CESTA: First Conclusions of the Technolangue MT Evaluation Campaign

O. Hamon(1,2), A. Popescu-Belis(3), K. Choukri(1), M. Dabbadie(4), A. Hartley(5), W. Mustafa El Hadi(4), M. Rajman(6), I. Timimi(4)

(1) ELDA - 55-57, rue Brillat Savarin, 75013 Paris – France
(2) LIPN UMR 7030 - Université Paris 13 & CNRS – 99 av. J.-B. Clément, 93430 Villetaneuse – France
(3) University of Geneva - 40 bvd du Pont d'Arve - CH-1211 GENEVA 4 – Switzerland
(4) IDIST / CESARTES - University of Lille 3 - rue du Barreau BP 149 - 59653 Villeneuve d'Ascq Cedex – France
(5) University of Leeds - Centre for Translation Studies Woodhouse Lane Leeds LS2 9JT – UK
(6) LIA - Ecole Polytechnique Fédérale de Lausanne Bât. INR - CH1015 Lausanne – Switzerland

E-mail: hamon@elda.org, andrei.popescu-belis@issco.unige.ch, choukri@elda.org, dabbadie@univ-lille3.fr, a.hartley@leeds.ac.uk, widad.mustafa@univ-lille3.fr, martin.rajman@epfl.ch, ismail.timimi@univ-lille3.fr

Abstract

This article outlines the evaluation protocol and provides the main results of the French Evaluation Campaign for Machine Translation Systems, CESTA. Following the initial objectives and evaluation plans, the evaluation metrics are briefly described: along with fluency and adequacy assessed by human judges, a number of recently proposed automated metrics are used. Two evaluation campaigns were organized, the first one in the general domain, and the second one in the medical domain. Up to six systems translating from English into French, and two systems translating from Arabic into French, took part in the campaign. The numerical results illustrate the differences between classes of systems, and provide interesting indications about the reliability of the automated metrics for French as a target language, both by comparison to human judges and using correlations between metrics. The corpora that were produced, as well as the information about the reliability of metrics, constitute reusable resources for MT evaluation.

1. Introduction

This paper describes the French national campaign for Machine Translation (MT) evaluation, named CESTA[1]. The paper introduces the evaluation protocol used for CESTA, inspired from (NIST, 2003), providing evaluation results using a number of well-known and experimental automated metrics. In addition, using scores from human evaluation of the same MT output, a number of meta-evaluations are also provided, testing the objectivity of automated evaluation metrics applied to French as target language.

The paper presents the CESTA evaluations campaign, its context and objectives, and the protocol that was used. We first summarize the results of the first campaign (Surcin et al., 2005) and the conclusions we drew from them. Then we describe in more detail the second evaluation campaign and present the results we obtained from the metrics, and their meta-evaluation. The conclusion attempts to sum up the CESTA evaluation campaigns and their contribution to the MT evaluation domain.

2. Context and Objectives

The CESTA project is a three-year campaign that started in January 2003 and will be ending during 2006. It has been funded by the French Ministry of Research and Education within the Technolangue framework (http://www.technolangue.net), and is integrated into the EVALDA evaluation platform.

The objectives of CESTA are manifold. The first is to provide a reliable evaluation protocol for MT systems. The one is to evaluate commercial and academic MT systems. Another important objective is to introduce experimental evaluation metrics (relying on semantics and syntax) and to "meta-evaluate" those metrics by comparing them to human evaluations. The targeted quality with respect to the FEMTI guidelines (Hovy, King & Popescu-Belis, 2002) is thus the principal MT functionality, the quality of the output text as a translation, with its two main aspects: accuracy (fidelity, or, henceforth, adequacy) and fluency (readability).

