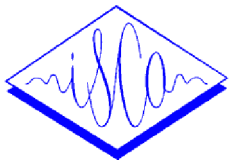

EXTENDING AN ON-LINE PARALLEL CORPUS MANAGEMENT SYSTEM TO HANDLE SPECIFIC TYPES OF STRUCTURED DOCUMENTS

Cong-Phap HUYNH, Christian BOITET, Georges FAFIOTTE
Laboratoire LIG, GETALP, GETA, Université Joseph Fourier
385, rue de la Bibliothèque, 38041 Grenoble, France
{Cong-Phap.Huynh, Christian.Boitet, Georges.Fafiotte}@imag.fr

ABSTRACT

Parallel bilingual or multilingual corpora are often handled as collections of segments without any specific document organization. We describe SECTra_w, a web-oriented system which has been used for online MT evaluations, and has recently been extended to handle multimodal documents such as French-Chinese/Vietnamese/Hindi/Tamil interpreted bilingual spontaneous dialogues, mainly spoken but also using some short texts, and multilingual written articles of an online encyclopedia annotated with UNL graphs.

Keywords: parallel corpora, translation memories, multiple annotations, multimodal dialogues, multilingual documents

INTRODUCTION

Very large parallel corpora aligned at the level of "translation segments" (sentences or titles) are used to develop Machine Translation (MT) systems, and to evaluate them. The segments are sometimes stored in various forms (with or without upper case characters, punctuations, segmentations for Asian writing systems) and with various annotations (POS, chunks). Corpora are often represented by collections of files and are not visible and even less directly usable by humans for checking, evaluating and experimenting.

We are developing SECTra_w, a web-oriented system for Evaluating, presenting, processing, enlarging and annotating Corpora of Translations. Several large parallel corpora, such as EuroParl, and the ERIM [Fafiotte 2003, 2004] corpus of French-Chinese/Vietnamese/Hindi/Tamil spoken interpreted bilingual spontaneous dialogues, have been imported into SECTra_w.1. MT subjective evaluation has also been performed through the web using SECTra_w.1.

Working on the ERIM corpus has shown the need of a software architecture permitting a wiki-like usage, to enable the study, distribution, and collaborative improvement and annotation of corpora. Also, for implementing functions such as replaying dialogues, it seems necessary to take into account the structure of the dialogues. Another example is a project of translating the Online Encyclopedia of Life Support Systems (EOLSS) from English into the five other Unesco official languages: each article is represented by two files, a standard .html file and a companion .unl file (list of multilingual segments annotated by UNL semantic graphs).

We are thus enlarging SECTra_w to manage parallel corpora of *segments* (such as translation memories) to handle the specific structure of *documents* from where they are drawn, and to define and implement processes on them according to that structure, if possible in a generic manner.

Another point is the importance of an adequate software architecture. Having started from BEYTrans [Bey 2006], an on-line system offering linguistic resources and tools to communities of volunteer translators, implemented in Xwiki, we have automatically got a good support for collaborative work, in particular users and rights management, while they would have been too difficult to add "après coup" to previous prototypes.

In the first section, we review the previous state of SECTra_w, and show its effective use for storing parallel text corpora as well as bilingual and bimodal interpreted spoken dialogues, and for supporting classical MT evaluation campaigns. In the following two sections, we will show how SECTra_w has been extended to handle the ERIM dialogues and the EOLSS/UNL documents, and develop specific functions directly based on their structures.

1. SECTRA_W.1: ONLINE COLLABORATIVE MANAGEMENT OF MULTILINGUAL CORPORA

1.1. Motivations

SECTra_w.1 has been developed first to handle via the web parallel corpora used to build and evaluate MT systems, that is, very large translation memories aligned at the segment level. For example, EuroParl contains about 20M words in each of 11 languages, or 80 K standard pages per language, and the CSTAR part of the BTEC corpus has 163 K segments, slightly more than 1M w in each of 5 languages, plus some recorded parts.

