DELIVERABLE

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

Deliverable 7.7 Final 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	20/02/13	J. Van de Walle	CL	First draft
2	4/03/13	R. van der Borgt	CL	Additions
3	22/03/13	J. Van de Walle	CL	Review
4	02/04/13	J.Tinsley	DCU	Final review, 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This deliverable contains original unpublished work except where clearly indicated otherwise.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—German, English—Japanese, English—Spanish, and English—Chinese 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.

The English—German and English—Japanese evaluations are carried out using held-out test data randomly selected from our parallel patent corpora. The English—Spanish and English—Chinese evaluations are carried out on a test set that consisted of recently harvested patent data that had no relation to our parallel patent corpora. For some of the automatic evaluations, test sets were segmented into sub-sets based on the IPC patent classification system. In doing this, the evaluation allows us to assess whether the translation systems perform better in some categories of patents (e.g. chemistry, engineering, etc.) than other.

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.

Evaluation Setup

The evaluation setup was described extensively in Deliverable D7.6, "First Report on the Intrinsic and Extrinsic Quality of MT". We will not repeat that description here. For details on the evaluation methodology, the test sets used, a description of the types of human and automatic evaluations performed, and information on the profile of the evaluators we refer to the Section "Evaluation Setup" in Deliverable D7.6, "First Report on the Intrinsic and Extrinsic Quality of MT".

One crucial difference, however, between the evaluation of language pairs English—Spanish and English—Chinese and all other language pairs we evaluated in this project is the composition of the test set. Whereas for all other language pairs, the test set was made up of a set of segments randomly selected and held out of our training corpora, the tests sets for language pairs English—Spanish and English—Chinese consisted of recent patent data that was harvested from the internet.

As the test sets that we used previously for the other languages were held out of our training data but not necessarily out of Google's, the expert reviewers requested that we create a new test set with recent patent data to minimize the chances that the test set data would have been part of Google's training data. It was assumed that this would allow for a fairer comparison between the evaluation scores of the PLuTO output and Google's. Since the evaluations for English—German and English—Japanese had already been carried out by the time the reviewers formulated their request, the scores that we report here for those language pairs are still based on test sets that were sampled from the parallel corpora.

In terms of the types of evaluations that have been carried out, there are also some significant differences with the evaluations that were previously reported for language pairs English—Portuguese and English—French. One of the evaluations that was performed for these language pairs was the usability evaluation, which consisted of an experiment in which we tried to simulate the patent searcher's use case by having evaluators assess the relevancy of a number of machine translated patents in relation to a hypothetical invention. We did not repeat this experiment for the language pairs that we report on in this deliverable for two reasons. First, we received feedback from the WON user group that the preparation of these experiments took up too much of their time. Second, the outcome of the experiments was indecisive and did not allow us to formulate any firm conclusions. For these reasons, it was decided not to repeat these experiments for the remaining language pairs.

In D7.6, "First report on the intrinsic and extrinsic quality of MT", we did not perform any productivity evaluations. In that deliverable, we announced that the results of the productivity evaluations would be reported in this deliverable. However, since it became clear during the discussions partners had regarding exploitation of the PLuTO service that the primary target group for exploitation would be the patent searchers and the use of machine translation in a search context, it was decided among the partners that productivity evaluations, which focus on the use of machine translation in a translation production context, would not be performed for all language pairs. It was decided that we would limit the productivity evaluations to one language pair involving a Western European language and one involving an Asian language. The language pairs selected for the productivity evaluation were French into English and Chinese into English. The results of these productivity evaluations are reported in section "Productivity Evaluations" on page 30 of this deliverable.

English—German MT System

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 German:

	PLuTO		Google		Systran	
	BLEU	METEOR	BLEU	METEOR	BLUE	METEOR
A (Human necessities)	55.53	53.85	37.84	39.74	24.49	29.47
B (Operations)	49.29	48.91	33.06	36.10	23.06	27.86
C (Chemistry)	59.16	57.43	41.53	42.93	25.87	31.11
D (Textiles)	52.86	52.05	35.35	38.06	23.15	28.57
E (Fixed constructions)	48.46	48.58	33.96	36.89	23.57	28.52
F (Mechanical engineering)	50.66	50.46	36.52	39.37	24.98	29.95
G (Physics)	52.31	51.52	31.76	35.27	22.50	28.05
H (Electricity)	53.54	52.69	32.86	36.14	23.32	29.15
Average	**52.73**	**51.94**	**35.36**	**38.06**	**23.87**	**29.09**

Table 1: Automatic evaluation scores English → German

Observations:

- Whereas the BLEU score tends to be slightly higher than the METEOR score for the PLuTO system, the reverse appears to be true for the Google and Systran systems. For rule-based systems such as Systran this is fairly common, as METEOR is known to allow for more lexical variation, but it striking that this also holds true for Google, which is a statistical system.
- The BLEU scores vary between 22.50 (Systran, Physics) and 55.53 (PLuTO, Human necessities).
- PLuTO outperforms Google and Systran in all domains with both metrics. Google comes second with BLEU scores that are 15 to 20 BLEU points lower than those of the PLuTO engines and Systran comes last, with BLEU scores that are about 10 BLEU points lower than those of Google.
- The PLuTO engine's performance varies according to the domain it is used in. It performs best with content in the Chemistry domain (IPC C; 59.16 BLEU/57.43 METEOR) and worst with content in the Fixed constructions domain (IPC E; 48.46 BLEU/48.58 METEOR).

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

	PLuTO		Google		Systran	
	BLEU	METEOR	BLEU	METEOR	BLUE	METEOR
A (Human necessities)	61.43	69.78	45.24	60.74	29.93	52.92
B (Operations)	55.13	65.22	40.00	56.45	28.34	50.48
C (Chemistry)	64.87	72.34	48.29	62.73	31.79	54.51
D (Textiles)	57.15	67.79	39.67	58.29	28.44	51.96
E (Fixed constructions)	53.23	63.72	38.67	55.25	28.69	50.36
F (Mechanical engineering)	55.11	66.92	41.10	58.95	30.88	53.71
G (Physics)	59.76	69.31	42.77	59.96	29.32	53.37
H (Electricity)	61.08	70.32	43.01	60.51	30.32	54.22
Average	**58.47**	**68.18**	**42.34**	**59.11**	**29.71**	**52.69**

Table 2: Automatic evaluation scores German → English

Observations:

- The METEOR scores are higher than the BLEU scores in all cases. The difference is the largest for Systran and the smallest for PLuTO.
- The BLEU scores vary between 28.44 (Systran, Textiles) and 64.87 (PLuTO, Chemistry).
- PLuTO seems to outperform Google and Systran in all domains. Google comes second and the difference with the PLuTO engines is bigger 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 Chemistry domain (IPC C; 64.87 BLEU/72.34 METEOR) and worst with content in the Fixed constructions domain (IPC E; 53.23 BLEU/63.72 METEOR).

