A Web-oriented System to Manage the Translation of an Online Encyclopedia

Using Classical MT and Deconversion from UNL

Christian Boitet[1,3] Cong-Phap Huynh[1,2] Hervé Blanchon[1,3] Hong-Thai Nguyen[1,3]

(1) GETALP, LIG (Laboratoire d'Informatique de Grenoble), UFR IMAG, France
(2) INP-G (Institut National Polytechnique de Grenoble)
(3) Université Joseph Fourier
{Christian.Boitet, Cong-Phap.Huynh, Herve.Blanchon, Hong-Thai.Nguyen}@imag.fr

***Abstract*. We start from a web-oriented system for evaluating, presenting, processing, enlarging and annotating corpora of translations, previously applied to a real MT evaluation task, involving classical subjective measures, objective n-gram-based scores, and objective post-edition-based task-related evaluation. We describe its recent extension to support the high-quality translation into French of the large on-line Encyclopedia of Life Support Systems (EOLSS) presented as documents each made of a web page and a companion UNL file, by applying contributive on-line human post-edition to results of Machine Translation systems and of UNL deconverters. Target language web pages are generated on the fly from source language ones, using the best target segments available in the database. 25 documents (about 220,000 words) of the EOLSS are now available in French, Spanish, Russian, Arabic and Japanese. MT followed by contributive incremental cheap or free post-edition is now proved to be a viable way of making difficult information available in many languages.**

Keywords: Machine Translation, human contributive post-edition, UNL, EOLSS, translation corpora, SECTra_w

INTRODUCTION

The first version of the SECTRa_w system was used in 2007 for an evaluation campaign in the TRANSAT project of France Telecom R&D. It then contained parts of several parallel corpora such as EuroParl, BTEC, etc., a small quantity of manually post-edited MT outputs, and some 30 hours of spoken interpreted task-related bilingual dialogues in several language pairs collected by the ERIM project [5]. Handling that corpus showed that a generic system for handling corpora should be adaptable to *multifile documents*, as a dialogue is represented by a main descriptor identifying the speakers, the language, and the turns, and each turn has its descriptor, identifying its sound file, and text files containing written transcriptions in one or more languages.

We have then extended that system to manage the translation of a part of the EOLSS[1] using classical MT and *deconversion* from UNL, in the framework of a research contract from UNDL-F[2], and again encountered a corpus of multifile documents: an article of EOLSS is represented by a web page, a folder of satellite files, and a *companion* `.unl` file. Numerous functionalities were added to SECTra_w for this project, in a generic way, so that they will be usable to support other translation jobs relative to large information sets disseminated by web sites. They involve a mixture of automatic techniques (MT and search in large translation memories) and human work performed online, using any browser. A long-term goal is to transform SECTra_w into a *corpus operating system*, programmable by final users (translation managers, linguists) with a specialized and simple command language and/or graphical interface.

In the following, we describe the particular structure of the EOLSS corpus, the general translation scenario, and how the main tasks are achieved: preparation by segmentation, import, production of translation candidates using MT systems and/or UNL deconverters, post-edition, and production of results and statistics.

I. STRUCTURE OF EOLSS/UNL CORPUS

We speak of "EOLSS/UNL" because we get EOLSS documents after they have been preprocessed by the UNDL Foundation, and not in their original form and format.

A. General structure

The EOLSS corpus consists of 6600 articles, written in English by specialists who are often not native English speakers. An article is about 30 standard pages long, so that EOLSS totals about 250,000 pages and 62.5 M words.

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

Figure 1: EOLSS article as it appears on the Web

The corresponding HTML code is:

```
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…
```

Figure 2: Corresponding .html file

[1] The online Encyclopedia Of Life Support Systems has been developed since 1996 using funds from the Dubai-based EOLSS Foundation, under the aegis of UNESCO.
[2] Universal Networking Digital Language Foundation.

The EOLSS/UNL-UnescoL project tackled 25 articles, about 220 K words, or 880 pages, in total 13673 segments (sentences or titles), as many UNL graphs, and a lexicon of about 15,000 simple or compound entries, half of them relative to various technical fields related to life support systems.

