# DELIVERABLE

| | |
|---|---|
| **Project Acronym:** | **PLuTO** |
| **Grant Agreement number:** | **250416** |
| **Project Title:** | **Patent Language Translations Online** |

# Deliverable 7.6 First Report on the Intrinsic and Extrinsic Quality of MT

**Authors:**

John Tinsley (DCU)
Joeri Van de Walle (CL)
Heidi Depraetere (CL)

| | | |
|---|---|---|
| Project co-funded by the European Commission within the ICT Policy Support Programme | | |
| Dissemination Level | | |
| P | Public | |
| C | Confidential, only for members of the consortium and the Commission Services | **x** |

# REVISION HISTORY AND STATEMENT OF ORIGINALITY

## *Revision History*

| Revision | Date | Author | Organisation | Description |
|---|---|---|---|---|
| 1 | 25/08/11 | J. Tinsley | DCU | Document creation |
| 2 | 07/09/11 | J. Van de Walle | CL | First draft |
| 3 | 08/09/11 | H. Depraetere | CL | Edited first draft |
| 4 | 13/10/11 | J. Van de Walle | CL | Incorporated status meeting feedback |
| 5 | 14/10/11 | J. Tinsley | DCU | Copy editing and formatting |
| | | | | |
| | | | | |
| | | | | |

**Statement of originality:**

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

# Table of Contents

# Executive Abstract

This deliverable provides a range of evaluation data detailing the performance of the English—Portuguese and English—French machine translation (MT) systems submitted as Deliverable 5.1. In addition to assessing the MT systems using automatic evaluation metrics such as BLEU and METEOR, a large-scale human evaluation is also carried out. MT system output is ranked from 1—5 based on the overall quality of translation, and the individual mistakes made are identified and classified in an error categorisation task.

On top of this standalone evaluation, the PLuTO MT systems are also benchmarked against leading commercial systems across two MT paradigms: Google Translator for statistical MT and Systran (Enterprise) for rule-based MT. A comparative analysis is carried out using both the automatic and human evaluation techniques described above.

All evaluations are carried out using held-out test data randomly selected from our parallel patent corpora. For the automatic evaluations, test sets were segmented into sub-sets based on the IPC patent classification system. In doing this, the evaluation would indicate in which categories of patents (e.g. chemistry, engineering, etc.) the translation systems were performing better.

Both automatic and human evaluations have shown that the PLuTO engines produce translations of a reasonable to good quality. The output of the PLuTO engines was preferred by all evaluators for all language pairs over that of Google Translate and Systran.

Further analysis revealed that there are quality differences across languages and IPC domains. These differences need to be explored further to identify areas that will allow us to improve translation quality further.

# 1. Evaluation Setup

## *1.1. Methodology*

MT quality is often being evaluated by means of automated metrics such as BLEU, METEOR, NIST and TER. Whereas these metrics are most certainly useful to help developers of MT systems measure the progress they are making, they do not necessarily stand in any direct correspondence to potential user satisfaction. To overcome this shortcoming of automated scoring we have opted for a mix of automatic and human evaluations in this project.

The focus of the automatically generated scores will be on measuring progress or degradation between the various versions of engines that will be built for each language pair, and that of the human evaluation will be on measuring the absolute translation quality and usability of the best performing engine.

At the same time, the availability of both automatically generated and human acquired scores will allow us to see if and how both scores correlate.

Figure 1 below provides an overview of all evaluations that have been performed to measure the quality of the machine translation output.

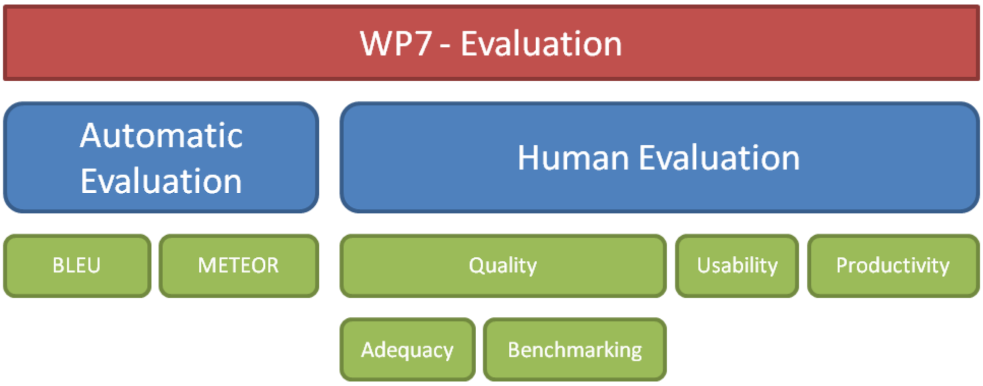


**Figure 1: Evaluation Overview**

## *1.2. Test Sets*

The test sets formed the basis for the evaluations. For the *automatic* evaluations, different sets were created for each of the eight main IPC classes A—H. For each IPC class a random sample of one thousand sentences was selected from the acquired training data, resulting in 8000 sentences to be evaluated in total. Sentences selected as part of the test sets were removed from the training data so as not to bias the evaluation results. The motivation behind splitting the data in this way was to evaluate the MT performance on the respective technical domains of patents, as well as to facilitate the domain adaptation experiments reported in Deliverable 5.1 at Month 12.

For the *human* evaluations, the size of the test sets was reduced to one hundred sentences per IPC class, resulting in a total of 800 sentences per language pair to be evaluated by the human evaluators. As the average segment length of the sentences in the set was on average about 30 words per segment, this reduction was necessary to keep the human evaluation within acceptable limits in terms of cost and time needed to complete them. Distribution of sentence length was taken into account when reducing the evaluation sets from 1000 sentences to 100.

For the productivity evaluation, which involves post-editing of PLuTO output and translation from scratch, the set of 800 sentences was further reduced to 400 sentences, again taking into account sentence length and IPC class.

## *1.3. Automatic Evaluation*

For the automatic evaluations, all segments in the test sets were translated with the trained engines. The resulting translations were then compared to the corresponding human reference translations using two different metrics: BLEU and METEOR[2]. This resulted in a score between 0 and 100 for each metric, with scores getting higher as translations were more similar to the reference translations.

## *1.4. Human Evaluation*

Three different types of human evaluations were carried out, each of which focused on a different aspect of the translation.

With the *quality evaluation* the focus was on the linguistic quality of the translations and how they compared to the output of other machine translation systems. With these evaluations we try to answer the question ‘how good is the translation?’.

With the *usability evaluation* the focus was on the utility of the translations to end users. With this evaluation we try to answer the question ‘how useful is the translation?’.

With the *productivity evaluation* the focus was on another utility aspect of the translations. With this evaluation we investigate in how far automated translation can speed up human translation by serving as a draft translation that the human translator can base his translation on. The question we are trying to answer with this type of evaluation is ‘is post-editing machine translation output quicker than translating from scratch and if so, how much quicker?’.

### 1.4.1. Quality evaluation

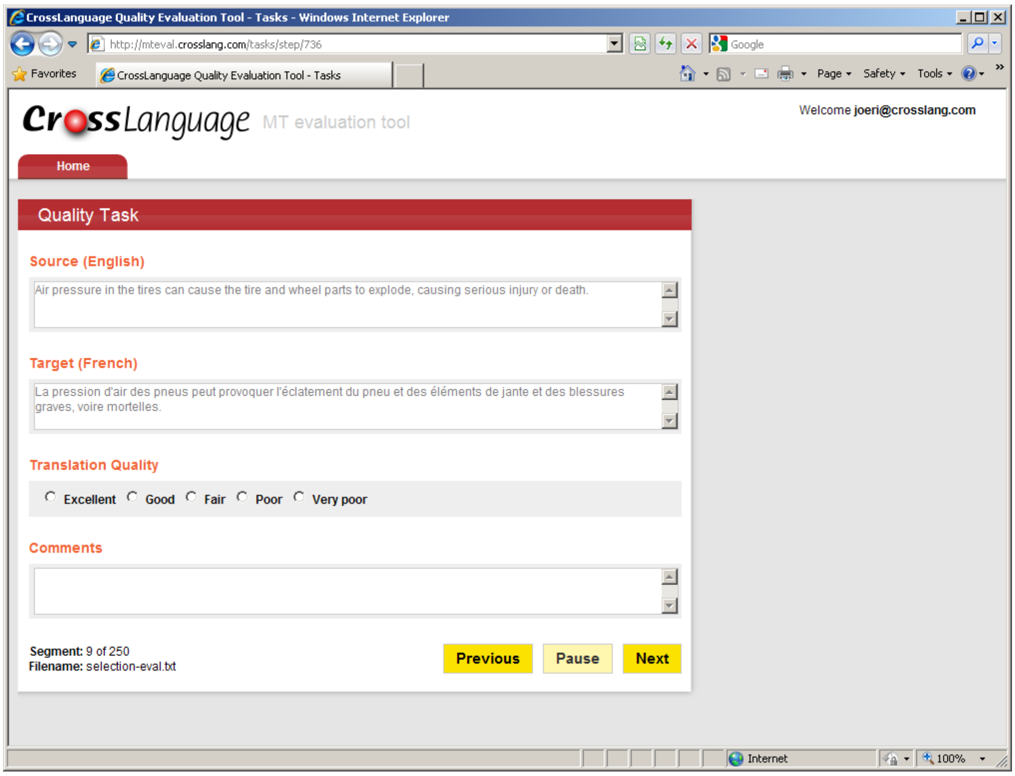
For the quality evaluations, evaluators were asked to carry out two evaluation tasks:

- an adequacy evaluation, and
- a ranking evaluation

For the *adequacy evaluation*, users were asked to evaluate the translation quality of each individual translated sentence in the human evaluation set by giving it a score from 1 (Very Poor) to 5 (Excellent). To help evaluators be as consistent as possible in their judgements, guidelines were provided as to how to decide on the appropriate value for a translation. See Appendix B for more detailed information on these guidelines.