Human Evaluation

Adequacy

Figure 1 below shows the adequacy scores obtained for translations from English into German:

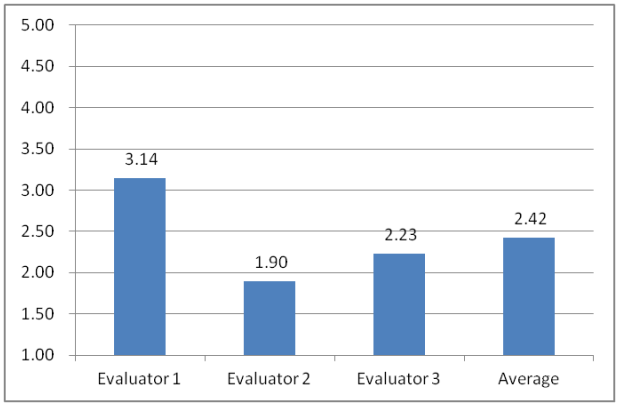

Figure 1: Human adequacy evaluation scores English → German

Observations:

- The different evaluators seem to have different opinions on the quality of the PLuTO output of the English into German engine. The scores differ as much as 1.24 on a scale of 1 to 5 (Evaluator 1: 3.14 vs. Evaluator 2: 1.90).
- Average score for the English into German language pair is 2.42, which is rather low.

Figure 2 below shows the adequacy scores obtained for translations from German into English:

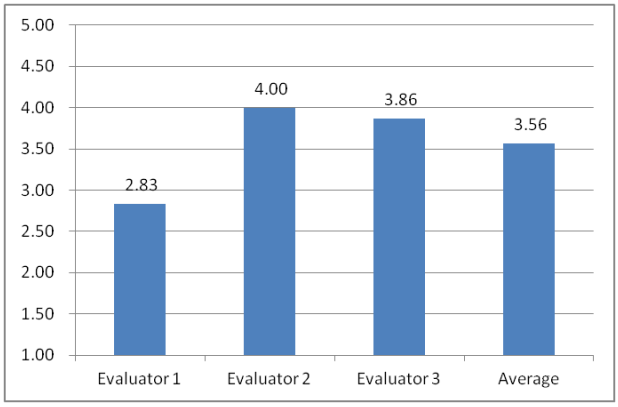

Figure 2: Human adequacy evaluation scores German → English

Observations:

- The different evaluators seem to have different opinions on the quality of the PLuTO output of the German into English engine. The scores differ as much as 1.17 on a scale of 1 to 5 (Evaluator 1: 2.83 vs. Evaluator 2: 4.00).
- Average score for the German into English language pair is 3.56, which is fairly high.

Benchmarking

Figure 3 below shows how evaluators have ranked the PLuTO English into German 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 translations 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.

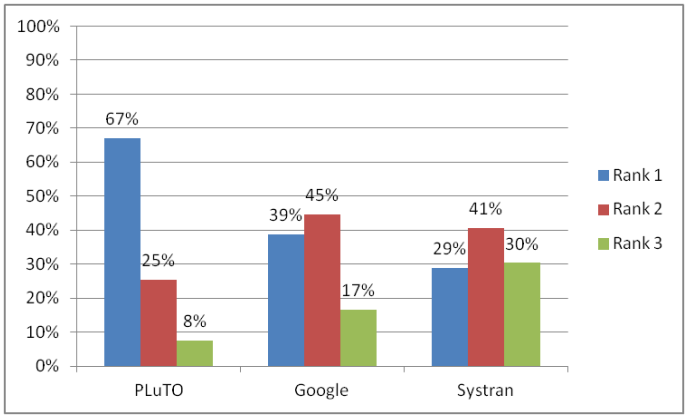

Figure 3: Human benchmarking evaluation English → German

Observations:

- Evaluators clearly seem to have a preference for the PLuTO output. It was selected as the best performing engine in 67% percent of the cases.
- Google output is the second best, before Systran output.
- These results confirm the findings of the automatic evaluation.

Figure 4 below shows how evaluators have ranked the PLuTO German into English output in comparison with the Google Translate and Systran output:

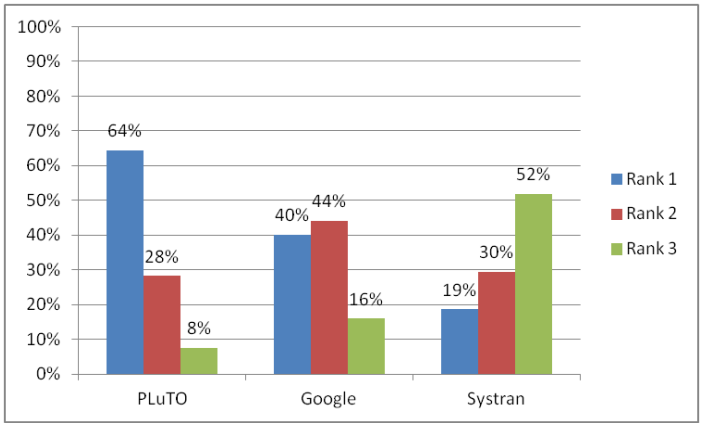

Figure 4: Human benchmarking evaluation German → English

Observations:

- Evaluators seem to have a clear preference for the PLuTO output, just 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.

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 5 below shows how errors were classified for the English into German language pair:

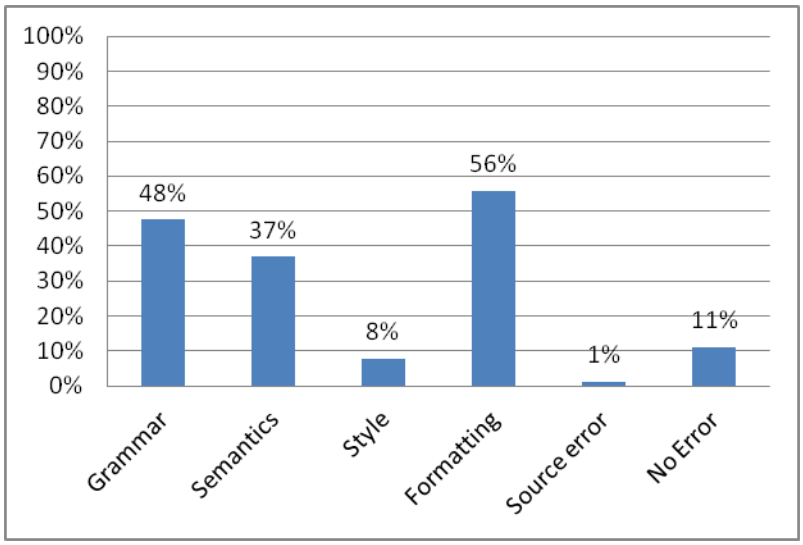

Figure 5: Error classification English → German

Observations:

- Only 11% of all segments in the evaluation set had no errors at all.
- The most common types of issues encountered in the output are formatting issues (56%) and grammatical issues (48%). See the examples below for more details.
- Semantic issues are also quite common (37%).