Each document is represented by two files, a standard Html file (.html), and a *companion* file in UNL format (.unl). There may also be a folder of satellite files (images, icons). Figure 1 gives an example from a document on the tsunamis, and Figure 3 a UNL graph.

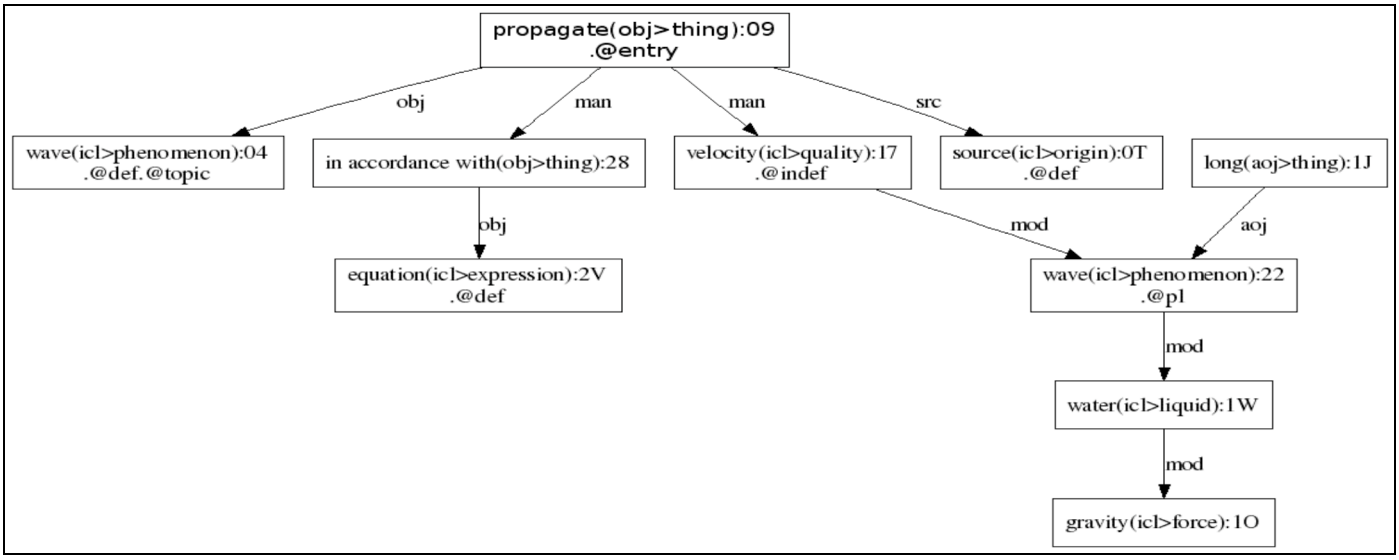

Figure 3: drawing of a UNL graph

Here is its form in the `.unl` companion file:

```
[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)
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 4: UNL format in the "companion" file

This project aimed at two major goals:

(1) produce HQ translations of 25 EOLSS articles provided by the UNDL Foundation, in the format above;

(2) do a feasibility study, in relation with UNESCO and UNDL-F, to test the applicability of the UNL-based architecture on the translation of EOLSS from English into the 5 other languages of UNESCO.

B. Remarks on `.html (.aspx)` files

The `.aspx` files are not always clean Xhtml, but sometimes invalid Html: an opening tag may have no corresponding closing tag, or the order of opening and corresponding closing tags is not correct. Therefore, before processing and working with `.aspx` files, we try to *normalize* them.

We also have to deal with difficulties of processing such as finding the boundaries of segments (smallest meaningful translation units such as sentences or titles) because of equations, figures, links, etc. inside sentences (see Figure 5).

