A Web Service Enabling Gradable Post-edition of Pre-translations Produced by Existing Translation Tools: Practical Use to Provide High-quality Translation of an Online Encyclopedia

Hervé Blanchon **Christian Boitet** **Cong-Phap Huynh**
Université de Grenoble
Équipe GETALP, Laboratoire LIG
385, rue de la bibliothèque, 38041 Grenoble, France
{Herve.Blanchon, Christian.Boitet, Cong-Phap.Huynh}@imag.fr

Abstract

SECTra_w is a Web-based system offering several services, such as supporting MT evaluation campaigns and online post-editing of MT results, to produce reference translations adapted to classical MT systems not built by machine learning from a parallel corpus. The service we are interested in here is the possibility for its users to import a document, or a set of documents (not only a list of preprocessed segments), and achieve high-quality translation by applying on-line human post-edition to results of Machine Translation systems. A community of contributing post-editors may carry out the on-line human post-edition. In this paper, we describe the use of SECTra_w to translate into French a set of 25 html documents (about 220,000 words) on water and ecology from the on-line Encyclopedia of Life Support Systems[1] (EOLSS) using a contributive on-line human post-edition framework.

1 Introduction

SECTra_w stands for « Système d'Exploitation Contributive de Corpus de Traductions sur le Web », in English, "Contributive Operating System of Translation Corpora on the Web". The adjective "operating" underlines the ultimate goal of producing a system that will be programmable and customizable by its final users. It is the successor of a previous environment we developed to support online volunteer translators communities [Bey, Y*., et al.* 2008].

SECTra_w system has been used for an internal evaluation campaign funded by France Telecom R&D from English to French in the tourism domain [Huynh, C.-P*., et al.* 2008]. This was a real MT evaluation task, including classical subjective measures, objective n-gram-based scores, and objective post-edition-based task-related evaluation.

SECTra_w was then extended to support a variety of corpus types, and tested for its capacity to handle large parallel corpora such as EuroParl [Koehn, P. 2005], BTEC [Takezawa, T*., et al.* 2002], a small quantity of manually post-edited MT outputs, and some 30 hours of interpreted task-related spoken bilingual dialogues in several language pairs collected by the ERIM project [Fafiotte, G. 2004].

Since February 2008, SECTra_w has been used in the French part of the EOLSS/UNL project of the UNDL Foundation, to support and manage the high quality translation of a part of the large EOLSS corpus. In this project, SECTra_w supports a contributive on-line environment for human post-edition applied to results of Machine Translation systems and of UNL deconverters[2].

After a brief presentation of SECTra_w, we introduce the EOLSS/UNL corpus and the workflow

[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] As the UNL part of the project does not add any significant information on the matter of the paper this part will be left out.

we used to translate English articles of the encyclopedia into French. In the last section, we report on results.

2 SECTra_w

The current version of SECTra_w (see Figure 7, Appendix A for an overview) supports two main tasks: MT system evaluation, and High Quality translation production through (collaborative) post-edition[3].

2.1 MT System evaluation

A corpus for an evaluation campaign is a collection of aligned rough source segments (one up to several sentences, a speech turn), candidate translations produced by Machine Translation systems, and collections of reference translations (gold standard translations).

SECTra_w was first developed to: (1) import, verify and correct a source segments corpus; (2) call various MT systems to get candidate translations; (3) allow collaborative post-editing of the candidate translations by human translators to produce other reference translations; (4) carry out online subjective evaluation (fluidity, adequacy) with a collection of human judges; (5) compute classical objective n-gram-based scores such as BLEU and NIST; (6) perform task-oriented evaluations by measuring an edit distance and/or the post-editing time.

SECTra_w could also import a complete evaluation data set (source corpus, candidate translations, reference translations) in order to carry out an evaluation campaign using all or some of the possible subjective and objective measures.