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
PRESSURE SENSITIVE ADHESIVES WITH A FIBROUS REINFORCING MATERIAL	Haftklebstoffe enthaltend ein faserförmiges Verstärkung Material	Formatting error
A hand-scrubbing brush as defined in Claim 2, characterized in that said gear mechanism is so designed that all of said rotational bodies rotate in the same direction, and that the direction of rotation of said motor (9) is reversible.	Handwasch Bürste nach Anspruch 2, dadurch gekennzeichnet, dass der Zahnrad Mechanismus derart ausgebildet ist, daß die Rotation Körper alle in die gleiche Richtung drehen, und daß die Drehrichtung des Motors (9) umkehrbar ist.	Formatting error
Device according to one of claims 1 to 16, this device being incorporated in a plastic box (51).	Vorrichtung nach einem der Ansprüche 1 bis 16, wobei diese Vorrichtung in einem Kunststoff Kasten (51) eingebaut ist.	Grammatical error

Source Segment	MT Target	Error Type
Use according to claim 1, wherein CCI-779 is to be administered at a dose of 10 to 100 mg per week.	Verwendung nach Anspruch 1, bei der CCI-779 zur Verabreichung in einer Dosis von 10 bis 100 mg pro Woche.	Grammatical error
Compounds of formula (V) according to claim 21: wherein R6, B, D, A, X, R1, R2, R3 and R4 are as defined in claim 1, for use as intemediates for the synthesis of compounds of formula (I).	Verbindungen der Formel (V) nach Anspruch 21, wobei: R6, A, B, D, X, R1, R2, R3 und R4 wie in Anspruch 1 definiert sind, zur Verwendung als intemediates für die Synthese der Verbindungen der Formel (I).	Semantic error
Method according to claim 1 or 2, characterized in that fibres are added to the concrete in an amount of at least 30 kg/m3, preferably between 40 and 80 kg/m3 of fibre concrete.	Verfahren nach Anspruch 1 oder 2, dadurch gekennzeichnet, daß man Fasern zusätzlich mit dem Beton in einer Menge von mindestens 30 kg/m3, vorzugsweise zwischen 40 und 80 kg/m3 aus Faserbeton.	Semantic error

Table 3: Error type examples for English → German

Figure 6 below shows how errors were classified for the German into English language pair:

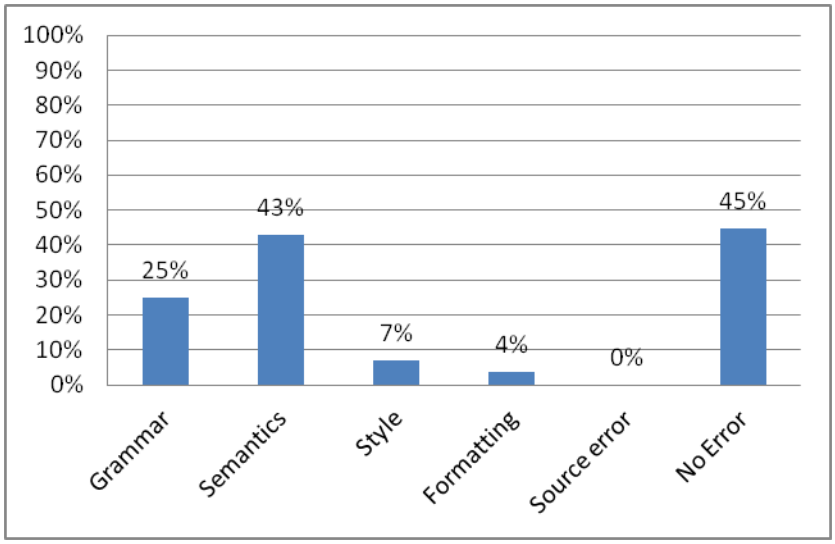

Figure 6: Error classification German → English

Observations:

- 45% 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 (43%), i.e. issues relating to the use of terminology, and grammatical issues (25%). See the examples below for more details.

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
Verfahren, nach Anspruch 1, bei dem die kritische, superkritische oder fast-kritische Flüssigkeit Schleppmittel aufweist.	A method according to claim 1, wherein said critical, supercritical or near critical fluid entraining agent.	Semantic error
Kosmetiktuchprodukt nach einem der vorangehenden Ansprüche, wobei die Mikroemulsion in einer Menge von etwa 0, 1 bis etwa 20 Gewichtsprozent der	A towelette product according to any one of the preceding claims, wherein the microemulsion in an amount from about 0. 1 to about 20 weight percent of the	Semantic error

Source Segment	MT Target	Error Type
Zusammensetzung, die das Substrat imprägniert, vorliegt.	composition impregnates the substrate.	
Einrichtung gemäß einem der Ansprüche 1 bis 16, wobei die Einrichtung in ein Kunststoffgehäuse (51) eingefügt ist.	Device according to any one of claims 1 to 16, wherein said device in a plastic housing (51) is inserted.	Grammatical error
Ein Verfahren nach Anspruch 3, dadurch gekennzeichnet, daß eine Verdampfung ausgeführt wird, indem das Salzwasser der Umgebungsluft ausgesetzt wird.	A method according to claim 3, characterized in that an evaporation is carried out by the salt water is exposed to the ambient air.	Grammatical error

Table 4: Error type examples for German → English

Discussion

For both the English into German system and the German 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, especially for translations from German into English, 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.

Another observation is that both language directions do not necessarily perform equally well. Translation from German 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 German limits the possible translation options for translation into English, whereas the reverse is true for translation from English into German: 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. At the high-end of the scale are domains such as Chemistry and Human necessities. At the low end we find Fixed constructions, Operations and Mechanical engineering. This behaviour seems to be consistent across language directions. 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 grammatical and semantic issues. English into German also struggles 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 build German compound nouns correctly might improve the output considerably: in the German output, many compound nouns are written as separate words.

English—Japanese MT System

Automatic Evaluation

Since for the English—Japanese MT system, we did not have enough data available for all eight IPC domains, the test set data was sampled randomly across domains. The test consisted of about 2,000 segments in total.