The second objective was to manage not only textual forms of various types for written corpora (raw, or preprocessed by word or chunk segmentation, punctuation suppression, or syntactic annotation), but also, for SLT (spoken language translation), the primary audio form, aligned with various transcriptions and annotations.

The third objective was to offer functionalities for study, evaluation, and processing. To study a corpus, one must measure it, browse it, replay it if it contains sound, with some filters (by language, person, etc.). Various corpus evaluation techniques must also be supported (subjective ones, based on human judgments, as well as objective ones, based on n-gram counts and on task-related measures).

Finally, we want to allow easy access to the data: far too often, results are reported by showing only tables of scores (BLEU, NIST, etc.), but no data — and it is now known how bad these scores correlate with human judgments and task-related objective measures such as understanding tests or end-to-end performance measure or post-edition effort (which we estimate using an edit distance).

1.2. SECTra_w.1 architecture

SECTra_w was first supposed to be developed as an extension of BEYTrans, itself programmed in Java over Xwiki. As that became too difficult because the two researchers involved were 12,000 Km away from each other, SECTra_w.1 was developed independently, using Xwiki, Ajax, Javascript, and Velocity. It took only 3 weeks before we could import large corpora and begin to use it for experiments and a real evaluation campaign. All text data, as well as the programs and the textual elements of the GUI, are stored in a MySQL database. The sound files are stored as such, with metadata and references to them in the database.

This first version met the objectives above: imported corpora were large and some of them multimodal (audio and text), interactive web-based MT evaluation was possible, as well as manual post-edition, all from any web navigator. An interesting addition is that the differences between raw MT results and post-editions may be shown in an intuitive way (insertions in red, deletions in overstriked blue).

1.3. Using SECTRa_w.1 for MT evaluation

1.3.1. Corpus import and call to MT systems

All corpora are converted to the Unicode UTF-8 encoding, and the original encoding is stored with other metadata. For importing parallel corpora, we use ad hoc scripts if the formats are very simple, or we first convert the files into simplified versions of the TMX XML format, CPM for any simple file (monolingual as the BTEC files, or bilingual as the Tanaka JE files), and then CPXM to consolidate all information relative to the same segment, or "polyphrase".

One can submit test corpora to MT systems directly from SECTra_w, or independently (and then import the results).

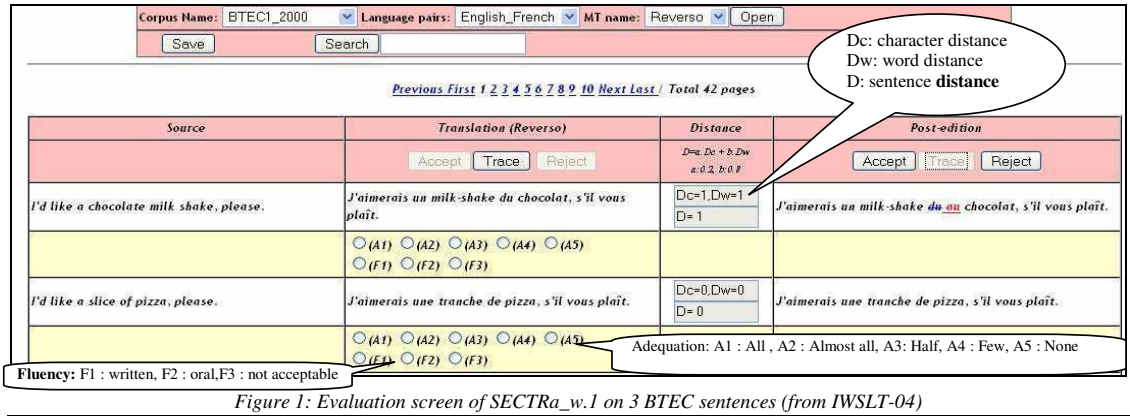

Figure 1: Evaluation screen of SECTRa_w.1 on 3 BTEC sentences (from IWSLT-04)

1.3.2. Subjective evaluation

The interface for subjective evaluation generalizes slightly those classically used by judges evaluating adequacy and fluidity (Figure 1). The number of possible choices (presented as radio buttons) is a parameter, as well as the help strings appearing in small balloons when the cursor hovers on a button. The following other features have been included and proved useful.