```
Technological layout includes ozonation
<!-- MP SYPH( --><script
id=mpch0012s1>MPSetChAttrs('ch0012','ch0',[
[6,1,-3,0,0],[8,1,-,0,0]]) </script> ![if
!ie]><span id=mpnnch0012ph
class=MPNNCode><img border=0  name=
mpch0012ph src= '&{DSMP.gEmptySrc};'
height=1
width='&{DSMP.gPlaceholderWidth};'></span>
MPSetChAttrs('ch0013','ch0',[[6,1,-3,0,0],
[8,1,-4,0,0],[10,1,-4,0,0],[],[],[],[25,2,-
11,0,0]]) </script> adsorption on the
granulated active carbon with the
disinfections by sodium hypochlorite.
   ...
```

Figure 5: Equations, links, etc. inside a segment

C. Remarks on .unl files

The UNL format (Uchida 2004) predates Xml as it was defined in 1996 (Figure 4). 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}`, `{ar}` for the translations into French, Spanish, etc., comments introduced by ";", and *out-of-text* symbols in an original segment replaced by `HTM1`, `HTM2`, etc. `{xxx}` tags may contain attributes like Xml attributes (without enclosing double quotes). It is possible to have several translations into some language, such as an automatic result or a post-edited version, and several UNL graphs, provided their attributes are different.

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* (a node having the `.@entry` attribute). 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 *anglo-semantic interlingua*.

D. General Scenario

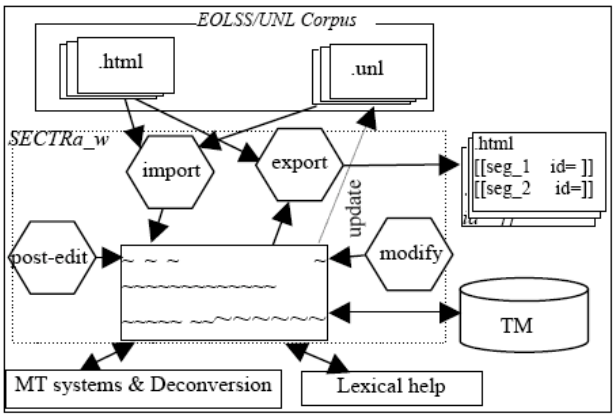

Figure 6: Scenario of using SECTra_w for EOLSS

Figure 6 shows the management of translation of the EOLSS/UNL corpus in SECTra_w. When a document (`.html` and `.unl` files) is imported, the `.html` file is first *segmented* into textual segments and HTML code. A corresponding *skeleton file* is also generated from the `.html` file. This file is very important for making sure the translated versions obtained are in the same format as the source one.

After the document is segmented and checked for correct alignment, its segments and UNL graphs are submitted to several MT systems (local and online), to Translation Memory (if large enough), and to available UNL deconverters.

An online collaborative post-edition editor visualizes the source segments and all their translations along with linguistic helps (at the document, segment, and lexical levels), allowing many post-editors to participate and produce HQ translations. The HQ translations are not only inserted in the skeleton file exactly at the position of their source segments, but also saved into the Translation Memory where they become instantly usable by all post-editors.

II. Preparation of Translation Work

A. Segmentation and import

The .html file is segmented into Html code and textual segments which must correspond to those bracketed by `{org}` and `{/org}` in the .unl file. But the segments in the .unl files are sometimes too short, that is, sub-segments from correct longer segments (see Figure 4 above). The source segments as well as their UNL graphs and translations are stored into the database of SECTra_w. A *skeleton* .html file is produced, with placeholders for the segments.

B. Segment representation

For each segment, SECTra_w/EOLSS stores:

- up to 3 forms for the source language: the literal content of the `.aspx` file, possibly a corrected from, the content of the `{org}` part in the `.unl` file, and the content of the following comment, if any.
- up to 2 forms for the UNL graphs: that found in the `.unl` file, and possibly a revised form.
