A Framework for Data Management for the Online Volunteer Translators' Aid System QRLex

Youcef Bey†‡, Kyo Kageura‡, and Christian Boitet†

†Laboratoire CLIPS-GETA-IMAG
Université Joseph Fourier
385, rue de la Bibliothèque.
Grenoble, France
youcef.bey@imag.fr

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

Abstract

A new framework for a system that aids online volunteer translators is proposed. As regards this proposal, first, the current status and conditions of online volunteer translators and their translation environments are examined, and general requirements for a system that would aid these translators are given. Our proposed approach for dealing with heterogeneous data, which involves providing a new XML structure that we have developed for maximizing efficiency and functionalities, is then described.

1. Introduction

For many years, specialists and researchers have become pessimistic in regards to the prospects for fully automatic translation capable of producing high quality translations equal to human translator. It is well known that the ALPAC report in 1966 evaluated the performances of Machine Translation (MT) systems negatively. Martin Kay stated:

"*...this happens when the attempt is made to mechanize the non-mechanical or something whose mechanistic substructure science has not yet been revealed...*" [6]

This situation has promoted a shift in emphasis of research from fully automated machine translation to computer aided human translation, which exploits the potential of computers to support human skills and intelligence [5]. Many industries have made a large investment in developing useful translation-aid tools, which has resulted in commercial Computer-Aided Translation (CAT) systems such as Translation Memory (TM) in various forms, dictionaries and terminology database techniques. However, these systems are not designed to be used by all translators. On the one hand, the commercial feature is a barrier for many translators. On the other hand, these tools do not provide content and functions that fully satisfy some translators. Online volunteer translators, to whom we specifically address our system, are among those excluded from commercial CAT systems. There is thus a real need to aid online volunteer translators and their communities by providing them with a free environment with a rich linguistic content and improved process and data management.

In section 2, we first outline the status and conditions of online volunteer translators and how they work in translation. Section 3 outlines the framework and technical modules we have defined on the basis of analyzing online translators' requirements and, in the process clarified the basic requirements for the data management module. In section 4, we present the XML structures that we defined for our data management module.

2. Aiding Online Volunteer Translators

We notice recently an important grow of volunteer translators who are translating thousands of documents in different fields and, thereby, showing the true way to break the language barrier. This is mainly due to the important role that the internet plays in allowing translators to join volunteer translation activities. For example, in the W3C consortium, there are 301 volunteer translators involved in translating thousands of specification documents covering approximately 41 languages. Documentation in the Mozilla project exists in 70 languages, and they are also translated by hundred of volunteer translators located in different countries. Other volunteer translator communities are involved in translation in non-identified projects; they, however, form translator groups without any orientation in advance.

According to our analysis of existing communities, volunteer translator communities consist mainly of two types:

- **Mission-oriented translators communities:** mission-oriented, strongly-coordinated group of volunteers involved in translating clearly defined sets of documents. These communities cover loosely technical documentation like translation of Linux documentation [15], W3C specifications [17], and open source Mozilla localization software [9].
- **Subject-oriented translators network communities:** individual translators who translate online documents such as news, analysis, and reports and make translations available in personal or group web pages [4] [11]. They form groups of translators without any orientation in advance, and they share similar opinions about events (anti-war humanitarian communities, report translation, news translation, humanitarian help, etc.).

For instance, almost all online translators show similar behavior. As for the first communities (henceforth "Linux communities"), volunteer translators in both the Traduc and Mozilla projects are invited to translate a list of documents available on web sites (of each project) in different formats (XML, SGML, HTML, HLP, plain text, etc.). Firstly, they check whether the relevant document has been translated; if not, they make a reservation and announce the beginning of a translation to other translators via a discussion list or email. To obtain the document to be translated, they download it directly from the CVS (Concurrent Version System) or ask the coordinator to send it via mail.

Once the relevant document is obtained, each translator has their individual translation environment, which is not similar to those of other translators (Figure 1). Indeed, the translation process is carried out in environments, which consist loosely of a set of tools: textual editor, dictionaries (electronic, paper version, or online), glossaries, terminology, and sometimes TM. In addition, we note the importance of the internet, which has become a precious linguistic resource for translators, who use it for recovering existing translation segments (quotations, collocations, technical terms, etc.).

Documents are translated in the respect of the original format. For example in the Traduc project, documents are structured in XML DocBook[1]; translators operate the translation process between XML markers. After the translation finished, they send the whole target document in the same structure to the coordinators.

