (*e.g.* expressions, idioms, collocations, etc.) have been previously translated by other volunteers.

(iii) *Internal documents detection:* this process is based on the result of Akin-I. After the initial detection of source and target documents in a specific community on the Web, the process cyclically returns to the Web, seeking additional translated documents. In the future, Akin-II, developed for crawling internal repositories after the initial identification of translation communities, will also be integrated.

The identified documents are subjected to (i) text extraction, (ii) segmentation and (iii) alignment [Walker *et al*. 2001]. The detection of sentences, or more precisely TUs, is achieved using the LingPipe tool [LingPipe 2006], which carries out sentence-boundary detection (detection of TUs) and linguistic unit detection (*e.g.* named entity detection). LingPipe can be trained to support additional languages (*e.g.* Chinese, Arabic, and French). Finally, a bi-text is constructed for each document, stored in TMX format, and put into the centralized database.

- **Database manager:** this module is the server of data to all modules of QRLex. All data flows are centralized in a relational database, which receives linguistic data in XML format from the structure manager module and serves the functionalities manager module and data control manager module. Structure data is analyzed using the DOM API[7] for the extraction of data from both XLD and TMX formats.
- **Data control manager:** open linguistic resource environments on the Web necessitate the intervention of human experts. In this case, the QRLex environment requires the interaction of linguistic experts or professional translators to increase the accuracy of data content and enhance the control of user interaction. This module is subdivided into two sub-modules:

  (i) *Data validation and enhancement:* the validation/enhancement process allows the environment to interact with linguists/translators or lexicographers via an interface to temporary data which has been put on the system for revision. Some such users have password permission to revise and update the

[7] All the internal dataflow is in XML and the whole environment exploits structured data via Document Object Modeling. For further information, refer to http://www.w3.org/DOM/.

data. The content is controlled by active translator communities[8] who continually maintain it and work actively to enhance specialized data for their fields of translation.

(ii) *Linguistic data control and administration:* the administration sub-module helps the administrator to control all access to the environment. He or she has the authority to suspend users (*e.g.* vandals) or to manage copyrights as appropriate. Furthermore, he or she can give access to information with hierarchical levels of privilege, so that users may have access to all of the data or only to parts of it. The authors emphasize that there is an administrator for each translator community, with the power to control the interactions of users with the environment and data.

- **Functionalities manager:** functionalities are the most important elements of CAT tools. Considering the needs of translators, the authors view functionalities as the most important criteria to be considered during tool development. This module is the main interface for the interaction between translators/users and QRLex. It offers the possibility to display, update and use data simultaneously during translation.

  The designers have thus developed QRedit, a specialized translation editor (http://hygrocybe.p.u-tokyo.ac.jp:8080/qredit_idiom), as a first attempt to allow volunteer translators to do translation with the possibility of exploiting existing dictionaries. Figure 13 gives a snapshot of the editor where the document is downloaded automatically and compiled with dictionary entries for increasing translation speed. Translators do not need to look anything up in a dictionary. In fact, all words in the source document (left side) are linked to their translation entries in the dictionaries. Translators have only to choose words to translate by moving the cursor to them and their translations are directly displayed in a *pop-up* window. When the translation is selected, it is automatically put in the right position in the target text area.

- **Akin System:** the detection of existing translation documents is carried out by the Akin system [AKIN 2006] [Tsuji *et al.* 2005], which detects English-Japanese translated documents using keywords (Figure 15). Integrating Akin into the QRLex framework allows:

  (i) *Avoiding the duplication of translation:* volunteers often check existing translations on the Web before starting the translation of a given document.

[8] The translator communities are active when the content is checked daily. This work phenomenon depends on the will of each community but often translators look for another translation and coordinate it together whenever it is possible.

They often search manually, but could make use of more efficient methods. Akin-I is intended to be called by the document manager module to check whether the relevant document has been translated on the Web or not, which avoids the translation of the same document on the Web.

(ii) *Recycling Web and community repositories:* Akin aims to prepare for the construction of TM by detecting and recycling the existing translation along with crawling the repositories of the translation communities on the Web.

(iii) *Detection of bi-segments:* Akin can be exploited at several levels. It can detect repositories of translation communities and documents at a high level, but also allows the detection of bi-segments (in source/target language) at a finer grain.

