SECTra_w.1 : an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora

Cong-Phap Huynh, Christian Boitet, Hervé Blanchon
Laboratoire LIG, GETALP, GETA, Université Joseph Fourier
385, rue de la Bibliothèque, 38041 Grenoble, France
{Cong-Phap.Huynh, Christian.Boitet, Herve.Blanchon}@imag.fr

Abstract

SECTra_w.1 is a web-oriented system mainly dedicated to the evaluation of MT systems. After importing a source corpus, and possibly reference translations, one can call various MT systems, store their results, and have a collection of human judges perform subjective evaluation online (fluidity, adequacy). It is also possible to perform objective, task-oriented evaluation by letting humans post-edit the MT results, using a web translation editor, and measuring an edit distance and/or the post-editing time. The post-edited results can be added to the set of reference translations, or constitute it if there were no references. SECTra_w.1 makes it possible to show not only tables of figures as results of an evaluation campaign, but also the real data (source, MT outputs, references, post-edited outputs), and to make the post-edition effort sensible by transforming the trace of the edit distance computation in an intuitive presentation, much like a "revision" presentation in Word. The system is written in java under Xwiki and uses the Ajax technique. It can handle large, multilingual and multimedia corpora: EuroParl, BTEC, ERIM (bilingual interpreted dialogues with audio and text), Unesco-B@bel, and a test corpus by France Telecom have been loaded together and used in tests.

Introduction

Recent MT evaluation campaigns have been criticized because only tables of figures such as scores (BLEU, NIST, ORANGE, METEOR…) are shown as results, while these n-gram-based measures have been shown not to correlate very well with human judgments, contrary to initial expectations [Callison-Burch & al. 2006]. Commercial MT systems have been consistently ranked low by these measures, while human judges ranked them quite high, or highest. Hence, it would be good to allow researchers (and others) to look at the real data and to "see for themselves". We also developed bilingual corpora of interpreted spoken bilingual dialogue (between French and Chinese, Vietnamese, Hindi and Tamil) and needed a web-oriented system to manage parallel multimodal corpora. The impetus to start developing such a system was actually a part of a research contract with FT R&D, where we had to organize a full fledged evaluation of MT systems (data collection, subjective and objective evaluation).

In this paper, we will first analyze what such a system should offer, and mention some further interesting problems that appear in this context. Then, we detail the three main functionalities of SECTra_w.1, namely visualizing, evaluating, and post-editing parallel corpora on the web.

1. Motivations & general architecture

To look at the real data and "see for oneself" is also needed if one considers the subjective evaluation parts of evaluation campaigns. They consist in letting humans (several per translation unit) issue judgments (of adequacy, fluidity, fidelity, etc.), but the agreement between judges is often not very good, and the procedure itself is often biased.

The first bias comes from showing judges a reference translation of an input instead of the input itself when judging adequacy (monolingual judges are cheaper than bilingual ones), and the second from showing to one judge the results of several MT systems in parallel, so that they compare them instead of grading them independently. A third problem is that, judges being expensive, only a very small fraction of the MT outputs can be judged subjectively. Hence, it would be good to make it possible for other people to continue to do subjective evaluation after the evaluation campaign, in a contributive, wiki way.

Finally, none of the above classical evaluation methods evaluate the "real" quality of MT systems, i.e. their utility for a certain task. The task of an MT system may be (1) to help humans understand texts in foreign languages (MT for watchers: intelligence, web browsing…), or (2) produce HQ translations (MT for translators, interactive MT for monolingual users), or (3) communicate (support of bilingual dialogues). These are the main reasons why we started the SECTra_w project.

SECTra_w.1 is a web-oriented system mainly dedicated to the evaluation of MT systems. As we concentrate now on the second task (produce HQ translations), SECTra_w.1 supports MT post-editing, and measures of the effort spent by a bilingual person to produce good (HQ) translations from the MT output, after having read the input.

After importing a source corpus, and possibly reference translations, one can call various MT systems, store their results, and have a collection of human judges perform subjective evaluation online

(fluidity, adequacy). It is also possible to perform objective, task-oriented evaluation by letting humans post-edit the MT results, using a web translation editor, and measuring an edit distance and/or the post-editing time. The post-edited results can be added to the set of reference translations, or constitute it if there were no references.