- Several judges can perform evaluation at the same time on the same part of the data, which appears as a web page of about 20 segments, thanks to Xwiki.
- A segment usually receives several evaluation scores from as many judges. These scores can be shown to users having enough access rights.
- A preliminary workflow tool is included, to define the judges and assign them sets of pages to evaluate.

1.3.3. Objective evaluation

Two kinds of objective evaluations are available:

- running scripts computing n-gram-based measures such as BLEU, NIST, WER, and an edit distance mixing distances at character and word levels.
- Letting humans post-edit (online) the MT output, and measuring the time taken, or estimating it from an edit distance between it and its post-edited form.

As our edit distance computation uses exchanges X as well as insertions I and deletions D, we replace each sequence InXm (resp. DnXm) by a sequence In+mDm (resp. Dn+mIm), to visualize the edit distance and hence the post-edition effort.

The post-edition interface is an extension of the subjective evaluation interface (Figure 1). To prevent accidental mistakes, only the "post-edition" column is editable.

2. ERIM CORPUS OF BILINGUAL SPOKEN DIALOGUES

2.1. Situation

The ERIM project first developed tools for volunteer non-professional interpretation over intranets or extranets, where participants communicate by mixing speech, short written messages (i.e., to spell a name), and sharing a "whiteboard".

It then turned towards collecting task-related spoken dialogues in the tourism domain, as an effort parallel to our research on spoken language translation (SLT) in the framework of the CSTAR and Nespole! projects.

Since 1999, we have collected with our partners (NLPR in China, Da Nang TU in Vietnam, IIT-Bombay in India) approximately 10h of French-Chinese, 10h30 of French-Vietnamese, 2h30 of French-Hindi, and 9h of French-Tamil. We say "approximately" because the formats of the dialogue and file descriptors have evolved, so that we did not yet import and measure all ERIM data into SECTra_w.

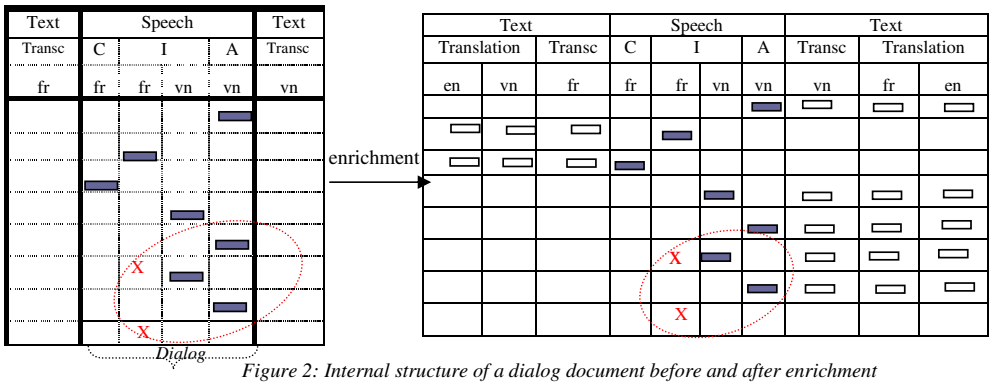

Figure 2: Internal structure of a dialog document before and after enrichment

The left part of Figure 2 shows the logical structure of the now available data. Interlocutor C is a French client, A is a Vietnamese agent (in a tourist office), and I is an interpreter. The time axis is vertical, going down. A segment is made of a speech turn, and possibly a short text, and the spoken part may be transcribed in text. Note that two consecutive turns may be in the same language in the case of a clarification sub-dialogue (marked X in the figure).

```
<?xml version="1.0" encoding="UTF-8" ?>
<Corpus nom=”ERIM” type=”ERIM"
structDc=”erim.dtd”>
<dialog num=”1” langues=”Vietnamese-French”>
<LOC nom=”C”/> <LOC nom=”I”/> <LOC nom=”A”/C>
<TP num=”1” Loc=”C”>
   <InfosTP lang=”Vn” durée=”13”
       Hdébut=”10:20:00” Hfin=”10:20:33”>
   </InfosTP>
 <son>TP1.wav</son>
 <minitexte>””</minitexte>
 <transc>”Xin chào”</transc>
 <tradu> <trad lang=”fr”>Bonjour</trad>
 </tradu>
</TP>
.....
</dialog>
.....
```