- for each target language, one or more *pre-translations*, outputs of MT systems.
- post-editions or direct human translations (at least *** quality, with score & metadata).

Each object associated to a segment has metadata indicating its producer (program or human), a *quality level* (from * to *****), and a *score* (from 0 to 20). As for levels,

'*' is for word by word translations;

'**' is for MT ouputs;

'***' is for post-editions or translations by humans knowing both languages;

'****' is for post-editions or translations by professional translators native speakers of the target language;

'*****' is for post-editions or translations done or blessed by bilinguals or translators certified by the organization disseminating the information (for EOLSS, we have none yet, but they would be translators employed by Unesco).

A priori scores are assigned in the profiles of the human contributors and of the MT systems. They can be modified by contributors during post-edition, or a posteriori by revisors or linguistic administrators. Typically, a bilingual science student would have (***, 11/20) if not versed in ecology, but s/he could change the score to 9/20 in case of doubt about a term, or 15/20 if s/he finds that the translation of that particular segment is particularly good.

Concerning MT systems, we currently fix some score after browsing through a sample of the MT outputs. An open and interesting research issue is to find good ways to compute scores reflecting the *usefulness for post-edition* of individual pre-translations.

The same source segment may appear at several places in several documents, and its translation may have to be different (even if the meaning is the same, the contexts can cause terminological divergences). Currently, we do as in IBM's TM/2 and consider *textual contexts,* equated with occurrences (context = place in some document), so that the different post-

editions of a segment (in a given target language) define a partition of the textual contexts. That should be refined, to allow users to personalize translations in certain contexts (as for menu items in end-user applications such as Notepad™).

C. Pre-processing and problems

Before a document is imported, both `.aspx` and `.unl` files are converted to UTF-8. The segmentation is guided by the segmentation done in the `.unl` file, but there are differences between segments contained in the `.unl` file and in the `.html` files because:

- *out-of-text* parts such as equations are often replaced by *special occurrences* such as HTM1, HTM2 in the `.unl` file.
- some segments in the `.unl` file have been slightly changed in other ways: some punctuations, words, or characters have been deleted or inserted, but the original text (presumably stored in a DB) has not been changed accordingly.
- some special symbols (space, quote, etc..) are represented differently. For example, the `.unl` file can contain "*42 000 'ecosystems'.*" and the .`html` file "*42 000 "ecosystems&quote;.",* with an error on an entity (`&quot;`).

Finally, the segmentation in the `.unl` file is often too fine-grained: a sentence of the form "text equation text" is one segment for translation, and should not be further split, although the equation could be replaced by a *special label* such as `$$_equ_23` to make automatic processing easier. But, in the `.unl` file, such a sentence will be split into 3 parts: 2 *infra-segments* around a piece of (not to be translated) code. As MT systems should be applied to whole segments, and deconversion to infra-segments (because the UNL graphs correspond to them), it is necessary to keep the whole segments and their infra-segments.