Two evaluation campaigns were carried out within the CESTA project. The first campaign aimed at drawing up an evaluation protocol in order to evaluate the systems on a general domain, without adapting their terminological resources. For the second campaign, the protocol was adapted and revised according to what we learnt from the first campaign. The evaluation was carried out on a specialized domain, allowing a domain adaptation phase, in order to compare the improvement of systems in translation quality in two cases: with and without terminological enrichment (Mustafa El Hadi et al., 2001, 2002; Babych et al., 2004). The domain has been chosen for the evaluation was the health/medical domain.

3. Evaluation Metrics Used for CESTA

3.1. Automated metrics

Five automated metrics were used in both evaluation campaigns. Three of them are well-known and by now widely accepted (though not without controversy) by the MT community, while the two others remain experimental. One of the goals of CESTA was also to study the reliability of these metrics.

BLEU (Bilingual Evaluation Understudy) is an automated metric first developed by IBM (Papineni et al., 2001), and subsequently tuned by the US National Institute of Standards (NIST). BLEU and its NIST version are based on a statistical comparison of the n-grams found in the candidate translation with one or more reference translations of the same text. The Weighted N-gram Model, or WNM (Babych and Hartley, 2004), is a

[1] CESTA stands for (in French): Campagne d'Évaluation des Systèmes de Traduction Automatique.

combination of BLEU and the Legitimate Translation Variation (LTV) metrics, which ascribes weights to words in the BLEU formulae depending on their frequency (computed using TF.IDF).

The X-Score metric (Rajman and Hartley, 2001) is based on the distribution of linguistic information within a text, such as morpho-syntactic categories, or syntactic relationships. The D-Score (Rajman and Hartley, 2001) measures the preservation of a text's semantic content throughout the translation process.

We also used internally two metrics that are more common in the ASR evaluation, namely mWER (Multi-references Word Error Rate) and mPER (Multi-references Position Independent Word Error Rate), in order to provide further points of comparison for a meta-evaluation of metrics.

3.2. Human Judgments

The CESTA evaluation included human judgment in order to enable the meta-evaluation of the automated metrics by comparing their scores with the ones assigned by humans. The selected criteria for human evaluations are fluency and adequacy, following the DARPA campaigns from the 1990s (White, O'Connell & O'Mara, 1994).

A dedicated web interface was created to input, store and process the human judgments. Each translated segment (typographical sentence) was evaluated by two judges, both for fluency and for adequacy. The segments shown to each judge were assigned at random, with a maximum of a hundred segments per judge to avoid overtiredness.

For *fluency*, the judges were asked to answer for each segment the question "Is this text written in good French?" by giving a score on a 5-point scale, from "native French" to "non understandable". For *adequacy* (fidelity), they were asked to compare the meaning of the evaluated segment to that of a reference translation and score adequacy on a 5-point scale from "whole meaning is present" to "nothing in common".

4. First CESTA Campaign

4.1. Overview

Starting with the first evaluation campaign (Surcin and al., 2005), both English-to-French and Arabic-to- French translation directions were introduced. Five commercial and academic systems registered for the first direction (English-to-French track): Comprendium S.L., RALI (University of Montréal), SDL International, Softissimo and Systran. Two systems participated in the second direction (Arabic-to-French track): CIMOS and Systran.

The texts belonged to the general domain: 15 documents from the *Journal of the European Community* (JOC) for the English-to-French track, and 16 documents from the *UNESCO 32nd General Conference* for the Arabic-to-French track. The two corpora, segmented at the sentence level, contained around 20,000 words each (source language). In order to mask the test corpora before giving the data to the developers of the systems, the test documents were randomly dispersed within masking corpora of some 200,000 words from the same lexical domain. The corpora were UTF-8 encoded and followed the NIST format (NIST, 2003).

For each test corpus four reference translations were available for automated metrics and for meta-evaluation purposes. One was the official translation of each text, produced within the originating organization; this was considered to be the most authoritative translation. Three other translations were done by translation agencies commissioned by the CESTA organizers, and were therefore high quality human translations. These four "reference" translations were used to compute the automatic scores, whereas the human judges were given reference segments that were selected from the authoritative translation by an additional linguist, aiming thus at the highest possible quality.