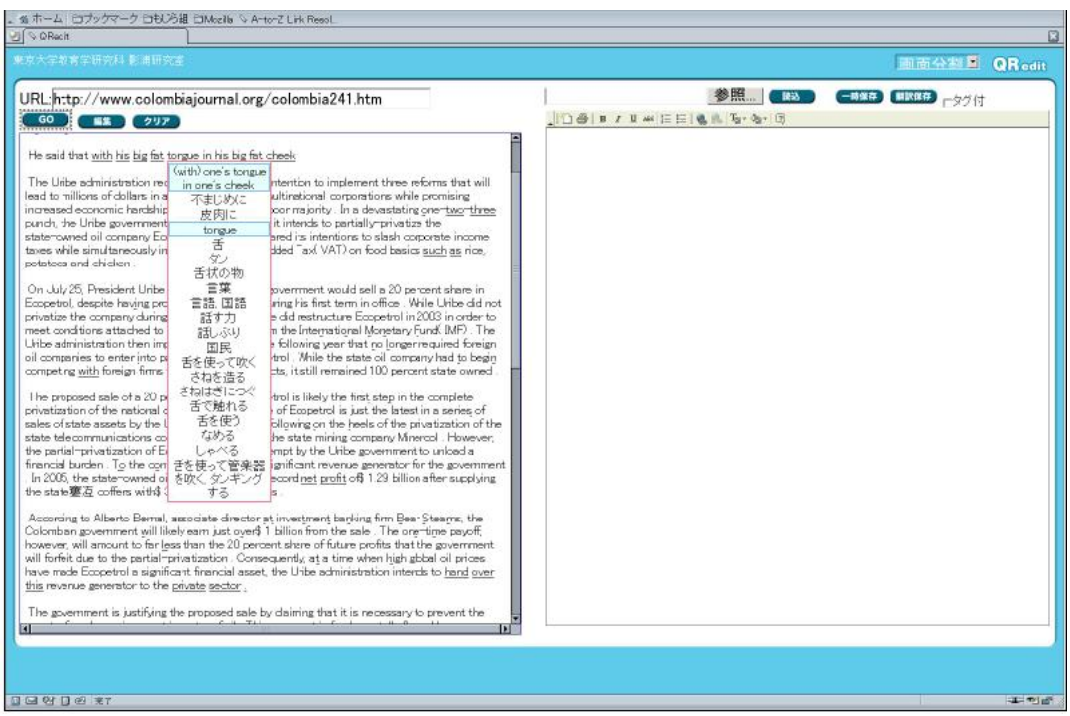


Figure 13. Screenshot of the QReditor

Keywords for search are translated using the Eijiro dictionary. They are translated into either English or to Japanese according to the direction of the desired search. The Akin search method thus differs from these of similar systems like STRAND [Resnik *et al*. 2003], which collect parallel corpora even for software or interface components.