III. MT AND DECONVERSION FROM UNL

Translational suggestions, or *pretranslations,* are outputs of MT systems and human transltions or pos-ediitons retrieved from the translation memory (exact matches only). Systran and Reverso have been used for EOLSS en_fr, but in principle more can and should be used. One pretranslation is chosen (by some crude rule at this point) to initialize the post-edition cell. Although the remaining pretranslations are very rarely looked up, their prove to be useful in those cases, and should be kept.

As far as deconverters are concerned, there is only one per language to date, and the reason to use them is to make comparisons on samples between the time to post-edit MT outputs and deconversion outputs. Besides producing HQ translations, with a working methodology and good tools, our second goal for the future is to improve the UNL approach until it becomes 2-3 times more efficient than the classical MT approach.

A. MT using classical MT systems

What to submit to MT systems?

- to web translators, preferably the *HTML source form,* because they are built to handle web pages.
- to MT systems able to use linguistic information attached to elements such as mathematical expressions or relations, icons, anchors…, *normalized forms* (such as in `.unl`, with *out-of-text* parts of sentences replaced by *special occurrences* such as `HTM1`, `HTM2`… or `$$_rel_1`, `$$_expr_2`…).

We submit to MT systems not only whole segments, but their infra-segments, if any, because some whole segments are in any case too long to be handled by available MT systems, and also because, in particular for the English-French pair, concatenating the MT outputs on the infra-segments of a segment may give an acceptable translation of that segment.

B. Deconversion from UNL

As we are not supposed to modify the UNL graphs and the segmentation appearing in the `.unl` files, we can get deconversion results only on `.unl` segments, which may be infra-segments.

From SECTra_w, we launch in the background a process to produce gif images of the new or modified UNL graphs, and another one to deconvert them in any language for which we have access to a deconverter. The results are stored in SECTra_w. If a graph is incorrect, the type of error is also stored. UNL graphs can also be revised, but that is not our task in the EOLSS/UNL project.

It is also possible to submit one ore more segments from the `.unl` file to both processes, from SECTra_w, or directly from a web service such as unldeco-FR, or its new extended form EMEU_w.

Figure 3 shows the drawing G. Sérasset's `unldeco` program produces for the segment in Figure 4.

IV. POST-EDITION

A. Management

The post-edition manager allows many users to work collaboratively at the same time on the same collection of data (segments, pages, document). For example, a document of 160 segments may have 25 sentences needing post-edition, and there may be two post-editors accessing this document. If the length of a page appearing in the post-edition window has been set to contain about 250 words (about 16 sentences in the case of EOLSS), the document will be divided in 10 (logical) pages. The post-edition manager ensures that 2 contributors never access the same segment at the same time, and warns them when they access the same page at the same time. It associates a red mark or background to the segments under process by somebody, and locks them temporarily. An orange mark or background is associated to a page containing a red segment (as well as its "free" segments). Other pages and segments are green (as for traffic lights).