[1] A rich XML Format used to produce readable HTML with OpenJade tool; for further information refer to : www.docbook.org

As for translation-aided tools proposed for aiding translation, *Linux community* makes certain resources available, which we resume it on the following tools:

- A set of local free dictionaries (In general not updated), glossaries, and links to other linguistic Web sites.
- A discussion list used for exchanging skills and solve most troubles faced during translation process.
- Control files for checking "Who does what?"; this is useful for the collaborative translation for the same document by several volunteer translators. Before starting a translation, users take a look at the most-recent end point of a translation, and restart the translation from there.
- CVS system for versioning document management.

Almost all linguistic tools are not maintained and updated; this due to the lack of automatic tools that coordinate and make tasks easier.

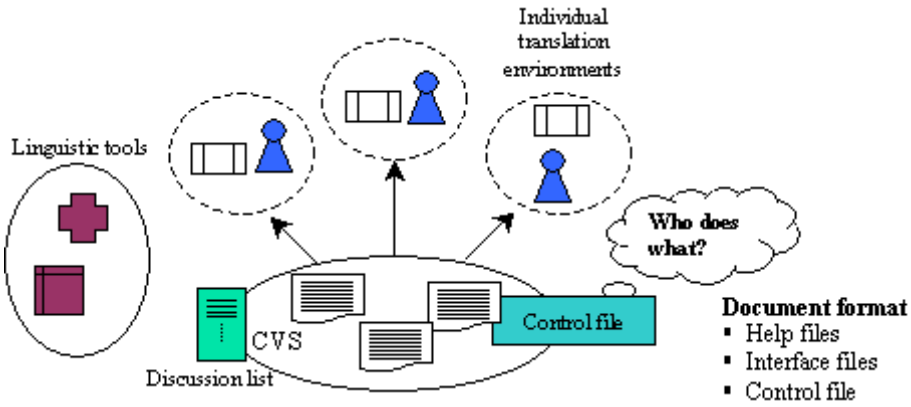


Figure 1: Translation method in Linux communities.

In the following section we clarify our design of QRLex framework according to translator needs and data management. On the one hand, we clarify different aspects to be taken in consideration when we attempt to design translators' aided tools, and give a detailed picture of modular architecture and structure data flow for dealing with linguistic and textual data in the QRLex framework.

3. Design: QRLex framework

By talking with volunteer translators as well as coordinators [18] and examining some existing translation aid systems [12] [14], we clarified a few essential general requirements:

- Contents of language reference tools cannot be separated from system functionality.
- Translators look for the information of (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 function and interfaces.

With respect to reference contents, therefore, our needs are as follows:

- Use the good reference contents whenever they are available.
- Enhance the contents when contents are not sufficient.
- Make the recyclable units available from existing relevant translated documents.

With respect to lookup functions, we need to maintain the unified function and interfaces for interaction with translators. With respect to language data management, we need to properly deal with relevant translated documents on the one hand and recyclable reference contents at various levels in a uniform way on the other. Compared to existing translation aid tools and related ideas, QRLex is much more content- and community-oriented.

Taking these general desiderata into consideration, we have defined a system that realizes QRLex framework by means of five functional modules (Figure 2), each of which covers specific tasks and deals with different types of data:

- **Structure manager:** A module that transform reference data and textual data to structured XML format. As for linguistic data, we have proposed a new XML structure in which we have compiled heterogeneous linguistic data including dictionaries, a Japanese pronunciation guide, technical terms and proper name resources. Therefore, this module preprocess linguistic data in various formats and creates XML XLD (XML Linguistic Data) for each resources, which are then stored in centralized database. In the same manner, source and corresponding translated documents are processed in the documents manager module and converted into structured LISA TMX standard (Translation Memory eXchange) [7].