The adequacy evaluation was only performed on the PLuTO output.

[2] Refer to Appendix A for a description of these metrics.


CrossLanguage Quality Evaluation Tool - Tasks - Windows Internet Explorer
http://mteval.crosslang.com/tasks/step/736
Google
Favorites
CrossLanguage Quality Evaluation Tool - Tasks
Page
Safety
Tools
CrossLanguage MT evaluation tool
Welcome joeri@crosslang.com
Home
Previous
Pause
Next
Internet
100%

Quality Task

Source (English)

Air pressure in the tires can cause the tire and wheel parts to explode, causing serious injury or death.

Target (French)

La pression d'air des pneus peut provoquer l'éclatement du pneu et des éléments de jante et des blessures graves, voire mortelles.

Translation Quality

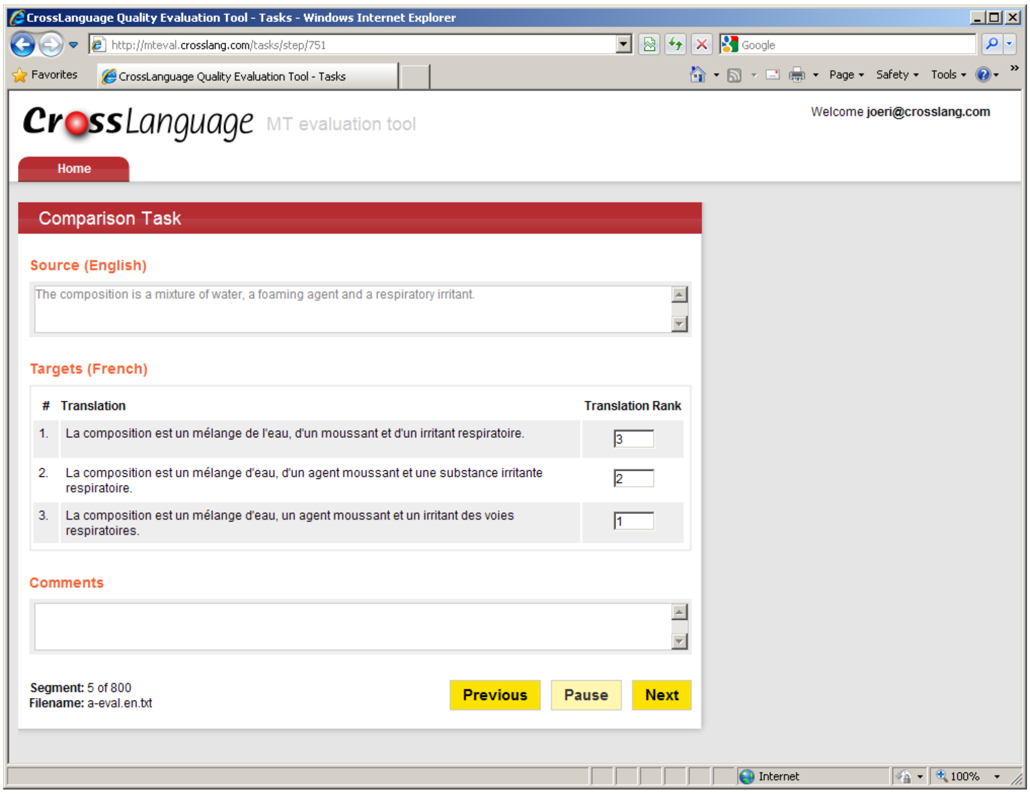
○ Excellent ○ Good ○ Fair ○ Poor ○ Very poor

Comments

**Segment:** 9 of 250
**Filename:** selection-eval.txt

Figure 2: Adequacy Evaluation Interface

For the *ranking evaluation*, users were asked to compare the output of three different machine translation systems: output of the PLuTO trained system, that of Google Translate, and that of a non-customised Systran system. For each sentence in the evaluation set the original sentence was shown with the corresponding translations generated by PLuTO, Google, and Systran. For each sentence the evaluators were asked to rank the translations in the order of perceived quality.


CrossLanguage Quality Evaluation Tool - Tasks - Windows Internet Explorer
http://mteval.crosslang.com/tasks/step/751
Google
Favorites
CrossLanguage Quality Evaluation Tool - Tasks
Page
Safety
Tools
CrossLanguage MT evaluation tool
Welcome joeri@crosslang.com
Home
Previous
Pause
Next
Internet
100%

Comparison Task

Source (English)

The composition is a mixture of water, a foaming agent and a respiratory irritant.

Targets (French)

| # | Translation | Translation Rank |
|---|---|---|
| 1. | La composition est un mélange de l'eau, d'un moussant et d'un irritant respiratoire. | 3 |
| 2. | La composition est un mélange d'eau, d'un agent moussant et une substance irritante respiratoire. | 2 |
| 3. | La composition est un mélange d'eau, un agent moussant et un irritant des voies respiratoires. | 1 |

Comments

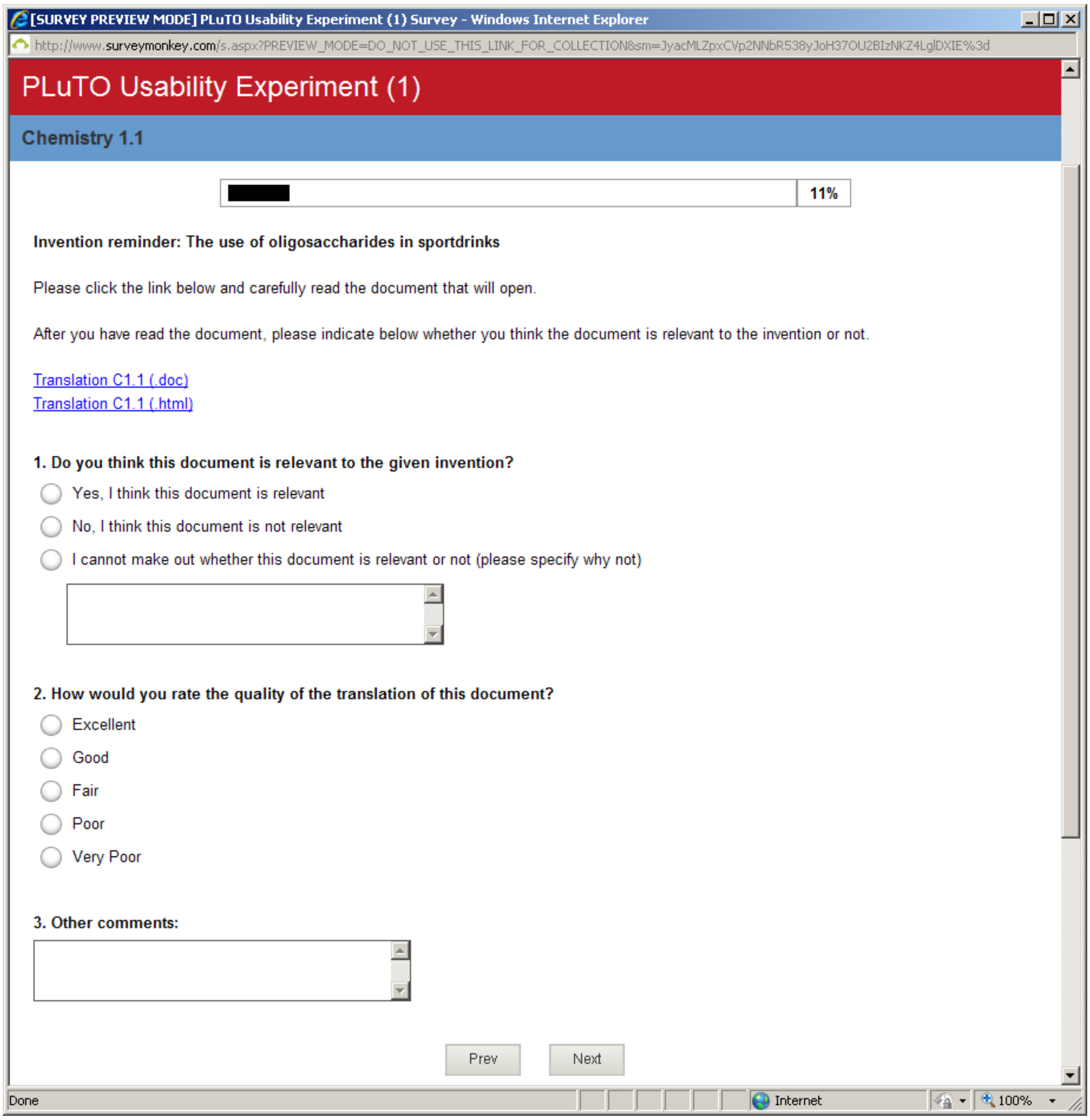
**Segment:** 5 of 800
**Filename:** a-eval.en.txt

Figure 3: Ranking Evaluation Interface

### 1.4.2. Usability evaluation

The usability evaluation was conducted as an online survey. The survey consisted of a test we have called the *usability experiment*, which we hoped would allow us to assess how useful the PLuTO service actually is. For the usability experiment, we put our self in the position of the patent information specialist that is confronted with an invention and has to search his database(s) for relevant prior art. That search will typically also return a number of documents that are written in a language that is not that of the searcher. In our experiment we presented the informant with 10 machine translated documents produced by the PLuTO engines and asked him to determine whether or not they were relevant to a given invention or not.