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

PLuTO		Google		Systran	
BLEU	**METEOR**	**BLEU**	**METEOR**	**BLUE**	**METEOR**
33.95	37.27	17.98	21.70	6.02	8.44

Table 5: Automatic scores English → Japanese

Observations:

- Overall, the METEOR scores are slightly higher than the BLEU scores.
- The BLEU scores vary between 33.95 (PLuTO) and 6.02 (Systran) (difference of 27.93).
- PLuTO clearly outperforms Google and Systran. Google comes second and Systran last with very low scores.

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

PLuTO		Google		Systran	
BLEU	**METEOR**	**BLEU**	**METEOR**	**BLUE**	**METEOR**
22.89	51.05	19.68	48.14	12.55	42.88

Table 6: Automatic scores Japanese → English

Observations:

- Overall, the METEOR scores are a lot higher than the BLEU scores. This might be explained by the fact that the same English words can often be translated into many different ways in Japanese. Since METEOR allows for more lexical variation this may explain why the METEOR scores are so high.
- The BLEU scores vary between 12.55 (Systran) and 22.89 (PLuTO) (difference of 10.34). Looking at the METEOR scores, the difference is smaller: 51.05 (PLuTO) versus 42.88 (Systran) (difference of 8.17).
- PLuTO again outperforms Google and Systran, but the difference between the different engines is a lot smaller than for the reverse language pair. Scores are fairly low for all engines.

Human Evaluation

Adequacy

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

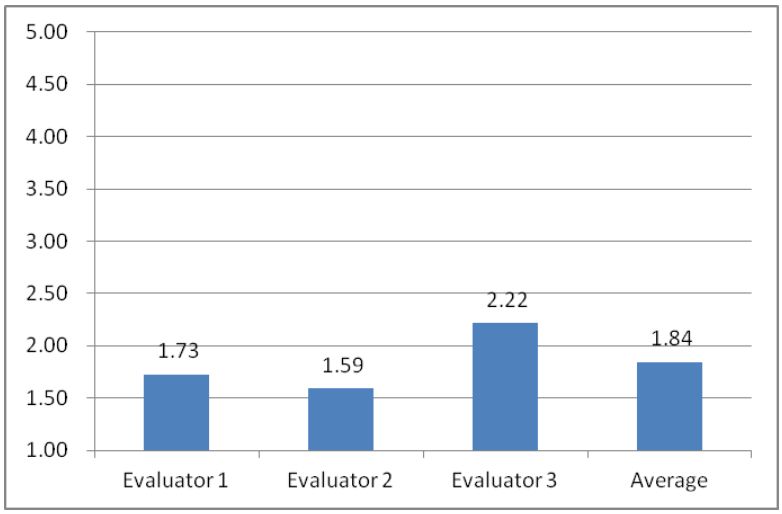

Figure 7: Human adequacy evaluation scores English → Japanese

Observations:

- There are slight differences between evaluators, but all evaluators seem to agree that the output is not very good.
- Average score for the English into Japanese language pair is 1.84, which is very low.

Figure 8 below shows the adequacy scores obtained for translations from Japanese into English:

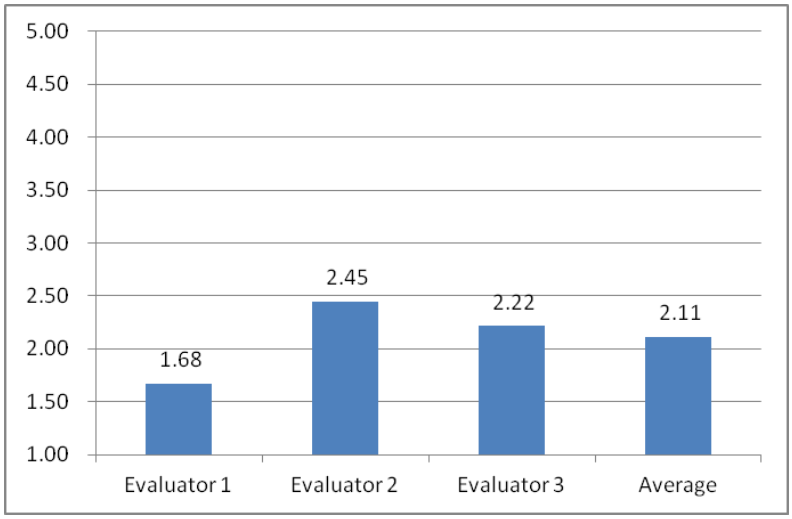

Figure 8: Human adequacy evaluation scores Japanese → English

Observations:

- Again, there are differences between evaluators, but again all evaluators seem to agree that the translations are not good.
- Average score for the Japanese into English language pair is 2.11, which is low.
- The human scores show the reverse tendency compared to the automatic scores: whereas the automatic scores indicate that the English into Japanese language direction is better, the human scores seem to say the opposite, although the difference between the two language directions is small (0.27).

Benchmarking

Figure 9 below shows how evaluators have ranked the PLuTO English into Japanese 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.

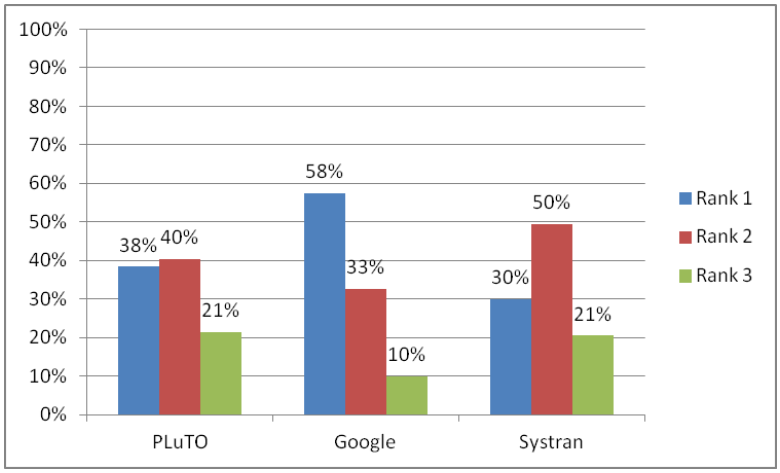

Figure 9: Human benchmarking evaluation English → Japanese

Observations:

- Evaluators clearly seem to have a preference for the Google output. It was selected as the best performing engine in 58% percent of the cases.
- PLuTO output comes in second and Systran output third, although the difference between the two engines is small.
- These results partly contradict the results of the automatic evaluation. While the automatic evaluation preferred the PLuTO output, the human evaluators seem to prefer the Google output.