2.2 (Collaborative) post-edition towards high-quality translation

Translating and extending translation corpora into new languages serves different purposes such as construction, enhancement, and evaluation of Machine Translation systems, multilingualization of websites and systems, etc. Therefore, there are more and more projects including the translation of corpora.

SECTra_w was thus further enhanced with the goal of supporting and facilitating the translation of corpora by allowing the creation and the organization of corpus translation projects easily and efficiently. The difference here is that a corpus is not a collection of rough segments, but a collection of documents, with simple or complex structures.

In order to conduct a corpus translation project in SECTra_w, the project leader (1) creates a translation project name along with users groups including accounts for human translators and project managers, (2) defines human translators' profiles[4], (3) imports the source corpus, (4) preprocesses the source corpus if necessary (by segmenting, converting, verifying, correcting it), (5) calls various Machine Translation systems to get translation suggestions, (6) assigns translation tasks to human translators if no suggestion is available or releases the corpus for collaboratively post-editing (translating), and finally (7) exports the results as files and/or makes them visible as web pages.

2.3 Current SECTra_w implementation[5]

In its current version, SECTra_w handles HTML documents. The goal is thus for example to take as input a web page in French and get a web page with the same layout in English.

A Web page contains HTML tags as well as plates (figures, tables, equations…). It is thus necessary to extract the *source segments*[6] from the source web page and give them identifiers (1). A *companion file* associated to each Web page describes the way source segments are produced. Other companion files contain also the identified plates of the page.

Source segments are then translated by a translation memory and external MT systems (2), and their perfect and/or draft translations are then inserted in the post-edition environment. Search in the translation memory is implemented through exact match. According to our experience and

[3] The reader may consult Jeffrey Allen's web site on that topic http://www.geocities.com/mtpostediting/.

[4] A profile may be set according to the translator's skills in both the source and target languages, and its knowledge of the domain terminology.

[5] In this section bracketed numbers correspond to those of the Figure 1.

[6] For us, a segment is a translation unit; it may be a title, a full sentence if it is not "sliced" by a plate (e.g. the second and third sentences [Figure 2] "Because … h^{-1}." and "Such … beach." are each considered as one segment), or part of a sentence when a plate is inserted (the first sentence "The … basin." is divided into two segments ["The … equation" and "where … basin."] because of the equation). Those two segments are called further infra-segments.

other available studies (cf. note 3), MT is always more useful than fuzzy match proposals.[7]

For each segment, one candidate translation (supposed to be "the best") is chosen by the system as initial value of its post-edition cell. If exact match do not provide any result, among MT draft results "the best" one is chosen using an *ad hoc* function (for example [Potet, M. 2009] uses a language model). For now, we simply use the score given by the administrator (after manual inspection of a sample of results) in the MT system profile.