In total, the experiment included four inventions, two in the field of chemistry and two in the field of mechanics and engineering. At the start of the experiment, informants were asked to indicate their field of expertise based on which they were directed to the appropriate inventions. The assumption was that the degree in which informants would be able to classify the machine translated documents correctly as relevant or non-relevant would be an indication of the actual practical use of the service.



[SURVEY PREVIEW MODE] PLuTO Usability Experiment (1) Survey - Windows Internet Explorer

http://www.surveymonkey.com/s.aspx?PREVIEW_MODE=DO_NOT_USE_THIS_LINK_FOR_COLLECTION&sm=JyacMLZpxCVp2NNbR538yJoH37OU2BIzNKZ4LglDXIE%3d

PLuTO Usability Experiment (1)

Chemistry 1.1

11%

**Invention reminder: The use of oligosaccharides in sportdrinks**

Please click the link below and carefully read the document that will open.

After you have read the document, please indicate below whether you think the document is relevant to the invention or not.

Translation C1.1 (.doc)
Translation C1.1 (.html)

**1. Do you think this document is relevant to the given invention?**

○ Yes, I think this document is relevant
○ No, I think this document is not relevant
○ I cannot make out whether this document is relevant or not (please specify why not)

**2. How would you rate the quality of the translation of this document?**

○ Excellent
○ Good
○ Fair
○ Poor
○ Very Poor

**3. Other comments:**

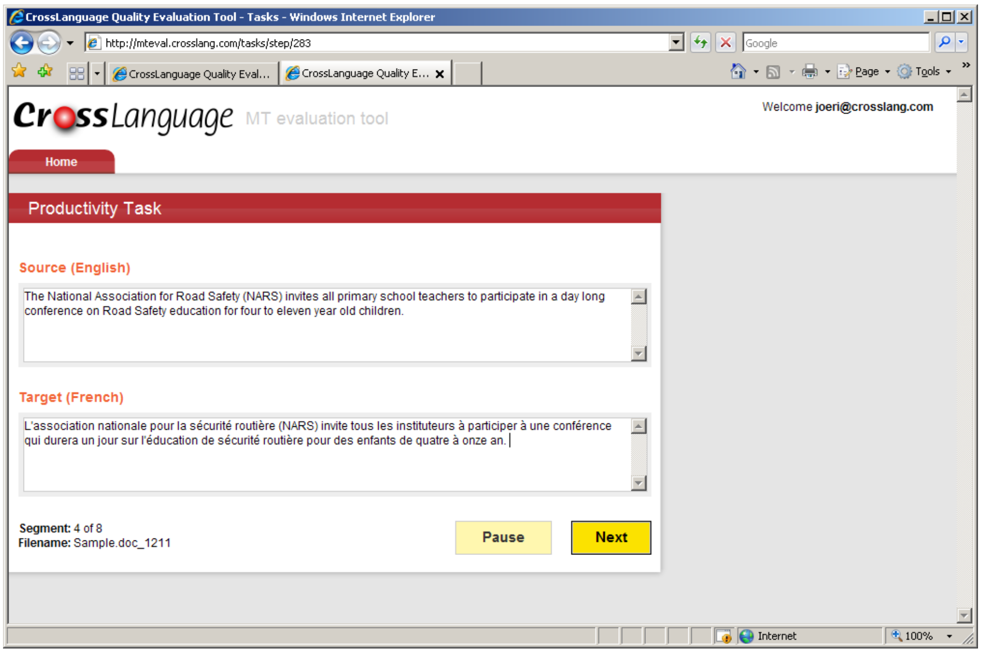
Figure 4: Usability Evaluation Interface

Of the 84 respondents that took part in the survey described in deliverable D7.2, “First report on survey’s results”, 26 indicated that they would be willing to participate in the usability experiment.

The usability evaluation was only performed on the PLuTO output for translation into English.

### 1.4.3. Productivity evaluation

With the productivity evaluation we tried to assess in how far productivity increases might be obtained by using automated translation as an aid to increase the speed of human translation. To evaluate this, evaluators were give a mix of segments pre-translated with the PLuTO system and segments without translation. Evaluators were asked to correct the translation output for those segments that had been pre-translated, and to create a translation from scratch for those segments that had not been processed by the PLuTO system.


CrossLanguage Quality Evaluation Tool - Tasks - Windows Internet Explorer
http://mteval.crosslang.com/tasks/step/283
Google
CrossLanguage Quality Eval...
CrossLanguage Quality E...
Page
Tools
CrossLanguage MT evaluation tool
Welcome joeri@crosslang.com
Home
Internet
100%

Productivity Task

Source (English)

The National Association for Road Safety (NARS) invites all primary school teachers to participate in a day long conference on Road Safety education for four to eleven year old children.

Target (French)

L'association nationale pour la sécurité routière (NARS) invite tous les instituteurs à participer à une conférence qui durera un jour sur l'éducation de sécurité routière pour des enfants de quatre à onze an.

**Segment:** 4 of 8
**Filename:** Sample.doc_1211

Pause Next

Figure 5: Productivity Evaluation Interface

In the background, the time the evaluators spent on editing the translation output or translating the segment was recorded. Recorded times then allowed us to calculate and compare the average throughput for segments in each of the categories.

The productivity evaluation was only performed on the PLuTO output for translation into English.

## *1.5. Evaluators*

To rule out subjectivity for as much as possible within the available budget, each evaluation was performed by three different evaluators.

All evaluators participating in the quality and productivity tests had experience with both patent content and machine translation. Evaluators were always native speakers of the target language with an excellent knowledge of the source language.

Evaluators for the usability evaluation were experienced patent information specialists with a good to excellent knowledge of the target language. Evaluators for the usability evaluation were members of the WON user group that volunteered for the evaluation.

# 2. English—Portuguese MT System

## 2.1. *Automatic Evaluation*

With the automatic evaluation, domain-specific test sets were used to calculate the translation scores. The overall score was obtained by taking the average of the domain-specific scores.

Table 1 below shows the automatic scores obtained for translations from English into Portuguese:

| | **PLuTO** | **Google** | **Systran** |
|---|---|---|---|
| **All** | 37.84 / 40.94 | 18.64 / 22.12 | 15.36 / 18.79 |
| | | | |
| **A (Human necessities)** | 32.42 / 35.29 | 17.71 / 20.96 | 13.42 / 16.40 |
| **B (Operations)** | 38.59 / 41.70 | 18.39 / 21.92 | 15.07 / 18.83 |
| **C (Chemistry)** | 26.26 / 28.78 | 14.54 / 16.66 | 11.1 / 13.09 |
| **D (Textiles)** | 39.42 / 42.67 | 20.24 / 24.10 | 16.8 / 20.58 |
| **E (Fixed constructions)** | 38.74 / 41.73 | 18.5 / 21.87 | 16.21 / 19.59 |
| **F (Mechanical engineering)** | 42.21 / 45.21 | 19.25 / 22.71 | 16.7 / 20.20 |
| **G (Physics)** | 44.51 / 47.79 | 20.6 / 24.38 | 16.41 / 20.12 |
| **H (Electricity)** | 40.40 / 43.87 | 20.17 / 24.37 | 17.56 / 21.60 |

**Table 1: Automatic evaluation scores English → Portuguese[3]**

Observations:

- Overall, the METEOR scores align quite well with the BLEU scores (with a few minor exceptions for Google and Systran).
- The BLEU scores vary between 11.1 (Systran, Chemistry) and 44.51 (PLuTO, Physics).
- PLuTO seems to outperform Google and Systran in all domains. Google comes second with about half the score of the PLuTO engines and Systran comes last, slightly below Google.
- The PLuTO engine's performance varies according to the domain it is used in. It seems to perform best with content in the Physics domain (IPC G; 44.51 BLEU/47.79 METEOR) and worst with content in the Chemistry domain (IPC C; 26.26 BLEU/28.78 METEOR).

Table 2 below shows the automatic scores obtained for translations from Portuguese into English:

| | **PLuTO** | **Google** | **Systran** |
|---|---|---|---|
| **All** | 42.55 / 56.14 | 25.29 / 49.98 | 12.93 / 37.93 |
| | | | |
| **A (Human necessities)** | 37.14 / 50.91 | 24.16 / 43.74 | 12.03 / 34.03 |
| **B (Operations)** | 44.32 / 58.26 | 24.65 / 48.18 | 12.59 / 39.94 |
| **C (Chemistry)** | 29.75 / 40.41 | 20.56 / 34.40 | 9.56 / 24.98 |
| **D (Textiles)** | 45.09 / 58.87 | 27.24 / 49.35 | 13.4 / 39.30 |
| **E (Fixed constructions)** | 43.38 / 58.25 | 24.67 / 48.72 | 13.63 / 40.31 |
| **F (Mechanical engineering)** | 47.57 / 60.46 | 25.86 / 49.58 | 13.7 / 40.80 |
| **G (Physics)** | 48.44 / 61.01 | 27.64 / 49.91 | 14.32 / 40.23 |
| **H (Electricity)** | 45.3 / 59.86 | 28.23 / 51.13 | 14.34 / 42.71 |

**Table 2: Automatic evaluation scores Portuguese → English[4]**

[3] Scores are BLEU / METEOR.

[4] Scores are BLEU / METEOR.

Observations:

- The METEOR scores align quite well with the BLEU scores for PLuTO and Google, but METEOR scores are considerably higher compared to BLEU scores for Systran.
- The BLEU scores vary between 9.56 (Systran, Chemistry) and 48.44 (PLuTO, Physics).
- PLuTO seems to outperform Google and Systran in all domains. Google comes second but the difference with the PLuTO engines is smaller than for the reverse language pair. Google widens the gap with Systran for this language pair.
- The PLuTO engine's performance varies according to the domain it is used in. It seems to perform best with content in the Physics domain (IPC G; 48.44 BLEU/61.01 METEOR) and worst with content in the Chemistry domain (IPC C; 26.26 BLEU/29.75 METEOR).