Figure 3 : Description of an ERIM dialogue's content

The right part of Figure 2 shows how the logical structure has to expand if one wants to add translations of the turns into the other language, and perhaps to English, for comparison purposes, using MT systems to produce draft translations (there is no French-Vietnamese MT system yet).

On the physical side, an ERIM corpus is a collection of dialogues collected under certain conditions (date, location, languages). These and other metadata, such as the location and number of dialogues and turns, are contained in XML descriptor files. Each dialogue has an XML descriptor. Data relative to turns are contained in sound files (.wav) and text files (.txt) for short messages and transcriptions, if any.

There have been previous efforts to enable the distribution, improvement, annotation of ERIM corpora or parts of them and to replay one or more dialogues, in totality or by language, speaker, etc. They have led to usable PC-based tools, but not to Web-based tools, because network-oriented features, collaborative features (up to wiki), and users and rights management must be planned from the beginning. SECTra_w.2 allows now to study that kind of corpus on the web, and to collaboratively annotate them.

2.2. Corpus study and measurement

It is possible to "replay" one or more dialogues, filtering them by language and/or interlocutor, and showing the associated short texts and transcriptions on demand.

We have included in SECTra_w.2 functions to compute and show quantitative information about the corpora it contains. A large part of that information depends on the structure of the documents in the considered corpus. For example, ERIM-related information concerns the dialogues, the turns, the languages, and possibly the 3 main actors (C, A, I).

Here is the information for the part of the French-Vietnamese dialogues so far uploaded into SECTra_w. The metadata files are the main source of the information.

# dialogs	15 (≈2.5 hours)	
	French	**Vietnamese**
# speech turns	336	336
# duration	70 mn	74.5 mn
Average by speech turn	12.5 s	13.29 s

Table 1: Information about F-V data in SECTra_w

2.3. Web-based annotation

Many kinds of annotations can be considered: transcriptions, translations of the transcriptions, comments of different types, and all annotations possible on texts, at various levels (morphology, syntax, semantics, pragmatics). SECTra_w.2 only supports the first 2 types at the time of writing.

The transcription environment inherits from the editing and replay environments. Each speech turn is repeated on demand, or with a certain frequency, while the user types the transcription, until s/he user goes to the next one (or another one).

Because speech recognition (ASR) is not yet freely available for the language pairs tackled so far, and in any case is bad over telephone lines or Voice/IP, we did not yet include the possibility to call an ASR for automatic pre-transcription.

As far as the translation of transcriptions is concerned, we simply reuse what has been developed for MT evaluation. External MT systems may be called if available, and the post-edition interface is adapted to the structure of the dialogues (see Figure 4).

The disposition and width of the columns can be changed according to the needs. For example, to compare the French transcriptions and translations of transcriptions, it is useful to place the corresponding columns near to one another.

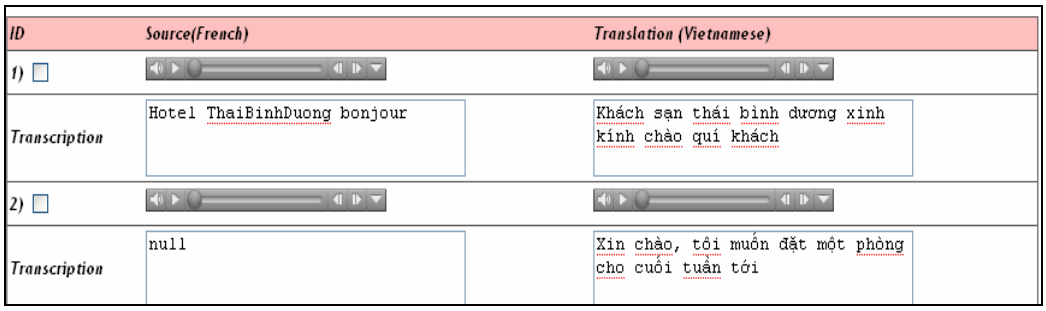

Figure 4: Replay and transcription interface

3. EOLSS/UNL CORPUS FOR A FEASIBILITY STUDY

EOLSS consists of 6600 articles, written in English by specialists since 1996. An article has about 30 standard pages, in total about 250,000 pages and 62.5 M words.