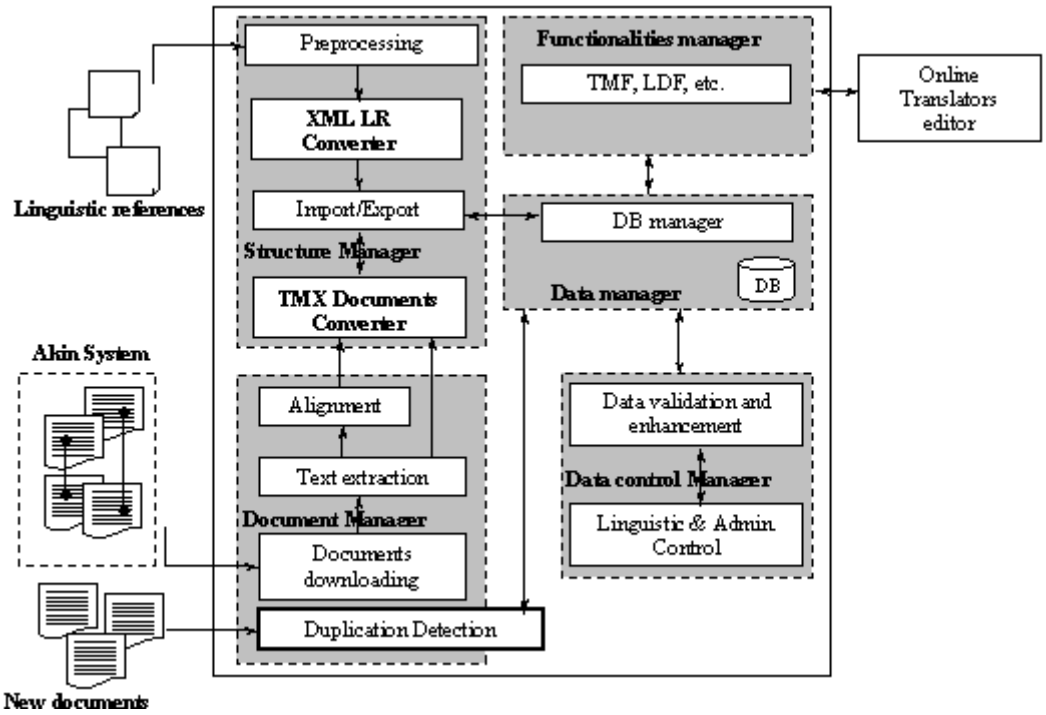


Figure 2: General framework for QRLex data management.

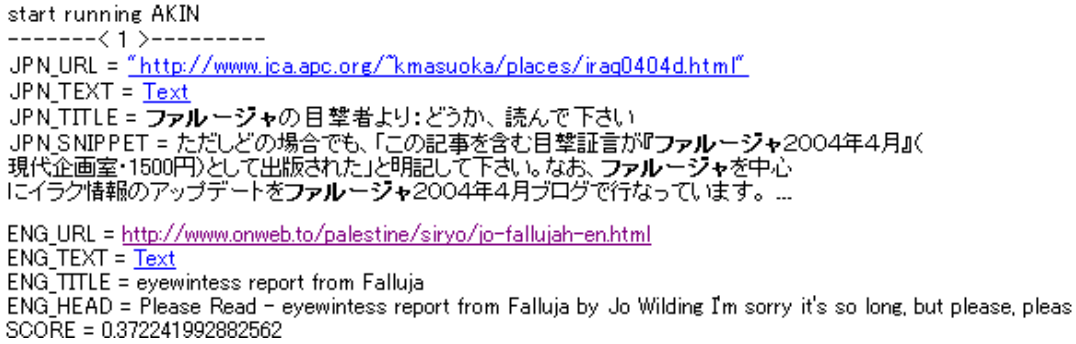
- **Documents manager:** This module is based on detection of existing documents in a documents repository of internal communities or research of translated documents on the Web. Existing translated documents are subjected to text extraction and alignment [3]. The final result is a set of Bitext, which will be stored in TMX format.

- **Database manager:** This module is the engine server of data to other QRLex modules. 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.

- **Data control manager:** It is well known that open linguistic resource environments on the Web necessitate the intervention of expert humans. In our case, the QRLex environment necessitates the interaction of linguistic expert or professional translators for more accurate data content and control of user interaction.

- **Functionalities manager:** In fact, talking about functionalities is the most important feature of CAT tools. Considering the content and needs of translators, we imagine functionalities as the most important criteria to be taken in consideration during the development process. We believe that a software designer does not possess a complete ability to imagine all translators' needs owing to the complexity of translation task and differing behaviors from one translator to another. Accordingly, many researchers have recently tried using empirical methods, such as direct observation, interviewing of volunteer translators and questionnaires before developing translation help tools [2][13]. In a QRLex environment, functionalities are developed according to online translator needs by direct interviewing and questionnaires as well as by using a rich linguistic data content.