With SECTra_w.1, it possible to show not only tables of figures as results of an evaluation campaign, but also the real data (source, MT outputs, references, post-edited outputs), and to make the post-edition effort sensible by transforming the trace of the edit distance computation in an intuitive presentation, much like a "revision" presentation in Word. It is also possible to recompute n-gram-based scores by using the post-edited translations as references, and/or by adding them to the already available references. However, that experiment has only been performed to date internally, with a small data set and off-the-shelf systems. We are looking forward to do it in a large evaluation campaign with many competing systems.

The system is written in java under Xwiki and uses the Ajax technique. It can handle large, multilingual and multimedia corpora: EuroParl, BTEC, ERIM (bilingual interpreted dialogues with audio and text), Unesco-B@bel, and a test corpus by France Telecom have been loaded together and used in tests.

During the development of SECTra_w.1, we encountered other interesting problems, and we plan to further develop SECTra_w to elucidate them.

First is the question of how to handle extremely large corpora, containing not only text, but audio, video, and various annotations such as POS tags, morpho-syntactic lattices, dependence and constituent trees, logical formulas, UNL hypergraphs, and correspondences (between successive representation levels in one language, and between similar levels across two languages). An interesting goal is to find an architecture to dynamically attach annotation platforms to SECTra_w.

Second, corpora such as phrasebooks do not contain only fixed sentences, but sentences with variables ("please give us $nb plates of $meat and $vegi."): one would like to develop systems to translate them and to store their instances and corresponding translations.

Third, a good corpus exploitation tool should offer a way of considering the "segments" of text as occurrences within documents. For that, it is necessary to store the context of production of each occurrence of a given segment. It may also be necessary to find ways to refer to the structures of the documents where segments occur. For example, we have integrated in SECTra_w.1 a "player" for the ERIM corpus of bilingual interpreted dialogues (French—Chinese, Vietnamese, Hindi, Tamil), and would like to integrate a generic functionality of that kind into a future version of SECTra_w.

2. Visualizing Parallel Corpora

That interface is not an autonomous java application; rather, it is realized by any web browser.

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) constitute a "polyphrase" and are presented in the same "row".
- No direct manipulation of the presentation parameters, but modifications of parameters (to be as "bare-bone" as possible, thus reducing development time and helping to concentrate on adding more useful features).

The interface to control the presentation of columns will be the same as for PIVAX, a web-oriented data base for heterogeneous MT systems (Nguyen 2007). As it was not finished at the time of writing, we show an example from PIVAX. The first screen shot shows columns in a certain order, with a movement button enhanced. After the user clicks on it, the column moves, as shown by the second screen shot.

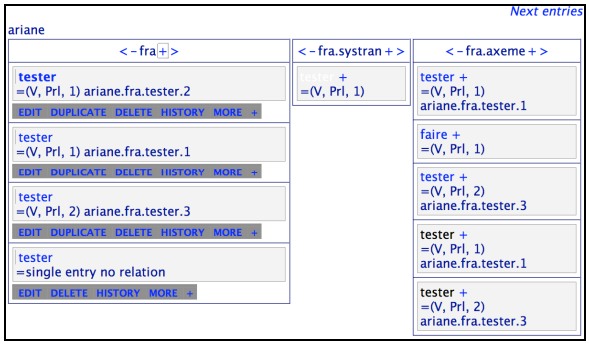

Figure 1: Column position control (1)

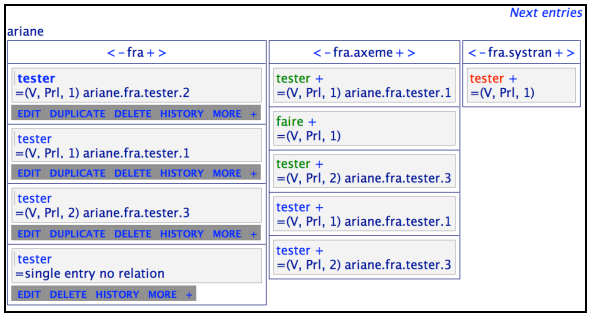

Figure 2: Column position control (2)

In the next SECTra_w example (Figure 3), the source text is at the left, and the post-edition column has been moved to the right with respect to its position in the post-edition interface, to bring a MT column near to the source text column. In the future, we will add the possibility to show some annotation under the source, such as its pronunciation, or an "active reading" presentation showing possible equivalents of words found in attached dictionaries.