As a feasibility study, and in relation with UNESCO and Unitar, UNDLF (Universal Networking Digital Language Foundation) has started a project to test the applicability of the UNL-based architecture on the translation of EOLSS into the 5 other languages of UNESCO.

The project tackles 25 articles, in total 13673 sentences, about 220 K words, or 880 pages, and a lexicon of about 15,000 simple or compound entries, half of them technical and relative to the various fields related to the life support systems.

3.1. Structure of the corpus

Each document is represented by two files, one standard Html file (.html), and one "companion" file in UNL format (.unl). Here is an example from a document on the tsunamis.

> The wave propagates from the source with a velocity of long gravity water waves in accordance with the equation
> CG = (g H)1/2, (1)
> where g is the acceleration due to gravity, and H is the depth of the basin.

Figure 5 : EOLSS article in English as it appears on the web

The UNL format (Uchida 2004) predates Xml as it was defined in 1996. It was originally built to be usable within raw text files as well as within Html files. It uses special tags such as [D] and [S] for document and sentence elements, and within a sentence {org} for the original text, {fr}, {sp}, {ru}, {cn} and {ar} for the translations into French, Spanish, etc., and comments introduced by ";". {xxx} tags may contain attributes like Xml attributes (without enclosing double quotes). It is possible to have more than one translation into some language, such as an automatic result or a post-edited version, and more than one

UNL graph, provided their attributes are different.

```
The wave propagates from the source with a
velocity of long gravity water waves in
accordance with the equation </p>
<p ALIGN="JUSTIFY"><font
size="4">C<sub>G</sub> = (g
H)<sup>1/2</sup>,  (1)</font></p>
<p ALIGN="JUSTIFY">where g is the
acceleration due to gravity, and H is the
depth of the basin.
```

Figure 6 : Corresponding .html file

The {unl} elements describe UNL (hyper-)graphs as lists of arcs. An arc bears a semantic relation such as man for "manner" or agt for "agent" and relates 2 nodes or 2 "scopes". A node bears a UW (universal word) and semantic attributes such as .@topic or .@indef. A scope is a subgraph made of a connex collection of arcs bearing the same "scope number" (:nn on each arc), and touching an "entry node" (having the .@entry attribute).

```
[S:44]{org}
The wave propagates from the source with a
velocity of long gravity water waves in
accordance with the equation{/org}
{unl}
obj(propagate(obj>thing):09.@entry,
       wave(icl>phenomenon):04.@topic.@def)
man(propagate(obj>thing):09.@entry,
       in accordance with(obj>thing):28)
man(propagate(obj>thing):09.@entry,
       velocity(icl>quality):17.@indef)
src(propagate(obj>thing):09.@entry,
       source(icl>origin):0T.@def)
obj(in accordance with(obj>thing):28,
       equation(icl>expression):2V.@def)
mod(velocity(icl>quality):17.@indef,
       wave(icl>phenomenon):22.@pl)
aoj(long(aoj>thing):1J,
       wave(icl>phenomenon):22.@pl)
mod(wave(icl>phenomenon):22.@pl,
       water(icl>liquid):1W)
mod(water(icl>liquid):1W,
       gravity(icl>force):1O)
{/unl}
[/S]; etc.
```

Figure 7 : "companion" files in UNL format

A UW such as `wave(icl>phenomenon)` corresponds to an "interlingual acception". It is made of a headword, usually an English word or term, such as `wave`, and a bracketed list of restrictions, such as `(icl>phenomenon)`.