## *2.2. Human Evaluation*

### 2.2.1. Adequacy

Figure 6 below shows the adequacy scores obtained for translations from English into Portuguese:

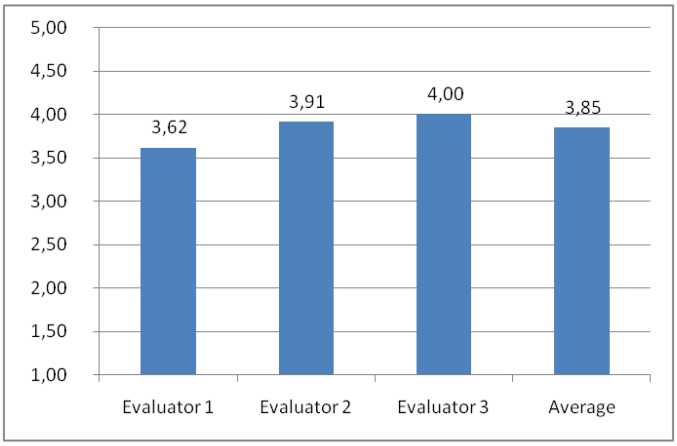


Figure 6: Human adequacy evaluation scores English → Portuguese

Observations:

- The different evaluators seem to have different opinions on the quality of the PLuTO output of the English into Portuguese engine. The scores differ as much as 0.38 (Evaluator 1: 3.62 vs. Evaluator 3: 4.00).
- Average score for the English into Portuguese language pair is 3.85, which is fairly high.

Figure 7 below shows the adequacy scores obtained for translations from Portuguese into English:

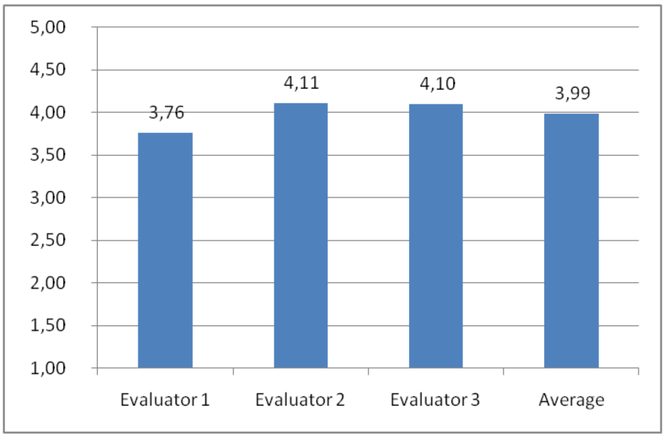


Figure 7: Human adequacy evaluation scores Portuguese → English

Observations:

- The different evaluators seem to have different opinions on the quality of the PLuTO output of the Portuguese into English engine. The scores differ as much as 0.35 (Evaluator 1: 3.76 vs. Evaluator 2: 4.11).
- Average score for the Portuguese into English language pair is 3.99, which is high.

## 2.2.2. Benchmarking

Figure 8 below shows how evaluators have ranked the PLuTO English into Portuguese output in comparison with the Google Translate and Systran output. Rank 1 indicates the number of times on the total amount of evaluated segments a segment was selected as being the best one. Rank 2 indicates the number of times it was chosen as second best and rank 3 indicates the number of times it was seen as the worst one.

In case of equal quality, evaluators were instructed to give the same rank. For instance, in case PLuTO and Google did equally well but better than Systran, ranks given would be 1 for PLuTO, 1 for Google, and 2 for Systran.

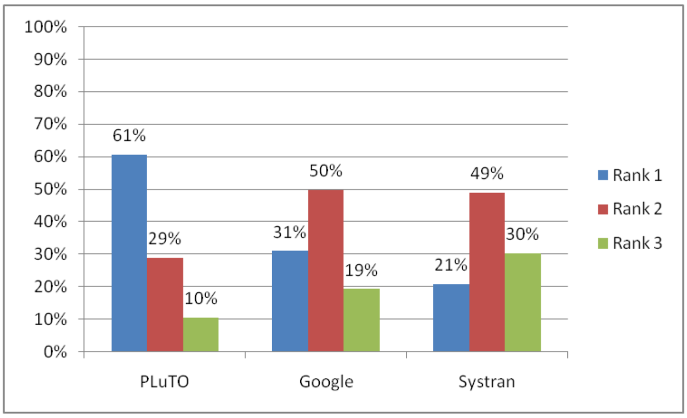


Figure 8: Human benchmarking evaluation English → Portuguese

Observations:

- Evaluators clearly seem to have a preference for the PLuTO output. It was selected as the best performing engine in 61% percent of the cases.
- Google and Systran output are close with a slight preference for Google.
- These results confirm the findings of the automatic evaluation.

Figure 9 below shows how evaluators have ranked the PLuTO Portuguese into English output in comparison with the Google Translate and Systran output:

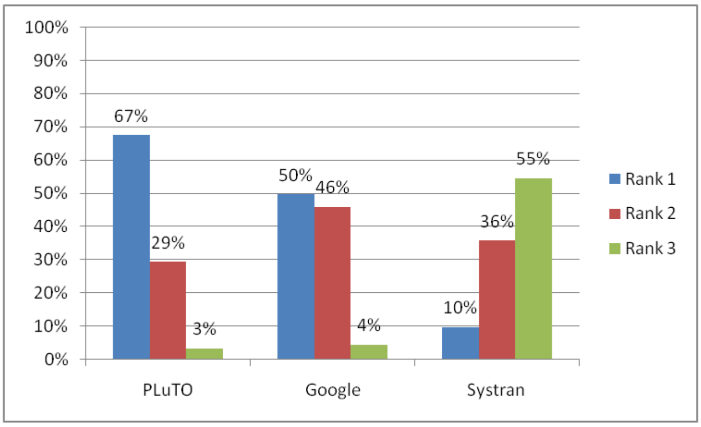


Figure 9: Human benchmarking evaluation Portuguese → English

Observations:

- Evaluators seem to have a preference for the PLuTO output, but the preference is not as outspoken as for the reverse language pair.
- Google, in second place, clearly beats Systran for this language pair.
- Again, these results seem to confirm the results of the automatic evaluation.

### 2.2.3. Error Analysis

One of the three evaluators per language pair that took the adequacy evaluation was also asked to categorise errors found in the PLuTO MT output.

Figure 10 below shows how errors were classified for the English into Portuguese language pair:

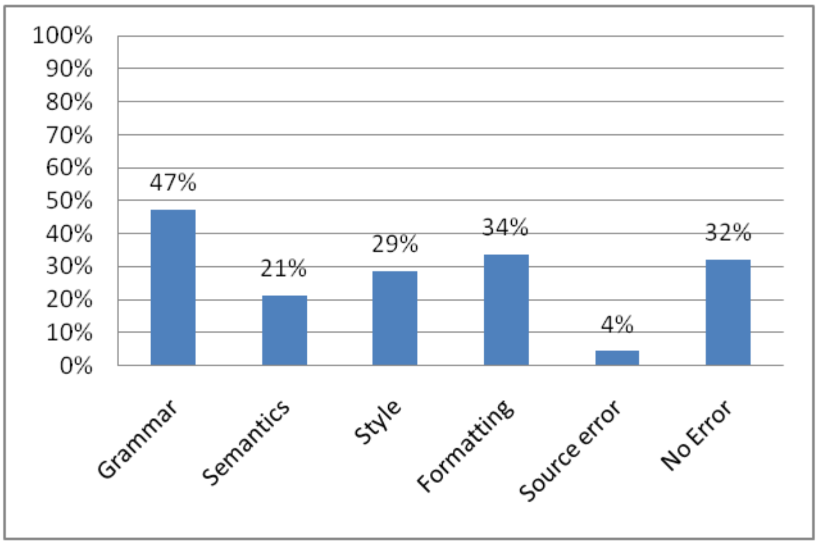


**Figure 10: Error classification English → Portuguese**

Observations:

- 32% of all segments in the evaluation set had no errors at all.
- The most common types of issues encountered in the output are grammatical issues (47%) and formatting issues (34%). See the examples below for more details.
- 4% of all segments in the evaluation set seem to have problems in the source text already.