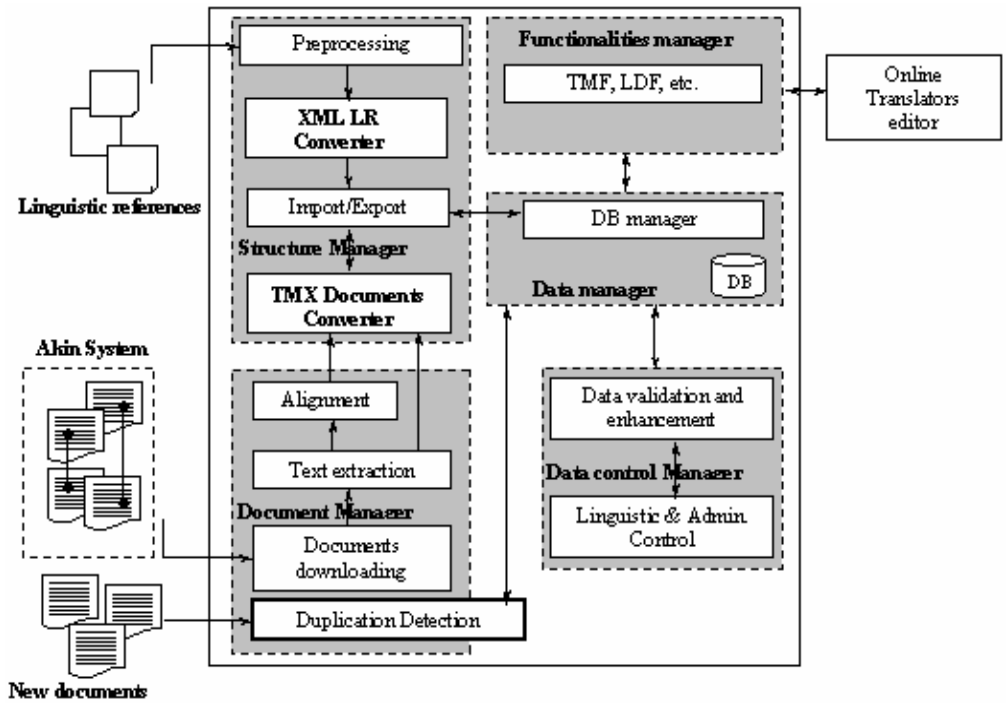
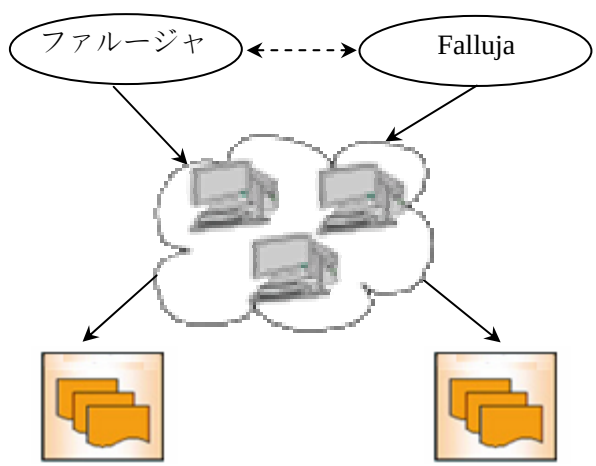


Figure 14. QRLex architecture for the data management



Source: Japanese documents Target: English documents

Figure 15. Detection of translation on the web using the Eijiro dictionary

Collecting parallel corpora can be helpful for translators, but may not be sufficient as filtering may be needed, and volunteers who seek translations from specific communities or translators may be dissatisfied. Hence, in this framework, collected documents tend to be relevant when they are collected from specific translation communities; if they are found elsewhere, translators may have to create their own translations.

Figure 16 gives an actual snapshot of results produced by Akin-1. The entries are numbered and selected according to the direction of keywords language (*e.g.* ファルージャ keyword). The URL of the source document in Japanese is displayed with its title and description. The target document is displayed in the same manner as the source document. The score (displayed at the bottom) shows the degree of matching between source and target document. It is calculated according to the number of words from the Japanese document which match translations in the English document.

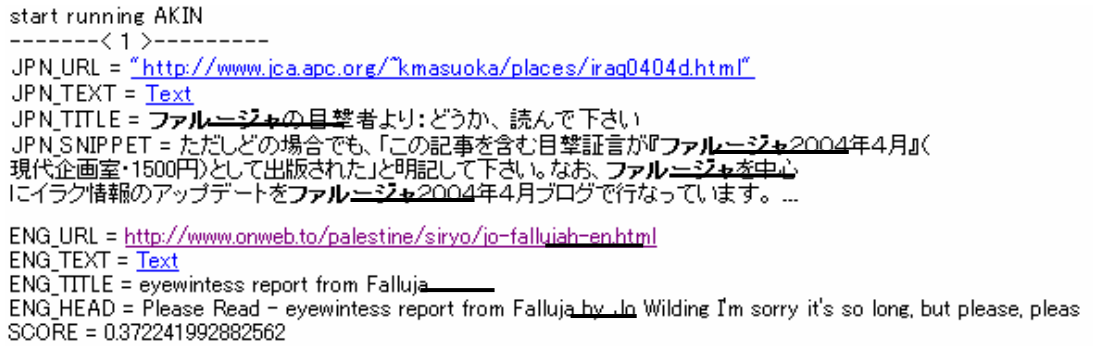


Figure 16. Detection of existing translation documents on the Web.

The Akin system, developed separately, aims to identify or detect existing documents which have been translated from English to Japanese. However, most translators disseminate their translations through specific locations on the Web, which serves as the repository of their respective communities. To take advantage of these internal community repositories, Akin has been improved to enable recycling of existing translated documents within each community rather than search of the entire Web.