4.2. Results

Table 1 below presents the fluency and adequacy results obtained thanks to the human evaluators.

Systems	Fluency	Adequacy
System 1-EN	0.459	0.561
System 2-EN	0.419	0.545
System 3-EN	0.353	0.489
System 4-EN	0.511	0.636
System 5-EN	0.503	0.608
System 1-AR	0.198	0.310
System 2-AR	0.083	0.166

Table 1: Human results for campaign #1

The results obtained automatically with measures indicated in Section 3.1 are presented in Table 2, which shows:

- scores obtained with the BLEU/NIST metric, with cumulative 4-grams computed regardless of the upper or lower case of the words (case-insensitive option);
- scores obtained with the WNM (F-measure) metric making use of the best reference translation;
- Pearson's correlation coefficient (last two lines) for each automated measure, with respect to the fluency and adequacy scores obtained by human judges, only for the English-to-French direction however.

Systems	BLEU	NIST	WNM F-measure
System 1-EN	0.438	9.640	0.676
System 2-EN	0.465	9.964	0.667
System 3-EN	0.375	9.022	0.654
System 4-EN	0.450	9.808	0.692
System 5-EN	0.572	11.025	0.691
System 1-AR	0.209	7.422	0.493
System 2-AR	0.086	4.787	0.515
Corr. fluency	0.68	0.71	0.994
Corr. adequacy	0.62	0.65	0.983

Table 2: Automatic results for BLEU, NIST and WNM

Table 3 below presents the automatic results obtained with the experimental automated metrics. The table shows:

- scores obtained with the X-Score;
- scores obtained with the D-Score;
- Pearson's correlation coefficient (last two lines) for these two metrics with respect to fluency and adequacy scores obtained by human assessment, only for the English-to-French direction.

Systems	X-Score	D-Score
System 1-EN	0.407	0.016
System 2-EN	0.394	0.019
System 3-EN	0.391	0.022
System 4-EN	0.418	0.014
System 5-EN	0.420	0.019
System 1-AR	0.383	0.016
System 2-AR	0.391	0.015
Corr. fluency	0.93	-0.81
Corr. adequacy	0.95	-0.82

Table 3: Automatic results for X-Score and D-Score

4.3. Discussion

In order to compare the automated metrics with the judgments produced by humans (meta-evaluation), a measure of the correlation with fluency and adequacy was computed. To abstract from the actual values of the scores, which are known to have little significance as absolute values (especially for BLEU and NIST), only the ranking of the systems was used. Table 4 presents the ranking of the systems according to all of the CESTA metrics, for the English-to-French direction. Each cell contains the ranking of the system indicated by each column. The last column provides the average correlation value with the ranking obtained using fluency and adequacy: these two rankings are the same (Table 4), which illustrates the correlation of these two parameters already observed (White, 2001).

Table 4 shows a strong correlation between the WNM metric (actually, its f-measure) and the human evaluation metrics. In addition, the BLEU and NIST metrics provide rankings that almost fully coincide with the human ones, with two exceptions: (a) system S2 appears to be favored by BLEU/NIST, since its rank increases from fourth to second; (b) the ranks of systems S4 and S5 are swapped. Overall, even if the BLEU/NIST metrics show acceptable correlation with human judgments, these results are clearly lower than those of the other statistical metric, WNM, which appears thus to be more reliable than BLEU/NIST on French target language as well (Babych & Hartley, 2004).

Furthermore, the first experimental metric, the X-Score, appeared to be very promising since it obtained good correlations with the human assessments. However, surprisingly, the D-Score has a strong *inverse* correlation – the explanation for this phenomenon is given elsewhere (Hamon & Rajman, 2006).