- **Akin System:** The detection of existing translation documents is carried out by Akin system [1], which detects English translated documents using keywords (Figure 3). In fact, integrating Akin into QRLex framework allows: (i) volunteer translators to avoid duplicating the same translation, (ii) translated units from the Web to be recycled for constructing QRLex Translation memory [16].
  - *Keywords:* ファルージャ
  - *Detection of translated document:*


start running AKIN
--------< 1 >----------
JPN_URL = "http://www.jca.apc.org/~kmasuoka/places/iraq0404d.html"
JPN_TEXT = Text
JPN_TITLE = ファルージャの目撃者より：どうか、読んで下さい
JPN_SNIPPET = ただしどの場合でも、「この記事を含む目撃証言が『ファルージャ2004年4月』(
現代企画室・1500円)として出版された」と明記して下さい。なお、ファルージャを中心
にイラク情報のアップデートをファルージャ2004年4月ブログで行なっています。...

ENG_URL = http://www.onweb.to/palestine/siryo/jo-fallujah-en.html
ENG_TEXT = Text
ENG_TITLE = eyewintess report from Falluja
ENG_HEAD = Please Read - eyewintess report from Falluja by Jo Wilding I'm sorry it's so long, but please, pleas
SCORE = 0.372241992882562

Figure 3: Detection of existing translated documents on the Web.

For these modules to work efficiently and effectively as intended, one of the important technical aspects is to define the format of the core language data for both reference and textual data, which will be elaborated in the next section.

4. Reference and document data management: XML definition

We choose to manage QRLex data in XML format for several reasons. Firstly, it is widely used as a document exchange format and for data storage and retrieval [10]. On the other hand, QRLex data is in a different text format; it is known that text format parsers are expensive. However, XML offer the possibility of easily making parsers using DOM (Document Object Modeling).

In the following paragraphs, we will explain the XML formats used to deal with reference data and textual data elements of various dictionaries, terminology lexicons and translated texts.

4.1.Reference data format:

Among existing reference data for which our management structure is defined, “Eijiro” and “Grand Concise” are two of the high quality English-Japanese unidirectional dictionaries widely used by many translators, “Nichigai” is specifically for proper names and “Medical Scientific Terms” resource is included to check the structure of terminological dictionaries. “Edict” is a free Japanese-English dictionary; we examined it for checking the directionality of the bilingual dictionaries (Table 1).

In fact, we notice that for each reference data there are a few requirements: (a) various levels of recyclable units should be dealt with within a unified framework, (b) existing high-quality contents should be properly accommodated and (c) unnecessary information contained in existing content should be properly excluded while necessary information common to any proper reference data and useful for translators should be incorporated. However, satisfying these requirements necessitates an internal XML structure for storing and exchanging content within different QRLex modules. After examination of existing XML standard formats for terminologies such as TBX (TermBase eXchange) and MARTIF (Machine-Readable Terminology Interchange Format), unfortunately, we found these formats did not satisfy this requirement [7][8]. On the other hand, existing high-quality reference data in electronic form take a variety of formats. We thus define the basic XML structure of the linguistic data by checking the data elements of various dictionaries and terminology lexicons. Figure 4 illustrates XLD (XML linguistic data) that we have developed for management of our heterogeneous reference data.

Table 1: Reference data in 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

The XLD format consists of three main parts (i) description element and attributes for the original linguistic resources and content, (ii) source entry elements, and (iii) target elements that contain expressions for explanation of source entries (Figure 5). Indeed, the XLD header describes the original version and the linguistic content (date of creation, authors, encoding, number of entries, source and target languages, etc.).

```
- <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>
```

Figure 4: Japanese "Nichigai" entries in XLD format.

```
<!-- 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 5: DTD (Document Type Definition) of XLD format.

Source and target elements contain additional information with a set of attributes for giving more clarification (Figure 5):

- xml:lang: source and target languages.
- additional-info: description of a source element in case we transform the relevant resource direction from, for example Japanese-English into English-Japanese.
- kata-pronunciation: an attribute containing the pronunciation in Katakana.

We have compiled all linguistic resources cited in Table 1 in the XLD structure. As for textual data, we proceed in the same manner, but we use XML TMX standard, which we found more useful for storing and managing recycled textual data from the Web.

4.2.Document (textual) data format

The document data structure should satisfy two requirements: (a) maximal facilitation of providing recyclable units and (b) unified management of translated documents. The first requirements come from translators, who strongly look for 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 in. Unlike reference data format, we found an existing standard framework TMX suitable for our aim. This standard is developed to simplify the storage of textual data from documents that contain formatting information such as HTML. As we need to deal with online documents, this format is suitable for our aim. This standard is defined by LISA comity, which allows a translation memory to be managed and both source and target sentences (and paragraphs) to be stored in multilingual format [7] [8]. Figure 6 illustrates an English-French-Spanish example using TMX format of translated documents done by volunteer translators in *paxhumana community* [11].