Of course, this "table-like" presentation has nothing to do with the logical structure of the data (at the "business level"), and with the way it is stored in the underlying database (at the "physical level").

Source	Translation (Reverso)	Distance	Post-edition
	Accept Trace Reject	D=a.Dchar+b.Dword a:0.2, b:0.8	Accept Trace Reject
?Hamburger and stew on the right side and salad, please.	Hamburger et ragoût à droite côté et salade, s'il vous plaît.	Dc=20,Dw=7 Dsent= 9.6	Un Hamburger et du ragoût à droite sur le côté et de la salade, s'il vous plaît.
That fried fish, one sausage with green peas, please.	Cela a frit du poisson, une saucisse avec les pois verts, s'il vous plaît.	Dc=25,Dw=8 Dsent= 11.4	Ce poisson frit, ~~Cela a frit du poisson,~~ une saucisse avec ~~les~~ des ~~pois~~ petits ~~verts,~~ pois, s'il vous plaît.
T-bone steak and sauerkraut and fried potatoes, please.	Steak avec un os en T et choucroute et a frit des pommes de terre, s'il vous plaît.	Dc=33,Dw=11 Dsent= 15.4	Du bifteck à l'os ~~Steak~~ et ~~avec~~ de ~~un~~ la ~~os en T et~~ choucroute et ~~a frit~~ des pommes de ~~terre,~~ terre frites, s'il vous plaît.
Roast chicken and two slices of ham on this side and spinach, please.	Poulet du rôti et deux tranches de jambon sur ce côté et épinards, s'il vous plaît.	Dc=8,Dw=2 Dsent= 3.2	Du Poulet du rôti et deux tranches de jambon sur ce côté et des épinards, s'il vous plaît.
I'd like breakfast, please.	J'aimerais petit déjeuner, s'il vous plaît.	Dc=3,Dw=1 Dsent= 1.4	J'aimerais un ~~J'aimerais~~ petit déjeuner, s'il vous plaît.

Figure 3: Visualization of a parallel corpus on the web

3. Classical MT Evaluation

3.1. Objective n-gram-based scores

We have integrated scripts provided by NIST to compute BLEU and NIST. WER is named "Dw" in the interface, while "Dc" is the character-based distance, and Dsent is a linear combination of both: Dsent = α Dc + β Dw (α and β are modifiable). In the future, we will use not only insertions, deletions and exchanges, but also shifts of blocks, and global changes (over a set of segments or a whole corpus).

In the default presentation, the first column contains the source segments, and the second is the post-edition column. Then comes a column for the MT output to be judged, with a radio button for each level of adequacy (A1—A5) and fluidity (F1—F5).

The fourth column is dedicated to post-edition. It is initialized by the MT output. The trace of the edit-distance computation is shown in that column: in red, inserted strings, in overstriked blue, erased strings.

3.2. Subjective evaluation

3.2.1. Principles

We follow the same presentation principles as above, and add a new one:

- Constant help: each grading button is "self-explaining" (an explanation of the score appears in a bubble when the cursor lies over it).

3.2.2. Examples

Figure 4 shows the default interface used by judges, and Figure 5 the interface used by the organizer (the scores given by all judges are shown).

Source	Translation (Reverso)	Distance	Post-edition
	Accept Trace Reject	D=a.Dcar+b.Dmot a:0.2, b:0.8	Accept Trace Reject
?Hamburger and stew on the right side and salad, please.	Hamburger et ragoût à droite côté et salade, s'il vous plaît.	Dc=20,Dw=7 Dsent= 9.6	Un hamburger et du ragoût à droite sur le côté et de la salade, s'il vous plaît.
	(A1) (A2) (A3) (A4) (A5) (F1) (F2) (F3)		
That fried fish, one sausage with green peas, please.	Cela a frit du poisson, une saucisse avec les pois verts, s'il vous plaît.	Dc=25,Dw=8 Dsent= 11.4	Ce poisson frit, une saucisse avec des petits pois, s'il vous plaît.
	(A1) (A2) (A3) (A4) (A5) (F1) (F2) (F3)		
T-bone steak and sauerkraut and fried potatoes, please.	Steak avec un os en T et choucroute et a frit des pommes de terre, s'il vous plaît.	Dc=33,Dw=11 Dsent= 15.4	Du bifteck à l'os et de la choucroute et des pommes de terre frites, s'il vous plaît.
	(A1) (A2) (A3) (A4) (A5) (F1) (F2) (F3)		