Metrics	System					Corr.
	S1	S2	S3	S4	S5	Flu&Ade
BLEU	4	2	5	3	1	0.5
NIST	4	2	5	3	1	0.5
WNM	3	4	5	1	2	1
X-Score	3	4	5	2	1	0.9
D-Score	4	2	1	5	2	-0.81
Fluency	3	4	5	1	2	-
Adequacy	3	4	5	1	2	-

Table 4: System ranking for campaign #1 and correlation of each automated metric with fluency and adequacy

5. Second CESTA Campaign

5.1. Objectives

For the second CESTA evaluation campaign, one of the objectives was to improve the reliability of the metrics and of the evaluation protocol with respect to the first evaluation campaign. Furthermore, another important objective was to observe the impact of domain adaptation, or terminological enrichment, on the overall output quality of the systems, given that the participants were informed that the texts to be translated would belong to the health/medical domain.

5.2. Participants

Five systems participated in the second evaluation campaign for the English-to-French direction, and only one for the Arabic-to-French direction. The participants were not exactly the same as those in the first evaluation campaign. Aachen University of Technology was the only participant in the Arabic-to-French direction, while for the English-to-French direction the systems and institutions were: Comprendium, RALI (University of Montréal), Softissimo, Systran and UPC (Technical University of Catalonia). As with the previous campaign, the results are presented below in an anonymized form, since the confidentiality agreement for the campaign prevents the disclosure of individual performances.

5.3. Data

The English domain-specific source corpus was composed of 16 documents from the bilingual Canadian governmental web site called "Health Canada" (http://www.hc-sc.gc.ca). The Arabic source corpus from the same domain was composed of 30 documents from the multilingual web sites of two international organizations and one NGO: UNICEF (http://www.unicef.org), the World Health Organisation (http://www.who.int) and Family Health International (http://www.fhi.org). Each of the two corpora contained about 20,000 words, corresponding to 917 segments for the English test set and 824 for the Arabic test set.

As in the first evaluation campaign, documents were segmented at the sentence level, and four reference translations were commissioned. The "official" translation originating from the same organization was no longer considered the authoritative translation, as its quality was criticized by a linguist consultant to CESTA. Therefore, one of the three other high quality human translations from translation agencies was used as a reference by human judges. All the three translations were added to the official one for use as references by automated metrics. The test corpora were once more randomly dispersed in a masking corpus about ten times larger.

5.4. Evaluation Protocol

To allow domain adaptation, a training corpus was distributed to the participants beforehand, for each translation direction. The corpus was representative of the test corpus that was finally distributed; therefore the participants could adapt their systems to the health domain before the evaluation took place.

As regards human evaluation, only a part of the entire output of the participating systems was evaluated – about a third of the output of each system – from lack of available judges. There were 4,608 evaluated segments, which were seen by 48 judges, i.e. 96 segments per evaluator. The human evaluation protocol is the same as above, and is described in (Surcin and al., 2005). Contrary to the first evaluation, adequacy and fluency evaluation were done in two separate stages, to avoid potential correlations due to their simultaneous assessment by human judges.

5.5. Results

As Table 1 did for the first evaluation campaign, Table 5 presents the human results obtained in the second evaluation campaign for fluency and adequacy. The human evaluators used the reference translation (the best of the three translation done by translation agencies) for English-to-French and for Arabic-to-French to grade each system's output independently. For meta-evaluation purposes, the official translation was tested as well.

Systems	Fluency	Adequacy
System 1-EN	0.547	0.596
System 2-EN	0.821	0.882
System 3-EN	0.575	0.609
System 4-EN	0.543	0.557
System 5-EN	0.320	0.460
Human-EN	0.888	0.800
System 1-AR	0.015	0.426
Human-AR	0.926	0.628

Table 5: Human results for campaign #2

Table 5 shows that system S2-EN obtains scores that are very close to the human reference, and in particular scores that are even better that those of the human reference for the adequacy score. The precise cause of this result has still to be found, but several hypotheses can be formulated. It is not impossible that the overall quality of the official translation (from the "Health Canada" web site) was judged unsatisfactory by the CESTA French-speaking judges, who preferred more often the output of one system, at least in terms of adequacy, and by comparison with the best agency translation. The official translation could also have been judged too free, for instance. Finally we find out that system S2-EN included some of the "Health Canada" files among the data that was used to set up the system, which could probably increase its score unduly.