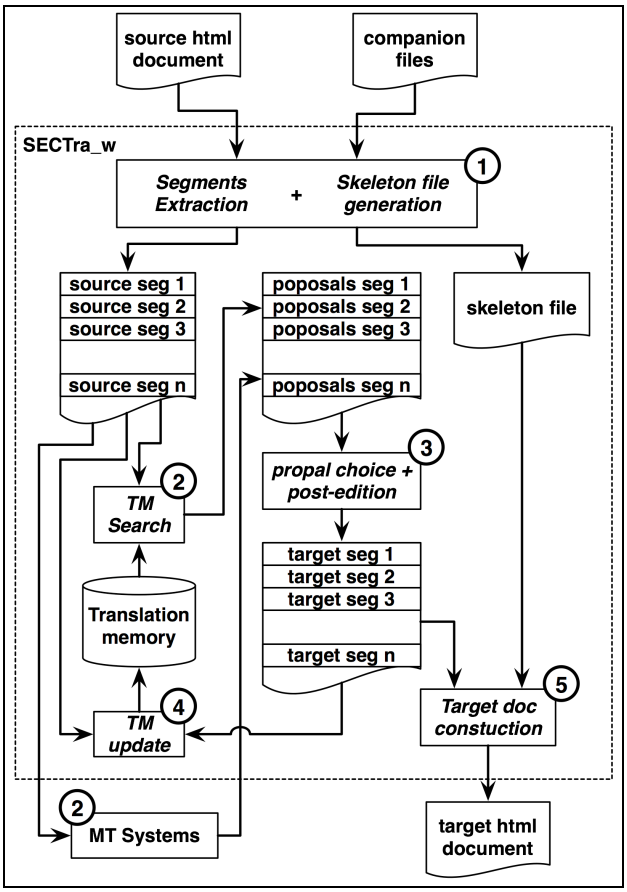

Figure 1: SECTra_w architecture

The user may choose any other candidate translation by simply clicking it (3). In this experiment, we have noticed: – that the users read the other alternative translations between 1 and 2% of the time just to check whether a better lexical equivalent would be available for one word or expression; – that the users reset the post-edition cell for 10% of those previous cases (logged by SECTra_w).

During post-edition, the Translation Memory is updated (4). Whenever it is requested, the translation may be inserted into a formatted target web page having exactly the same layout as the source page (5). The formatted target web page is constructed through a *skeleton file*, using the source web page, the plates and the segments identifiers.

3 The EOLSS/UNL task

The whole EOLSS encyclopedia consists of 6600 articles[8], written in English by specialists who are often not native English speakers.

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. Because** the average depth of the world ocean is 4 km, the typical velocity of tsunami in the ocean is 200 m s^{-1} or 720 km $\mathbf{h^{-1}}$**.**

Such a wave, propagating with the velocity of an airplane, may traverse the Pacific ocean in 10-12 hours and bring down a wall of water 10 m high with a velocity of more than 70 km h^{-1} upon a calm ocean **beach.**

Figure 2: excerpt of an EOLSS article as it appears on the Web (segments boundaries are bold-faced)

3.1 The EOLSS/UNL corpus

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

The project aimed at two major goals: (1) produce high quality translations of 25 EOLSS articles[9] relative to various technical fields related to life support systems provided by the UNDL Foundation; (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[10].

[7] With MT + fuzzy match we do not save more than 50% as far as translation time is concerned compared with from scratch human translation. While with Systran we always save at least 50% of the time (55% on a regular basis and 65% for one of the authors for the post-edition of 7000 segments in this experiment).

[8] An article is about 30 standard pages long. EOLSS totals about 250,000 pages and 62.5 M words.

[9] A total of about 220 K words, or 880 pages, 13673 segments (sentences or titles), as many UNL graphs, and a lexicon of about 15,000 simple or compound entries (half of them).

[10] In this paper we focus only on the first goal.

3.2 General structure of an article

Each article is represented by two files:

- an .html file,

```
<p ALIGN ="JUSTIFY">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.  Because the average depth of the world ocean is 4 km, the typical velocity of tsunami in the ocean is 200 m s<sup>-1</sup> or 720 km h<sup>-1</sup>.</p> <p ALIGN="JUSTIFY">Such a wave, propagating with the velocity of an airplane, may traverse the Pacific ocean in 10-12 hours and bring down a wall of water 10 m high with a velocity of more than 70 km h <sup>-1</sup> upon a calm ocean beach.</p>
```

Figure 3: .html file corresponding to Figure 2

- and a *companion* file in UNL format (.unl). This companion document contains UNL segments and plates used to construct the skeleton file.