The UNL graph associated to a sentence in some natural language is a semantic structure of an equivalent English sentence. That is why the UNL language of semantic graphs may be aptly called an "anglosemantic" interlingua.

3.2. Scenario of work and associated processes

In this project, SECTra_w.2 is used to centralize all the work, with the scenario illustrated by Figure 8.

Import. The .html file is segmented into Html code and textual segments which must correspond to those in the .unl file. The source segments as well as their UNL graphs and translations (if any) are stored into SECTra_w. A "skeleton" Html file is produced, with placeholders for the segments.

Document verification. Once a document is imported, it is checked for correct alignment, and for possible small differences in the strings (e.g., the commented string after an {org} (source string) may contain special occurrences such as TPN1 to stand for a measure unit such as `m/s`).

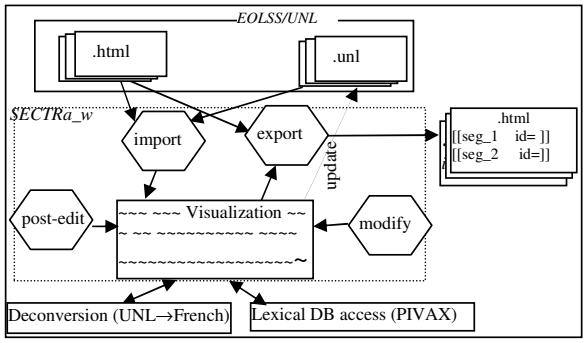

Figure 8: Scenario of using SECTra_w for EOLSS/UNL

UNL graph drawing. A process is launched in the background to produce gif images of the new or modified UNL graphs. Figure 9 shows the drawing G. Sérasset's `unldeco` program produces for the first segment.

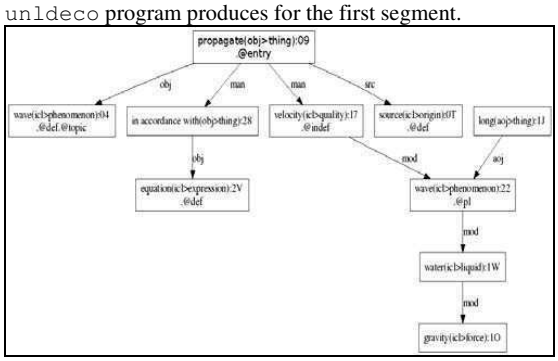

Figure 9: UNL graph drawing

Vocabulary check. The vocabulary of the new segments is checked against the available dictionaries, stored in the PIVAX lexical database (Nguyen &al. 2007) and the list of unfound items is updated so that developers can work on it.

Translation and deconversion. New source segments and their UNL graphs are submitted to available MT systems and deconverters. Post-edition efforts are compared. The goal is to improve the UNL approach until it becomes 2-3

times more efficient than the classical MT approach.

Post-edition. The post-edition interface is the same as presented in Figure 1 and can be accessed either directly, or by viewing an enriched Html form of the document, selecting a passage, and asking to post-edit it. The Html form is updated when changes are made, so that the current version in the target language is visible. We plan to add a dictionary pane to the post-edition screen, so that post-editors can use and enrich the dictionaries while they work.

Visualizing documents in parallel. Various parallel presentations are provided, in table layout (variations of the post-edition presentation), or in document layout (source and translation(s) side by side, or one above the other).

Export of results in UNL format. On demand, at specified intervals, or when changes have been made, an image .unl file is created by inserting into the original .unl file the results of translations, deconversions and post-editions, with appropriate metadata in the attributes of elements, as well as comments for reporting errors found in the UNL graphs and other problems such that domain-oriented headwords unfound in terminological databases.

Multilingual access service. An *iMAG* (interactive Multilingual Access Gateway) is being built over SECTra_w to allow browsing a website with EOLSS-related information in the Unesco languages. The difference with classical web MT servers is that this iMAG is dedicated to an EOLSS-specific text type and vocabulary.