SECTra_w always displays the percentage of post-edited sentences in a document, and updates it when a user completes a post-edition.

The post-edition manager also handles information such as author's name, start time, finish time, total duration, status, changed characters and words, and other measures of the post-edition effort and cost.

Figure 7: SECTra_w post-edition window. Source text on the left, MT pretranslations on the right

Figure 8: SECTra_w post-edition window showing the post-edition effort as insertions and deletions of substrings

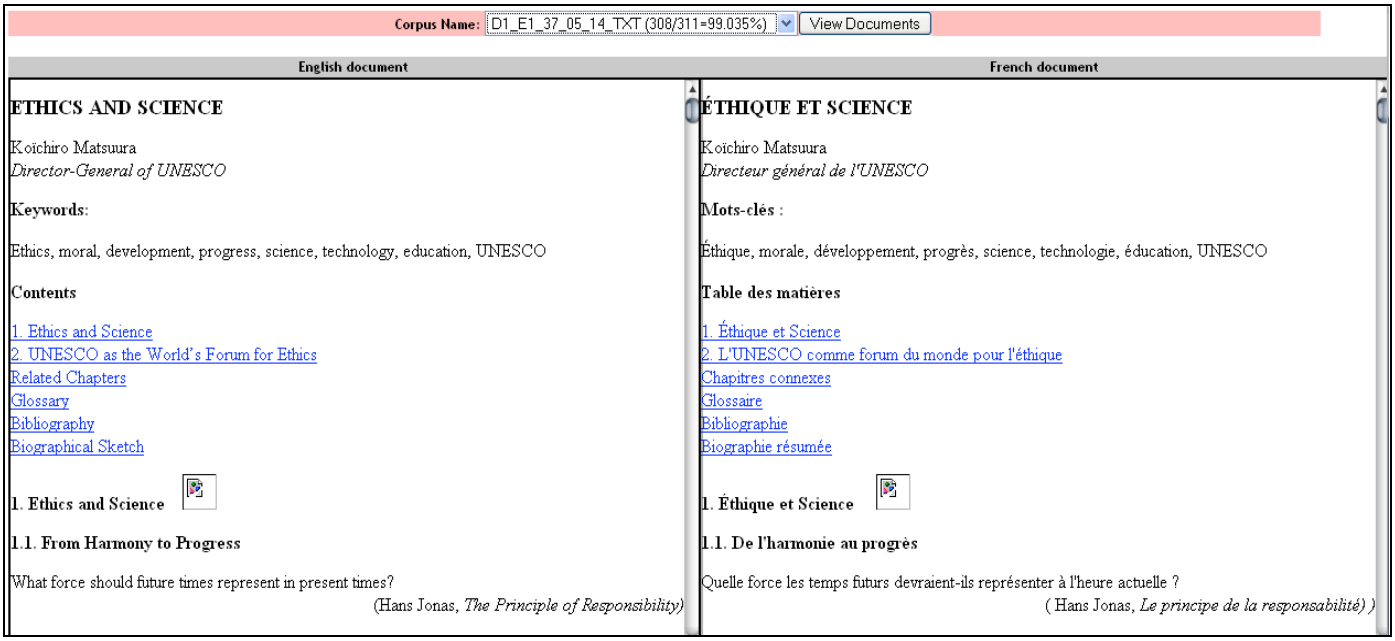

Figure 9: Source and target document parallel visualization

There are several classical possible measures.

- In the profession, translators are paid by words or by pages (1 standard page of English has 250 words), with rates corresponding to the time taken, itself linked to the difficulty of the task (language pair, complexity of syntax, difficulty of terminology, proportion of examples found in the translation memory for each bracket of matching ratio, e.g. [0%..74%], [75%..89%], [90%..100%]).
- The simplest and most reliable measure is the post-editing time[3], impossible to measure reliably when post-edition is done on the web. However, it can be estimated *a posteriori,* by tuning the coefficients and weights of a mixed edit distance between the MT output and the final post-edited result.

B. Editor layout

We follow the following presentation principles.

- *Verticality*: all objects of the same type should appear in the same column.
- *Horizontality*: all objects linked with the same source segment (possibly including its corrections) are presented in the same row.
- *Locality*: main functions always reside in the same area. Post-edition happens in the upper pane, where everything concerning segments appears (source text, post-edited text, MT results, suggestions from the TM). Dictionary-related information and activity is located in the lower pane (interface with a lexical database containing information dedicated to the corpus at hand and modifyable by contributors).
  Objects or important zones should be kept at the same place and with approximately the same size. Accordingly, the current segment should not move down when the translator clicks to go to the next one. Rather, the next one should move up[4].
- *Proactivity*: the system should propose suggestions for translations of a segment and its words or expressions immediately when the user clicks on it. Hence, MT as well as search in the TM and in dictionaries should happen (and happens) before, in the background, and be available without any explicit action of the user.

The *post-edition interface* can be accessed either directly, or by viewing an Html form of the translated document, shown side by side with the original, selecting a passage, and asking to post-edit it.

The *side by side Html form* is shown in a separate tab and can be updated by clicking on *refresh*, so that *the effects of changes are immediately visible*.