3.3 Remarks on the .unl file

The UNL format [Uchida, H. 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, and within a sentence {org} for the original text, {fr}, {sp}, etc. for the translations into French, Spanish, etc., comments introduced by ";", and out-of-text symbols in an original segment replaced by HTM1, HTM2, etc.

```
[S:44]
{org} The … equation {/org}
{unl} … {/unl} ;unl graph representing the sentence
[/S]
[S:45]
{org} where … basin. {/org}
{unl} … {/unl}
[/S]
[S:46]
{org} Because … is 200 HTM1 or 720 HTM2. {/org}
{unl} … {/unl}
[/S]
[S:47]
{org} Such a wave, … 70 HTM1 … ocean beach. {/org}
{unl} … {/unl}
[/S]
```

Figure 4: the normalized .unl file from Figure 2

4 The EOLSS/UNL workflow

In order to carry out the translation task, we first imported the segmented articles into SECTra_w. Then we collected MT translation candidates for each of them. These translation candidates were then manually post-edited to produce (1) high quality translations of each segment, and (2) translations of the EOLSS article themselves.

We involved a group of student translators, one of the authors and a linguist of our team. High quality was judged by checking the results as a standard professional mean.

4.1 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 (cf. Figure 4). The source segments are stored into the database of SECTra_w. The segmented representation of our example is the following:

```
$$_seg_1: The wave …accordance with the equation
$$_equ_1: C<sub>G</sub> = (g H)<sup>1/2 </sup>
$$_seg_2: where g …depth of the basin.
$$_seg_3: Because …, the typical velocity of tsunami in the ocean is 200 m s<sup>-1</sup> or 720 km h<sup>-1</sup>.
$$_seg_3: Such a wave, … more than 70 km h <sup>-1</sup> upon a calm ocean beach.
```

Figure 5: segmentation of the original .html file using the information contained in the .unl file.

A *skeleton* HTML file is produced, with placeholders for the segments.

```
<p ALIGN ="JUSTIFY">$$_seg_1</p><p ALIGN="JUSTIFY"><font size="4">
$$_equ_1</font>
</p><p ALIGN ="JUSTIFY">$$_seg_2</p> <p ALIGN="JUSTIFY">$$_seg_3</p>
```

Figure 6: Skeleton file associated with our example document

4.2 Segment representation

For each segment, SECTra_w stores:

- 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:

- '*' for word by word translations;
- '**' for MT outputs;
- '***' for post-editions or translations by humans knowing both languages;
- '****' for post-editions or translations by professional translators native speakers of the target language;
- '*****' for post-editions or translations done or certified by bilinguals or translators certified by the organization disseminating the information.

A priori scores are assigned in the profiles of the human contributors and of the MT systems. 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 pretranslations.

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[11] 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++™).

4.3 Off-the-shelf MT systems' pretranslation

Translational suggestions, or pretranslations, are outputs of MT systems and human translations or post-editions retrieved from the translation memory (exact matches only).

Systran and Reverso have been used for EOLSS, 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 some cases, and should be kept.

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 bold faced in Figure 4).

We submit to MT systems not only whole segments, but their infra-segments[12], if any, because some whole segments are in some cases 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.

4.4 Post-Edition

4.4.1 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

[11] TM/2 is still used by IBM and its subcontractors to translate more then 20M words/year towards more than 25 languages.

[12] Cf. note 6.

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.

There are several classical possible measures [King, M.*, et al.* 2003]. We would like to point out the following points:

- 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, 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.

4.4.2 Editor layout

We follow the following presentation principles (Figure 9, Appendix C).

- *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).
- *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.

Post-edition efforts can be visualized (Figure 8, Appendix B) by the user in the Post-Edit column of the interface through the trace button (Figure 7, Appendix A).

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 (Figure 10, Appendix D), 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 *effects of changes are immediately visible.*

5 Conclusion, results and perspectives

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.

About 40 volunteers (French native speakers from our lab, several students in professional translation, and some junior university science students knowing English well enough) have improved results of MT systems through contributive on-line human post-edition for 5 to 10 to 100 hours. Target language web pages are generated on the fly from source language pages, using the best target segments available.

In the mean time, we have conclusively shown that high quality 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 high quality 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).

References

Bey, Y., Kageura, K. and Boitet, C. (2008) *BEYTrans: A Wiki-based Environment for Helping Online Volunteer Translators*. in Yuste Rodrigo, E. (ed.), Topics in Language Resources for Translation and Localisation. pp. 135-150.