Figure 10 below shows how evaluators have ranked the PLuTO Japanese into English output in comparison with the Google Translate and Systran output:

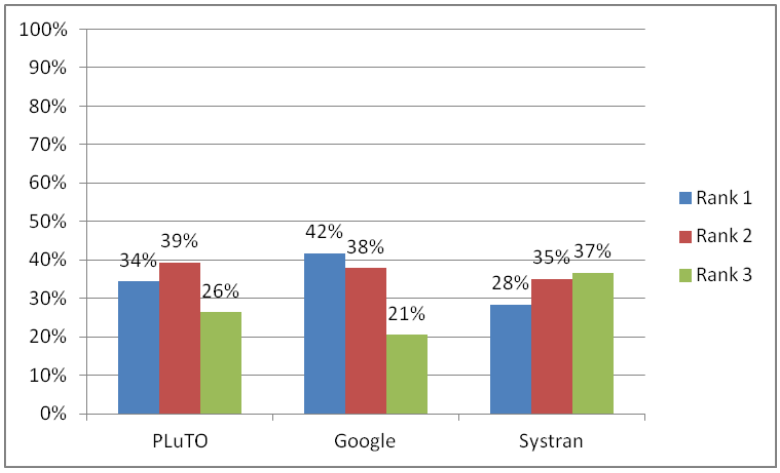

Figure 10: Human benchmarking evaluation Japanese → English

Observations:

- Evaluators seem to have a slight preference for the Google output. It was selected as the best in 42% of the cases. The difference between the various engines is a lot smaller than for the other language pairs, though. No engine seems to be clearly better than another.
- At first sight, the benchmarking results seem to contradict the automatic scores: whereas the automatic scores showed better results for PLuTO than for Google, the benchmarking result show the opposite. However, the automatic scores were generally fairly low for all language pairs and the differences between them were small. In that respect, one could say the automatic evaluation results are in line with the human benchmarking results: both show that there is no clear winner.

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

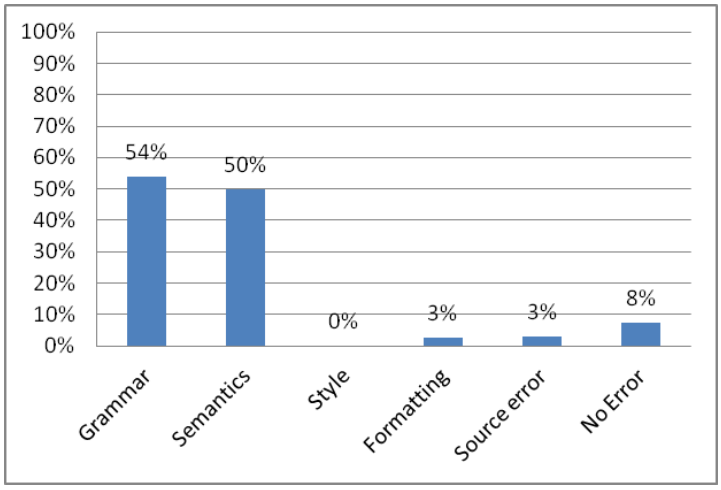

Figure 11: Error classification English → Japanese

Observations:

- Only 8% of all segments in the evaluation set had no errors at all. This figure is low, which is not surprising, considering the low automatic and human evaluation scores.
- The most common types of issues encountered in the output are grammatical issues (54%) and semantic issues (50%). See the examples below for more details.
- Only 3% of all segments in the evaluation set seem to suffer from formatting problems. This can possibly explained by the high error rates in the Grammar and Semantics categories: Errors in those categories are a lot more significant and the evaluator may therefore have focussed on those first.

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
The burn-in test can directly evaluate a dielectric film of an actual device, and the test reveals every defect cause including migration of an aluminum interconnection by applying high temperature and high	バーンインテストの実際のデバイスの誘電体膜を直接評価でき、テストの結果、アルミニウム配線のマイグレーションが各欠陥を含む高温高電界ストレスを印加	Grammatical error

Source Segment	MT Target	Error Type
electric field stresses.	した。	
Fig. 9 is a circuit diagram showing a conventional signal amplifying circuit of the MR element.	図９は、従来の信号増幅回路を示す回路図で、ＭＲ素子である。	Grammatical error
Fig. 2 shows an equivalent circuit of a single electron tunneling transistor.	図２は、単一電子トンネルトランジスタの等価回路を示す。	Semantic error
The fitting projection 60 allows fitting when the pivot plate portion 41 is installed from the right and left, and further facilitates positioning of the pivot plate portion 41.	嵌合凸部６０は、回動板部４１が設置されているときに嵌合することができるから、さらに左右のピボットプレート部４１の位置決めが容易である。	Semantic error

Table 7: Error type examples for English → Japanese

Figure 12 below shows how errors were classified for the Japanese into English language pair:

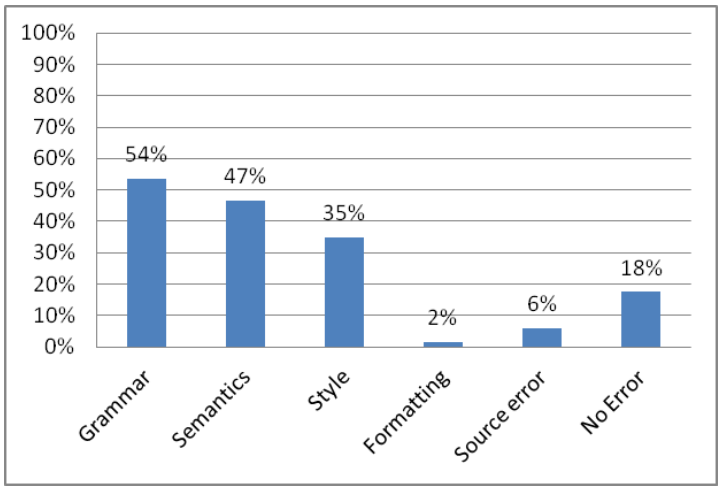

Figure 12: Error classification Japanese → English

Observations:

- 18% of all segments in the evaluation set had no errors at all, which is fairly high, considering the low scores that were generally given to the output.
- The most common types of issues encountered in the output are grammatical issues (54%), semantic issues (47%) and style issues (35%). See the examples below for more details.
- 6% of all segments in the evaluation set seemed to have issues in the source text already.