As for the overall ranking obtained through human judges, for the English-to-French direction, system 2 clearly stands out as much "better" than the others in terms of output quality, at a level which is comparable to human translations (this surprising fact is yet to be fully analyzed). Three systems – 1, 3 and 4 – generate output of comparable quality, followed at a perceptible distance by system 5. The rankings according to fluency and adequacy are the same.

For the Arabic-to-French direction, even if it is difficult to draw conclusions with only one system, system S1-AR seems very far from the fluency score obtained by the reference translation, but on the contrary is fairly close to their adequacy score, which means that a significant proportion of the content of the segments is judged conveyed by the MT output, but the overall well-formedness and style are perceived as very weak, especially in direct comparison with a human translation alone.

For adequacy, the inter-rater correlation is 0.637, whereas for fluency, the correlation is 0.585. Moreover, the correlation between fluency and adequacy judgments is to 0.386, which is quite a low value that contrasts with the ones obtained in the first campaign. The inter-rater agreement is thus not very good, but this could be due to the health domain, which is more difficult to evaluate: although attention was paid that the judges come from the medical domain, their linguistic appreciation of the translations could have been more biased than that of the judges of the first campaign.

As Table 2 did for the first evaluation campaign, Table 6 below shows the results of the automated BLEU/NIST and WNM metrics and their correlation with fluency and adequacy.

Systems	BLEU	NIST	WNM F-measure
System 1-EN	0.378	9.047	1.435
System 2-EN	0.896	14.395	1.150
System 3-EN	0.384	9.174	1.255
System 4-EN	0.398	9.349	1.198
System 5-EN	0.339	8.454	1.346
System 1-AR	0.423	9.082	1.124
Corr. fluency	0.59	0.55	0.63
Corr. adequacy	0.79	0.75	0.60

Table 6: Automatic results for BLEU, NIST and WNM

The scores of the BLEU and NIST automated metrics are better correlated with adequacy than in the first evaluation campaign (cf. Table 2), but less with fluency. As for the WNM f-measure, its correlation is much lower than in the first campaign, though still at an acceptable level (this surprising result still awaits analysis). System S2-EN again stands out, as with the human evaluation, scoring much better than all the other systems, for the reasons mentioned above related to the data used for set up. Again, there is a group of three systems which are close, and the same system far behind.

As Table 3 did for the first evaluation campaign, Table 7 presents the results for the experimental metrics X-Score and D-Score.

Systems	X-Score	D-Score
System 1-EN	0.387	0.014
System 2-EN	0.352	0.016
System 3-EN	0.377	0.014
System 4-EN	0.349	0.015
System 5-EN	0.360	0.017
System 1-AR	0.362	0.014
Corr. fluency	-0.17	-0.25
Corr. adequacy	-0.22	-0.02

Table 7: Automatic results for X-Score and D-Score

If we take into consideration the first evaluation campaign, the results of the two metrics are very surprising. First the correlations with human judges are very low – but this was also the case for the D-Score in the first campaign. An explanation for the X-Score is that the learning corpus (for fluency) required for evaluation is the same as the one used before. We could certainly reduce this problem using a corpus of the same domain used in the second evaluation campaign. Therefore the X-Score appears to be sensitive to the domain of the corpus, which could be due to the fact that grammatical relations are different from a domain to another.

Beyond this, the scores are consistent for both metrics in contrast to the previous evaluation, automated and human. These results probably indicate that the two experimental metrics studied by CESTA evaluate other quality-related parameters of the translations than the two measured here by the human judges.