Fafiotte, G. (2004) *Buliding and Sharing Multilingual Speech Ressources using ERIM Generic Platform*. Proc. COLING 2004 – Worikshop on Multilingual Linguistic Ressources. Geneva, Switzerland. August 28, 2004. 8p.

Huynh, C.-P., Boitet, C. and Blanchon, H. (2008) *SECTra_w.1: an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora*. Proc. LREC'08. Marrakech, Morocco. 28-30 May, 2008. 6 p.

King, M., Popescu-Belis, A. and Hovy, E. (2003) *FEMTI: creating and using a framework for MT evaluation*. Proc. MT Summit IX. New Orleans, USA. September 23-27, 2003. 8 p.

Koehn, P. (2005) *Europarl: A Parallel Corpus for Statistical Machine Translation*. Proc. MT Summit X. Phuket, Thailand. September 13-15, 2005. vol. **1/1**: pp. 79-86.

Potet, M. (2009) *Méta-moteur de traduction automatique : proposition d'une métrique pour le classement de traductions*. Proc. RECITAL 2009. Senlis, France. 24-26 juin 2009. 10 p.

Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H. and Yamamoto, S. (2002) *Towards a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World*. Proc. LREC-2002. Las Palmas, Spain. May 29-31, 2002. vol. **1/3**: pp. 147-152.

Uchida, H. (2004) *The Universal Networking Language (UNL), Specifications, Version 3, Edition 3*. UNL Center. 43 p.