```
<?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="es">
          <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 6: Source and translated document in TMX format

According to the needs of online volunteer translators (explained before in this paper) and the whole design of QRLex for managing linguistic data in heterogeneous formats, structuring

linguistic data in XML formats makes it easy to construct a parser and to develop improved functionalities. At the management level, all importing data (references data or textual data) and the internal flow between modules will be stored in XLD and TMX format.

5. Conclusion

We propose that new aspects to be taken in consideration before any design for online CAT tools. We examined translators' needs, firstly by analyzing various scenarios of translation of existing online translator communities, and after that, by interviewing online translators. This work has clarified and affected our imagination regarding the conceptualization of a new framework based on two aspects: (i) a rich content and (ii) improved functionalities. As for content, almost all translators ask for a rich content in various formats, dictionaries, glossaries, and translation memories.

We have developed an XLD format for compiling heterogeneous linguistic data for storing usable free dictionaries and allowing importation of new linguistic resources to centralized relational database of QRLex. On the other hand, a TM constitutes a precious linguistic resource which almost all translators need for accelerating and improving quality of document translation on the web. It will be conscientiously created by recycling existing translated documents on translator community web sites or on the Web by crawling and exploiting existing Web search engines like Akin system.

From conceptual viewpoints, storing rich heterogeneous linguistic data, translation memory, and adding improved functionalities in integrated computer-aided translation environment is the more important aspect that volunteer translator communities ask for. We thus propose our general architecture of online aid system and are developing separately each module for the whole future concretization of QRLex.

In the near future, firstly we envisage joining online translators from different communities in collaborative translation tasks to solve difficult problems during the translation process and gathering effort for producing high quality translations, not by one translator in at one times but by several translators, several times. On the other hand, we envisage to add into the main framework another module for allowing collaborative edition of documents, based on the new collaborative technology systems like Wiki system [19].

6. Acknowledgements

This work is supported by the grant-in-aid (A) 17200018 “construction of online multilingual reference tools for aiding translators” by Japan Society of the Promotion of Sciences (JSPS).

7. References

[1] Akin system, http://apple.cs.nyu.edu/akin/.

[2] Eneko, E. et al. 2000. A Methodology for Building Translator-oriented Dictionary System. *Machine Translation*, 15, pp. 295-310.

[3] Gupta, S. and G. Kaiser. 2005. Extracting Content from Accessible Web Pages. *ACM Series,Vol.88 archive Proceedings of the 2005 International Cross-Disciplinary Workshop on Web Accessibility*, pp. 26-30.

[4] Human Rights: English to Japanese news translation, http://teanotwar.blogtribe.org.

[5] Hutchins, J. 1998. The Origins of the Translator’s Workstation. *Machine Translation*, 13, pp. 287-307.

[6] Key, M. 1997. The Proper Place of Men and Machines in Language Translation. *Machine Translation*, 12, pp. 3-23.

[7] Localization Industry Standards Association (LISA), TMX and TBX Standards, http://www.lisa.com/.
[8] Melby, A. K. The CLS framework. *WWW documents*, http://www.ttt.org/clsframe/.
[9] Mozilla project, http://frenchmozilla.online.fr/ and http://frenchmozilla.sourceforge.net/.
[10] Neumuller, M. 2002. Compact Data Structures for Querying XML. *EDBT 2002 PhD Workshop, EDBT*, pp. 127-130.
[11] Paxhumana, translation of various humanitarian reports in French, English, German, Spanish, http://paxhumana.info.
[12] Similis, http://www.lingua-et-machina.com/.
[13] Teoh, I.-E.-H. and E.-K. Tang. 2004. User Model for Prototyping Computer-Aided Translation System. *Proceedings of the 7th International Conference on WWCS (Work with Computing Systems)*.
[14] TRADOS, http://www.trados.com/.
[15] Traduc project, linux documentation translation, http://wiki.traduc.org/.
[16] Tsuji, K. and al. 2005. Evaluation of the Usefulness of Search Engines in Validating Proper Name Transliterations. *11th Conference of Natural Language Processing Society of Japan*, pp. 352-355.
[17] W3c, specification translation, http://www.w3.org/Consortium/Translation.
[18] We interviewed six online volunteer translators (three of them are professional translators who also do online volunteer translation; the other three are non-professionals), and talked with the core person of the French Linux localization project. In addition, the second author is a semi-professional translator, who has so far published eight translated books, various articles and involved in volunteer translation work for nearly 10 years in the field of human rights with special reference to East Timor.
[19] Wiki collaborative translation, http://www.translationwiki.net/index2.php