Dc: character distance, **Dw**: word distance
D: sentence distance

Adequation: A1 : All , A2 : Almost all, A3: Half, A4 : Few, A5 : None

Fluency: F1 : written, F2 : oral,F3 : not acceptable

Figure 4: Interface for subjective evaluation

Source	Translation (Reverso)	Distance	Post-edition			
	Accept Trace Reject	D=a.Dchar+b.Dword a:0.2, b:0.8	Accept Trace Reject	HERVE	GEORGES	ACHILLE
?Hamburger and stew on the right side and salad, please.	Hamburger et ragoût à droite côté et salade, s'il vous plaît.	Dc=20,Dw=7 Dsent= 9.6	Un Hamburger et du ragoût à droite sur le côté et de la salade, s'il vous plaît.	(A2) (F2)	(A2) (F3)	(A1) (F3)
That fried fish, one sausage with green peas, please.	Cela a frit du poisson, une saucisse avec les pois verts, s'il vous plaît.	Dc=25,Dw=8 Dsent= 11.4	Ce poisson frit, ~~Cela a frit du poisson,~~ une saucisse avec ~~les~~ des ~~pois~~ petits ~~verts,~~ pois, s'il vous plaît.	(A3) (F3)	(A2) (F3)	(A2) (F3)
T-bone steak and sauerkraut and fried potatoes, please.	Steak avec un os en T et choucroute et a frit des pommes de terre, s'il vous plaît.	Dc=33,Dw=11 Dsent= 15.4	Du bifteck à l'os ~~Steak~~ et ~~avec~~ de ~~un~~ la ~~os en T et~~ choucroute et ~~a frit~~ des pommes de ~~terre,~~ terre frites, s'il vous plaît.	(A3) (F3)	(A2) (F3)	(A3) (F3)
Roast chicken and two slices of ham on this side and spinach, please.	Poulet du rôti et deux tranches de jambon sur ce côté et épinards, s'il vous plaît.	Dc=8,Dw=2 Dsent= 3.2	Du Poulet du rôti et deux tranches de jambon sur ce côté et des épinards, s'il vous plaît.	(A2) (F2)	(A2) (F2)	(A2) (F3)
I'd like breakfast, please.	J'aimerais petit déjeuner, s'il vous plaît.	Dc=3,Dw=1 Dsent= 1.4	J'aimerais un ~~J'aimerais~~ petit déjeuner, s'il vous plaît.	(A1) (F2)	(A1) (F1)	(A1) (F1)

Figure 5: Visualization of results of subjective evaluation

3.2.3. Organization of an experiment

In the current implementation,

- every candidate translation is judged by 3 judges,
- the distribution and control of work is done through a "mini-workflow".

3.2.4. Experiment and validation

We have tested SECTra_w.1 in an internal evaluation campaign funded by FT R&D, on two English-French corpora constructed from English segments from the BTEC corpus [Kikul, G., et al. 2004, Takezawa, T., et al.2002], distributed by ATR for IWSLT-06, from their translations by 2 commercial and 1 research MT systems, and from post-editions of MT results.

For lack of time and human resources, we did not perform again the evaluations we did for the IWSLT-06 evaluation campaign. However, we measured the average time taken by judges to perform subjective evaluation on our (admittedly different) data, and that time diminished by a factor of about 5.

As this factor is almost the same as that observed in our lab during IWSLT-06, when we asked Cao WenJie, a Chinese PhD researcher, to stop comparing MT outputs before rating them, when she evaluated adequacy on the Chinese-English test corpus, we attribute that result to the new presentation, where a judge never sees more than 1 MT output for a given segment in the same screen.