Appendix A. The whole SECTra_w Interface

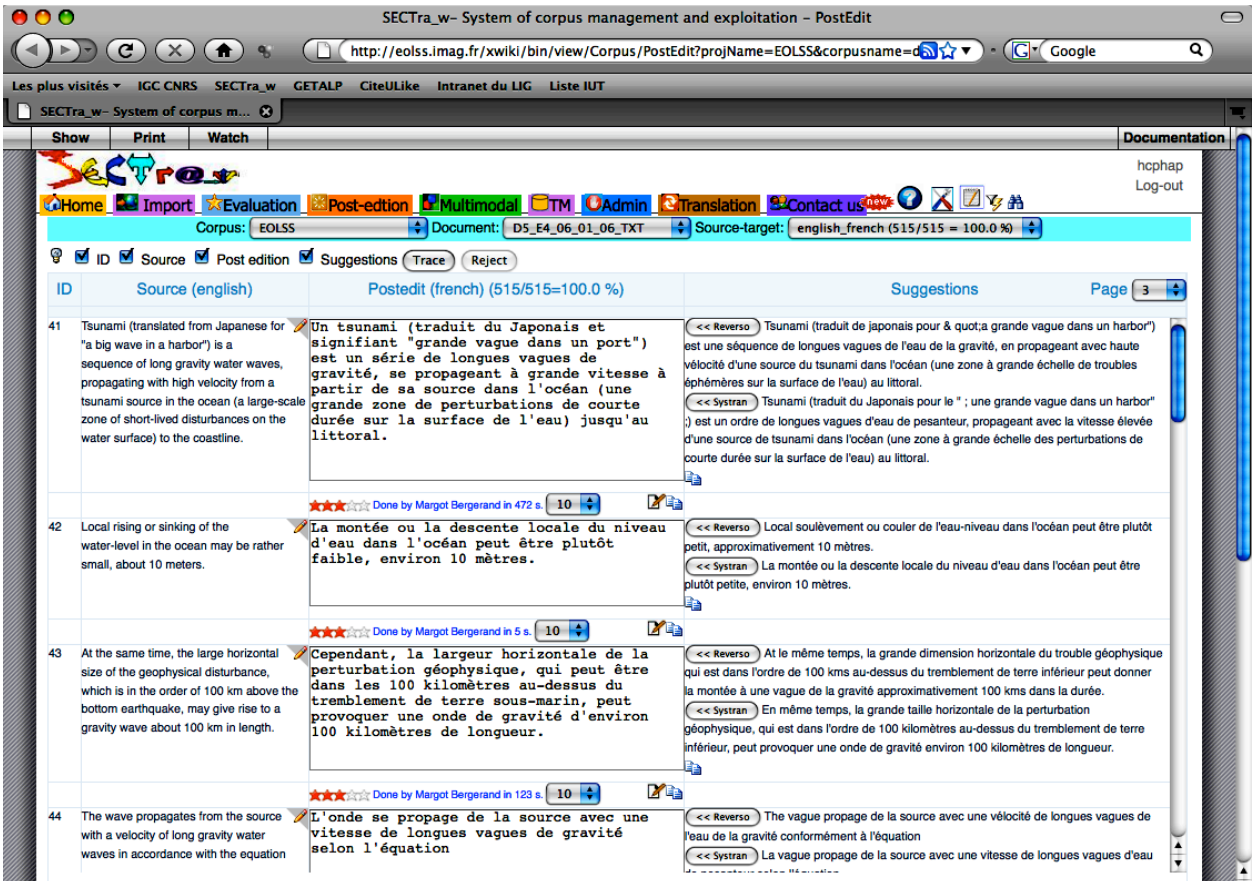

Figure 7: The SECTra_w interface

User "hcphap" is connected on the "EOLSS" corpus, browsing page "3" of document "D5_E4_06_01_06_TXT", source language is "English", Target Language is "French". The document consists of 515 segments that have all been translated and post-edited ("515/515=100%").