Linguistic helps are now provided in several forms:

- at document level, by parallel views of the source language and the target language files (Figure 9).
- at segment level, by showing several *MT pretranslations,* if available (Systran and Reverso have been used), and exact matches in the TM[5].
- at lexical level, by a pane through which post-editors can consult an interface with a lexical database specialized to the EOLSS corpus, stored in a PIVAX lexical database (Nguyen &al. 2007).
- unpostedited segments can be easily *filtered* and proposed to contributors.

[3] See Jeff Allen's (http://www.geocities.com/mtpostediting) web site on MT postediting web site for references and experiments.

[4] Due to technical difficulties, this feature is not yet implemented.

[5] Suggestions from the Translation Memory (TM) are included, but the interface should show *analogical rectangles* (example found, its translation, source segment, and one or more translation suggestions).

V. Production of Results and Statistics

A. Export of results

Results produced are:

- the original `.unl` files enriched with translations (& associated metadata),
- the target language .aspx files corresponding to the source language files, in the same format.

In general, there can be 4 kinds of target language segments, obtained by (1) MT, (2) post-edition of MT result (s), (3) deconversion from the UNL graph, (4) post-edition of deconversion results. For handling detected errors, the original content of `.unl` files should not be altered, only additions are permitted.

If a segmentation is found to be wrong, e.g. 3 sub-segments instead of 1 segment, the translation of the segment should be put as translation of the 1st sub-segment, with special metadata, and no post-edited translation appearing in the 2 other sub-segments. However, MT should also applied on sub-segments, and results included, in *raw* (not post-edited) form. If a UNL graph is found to be wrong (there are many errors in the current version), a comment to that effect may be inserted after it, before the `{/unl}` tag.

Statistics are also produced, for information purposes: words, characters, segments, documents, average, processing times (MT, deconversion, post-edition of each type), etc., for each contributor.

B. Current experimentation

1) MT and human post-edition

About 30 French native speakers from our lab have experimented the tool, for 5 to 10 to 100 hours. Several students in professional translation also joined, as well as some junior university science students knowing English well enough.

2) Lexical database

There are actually 3 "lexical collections": *collected entries, proposed entries, normalized entries.*

- We have collected and put in a PIVAX database about 120 Mb of data related to the UW headwords of the EOLSS documents, from freely available sources such as IATE.
- The *proposed entries* are those which have been proposed directly by posteditors, or extracted from the <source, postedition> pairs (we give them different scores, like 0.5 and 0.3). They are stored in PIVAX-EOLSS, an instance of the PIVAX lexical database (also a web service). As said above, SECTra_w has yet to offer a mini-interface to interact with PIVAX, but it is possible to use the PIVAX full interface in another tab.