Several research groups have taken part in the QRLex project, each exploiting its special skills. The server will be set up at the University of Okayama under the direction of Dr. Koishi, who is also working with his team on the automatic compilation of Japanese-French and Japanese-English terminology. The University of the Okayama is contributing help for the construction of bilingual Japanese-French linguistic resources, for alignment of resource, and for other related tasks.

5. Conclusion

The authors have proposed a new framework for a system which will aid online volunteer translators to perform translation on their PCs while sharing resources and tools and communicating via a Web site. The current status and conditions of online volunteer translators and their translation practices and tools have been examined, and related work has been discussed. The researchers examined translators' needs, first by analyzing various translation scenarios within existing online translator communities and existing environments, and subsequently by interviewing online translators. This work has clarified and modified the authors' views regarding the design of a new framework emphasizing two aspects: (i) a rich content and (ii) improved functionalities.

The system's general requirements have been derived from these main points of emphasis. Most translators request rich content in various formats, such as dictionaries, glossaries, and translation memories. The developers have accordingly developed the XLD format for compiling heterogeneous linguistic data, *e.g.* for storing usable free dictionaries, and for allowing importation of new linguistic resources to centralized relational databases within the QRLex system. At the same time, a translation memory (TM) constitutes a precious linguistic resource which most translators need to accelerate translation and improve its quality. TM will be constantly developed by recycling the documents translated on translator community Web sites or documents found on the Web by specialized search utilities like the Akin system.

From a conceptual point of view, volunteer translator communities' principal demands are for (1) storing and accessing rich heterogeneous linguistic data; (2) building large and adequate translation memories; and (3) adding improved functionalities in integrated computer-aided translation environments. The authors have thus proposed a new general architecture for online translation aid systems. They are currently developing QRLex system modules separately and intend to integrate them next year into a first working version, to be used by several online volunteer translator communities.

In parallel, the authors are working on TRANSBey [Bey *et al.* 2006], geared toward fully online translators such as contributors to tranlationwiki.net. The intent is to enable online translators to collaborate to solve difficult problems during the translation process so as to jointly produce high quality translations. A document would not necessarily be translated once and for all by a unique translator, but could instead be translated by several translators, and certain passages might be translated several times. For this purpose, the developers will have to design and implement another module (again using the Wiki technology again), a Web-oriented translation editor usable through any navigator and allowing the online collaborative edition of documents.

Acknowledgements

The authors are very grateful to Prof. Akiko Aizawa for her advice and help during their stay at NII (National Institute of Informatics, Tokyo, Japan). Special thanks go to Dr. Christophe Chenon (IBM, France) for his valuable suggestions and orientation and to Dr. Mark Seligman (Spoken Translation, Inc., Berkeley, USA) who kindly improved the English level of this paper.

Last, but not least, this work is partly supported by grant-in-aid (A) 17200018 "construction of online multilingual reference tools for aiding translators" of JSPS (Japan Society for the Promotion of Sciences).

References

Agirre, E., X. Arregi, X. Artola, A. Díaz de Ilarraza and K. Sarasola, "Lexical KnowledgeRepresentation in an Intelligent Dictionary Help System," In *Proceedings COLING'94*, 1994, Kyoto, Japan, pp. 544-550.

Agirre, E., X. Arregi, X. Artola, A Díaz de Ilarraza., K. Sarasola and A. Soroa, "Constructing and intelligent dictionary help system," *Natural Language Engineering, Cambridge University Press*, 3, 1996, pp. 229-252.

Agirre, E., X. Arregi, X. Artola, A. Díaz de Ilarraza, K. Sarasola and A. Soroa, "A Methodology for Building Translator-oriented Dictionary Systems," *Machine Translation*, 15, 2000, pp. 295-310.

AKIN, Detection of Translated Documents, version 1.0.0, http://apple.cs.nyu.edu/akin/, 2005.

Allen, J., "Postediting: an integrated part of a translation software program (Reverso Pro 4)," *Language International Magazine*, 13(2), 2001, pp. 26-29.

Augar, N., R. Raitman and W. Zhou, "Teaching and Learning Online with Wikis School of Information Technology," In *Proceedings of the 21st Australasian Society of Computers in Learning in Tertiary Education Conference*, 2004, Australia, pp. 95-104.