5.6. Meta-evaluation

As Table 4 did for the first evaluation campaign, Table 8 presents the ranking of the systems, in order to determine the differences between the systems, as well as the correlation of the automated metrics with the human judgments.

First, we observe that the rankings for both fluency and adequacy are the same, as in the first campaign, even if the evaluation protocol has slightly changed between campaigns – in principle, its accuracy has improved. However, the agreement of these scores with the automated metrics is perceptibly lower than in the first campaign, apart again from the BLEU/NIST metrics, which are even better correlated than before.

Metrics	System					Corr.
	S1	S2	S3	S4	S5	Flu/Ade
BLEU	4	1	3	2	5	0.7
NIST	4	1	3	2	5	0.7
WNM	1	5	4	3	2	-0.7
X-Score	1	4	2	5	3	0.1
D-Score	4	2	5	3	1	-0.4
Fluency	3	1	2	4	5	-
Adequacy	3	1	2	4	5	-

Table 8: System ranking for campaign #2

These results show again that the second system is clearly above the other systems and the fifth system clearly below, while the three other systems are close.

5.7. Impact of Domain Adaptation on Quality

Four systems participated in the two evaluation campaigns. Tables 9 and 10 attempt to show the impact of domain adaptation (or terminological enrichment) on the quality of MT output. These tables show the following results, separately for human and automated evaluations (the systems are designated using letters since the numbers are not the same in the two campaigns):

- the results of the four same systems in both evaluations;
- fluency and adequacy scores for evaluation campaign #1;
- fluency and adequacy scores for evaluation campaign #2.

Systems	Flu. #1	Flu. #2	Ade. #1	Ade. #2
System A-EN	0.459	0.547	0.561	0.596
System B-EN	0.419	0.821	0.545	0.882
System C-EN	0.511	0.575	0.636	0.609
System D-EN	0.503	0.543	0.608	0.557

Table 9: Impact of domain adaptation – human evaluation

Systems	BLEU #1	BLEU #2	X-Score #1	X-Score #2
System A-EN	0.438	0.378	0.407	0.387
System B-EN	0.465	0.896	0.394	0.352
System C-EN	0.450	0.384	0.418	0.377
System D-EN	0.572	0.398	0.420	0.349

Table 10: Impact of domain adaptation –
automated evaluation

The scores of the human judges clearly show an improvement in the overall output quality. However, the scores of the automated metrics are quite the opposite, as they decrease in most cases (except for system S2). The explanation lies probably in the different nature of the corpora that were used for testing, a difference in genre that does not allow for comparisons between campaigns (cf. Babych et al., 2005).

6. Conclusions and Prospects

The numerous figures resulting from the second evaluation have to be still analyzed, but the first analyses provided above already look promising and prompted much discussion.

A protocol to evaluate MT systems has been established by CESTA and can be reused easily to perform other MT evaluations. Furthermore two new metrics are now available, though these are still experimental and need to be studied in depth before full validation. For instance, one objective is to establish a fluency corpus for the medical domain and the X-Score.

To complete the second evaluation campaign, a third round of tests should be carried out using versions of the MT systems before domain adaptation on the data from the second campaign, so that the real merits of domain adaptation can be assessed. This prospect, however, extends beyond the limits of the CESTA campaign.

The results will be discussed within an upcoming workshop and the conclusions will be published in a final report. All the data (corpora, reference translations, results, etc.) will be released by ELDA as an evaluation package, available for public use. In the long term, we would like to introduce an MT evaluation platform which will allow MT systems to compute their own evaluations.

Acknowledgements

We would like to thank all the participants in both CESTA campaigns. We are grateful to Philippe Langlais of RALI for additional insights. We are also very grateful to all the human evaluators for their work.

References

Babych B., Hartley A. (2004). Extending the BLEU MT Evaluation Method with Frequency Weightings. In *ACL 2004 Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, July 2004, pp. 622-629.