4. Task-related Objective Evaluation

4.1. Task and Measure Chosen

Measures involving humans but no human judgments, such as task-related measures, are objective, not subjective. There are 3 main tasks of MT, namely helping humans produce high quality translations, understand texts in foreign languages, and communicate. We have chosen the first task, as it seems easier to rate than the others. As a matter of fact, there are classical possible measures:

- in the profession, translators are paid by words or by pages (1 standard page 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 (see Jeff Allen's mtpostediting web site on geocities for references and experiments).
- post-editing time can quite reliably be estimated a posteriori, by measuring an adequate edit distance between the MT output and the final post-edited result.

Remark: measures such as WER and mWER (used in many recent MT evaluation campaigns) are edit distances, but are not related to the task performed because they compare MT outputs with "reference translations", and not with post-ed-editions. That is because the set of (good) possible translations of a sentence (even if short) is very big, and quite sparse for any edit distance — contrary to the set of possible written transcriptions of a spoken utterance. Hence, they do not qualify as task-related measures.

4.2. Interface Principles

We follow the same principles as above (verticality, horizontality, constant help, bare-bone presentation), and add a few others:

- 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 (lexical information merged from available

sources, and minimal interface with an online lexical database dedicated to the corpus at hand).

- Layout: objects or important zones should be kept at the same place and with approximately the same size. Accordingly, the "current segment" (polyphrase) does not move down when the translator clicks to go to the next one. Rather, the next one moves up. We also try to show approximately twice as many previous segments as succeeding ones e.g., 6 before and 3 after), which corresponds to the "golden ratio".
- 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 before, in the background, and be available without any explicit action of the user.

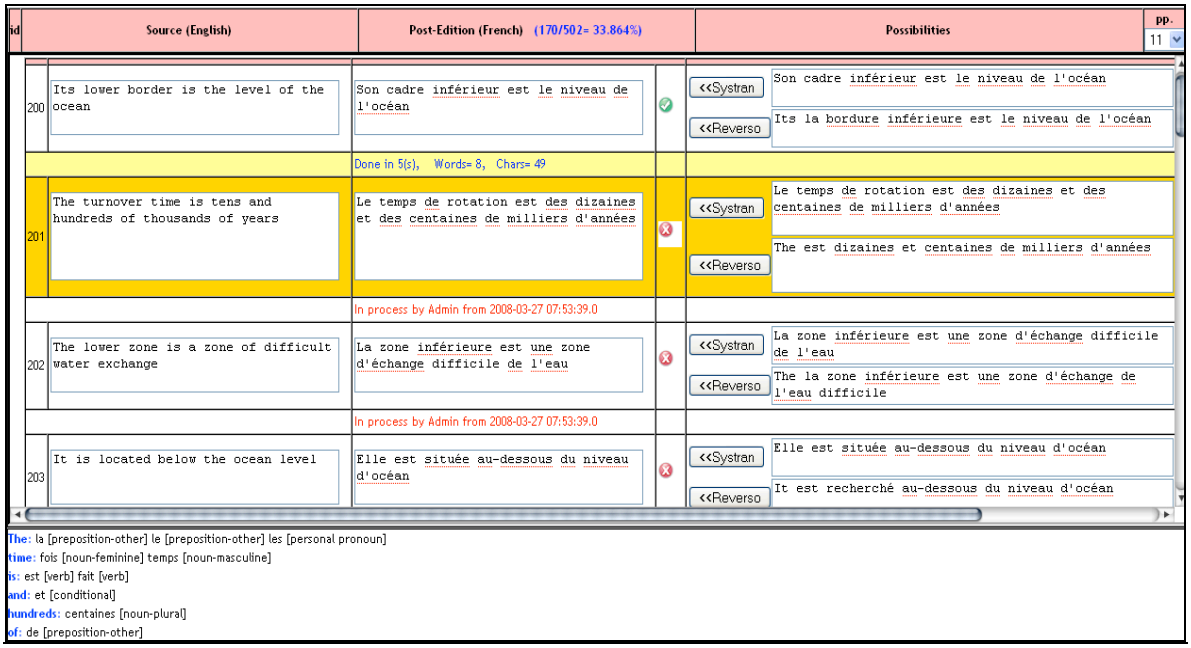

Figure 6: Interface of task-related objective evaluation

4.3. Current Interface

In SECTra_w.1, translation suggestions are provided only by calls to MT systems. We have only implemented the search for exact matches in the TM. A good fuzzy search is actually quite complex to program, and we hope to integrate in the future that of Similis™, based on the model of "layered translation memories" (Planas 1998, 2000).