Table 3 below shows a few examples of the most common problems found in the MT output. The sample segments have a least the error that is mentioned in the Error Type column, but may include errors of another nature, too.

| Source Segment | MT Target | Error Type |
|---|---|---|
| 1) , 49. Another strategy has modified the peptide backbone of GRF by the incorporation of peptide bond isoteres in the N-terminal region. | 1) 49. Outra estratégia modificou o esqueleto peptídico de GRF através da incorporação de isoteros de péptidos ligados na região de terminal-n. | Grammatical error |
| Transcription and translation of the DNA sequence under control of the regulatory sequences causes expression of the tachyplesin sequence at levels which provide an anti-fungal amount of the peptide in the tissues of the plant which are normally infected by fungal pathogens. | A transcrição e tradução da sequência de ADN, sob o controlo das sequências regulatórias originam a expressão da sequência de taquiplesina a níveis que proporcionam uma quantidade de péptido anti -fúngico nos tecidos da planta que são normalmente infectados por fungos patogénicos. | Grammatical error |

| **Source Segment** | **MT Target** | **Error Type** |
|---|---|---|
| The constant meshed engagement between the elongated gears 154, 156 and the underlying gear trains (see Fig. 6) assures that servo-sensor 270 continuously monitors the position of backstop 178 and pusher element 172. | O engate engrenado constante entre as engrenagens alongadas 154, 156 e a subj acente trens de engrenagem (ver Fig. 6) assegura que servo-sensor 270 monitoriza continuamente a posição do batente (178 e o elemento propulsor 172. | Formatting error |
| Furthermore, the novel compounds of the formula III may be prepared by the following process, as summarized in Synthetic Charts V to VII, wherein P1, P2, P3, P4, P5, P6, P7, P8, Pa, Pb, Pc and Pd are protective groups, Ra is lower alkyl and Rb and Rc are the same as above. | Além disso, os novos compostos da fórmula III podem ser preparados pelo seguinte processo, tal como resumido em synthetic esquemas v a vii, em que P1, P2, P3, P4, P5, P6, P7, P8, Pa, pb, PC e Pd são grupos de protecção, ra é alquilo inferior e Rb e rc são os mesmos que acima. | Formatting error |
| Additional amounts of 4-N,N-dimethylaminopyridine (809 mg) and acetic anhydride (1.25 ml) were added to the reaction mixture and stirring was continued overnight. | Quantidades adicionais de 4-N, n-dimethylaminopyridine (809 mg) e anidrido acético (1.25 ml) foram adicionados à mistura da reacção e a agitação foi continuada durante a noite. | Source error |
| Lung diffusion capacity was determined by the carbon monoxide single breath method described in Forster et al, J. | Capacidade de difusão do pulmão foi determinada pelo método do sopro único de monóxido de carbono descrito em forster et ai, j. | Source error |

Table 3: Error type examples for English → Portuguese

Figure 11 below shows how errors were classified for the Portuguese into English language pair:

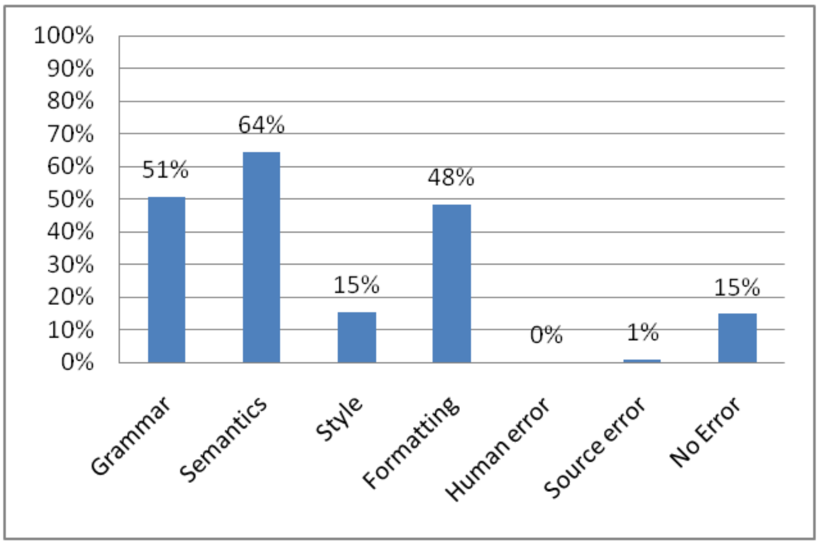


Figure 11: Error classification Portuguese → English

Observations:

- Only 15% of all segments in the evaluation set had no errors at all.
- The most common types of issues encountered in the output are semantic issues (64%), i.e. issues relating to the use of terminology, and grammatical issues (51%). See the examples below for more details.
- Formatting issues, i.e. issues relating to capitalisation, punctuation, date and number formats, etc. are also quite common (48%).

Table 4 below shows a few examples of the most common problems found in the MT output. The sample segments have a least the error that is mentioned in the Error Type column, but may include errors of another nature, too.

| Source Segment | MT Target | Error Type |
|---|---|---|
| Os pós assim obtidos foram misturados com silica gel de grão fino (Aerosil*, 200 g) e introduzidos em cápsulas de gelatina dura N° 3 (100) para dar cápsulas entéricas que contêm 0,5 mg da 1 3,1 4-di-hidro-15-ceto16,1 6-difluoro-20 etil-PGE2 por cápsula. | The powders thus obtained were mixed with fine-grain silica gel (Aerosil *, 200 g) and filled into hard gelatine capsules no. 3 (100) to give enteric capsules which contain 0,5 mg 1 3,1 4-di-hidro-15-ceto16,1 6-difluoro-20 etil-PGE2 per capsule. | Semantic error |
| Esta mistura pode ser constituída por vários conjugados de pesos moleculares diferentes. | This mixture may comprise various conjugates of different molecular weights. | Semantic error |
| Duas amostras da mistura da reacção conjugada (uma representando uma amostra di alisada com H2O da outra) foram armazenadas a 3-8 C até utilização. | Two conjugate reaction mixture samples (one representing a with H2O dialyzed sample of the other) were stored at 3-8 C until use. | Grammatical error |
| O tempo particular em que cada fármaco deve ser libertado varia significativamente com cada fármaco e depende da sua farmacocinética especifica. | The particular time in which each drug should be released vary significantly with each drug and depends on its specific pharmacokinetics. | Grammatical error |
| Compostos que podem ser utilizados como agentes de revestimento são, por exemplo, hidroxipropilmetilcelulose, etilcelulose, hidroximetilcelulose, hidroxipropilcelulose, polioxietilenoglicol, Tween 80, Pluorinc F68, e pigmentos tais como o óxido de titânio e o óxido férrico. | Compounds which may be used as coating agents are, for example, hydroxypropylmethylcellulose, ethylcellulose, hydroxymethylcellulose, hydroxypropylcellulose, polyoxyethylene glycol, Tween 80, pluorinc F68, and pigments such as titanium oxide and ferric oxide. | Formatting error |
| J = 14,4 1.4 Hz); 2.81 (IR dupleto, J= 1,4 Hz); 3.90 (IR dupleto, J= 11.0 Hz); 4.48 (IR dupleto, J= 11,0 Hz); 7.30 (IR dupleto de dupletos, J 8,6 2.2 Hz); 7.38 (IR dupleto, J = 2.2 Hz); 7.78 (IR dupleto, J = 8.6 Hz). | J = 14,4 1.4 Hz); 2.81 (1H, doublet, J = 1,4 Hz); 3.90 (1H, doublet, J = 11.0 Hz); 4.48 (1H, doublet, J = 11,0 Hz); 7.30 (1H, doublet of doublets, J 8,6 2.2 Hz); 7.38 (1H, doublet, J = 2.2 Hz); 7.78 (1H, doublet, J = 8.6 Hz). | Formatting error |

Table 4: Error type examples for Portuguese → English

### 2.2.4. Usability Evaluation

The usability evaluation for this language pair is still ongoing. Results will be published in an updated version of this deliverable before M24.

### 2.2.5. Productivity Evaluation

The productivity evaluation will be performed once the Translation Memory component from deliverable D4.1, “TM data resources”, for this language pair has been finalised. Results will be published in deliverable D7.7, “Final report on the intrinsic and extrinsic quality of MT”.

## *2.3. Discussion*

For both the English into Portuguese system and the Portuguese into English system automatic scores seem to correlate well with the results of the human quality evaluations. For both language directions the PLuTO output clearly seems to come out on top.

Although the engines rank the same in the automatic and the human evaluations, the human appreciation of the output is higher than one would expect based on the automatic scores. This may, at least partly, be explained by the fact that the evaluation guidelines were focusing on the adequacy of the translation (‘in how far is the meaning present in the source sentence preserved in the translation?’), rather than on absolute quality. The results of the usability evaluation, that will be published in deliverable D7.7, “Final report on the intrinsic and extrinsic quality of MT”, may provide additional proof of this.

Another observation is that both language directions do not necessarily perform equally well. Translation from Portuguese into English generally seems to score higher than the reverse language pair. We assume that this can, to a large extent, be attributed to the different nature of the languages. For instance, gender agreement in Portuguese limits the possible translation options for translation into English, whereas the reverse is true for translation from English into Portuguese: the absence of gender information in, for example, articles causes ambiguity that the translation engine has to deal with.

The automatic evaluations also seem to hint at domain-related quality differences. It seems like some domains lend themselves better for translation than others do. At the high end of the scale are domains such as Mechanical engineering, Physics, and Electricity. At the low end we find Human necessities and, especially, Chemistry. This behaviour seems to be consistent across language directions. The lower scores for translations in the Chemistry domain might be explained by the abundant presence of chemical structures in chemical patent text, which pose a specific challenge for the translation systems. Further investigation is required to determine what other factors might be causing these differences in quality between the different domains.

A conclusion that can be drawn from the error analysis is that both language pairs seem to struggle with formatting issues. Whereas grammatical and semantic issues are typically more difficult to solve, we may be able to achieve some quick wins by resolving these formatting issues (or at least some of them). For instance, finding a way to identify and protect names of chemical compounds might improve the output considerably.

# 3. English—French MT System

## *3.1. Automatic Evaluation*

With the automatic evaluation, domain-specific test sets were used to calculate the translation scores. The overall score was obtained by taking the average of the domain-specific scores.