- The *normalized entries* are already used entries that have been adopted as the would-be norm for this context, and are used to build the deconversion dictionary. They are given weights higher than some threshold (e.g., 0.8). In fact, proposed and normalized entries are stored in the same PIVAX instance, only weights differ.

As for MT results, the lexical interface should be *proactive*, meaning that results of dictionary look-up will be precomputed and stored with each segment. That part is still under development: lexical information has been harvested for the whle domain of water and ecology, but segment-specific small dictionaries are still to be implemented.

Conclusion

We have described SECTra_w, a web-oriented **S**ystem for **E**xploiting (evaluating, presenting, processing, enlarging and annotating) **C**orpora of **Tra**nslations on the **w**eb, and in more detail its extension and use to support high-quality translation of a small part of EOLSS, a large on-line encyclopedia, where each document is made of a web page, its satellite files, and a companion UNL document.

Results of MT systems and/or outputs of UNL deconverters have been improved through contributive on-line human post-edition by about 40 volunteers. Target language web pages are generated on the fly from source language pages, using the best target segments available.

Concerning UNL, the hope is to improve the UNL deconverters and the semi-automatic enconversion process until using UNL will be significantly more efficient, at least with 5 target languages.

In the mean time, we have conclusively shown that HQ translations can be obtained using commercial MT and contributive post-edition done on the web, for the most part on a voluntary basis, thus making HQ multilingual access to interesting but often arduous information possible.

Acknowledgments

The work reported here has mainly been supported by a research contract from the UNDL Foundation and by a MIRA PhD grant from the RRA (Région Rhône-Alpes). We would also like to thanks 3 anonymous reviewers for their pertinent comments, which have been taken into account as much as possible given the space and time constraints.

References

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

[2] Blanchon H., Boitet C., and Choumane A. 2006. *Traduction automatisée fondée sur le dialogue et documents auto-explicatifs: bilan du projet LIDIA*. International journal Traitement Automatique des Langues (TAL). 2006, vol. 47(3), 30 pages.

[3] 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.

[4] Choumane A., Blanchon H., Roisin C. 2005. *Integrating translation services within a structured editor*. Proceedings of the ACM Symposium on Document Engineering (DocEng 2005). Bristol, United Kingdom. November, 02-04, 2005. vol. 1/1: pp. 165-167.

[5] Fafiotte, G. (2004). *Building and sharing multilingual speech resources, using ERIM generic platforms.* COLING-MLR 2004, Geneva, Switzerland.

[6] 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 (7th International Symposium on Natural Language Processing), Bangkok, Thailand, 6 p.

[7] 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.

[8] 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

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

APPENDIX: EXAMPLE ALLOWING TO ESTIMATE THE TRANSLATION QUALITY

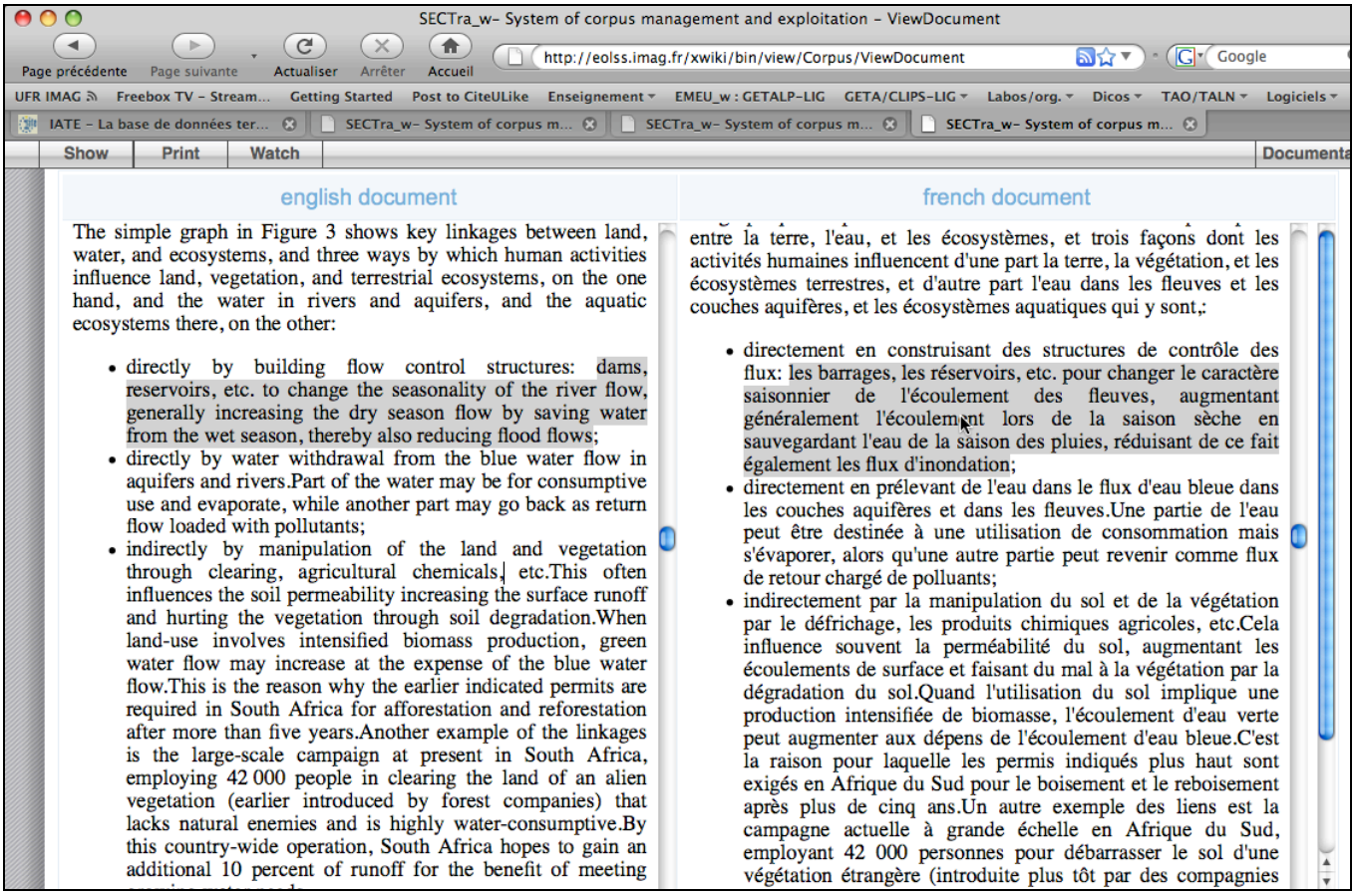

Figure 10 : Source-target parallel view of a document. Clicking on a segment opens the post-edition window at that point.