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
この機能は、主メモリセルアレイ１０３１の欠陥メモリセルの救済に利用されている。	This function is of the main memory cell array 10 31 is used to repair the defective memory cell.	Grammatical error
したがって、ウインドも動作しない。	Therefore, the window do not operate.	Grammatical error
以上，本発明の好適な実施の形態について，添付図面を参照しながら説明したが，本発明はかかる構成に限定されるものではない。	Thus, preferred embodiments of the present invention is described with reference to the accompanying drawings, but the present invention is not limited to such a configuration.	Semantic error
なお、孔版原紙Mが排版容器４０内に	The stencil sheet M is in plate delivery	Semantic error

Source Segment	MT Target	Error Type
搬送されたか否かは、排版センサ５３での投光部と受光部の間の光が再び投受光されることにより検知できる。	container 40 whether or not the conveyed to the discharge plate sensor 53 between light emitting portion and a light-receiving portion of the light is again by projecting and receiving light can be detected.	
次いで，本実施の形態の特徴である工程に入る。	Next, the present embodiment as a feature of the process is started.	Style error
更に、必要に応じて、老化防止剤や比重調整の充填剤として酸化亜鉛や硫酸バリウム等を配合することができる。	Further, if necessary adjusting specific gravity and the antioxidant and the filler such as zinc oxide or barium sulfate as can be incorporated.	Style error

Table 8: Error type examples for French → English

Discussion

At first sight, automatic and human evaluation results seem to be contradictory when it comes to indicating which engine is performing best: the automatic scores indicate that the PLuTO translations are closer to the reference translations than the translations from the other engines, but the benchmarking evaluation shows that the human evaluators seem to prefer the Google output over the PLuTO output.

Also when it comes to determining which language direction performs best, English into Japanese or Japanese into English, automatic and human scores seem to contradict each other: the automatic scores seem to indicate that the English into Japanese system is producing the best translations, while the human evaluations seem to suggest that the Japanese into English system is better.

Although there are differences between the automatic and human evaluations, the differences appear to be small, and all scores, be it automatic or human are very low. We would therefore be inclined to conclude that the differences are insignificant. The bottom line seems to be that neither of the language directions produces output that is good enough to be usable.

The error analysis underpins this analysis. Considering that over 50% of the evaluated sentences show grammatical problems and 50% also show issues of a semantic nature, it should come as no surprise that the translations score low on adequacy and fail to outperform competing MT systems.

On the other hand, we find that Google and Systran, too, are struggling with the English—Japanese language pair. There is no clear preference for any system for any of the language directions. This is not surprising: English—Japanese is generally acknowledged to be a difficult language pair for machine translation. The scores reported by the Kyoto Free Translation Task[1], for example, are in the same range (or even lower) than the scores we were able to obtain for this language pair.

[1] See http://www.phontron.com/kftt.

English—Spanish MT System

Automatic Evaluation

With the automatic evaluation for English—Spanish, no domain-specific test were used; only general scores are available. As explained above, new evaluation sets were created for this language pair, on the basis of more recent patent material. Unfortunately, sufficient material was not available for all domains, so only a general score was calculated, based on recent data in different domains.

Table 9 below shows the automatic scores obtained for translations from English into Spanish.

PLuTO		Google		Systran	
BLEU	**METEOR**	**BLEU**	**METEOR**	**BLUE**	**METEOR**
31.5	57.0	37.2	62.0	25.8	52.0

Table 9: Automatic scores English → Spanish

Observations:

- Overall, the METEOR scores are much higher than the BLEU scores. For the PLuTO output, this might be explained by the fact that the English—Spanish system was trained primarily with general domain data. The more general translations for specific terms might be forgiven by the METEOR metric but not by the BLEU metric. This does not explain why we see the same tendency with the other engines, though.
- The BLEU scores vary between 37.2 (Google) and 25.8 (Systran) (difference of 11.4); the METEOR scores vary between 62 (Google) and 52 (Systran) (difference of 10).
- Google seems to outperform PLuTO and Systran. PLuTO comes second, still clearly before Systran.

Table 10 below shows the automatic scores obtained for translations from Spanish into English:

PLuTO		Google		Systran	
BLEU	**METEOR**	**BLEU**	**METEOR**	**BLUE**	**METEOR**
32.0	34.0	35.0	36.0	26.81	32.0

Table 10: Automatic scores Spanish → English

Observations:

- Overall, the METEOR scores are somewhat higher than the BLEU scores, but the difference is not as outspoken as for the reverse language pair.
- The BLEU scores vary between 26.81 (Systran) and 35 (Google), (difference of 8.19). Looking at the METEOR scores, the difference is smaller: 36 (Google) versus 32 (Systran) (difference of 4).
- Google seems to outperform PLuTO and Systran. PLuTO comes second, still before Systran.

Human Evaluation

Adequacy

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

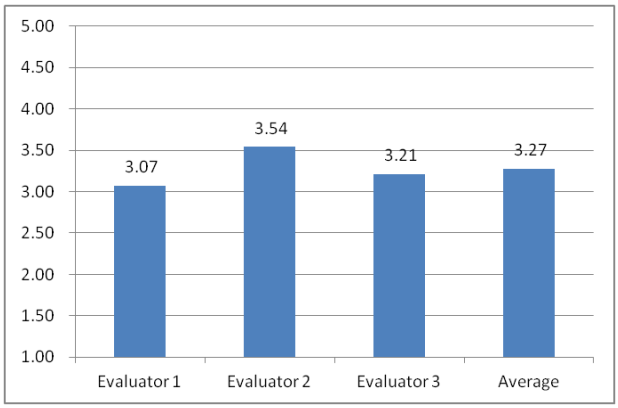

Figure 13: Human adequacy evaluation scores English → Spanish

Observations:

- There is some variation in the scores of the different evaluators, but generally speaking they show that the quality of the PLuTO output of the English into Spanish engine is fairly average. The score difference between the most positive and the most negative evaluator is 0.47.
- Average score for the English into Spanish language pair is 3.27, which is average to low.

Figure 14 below shows the adequacy scores obtained for translations from Spanish into English:

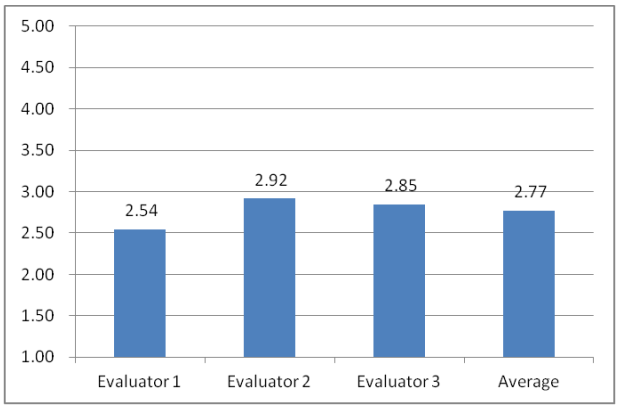

Figure 14: Human adequacy evaluation scores Spanish → English

Observations:

- There is some variation in the scores of the different evaluators, but generally speaking they show that the quality of the PLuTO output of the Spanish into English engine is rather poor. The score difference between the most positive and the most negative evaluator is 0.38 (Evaluator 1: 2.54 vs. Evaluator 2: 2.92).
- Average score for the Spanish into English language pair is 2.77, which is low.