Table 5 below shows the automatic scores obtained for translations from English into French.

| | **PLuTO** | **Google** | **Systran** |
|---|---|---|---|
| **All** | 56.28 / 65.45 | 43.32 / 57.58 | 32.92 / 51.09 |
| | | | |
| **A (Human necessities)** | 56.21 / 65.45 | 42.67 / 57.00 | 31.62 / 50.12 |
| **B (Operations)** | 55.57 / 65.76 | 44.58 / 58.39 | 33.82 / 51.77 |
| **C (Chemistry)** | 60.9 / 69.18 | 45.92 / 59.82 | 31.72 / 51.37 |
| **D (Textiles)** | 58.00 / 66.93 | 44.80 / 58.62 | 33.09 / 51.53 |
| **E (Fixed constructions)** | 52.64 / 62.71 | 41.93 / 55.75 | 32.30 / 49.50 |
| **F (Mechanical engineering)** | 56.69 / 66.35 | 45.34 / 58.97 | 35.00 / 52.40 |
| **G (Physics)** | 54.74 / 65.32 | 40.24 / 55.77 | 32.69 / 51.02 |
| **H (Electricity)** | 55.18 / 65.61 | 40.96 / 56.49 | 32.40 / 50.92 |

Table 5: Automatic scores English → French[5]

Observations:

- Overall, the METEOR scores align quite well with the BLEU scores (with a few minor exceptions).
- The BLEU scores vary between 31.62 (Systran, Human necessities) and 60.9 (PLuTO, Chemistry) (difference of 29.28). Looking at the METEOR scores, the difference is a lot smaller: 49.50 (Systran, Fixed constructions) versus 69.18 (PLuTO, Chemistry) (difference of 19.68).
- PLuTO seems to outperform Google and Systran in all domains. Google comes second with acceptable scores and Systran is clearly lagging behind.
- The PLuTO engine's performance varies according to the domain it is used in, but the variation is a lot less outspoken than it was for the English--Portuguese language pairs. Where the maximum difference for those language pairs was for both around 18 points, the difference for English into French is only 8.26. Best results are obtained in the Chemistry domain (IPC C; 60.9 BLEU/69.18 METEOR) and worst ones in the Fixed constructions domain (IPC E; 52.64 BLEU/62.71 METEOR). Even for the domain in which the worst scores are obtained, the scores are still quite high.

Table 6 below shows the automatic scores obtained for translations from French into English:

| | PLuTO | Google | Systran |
|---|---|---|---|
| **All** | 56.92 / 67.44 | 42.52 / 59.65 | 28.90 / 53.67 |
| | | | |
| **A (Human necessities)** | 58.35 / 68.22 | 43.60 / 60.58 | 28.05 / 53.46 |
| **B (Operations)** | 55.03 / 66.95 | 42.29 / 59.84 | 30.45 / 54.53 |
| **C (Chemistry)** | 62.01 / 70.03 | 46.66 / 61.81 | 29.92 / 54.44 |
| **D (Textiles)** | 56.51 / 67.03 | 42.53 / 59.35 | 24.49 / 53.54 |
| **E (Fixed constructions)** | 53.85 / 64.73 | 40.27 / 57.29 | 30.12 / 52.99 |
| **F (Mechanical engineering)** | 57.21 / 67.77 | 43.28 / 60.36 | 31.28 / 55.35 |
| **G (Physics)** | 56.21 / 67.90 | 40.74 / 59.51 | 25.55 / 53.11 |
| **H (Electricity)** | 56.32 / 67.53 | 41.36 / 58.89 | 25.89 / 51.91 |

Table 6: Automatic scores French → English[6]

Observations:

- Overall, the METEOR scores align quite well with the BLEU scores (except for Systran).
- The BLEU scores vary between 24.49 (Systran, Textiles) and 62.01 (PLuTO, Chemistry). (difference of 37.52). Looking at the METEOR scores, the difference is a lot smaller: 51.91 (Systran, Electricity) versus 70.03 (PLuTO, Chemistry) (difference of 18.12).
- PLuTO seems to outperform Google and Systran in all domains. Google comes second with acceptable scores and Systran is clearly lagging behind.
- The PLuTO engine's performance varies according to the domain it is used in, but, just as with the English into French language pair, the variation is a lot less outspoken than it was for the English--Portuguese language pairs. Where the maximum difference for those language pairs was for both around 18 points, the difference for French into English is only 8.16. Best results are obtained in the Chemistry domain (IPC C; 62.01 BLEU/70.03 METEOR) and worst ones in the Fixed constructions domain (IPC E; 53.85 BLEU/64.73 METEOR). Even for the domain in which the worst scores are obtained, the scores are still quite high.

[5] Scores are BLEU / METEOR.

[6] Scores are BLEU / METEOR.

## 3.2. *Human Evaluation*

### 3.2.1. Ranking

Figure 12 below shows the adequacy scores obtained for translations from English into French:

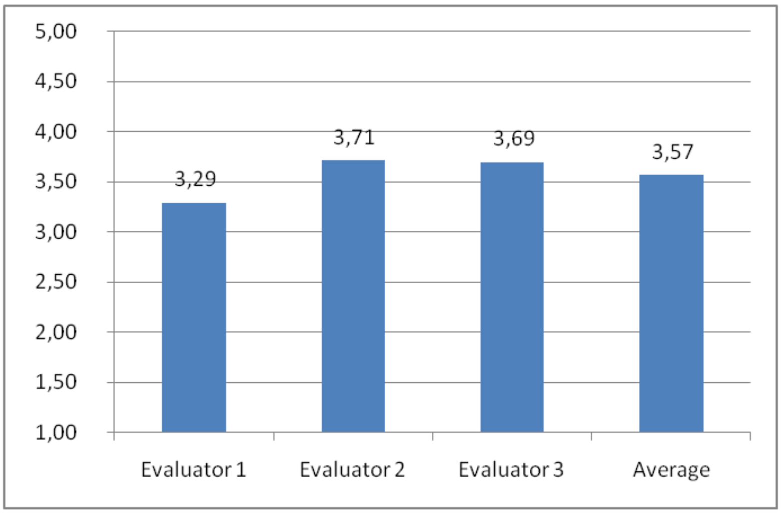


Figure 12: Human adequacy evaluation scores English → French

Observations:

- The different evaluators seem to have different opinions on the quality of the PLuTO output of the English into French engine. The scores differ as much as 0.42 (Evaluator 1: 3.29 vs. Evaluator 2: 3.71).
- Average score for the English into French language pair is 3.57, which is not so high.

Figure 13 below shows the adequacy scores obtained for translations from French into English:

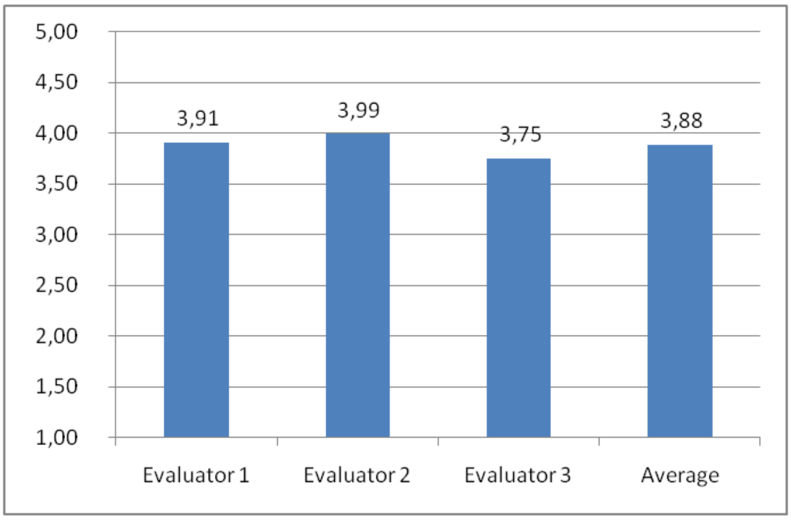


Figure 13: Human adequacy evaluation scores French → English

Observations:

- The different evaluators seem to agree on the quality of the PLuTO output of the French into English engine. The difference between the highest and the lowest score is 0.24 (Evaluator 3: 3.75 vs. Evaluator 2: 3.99), but a minimal difference between evaluators is always to be expected.
- Average score for the French into English language pair is 3.88, which is fairly high.

### 3.2.2. Benchmarking

Figure 14 below shows how evaluators have ranked the PLuTO English into French output in comparison with the Google Translate and Systran output. Rank 1 indicates the number of times on the total amount of evaluated segments a segment was selected as being the best one. Rank 2 indicates the number of times it was chosen as second best and rank 3 indicates the number of times it was seen as the worst one.

In case of equal quality, evaluators were instructed to give the same rank. For instance, in case PLuTO and Google did equally well but better than Systran, ranks given would be 1 for PLuTO, 1 for Google, and 2 for Systran.

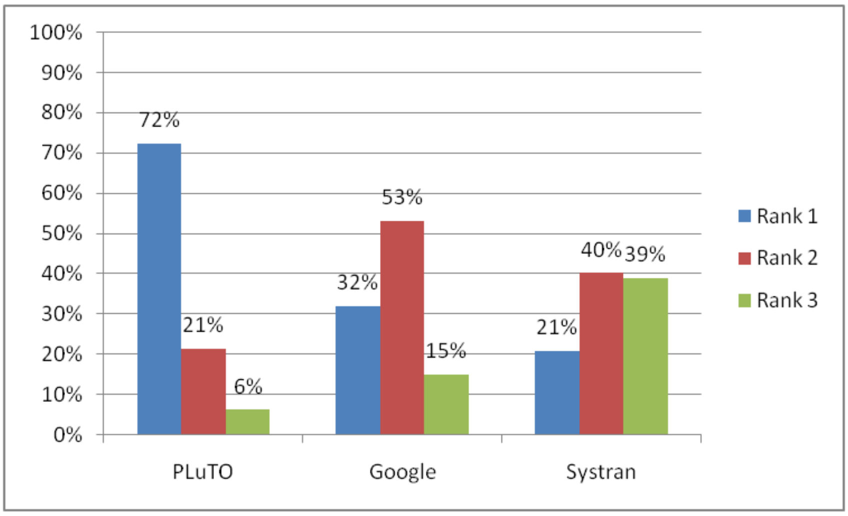


Figure 14: Human benchmarking evaluation English → French

Observations:

- Evaluators clearly seem to have a preference for the PLuTO output. It was selected as the best performing engine in 72% percent of the cases.
- Google and Systran output are close with a slight preference for Google.
- These results confirm the findings of the automatic evaluation. The METEOR scores seem to correspond better with the human evaluation results than the BLEU scores.