Appendix B. Post-edition effort visualization

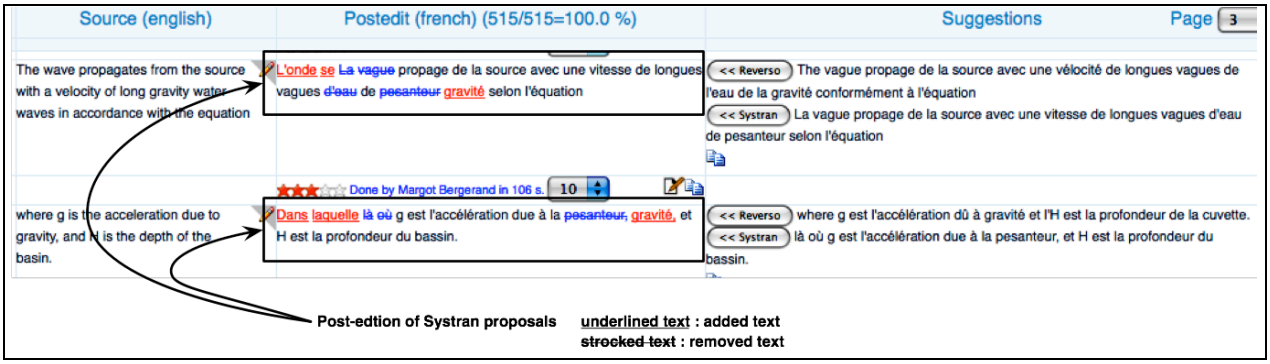

Figure 8: SECTra_w post-edition window showing the post-edition effort

Appendix C. Post-edition interface

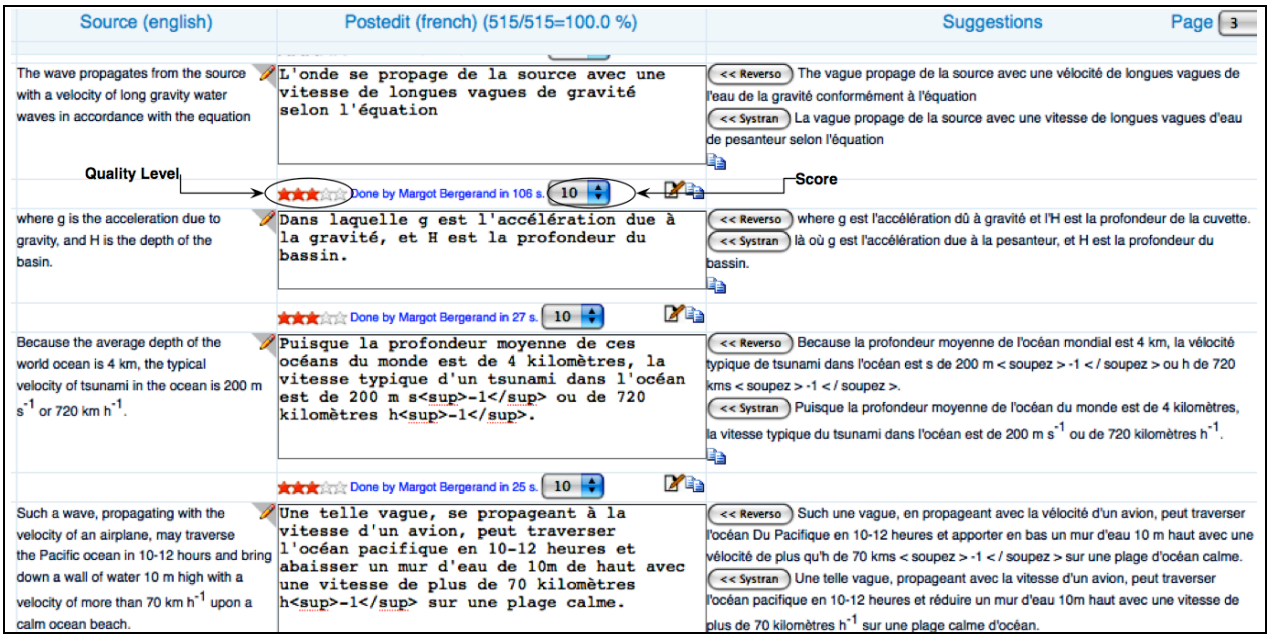

Figure 9: SECTra_w post-edition window. Source segments on the left, MT pre-translations on the right

Appendix D. Aligned English source and French target documents

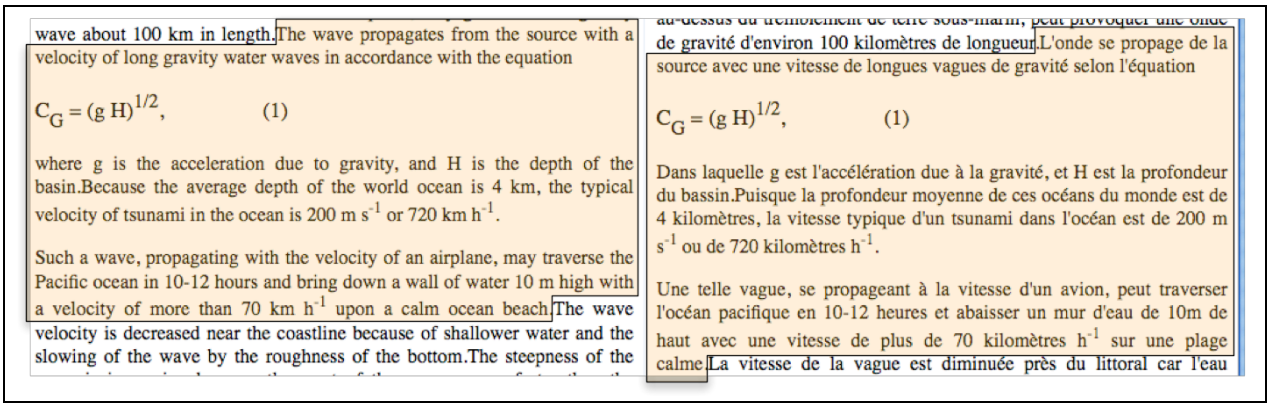

Figure 10: Source and target documents parallel visualization