CONCLUSION

Parallel bilingual or multilingual corpora are often handled as collections of segments without specific document organization. We have described SECTra_w, a web-oriented system which has been used for online MT evaluations, and has recently been extended to handle multimodal documents such as French-Chinese and French-Vietnamese interpreted bilingual spontaneous dialogues, mainly spoken but also using some short texts, and multilingual written articles of the EOLLS encyclopedia annotated with UNL graphs.

Our aim for the future is to develop SECTra_w so that it becomes an "operating platform" for translation-oriented corpora, where new functions such as various types of annotations or expansions can be implemented either by communicating with other web services, or by building them directly over SECTra_w, using a suitable API. We are studying how to define and implement new interfaces and functions in a generic way, using the structure of documents in a given corpus as parameter.

We also aim at scaling up to handle huge corpora, which need terabytes of storage if they include sound, images, video, and multiple annotations (less space-hungry but leading to heavy computations).

ACKNOWLEDGMENTS

The work reported here has mainly been supported by a PhD MIRA grant from the RRA (Région Rhône-Alpes). We would also like to thanks two anonymous reviewers for many pertinent comments, which have been taken into account as much as possible given the space constraint.

REFERENCES

[1] Bey Y., Boitet C., Kageura K. (2006). *The TRANSBey Prototype: An Online Collaborative Wiki-Based CAT Environment for Volunteer Translators*. Proc. LR4Trans-III, E. Yuste (ed.), LREC-06, Genoa, Italy, pp. 49-54.

[2] Blanchon H., Boitet C. & Besacier L. (2004). *Evaluation of Spoken Dialogue Translation Systems: Trends, Results, Problems and Proposals*. Proc. COLING-04, 7 p.

[3] Blanchon H., Boitet C., Brunet-Manquat F., Tomokiyo M., Hamon A., Hung V. T. & al. (2004). *Towards Fairer Evaluations of Commercial MT Systems on Basic Travel Expressions Corpora.* Proc. IWSLT-04, pp. 21-26.

[4] Boitet C. & Blanchon H. (1994) *Multilingual Dialogue-Based MT for Monolingual Authors: the LIDIA Project and a First Mockup.* Machine Translation 9/2, pp 99—132.

[5] Boitet C., Bey Y., Tomokiyo M., Cao W. & Blanchon H. (2006). *IWSLT-06: Experiments with Commercial MT Systems and Lessons from Subjective Evaluations.* Proc. IWSLT-06, pp. 23—30.

[6] Fafiotte G., Boitet C. (2003). *ERIM, a platform for supporting and collecting multimodal spontaneous bilingual dialogues,* IEEE NLP-KE2003, Beijing, 26-29/10/2003.

[7] Fafiotte G., Boitet C., Seligman M. & Zong C. (2004). *Collecting Bilingual Dialogues using a Web-Based Platform for the Study of Interpretation.* Proc. LREC-04, 9 p.

[8] Koehn P. (2003). *Europarl: A Multilingual Corpus for Evaluation of Machine Translation*. 2 p. http://people.csail.mit.edu/koehn/publications/europarl/

[9] Kraif O. (2006). *Corpus multilingues — multilingual corpora.* 22/11/06, 3 p. http://w3.u-grenoble3.fr/kraif/

[10] Nguyen H-T., Boitet C., Sérasset G. (2007). *PIVAX, an online contributive lexical database for heterogeneous MT systems using a lexical pivot,* Proc. SNLP-07, 6 p.

[11] Tsai W-J. (2004). *La coédition langue-UNL pour partager la révision entre langues d'un document multilingue*, Ph.D Thesis, Université Joseph Fourier, 310 p.

[12] Uchida H. (2004) *The Universal Networking Language (UNL) Specifications Version 3 Edition 3.* UNL Center, UNDL Foundation, December 2004, 43 p. http://www.undl.org/unlsys/unl/UNLSpecs33.pdf

[13] Bowker L. (2000). *Towards a methodology for exploiting specialized target language corpora as translation resources.* International Journal of Corpus Linguistics, 5(1), pp. 17-52.

[14] XWiki Enterprise. http://en.wikipedia.org/wiki/XWiki, 11/02/2008.