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

Bey, Y., K. Kageura and C. Boitet, "A Framework for Data Management for the Online Volunteer Translators' Aid System QRLex," In *Proceedings of the 19th Pacific Asia Conference on Language, Information and Computation,* 2005, Taipei, Taiwan, pp. 51-60.

Boitet, C., "New architectures for "democratic" tunable quality MT systems," In *Proceeding of Pacific Association for Computational Linguistics, H. Sakaki, (ed.)*, 2005, Meisei daigaku, Hino campus, Tokyo, Japan, pp. 33-57.

Bowker, L., *Computer-Aided Translation Technology: A Practical Introduction*, Didactics of Translation Series, University of Ottawa Press, 2002.

HTMLArea, In-browser WYSIWYG editor for XWiki, http://www.htmlarea.com/, 2006.

Hutchins, J., "The Origins of the Translator's Workstation," *Machine Translation*, 13, 1998, pp. 287-307.

Hutchins, J., Machine Translation and Computer-Based Translation Tools: What's Available and How It's Used, http://ourworld.compuserve.com/homepages/WJHutchins/, 2003.

Kay, M., "The Proper Place of Men and Machines in Language Translation," *Machine Translation*, 12, 1997, pp. 3-23.

IBM LOCALIZATION, XML in localization: a practical analysis, http://www-106.ibm.com/developerworks/xml/library/x-localis/#example1, 2005.

LingPipe, Linguistic Toolkit: Sentence-Boundary Detection, Named-Entity Extraction, Language Modeling, Multi-Class Classification, http://alias-i.com/lingpipe/demo.html, 2006.

LISA, Localization Industry Standards Association, http://www.lisa.com/, 2006.

Mangeot, M., "An XML Markup Language Framework for Lexical Databases Environments: the Dictionary Markup Language," In *Proceeding of International Standards of Terminology and Language Resources Management, LREC 2002 workshop*, 2002, Las Palmas, Islas Canarias, Spain, pp. 37-44.

Martin, W., "User-orientation in Dictionaries: 9 Propositions," In *Proceedings BudaLEX´88*, 1988, Budapest: akadémiai kiadó, Hungary, pp. 393-399.

MOZILLA, Open Software Localization, http://frenchmozilla.online.fr/, 2005.

MT POST-EDITING, MT postediting and translators tools reviews. http://www.geocities.com/mtpostediting/, 2006.

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

Queens, F. and U. Recker-Hamm, "A Net-Based Toolkit for Collaborative Editing and Publishing of Dictionaries," *Literary and Linguistic Computing Advance Access, Oxford Journal, Lit Linguist Computing*, 20(1), 2005, pp. 165-175.

REDOX, Wiki Engine of XWiki, http://radeox.org/space/start, 2005.

Saha, G.K.A., "Novel 3-Tiers XML Schematic Approach for Web Page Translation," In *ACM IT Magazine and Forum*, 6(43), 2005, http://www.acm.org/ubiquity/views/v6i43_saha.html.

Schwartz, L., "Educational Wikis: Features and Selection Criteria," In *the International Review of Research in Open and Distance Learning*, technical report R27/0311, Athabasca University, Canada's Open University, 2004.

SIMILIS, Second Generation of Translation Memory Tool, http://www.lingua-et-machina.com/, 2005.

TEANOTWAR, Human Rights Documents Translation, English to Japanese News Translation, http://teanotwar.blogtribe.org/, 2005.

TRADOS, Translation Memory Tool, http://www.trados.com/, 2005.

TRADUCT, Linux Documentation Translation, http://wiki.traduc.org/, 2005.

Tsuji, K., Sato S. and Kageura K., "Evaluation of the Usefulness of Search Engines in Validating Proper Name Transliterations," In *Proceeding of 11th Conference of Natural Language Processing Society of Japan*, 2005, Japan, pp. 352-355.

W3C, Specification Translation, http://www.w3.org/Consortium/Translation, 2005.

Walker, D.J., Clements D.E., Darwin M. and Amtrup W., "Sentence Boundary Detection: A Comparison of Paradigms for Improving MT Quality," In *Proceedings of the 8th Machine Translation Summit*, 2001, Santiago de Compostela, Spain, pp. 369-372.

WIKITRANSLATION, Collaborative Wiki-Based Translation, http://www.translationwiki.com, 2006.

XWIKI, Open Source Java-Based Wiki, http://www.xwiki.com/, 2006.