Babych B., Hartley A., Elliott, D. (2005). Estimating the predictive power of n-gram evaluation metrics across languages and text types. In *Proceedings of MT Summit X*, pp. 12-16 September, 2005, Phuket, Thailand, pp 412-418.

Dabbadie, M**.,** Mustafa El Hadi W., Timimi, I. (2004). CESTA: The European MT Evaluation Campaign. In *Multilingual Computing & Technology*, Vol. 15, issue 5, p. 10-12, Sandpoint, Idaho, 2004.

Hamon, O., Rajman, M. (2006). X-Score: Automatic Evaluation of Machine Translation Grammaticality. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genova, Italy, May 2006, in press.

Hovy E., King M., Popescu-Belis A. (2002). Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, vol. 17, n. 1, p.43-75.

ISO 1999. Standard ISO/IEC 9126. Part 1: Information Technology – Software Engineering – Quality characteristics and sub-characteristics. Software Quality Characteristics and Metrics. Part 2: Information Technology – Software Engineering – Software Products Quality: External Metrics.

Mustafa El Hadi W., Timimi I., Dabbadie M. (2001). Setting a Methodology for Machine Translation Evaluation. In *Proceedings of the 4th ISLE Workshop on MT Evaluation*, MT Summit VIII, Santiago de Compostela, September 2001, pp. 49-54.

Mustafa El Hadi W., Timimi I., Dabbadie M. (2002). Terminological Enrichment for non-Interactive MT Evaluation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas de Gran Canarias, May 2002, pp. 1878-1884.

Mustafa El Hadi W., Dabbadie M., Timimi I., Rajman M., Langlais P., Hartley A., Popescu-Belis A. (2004). CESTA – Machine Translation Evaluation Campaign. In *Proceedings of the LR4Trans Workshop of the 20th International Conference on Computational Linguistics*, COLING'2004, Geneva, August 2004, pp. 8-17.

NIST, 2003. The 2004 NIST Machine Translation Evaluation Plan (MT-04), v2.1. http://www.nist.gov/speech/tests/mt.

Papineni K., Roukos S., Ward T. and Zhu W.-J. (2001). Bleu: a Method for Automatic Evaluation of Machine Translation, IBM Research Report RC22176 (W0109-022). In *Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics (ACL)*, Philadelphia, July 2002, pp. 311-318.

Rajman M., Hartley A. (2001). Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores. In *Proceedings of the Fourth Workshop on MT Evaluation*, MT Summit VIII, Santiago de Compostela, September 2001, pp. 29-34.

Schmid H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, 1994.

Surcin S., Hamon O. , Hartley A., Rajman M., Popescu-Belis A., Mustafa El Hadi W., Timimi I., Dabbadie M., Choukri K. (2005). Evaluation of Machine Translation with Predictive Metrics beyond BLEU/NIST: CESTA Evaluation Campaign #1. In *Proceedings of MT Summit X*, pp. 12-16 September, 2005, Phuket, Thailand, pp 117-124.

Thompson H., Brew C. (1994). Automatic Evaluation of Computer Generated Text. In *Proceedings of the ARPA/ISTO Workshop on Human Language Technology*, 1994, pp. 104-109.

White, J. S. (2001). Predicting Intelligibility from Fidelity in MT Evaluation'. In *Proceedings of the MT Summit VIII Workshop on MT Evaluation "Who did what to whom?"*, Santiago de Compostela, Spain, pp. 35-38.

White, J.S., T. O'Connell, F. O'Mara. (1994). The DARPA Machine Translation Evaluation Methodologies: Evolution, Lessons and Future Approaches. In *Proceedings of the first Conference of the Association for Machine Translation in the Americas*. Columbia, USA.

Zhang Y., Vogel S., Waibel A. (2004). Interpreting BLEU/NIST Scores: How Much Improvement? Do We Need to Have a Better System? In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, May 2004, pp. 2051-2054.