Figure 15 below shows how evaluators have ranked the PLuTO French into English output in comparison with the Google Translate and Systran output:

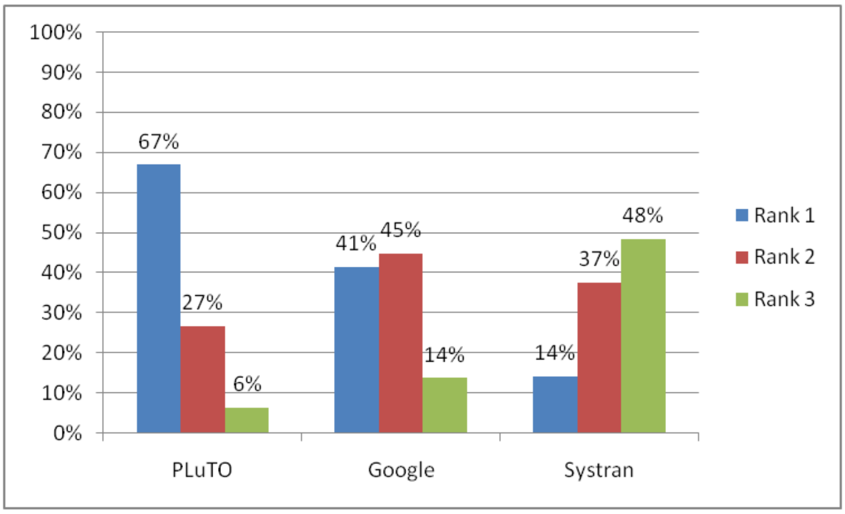


Figure 15: Human benchmarking evaluation French → English

Observations:

- Evaluators seem to have a clear preference for the PLuTO output. It was selected as the best in 67% of the cases.
- Google, in second place, clearly beats Systran for this language pair.
- Again, these results seem to confirm the results of the automatic evaluation, but contrary to what we found for the English into French language pair, for this language pair the human ranking results seem to agree better with the BLEU scores than with the METEOR scores.

### 3.2.3. Error Analysis

As for the English--Portuguese language pairs, one of the three evaluators per language pair that took the adequacy evaluation was also asked to categorise errors found in the PLuTO MT output.

Figure 16 below shows how errors were classified for the English into French language pair:

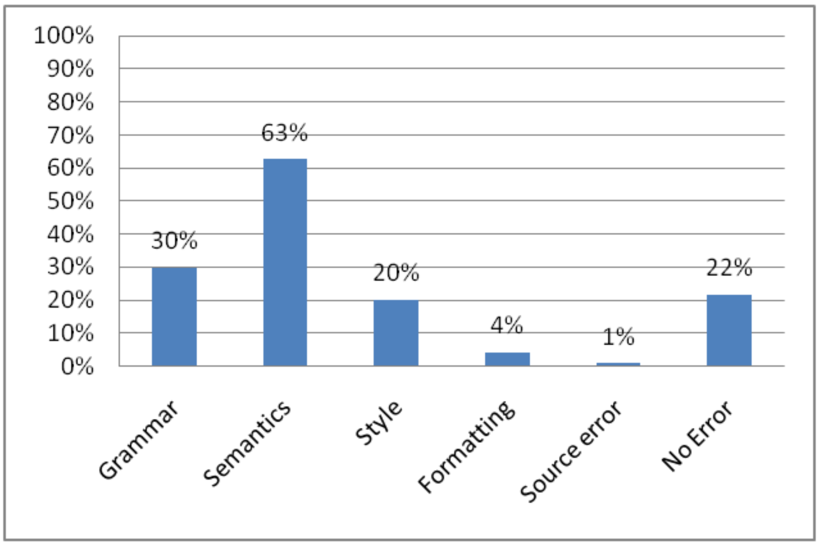


Figure 16: Error classification English → French

Observations:

- 22% of all segments in the evaluation set had no errors at all.
- The most common types of issues encountered in the output are semantic issues (63%) and grammatical issues (30%). See the examples below for more details.
- Only 4% of all segments in the evaluation set seem to suffer from formatting problems.

Table 7 below shows a few examples of the most common problems found in the MT output. The sample segments have a least the error that is mentioned in the Error Type column, but may include errors of another nature, too.

| **Source Segment** | **MT Target** | **Error Type** |
|---|---|---|
| As the lactic acid bacteria, it is preferable to use Lactobacillus sake D-1001 (NIBH Deposition No.11708, IPOD FERM BP-08544) together with Lactococcus lactis IFO | En tant que bactérie lactique, il est préférable d'utiliser Lactobacillus sake d-1001 (nibh dépôt no.11708, IPOD FERM BP-08544) avec Lactococcus lactis ifo | Semantic error |
| The method according to claim 4 or 5, wherein the core contains 50.0 mg cyclophosphamide (53.5 mg cyclophosphamide monohydrate), 39.0 mg lactose monohydrate, 20.0 mg nonpreswollen corn starch, 40.0 mg microfine cellulose, 2.0 mg highly dispersed | Procédé selon la revendication 4 ou 5, dans lequel le noyau contient 50,0 mg de cyclophosphamide (53,5 mg de monohydrate de cyclophosphamide), 39,0 mg de monohydrate de lactose, 20,0 mg d'amidon de maïs nonpreswollen, 40,0 mg de cellulose microfine, 2,0 mg de silice | Semantic error |

| Source Segment | MT Target | Error Type |
|---|---|---|
| silica, 4.0 mg talcum, and 1.5 mg magnesium stearate. | hautement dispersée, 4,0 mg de talc et 1,5 mg de stéarate de magnésium. | |
| The torque control device according to claim 6), characterized in that said speed detecting means comprise an incremental bi-directional encoder (7). | Le dispositif de contrôle du couple selon la revendication 6), caractérisé en ce que ledit moyen de détection de vitesse comprennent un codeur incrémental bidirectionnel (7). | Grammatical error |
| The thermal insulation board (6) of at least one of claims 1 to 8, characterized in that it has a thickness from 20 to 100 mm, preferably 30 to 50 mm. | La plaque d'isolation thermique (6) selon au moins une des revendications 1 à 8, caractérisé en ce qu'il présente une épaisseur de 20 à 100 mm, de préférence de 30 à 50 mm. | Grammatical error |

Table 7: Error type examples for English → French

Figure 17 below shows how errors were classified for the French into English language pair:

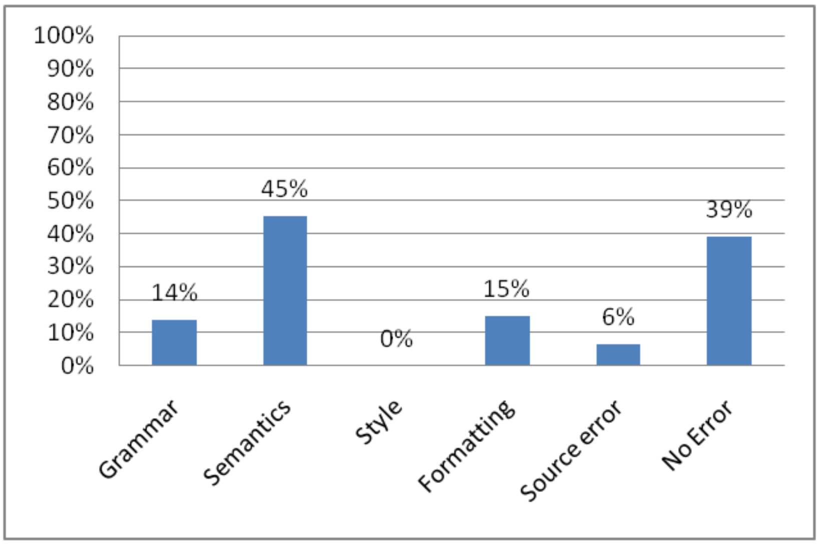


Figure 17: Error classification French → English

Observations:

- 39% of all segments in the evaluation set had no errors at all. This is a lot.
- The most common types of issues encountered in the output are semantic issues (45%) and formatting (15%) and grammatical (14%) issues. See the examples below for more details.
- 6% of all segments in the evaluation set seemed to have issues in the source text.