As far as lexical help is concerned, the left side of the lower pane shows a list of equivalents from one or more online dictionaries or terminological databases, and the right part is a small form to access an instance of the PIVAX database [Nguyen 2007].

4.4. Experimentation

A first experimentation has been performed in November 2007, without integrated language resources (dictionaries, terminologies, glossaries).

Work is in process in the framework of the EOLSS project. This project aims at translating 25 articles of the online Encyclopedia of Life Support Systems (EOLSS) — about 220 K words, or 880 K standard pages. In that context, access to some dictionaries and to PIVAX from the post-edition interface has been developed.

5. Conclusion

SECTra_w.1 is a web-oriented system for managing multilingual corpora on the web. In its current state, it is mainly dedicated to the evaluation of MT systems. After importing a source corpus, and possibly reference translations, one can call various MT systems, store their results, and have a collection of human judges perform subjective evaluation online (fluidity, adequacy). It is also possible to perform objective, task-oriented evaluation by letting humans post-edit the MT results, using a web translation editor, and measuring an edit distance (and/or the post-editing time). The post-edited results can be added to the set of reference translations, or constitute it if there were no references.

A first experiment has been performed on a real evaluation task, without integrated dictionary help. Another one is in progress in the context of the EOLSS/UnescoL project, on a larger scale, with integrated dictionary help.

Further research will concern ways to handle special types of documents (structured "multifile" documents), multimodal documents (to manage really large-scale data), and processing of annotations of various types (morphosyntactic lattices, decorated trees, UNL hypergraphs, etc.)

6. Acknowledgements

This work has been partly supported by a MIRA scholarship of the Rhône-Alpes Region and by the research contract TRANSAT awarded to our laboratory by FranceTelecom R&D.

7. References

Bey Y., Boitet C., Kageura K. (2006), *The TRANSBey Prototype: An Online Collaborative Wiki-Based CAT Environment for Volunteer Translators*, In Proceedings of the 3rd International Workshop on Language Resources for Translation Work, Research & Training (LR4Trans-III), E. Yuste, (ed.), LREC-06, Genoa, Italy, pp. 49-54.

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.

Blanchon H., Boitet C. & Besacier L. (2004) *Evaluation of Spoken Dialogue Translation Systems: Trends, Results, Problems and Proposals*. Proc. COLING-04, Genève, 23-27/8/04, ACL, 7 p.

Blanchon H., Boitet C., Brunet-Manquat F., Tomokyo M., Hamon A., Hung V. T. & al. (2004) *Towards Fairer Evaluations of Commercial MT Systems on Basic Travel Expressions Corpora.* Proc. IWSLT-04 (International Workshop on Spoken Language Translation), Kyoto, Japan, September 30 - October 1, ATR, pp. 21-26.

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 (International Workshop on Spoken Language Translation), Kyoto, 27-28/11/06, ATR, pp. 23—30.

Boitet C., Blanchon H. (1994), *Multilingual Dialogue-Based MT for Monolingual Authors,* Machine Translation vol. 9(2).

Callison-Burch C., Osborne M. & Koehn P. (2006) *Re-evaluating the Role of BLEU in Machine Translation Research.* Proc. EACL-06, Trento, April 3-7, 2006, ITC/irst, ed., 8 p.

Fafiotte G., Boitet C. (1994), *ERIM, a platform for supporting and collecting multimodal spontaneous bilingual dialogues,* IEEE NLP-KE2003, Beijing, 26-29/10/2003.

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 (Language Resources and Evaluation Conference), Lisbonne, 24-28/5/04, 9 p.

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

Kraif O. (2006) *Corpus multilingues — multilingual corpora.* 22/11/06, 3 p. http://w3.u-grenoble3.fr/kraif/index.php-?option=com_content&task=view&id=20&Itemid=36

NIST (2001) *Automatic Evaluation of Language Translation using N-gram Co-occurrence Statistics.* 8 p. http://www.nist.gov/speech/tests/mt/doc/