Table 8 below shows a few examples of the most common problems found in the MT output. The sample segments have a least the error that is mentioned in the Error Type column, but may include errors of another nature, too.

| Source Segment | MT Target | Error Type |
|---|---|---|
| Le sulfate de magnésium peut être introduit pour conserver la stabilité structurale du matériau traité, avec pour effet de le maintenir sensiblement sur place. | Magnesium sulfate may be introduced to maintain the structural stability of the treated material, thereby substantially maintaining the on-site. | Semantic error |
| Procédé suivant la revendication 1, caractérisé en ce qu'on assure le balayage d'une transversale à la direction générale d'observation (D) de l'individu par l'image d'une fente (22) donnant accès au | Method according to claim 1, characterized in that provision is made for the scanning of a line transverse to the general direction of observation (d) of the person in the image of a slit (22) giving access to the detection | Semantic error |

| Source Segment | MT Target | Error Type |
|---|---|---|
| récepteur de détection | receiver | |
| Ces compositions peuvent généralement comprendre un antagoniste de SNRI-NMDA à double action (par exemple de la bicifadine et/ou du milnacipran). | The compositions may generally include a dual-acting snri-nmda antagonist (for example the bicifadine and/or milnacipran). | Formatting error |
| Milieu selon la revendication 5, caractérisé en ce que l'indicateur de pH est choisi parmi les composants suivants, à savoir rouge de phénol, bleu de bromothymol, pourpre de bromocrésol, rouge neutre. | Medium according to claim 5, characterized in that the ph indicator is selected from the following compounds, namely phenol red, bromothymol blue, bromocresol purple, neutral red. | Formatting error |
| Il en résulte que cette souche de Lactobacillus et ses substances antibactériennes conviennent particulièrement à la mise au point au sens large de nouveaux médicaments, d'aliments fonctionnels, d'additifs diététiques, et analogues. | As a result, said Lactobacillus strain and its antibacterial substances are particularly suitable for the development in the broadest sense, new drugs, functional food, dietary additives, and the like. | Grammatical error |
| Ladite invention a également trait à des pompes sans stents destinées à être insérées entre la veine et l'aorte, entre la veine cave et l'artère pulmonaire, et conçues pour être utilisées en chirurgie cardiaque. | The invention also relates to pumps without stents for insertion between the vein and the aorta between the vena cava and the pulmonary artery, and suitable for use in cardiac surgery. | Grammatical error |

Table 8: Error type examples for French → English

### 3.2.4. Usability Evaluation

The usability evaluation for this language pair is still ongoing. Results will be published in an updated version of this deliverable before M24.

### 3.2.5. Productivity Evaluation

The productivity evaluation will be performed once the Translation Memory component from deliverable D4.1, “TM data resources”, for this language pair has been finalised. Results will be published in deliverable D7.7, “Final report on the intrinsic and extrinsic quality of MT”.

## *3.3. Discussion*

Generally speaking, the human evaluation results seem to confirm the scores of the automatic evaluation: both evaluations indicate that PLuTO is the best performing engine. Google is a distinct second and Systran comes third. However, looking at the results more closely, there are a few more observations to be made. Firstly, there is the apparent discrepancy between the English--Portuguese and the English--French engines. Although the automatic scores for the English--French engines are higher, those for the human evaluations are lower. As mentioned when discussing the English--Portuguese results, we feel the Portuguese results are exceptionally high. The English--French results may just be more realistic. A second observation is that there seems to be a difference as to how the automated metrics correspond to the human evaluations. For the English into French system, the METEOR metric seems to correspond better to the human judgements; for the French into English system the BLEU scores seem to correlate better with the human evaluations.

Similar to what we observed with the English--Portuguese systems, both language directions do not seem to perform equally well. The difference between the English--French language directions seems to be smaller than that between the English--Portuguese language pairs, though, and that may explain why the tendency is not completely consistent across all evaluations. The automatic metrics show hardly any difference. Average BLEU score across domains for English into French is

56.93; average METEOR score is 67.51. For French into English the scores are 56.26 (BLEU) and 65.86 (METEOR). The human adequacy evaluation shows a bigger difference: 3.57 for English into French versus 3.88 for French into English. This difference may be attributed to one evaluator who seems to have been particularly harsh in his judgement of the English into French language pair (3.29).

Looking at the differences between domains we noticed that they still appear, but are less pronounced as for the English--Portuguese language pairs. Remarkably enough, the domain that yielded the lowest scores for the English--Portuguese language pairs, Chemistry, seems to produce the best results for the English--French language pairs. Further investigation will have to reveal why.

# 4. Discussion and Conclusions

Looking at the evaluation results obtained so far, it is clear that different language combinations reveal different tendencies. Sometimes these tendencies are consistent within a language pair but not when compared to other language pairs. At other times, the target language seems to be the driving factor.

For example, for the English--French language pairs, the automatic scores seem to correlate well with the human evaluations. Both are rather high. This holds true for the scores obtained for translations from English into French as well as for those from French into English. For the English--Portuguese scores this is not the case: the automatic scores are not bad, but not particularly high either. The human evaluations are exceptionally high compared to the automatic scores. It is hard to say at this point what might be the reason for these differences. We hope the outcome of the usability evaluations will help us determine whether these differences are significant or not.

When we look at translations from or into English, it appears translations into English are generally speaking better received than translations from English. The highest scores, both automatic and human, were obtained for translation from Portuguese into English and for French into English. Looking at the competition, we also see that the PLuTO's lead, particularly in relation to Google, is reduced for translations into English. So it seems like other systems, too, are performing better for translation into English. This means competition is likely to be stronger for translation into English.

Another observation that cannot be ignored is that there are clearly quality differences between translations in different domains. Although differences have been observed for all language directions, the domains that seem to yield better translations are not consistent across languages. Content in the chemistry domain seems to produce low quality translations for language directions that involve Portuguese as either source or target language, whereas the same domain seems to yield the best quality translations for language pairs that involve French. Time will need to be invested to learn more about where this behaviour comes from. A better insight into why certain domains score better or worse for a certain language pair may provide valuable hints to boost translation quality for domains that are currently not scoring well. Availability of data for a domain is a factor that may have played a role.

Another path that may be explored for further improving translation quality is to look at the type of issues that were encountered in the translations. As is to be expected, grammatical and semantic issues are fairly common. However, problems of a semantic or grammatical nature are hard to resolve. Quick wins may be obtained by focussing on formatting issues, which also seem to occur frequently. Formatting issues may be easier to spot and resolve by implementing pre- and post-processing routines around the translation process. For instance, names of chemical compounds or proper names may be detected prior to translation and protected throughout the translation process.

In conclusion, we think it is safe to say that the evaluations performed to date on the PLuTO language pairs that have already been completed, confirm that, generally speaking, all engines produce translations of acceptable quality. Comparisons with leading competing translation systems have shown that the PLuTO engines are able to produce translations of at least the same (if not better) quality than those produced by the established systems. We will continue to monitor the quality of the PLuTO translations against that of the competition so as to make sure we keep this competitive edge.

This does not mean that there is no more room for improvement. Differences in quality between translations for different domains will be investigated to see if any improvements can be made to boost the quality of translations for those domains that are currently lagging behind. We will also explore potential benefits that may be obtained from plugging in pre- and post-processing steps into the translation workflow to reduce the number of formal issues observed in the translations.

# Appendix A: Metrics used for Automatic Evaluation

## BLEU

From wikipedia.org:

**BLEU** (**Bilingual Evaluation Understudy**) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and that of a human: "the closer a machine translation is to a professional human translation, the better it is". BLEU was one of the first metrics to achieve a high correlation with human judgements of quality, and remains one of the most popular.

Scores are calculated for individual translated segments—generally sentences—by comparing them with a set of good quality reference translations. Those scores are then averaged over the whole corpus to reach an estimate of the translation's overall quality. Intelligibility or grammatical correctness is not taken into account.

BLEU is designed to approximate human judgement at a corpus level, and performs badly if used to evaluate the quality of individual sentences.

BLEU's output is always a number between 0 and 1. This value indicates how similar the candidate and reference texts are, with values closer to 1 representing more similar texts.

Academic reference:

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation" in *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics* pp. 311–318.

## METEOR

From wikipedia.org:

**METEOR** (**Metric for Evaluation of Translation with Explicit ORdering**) is a metric for the evaluation of machine translation output. The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. It also has several features that are not found in other metrics, such as stemming and synonymy matching, along with the standard exact word matching. The metric was designed to fix some of the problems found in the more popular BLEU metric, and also produce good correlation with human judgement at the sentence or segment level. This differs from the BLEU metric in that BLEU seeks correlation at the corpus level.

Results have been presented which give correlation of up to 0.964 with human judgement at the corpus level, compared to BLEU's achievement of 0.817 on the same data set. At the sentence level, the maximum correlation with human judgement achieved was 0.403.

Academic reference:

Banerjee, S. and Lavie, A. (2005) "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments" in *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005*

# Appendix B: Human Evaluation Guidelines

## Adequacy Evaluation

The table below list the values evaluators could choose from to label translation quality. The table also explains how each of the values should be interpreted.

| Values | Description |
|---|---|
| **Excellent (5)** | Read the MT output first. Then read the source text (ST). **All** meaning expressed in source fragment appears in the translation fragment. Your **understanding is not improved** by reading the ST because the MT output is satisfactory and would not need to be modified (**grammatically correct/proper terminology is used**/maybe not stylistically perfect but fulfils the main objective, i.e. transferring accurately all information). |
| **Good (4)** | Read the MT output first. Then read the source text. **Most** meaning expressed in source fragment appears in the translation fragment. Your **understanding is not improved** by reading the ST even though the MT output contains **minor grammatical mistakes** (word order/punctuation errors/word formation/morphology). You would **not need to refer to the ST** to correct these mistakes. |
| **Fair (3)** | Read the MT output first. Then read the source text. **Much** meaning expressed in source fragment appears in the translation fragment. However, your **understanding is improved** by reading the ST allowing you to correct **minor grammatical mistakes** in the MT output (word order/punctuation errors/word formation/morphology). You would **need to refer to the ST** to correct these mistakes. |
| **Poor (2)** | Read the MT output first. Then read the source text. **Little** meaning expressed in source fragment appears in the translation fragment. Your **understanding is improved considerably** by reading the ST, due to **significant errors** in the MT output (textual and syntactical coherence/textual pragmatics/word formation/morphology). You would have to **re-read the ST a few times to correct** these errors in the MT output. |
| **Very poor (1)** | Read the MT output first. Then read the source text. **None** of the meaning expressed in source fragment appears in the translation fragment. Your **understanding only derives from reading the ST,** as you could not understand the MT output. It contained serious errors in any of the categories listed above, including wrong POS. You could only produce a translation by dismissing most of the MT output and/or re-translating from scratch. |

Table 9: Adequacy Evaluation Guidelines