Benchmarking

Figure 15 below shows how evaluators have ranked the PLuTO English into Spanish 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.

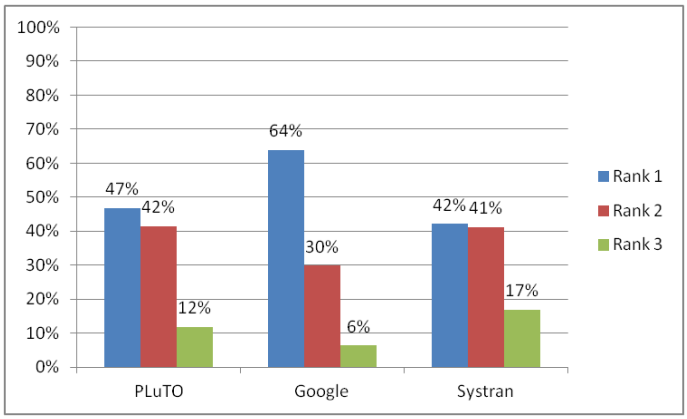

Figure 15: Human benchmarking evaluation English → Spanish

Observations:

- Evaluators clearly seem to have a preference for the Google output. It was selected as the best performing engine in 64% percent of the cases.
- PLuTO output is selected as the best in 47% of the cases, but the difference with Systran is minimal.
- These results confirm the findings of the automatic evaluation.

Figure 16 below shows how evaluators have ranked the PLuTO Spanish into English output in comparison with the Google Translate and Systran output:

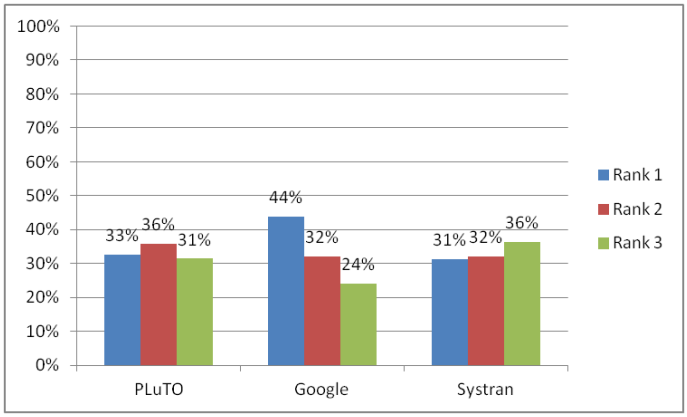

Figure 16: Human benchmarking evaluation Spanish → English

Observations:

- Evaluators seem to have a slight preference for the Google output, but the preference is not very outspoken.
- PLuTO output comes out only marginally better than Systran output.
- Again, these results seem to confirm the results of the automatic evaluation, which showed the same tendency. The scores are close. None of the systems seems to be doing a particularly good job.

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

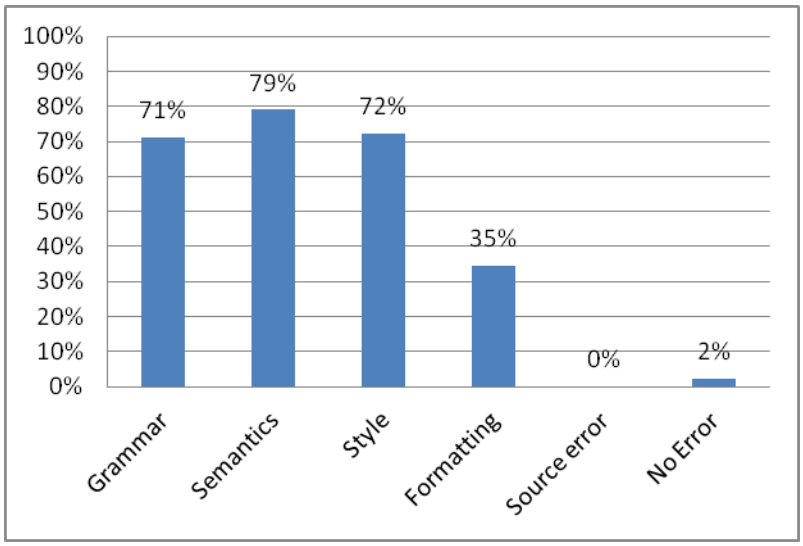

Figure 17: Error classification English → Spanish

Observations:

- Only 2% of all segments in the evaluation set had no errors at all.
- The most common types of issues encountered in the output were semantic issues (79%), style issues (72%), and grammar issues (71%). See the examples below for more details.
- The high rate of semantic and grammatical issues explains the low scores of the human evaluations.

Table 11 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
It also signals when the distal end is oriented correctly for insertion.	También es una señal cuando el extremo distal se orienta correctamente para inserción.	Semantic error
Said set comprises a pivoting handle (18) and a skid (13) for clearing obstacles and uneven parts.	Dicho conjunto comprende un mango pivotante (18) y un patín (13) para remover obstáculos y partes desiguales.	Semantic error
The frame (3) comprises longitudinal beams (30-30) and cross beams (31-31) in the form of sections, creating a self-supporting structure.	El bastidor (3) comprende unas barras longitudinales (30-30) y vigas transversales (31-31) en forma de segmentos, creando una estructura autoportante.	Style error
A biodegradable coffin that has a box and a flexible bag.	Un ataúd biodegradable que tiene una caja y una bolsa flexible.	Style error
Hence, the invention provides a mechanical interlocking system which provides optimum operation and minimises the number of components required.	Por tanto, la invención proporciona un sistema de interbloqueo mecanico que proporciona un funcionamiento optimo y minimiza el numero de componentes requeridos.	Grammatical error
The device, which is built into the structure of the machine, can be used to exert an instantaneous balanced force, the magnitude of which varies uniformly.	El dispositivo, que esta montado en la estructura de la máquina, pudiendo utilizarse para ejercer una fuerza equilibrada instantaneo, cuya magnitud varía de manera uniforme.	Grammatical error

Table 11: Error type examples for English → Spanish

Figure 18 below shows how errors were classified for the Spanish into English language pair:

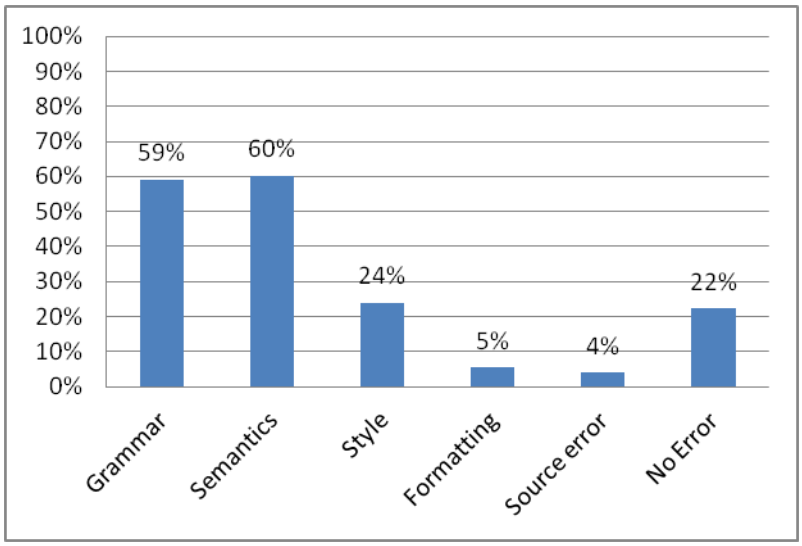

Figure 18: Error classification Spanish → English

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 (60%) and grammatical issues (59%). See the examples below for more details.
- Style issues are also quite common (24%).

Table 12 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
Maceta para plantas, formada por dos recipientes independientes de igual contorno, que son acoplables uno dentro de otro.	Pot for plants, formed by two independent compartments of equal contour, which can be coupled together within one another.	Semantic error
El aplicador incluye un mango (4) con un primer tramo fijo (7) articulado mediante un pivote (9) con un segundo tramo (8) telescópico.	The applicator includes a shaft (4) with a first fixed tranche (7) which is articulated through a driving pin (9) with a second section (8) telescopic.	Semantic error
De aplicación en los sectores alimentario, farmacéutico y cosmético y en el de la nanotecnología.	Of implementation in the sectors food, pharmaceutical and cosmetic and in the of nanotechnology.	Grammatical error
Así mismo, también se refiere a un método de determinación de angiogénesis patológica o de su progresión.	Likewise, also relates to a method of determining pathological angiogenesis or their progress.	Grammatical error
En la caja las esquinas están constituidas por columnas de apoyo.	In the housing the corners are constituted by supporting columns.	Style error
Además, un extensor magnético configurado para guiar el flujo magnético y evitar el escape de flujo está dispuesto sobre una superficie interna de la horquilla del rotor.	In addition, a spanner magnetic configured to guide the magnetic flux and prevent the escape of flow is disposed on an inner surface of the fork of the rotor.	Style error

Table 12: Error type examples for Spanish → English

Discussion

For both the English into Spanish system and the Spanish into English system automatic scores seem to correlate fairly well with the results of the human quality evaluations. Both the automatic and the human evaluation indicate that the quality of the Google output is better than that of PLuTO and Systran. However, whereas there is little difference between the automatic scores for both language directions (31.5 BLEU for the PLuTO English into Spanish output vs. 32.0 BLEU for the PLuTO Spanish into English output) we see a more significant difference in the human appreciation of the language directions (3.27 for the English into Spanish output vs. 2.77 for the Spanish into English output). It is not immediately clear what might explain this difference.

The benchmarking evaluation, on the other hand, does show the same trend in both the automatic and human evaluations: evaluators' preference for Google seems to be more outspoken for language direction English into Spanish than for Spanish into English. This shows both in the automatic scores (5.7 BLEU points difference for English into Spanish vs. 3.0 BLEU points difference for Spanish into English) and in the human evaluations (17% points difference for English into Spanish vs. 11% points difference for Spanish into English).

Although the difference is consistent, it is not that big. Comparing the scores we would say that the Google output might just be good enough to be usable, whereas the PLuTO output may just fall short, although it is probably a border case decision for both systems. Looking at the error types, it is clear that both language directions suffer from excessive grammatical and semantic mistakes. This would seem to corroborate the interpretation that both language directions need more work before they can actually be of use.

One of the explanations for the rather low scores of the PLuTO systems might be the quality of the training data. Whereas for the other languages pairs, the consortium possessed large amounts of data in the patent domain, this was not the case for the English—Spanish language pair. Only a limit amount of the data that was available to train the system was truly patent data. The vast majority of the data, however, was general domain data. It is a known fact that the degree of agreement between the training data and the data the user intends to translate with the trained system, has a direct impact on the output quality of the trained system. The better the training data resembles the input data, the better the output quality will be. This is clearly a criterion that was not met for this language pair; hence the lower evaluation scores.

English—Chinese MT System

Automatic Evaluation

As with the automatic evaluation for English—Spanish, no domain-specific test were used for the automatic evaluations of the English—Chinese language pair. Only general scores are available. As explained above, new evaluation sets were created for this language pair, on the basis of more recent patent material. Unfortunately, sufficient material was not available for all domains, so only a general score was calculated, based on recent data in different domains.

Table 13 below shows the automatic scores obtained for translations from English into Chinese.

PLuTO		Google		Systran	
BLEU	**METEOR**	**BLEU**	**METEOR**	**BLUE**	**METEOR**
17.4	40.0	20.6	45.0	8.4	26.0

Table 13: Automatic scores English → Chinese

Observations:

- Overall, the METEOR scores are much higher than the BLEU scores. The same trend was observed for language direction English into Spanish and Japanese into English.
- The BLEU scores vary between 20.6 (Google) and 8.4 (Systran) (difference of 12.2); the METEOR scores vary between 45.0 (Google) and 26.0 (Systran) (difference of 19.0).
- Google seems to outperform PLuTO and Systran, but none of the BLEU scores are very high.

Table 14 below shows the automatic scores obtained for translations from Chinese into English:

PLuTO		Google		Systran	
BLEU	**METEOR**	**BLEU**	**METEOR**	**BLUE**	**METEOR**
18.8	31.0	23.0	32.0	10.9	25.0

Table 14: Automatic scores Chinese → English

Observations:

- Overall, the METEOR scores are again higher than the BLEU scores, but the difference is not as outspoken as for the reverse language pair.
- The BLEU scores vary between 23.0 (Google) and 10.9 (Systran), (difference of 12.1). Looking at the METEOR scores, the difference is smaller: 32.0 (Google) versus 25.0 (Systran) (difference of 7).
- Again, Google seems to outperform PLuTO and Systran, but scores are roughly in the same range as for the reverse language pair. In other words: none of the engines scores very well.

Human Evaluation

Adequacy

Figure 19 below shows the adequacy scores obtained for translations from English into Chinese:

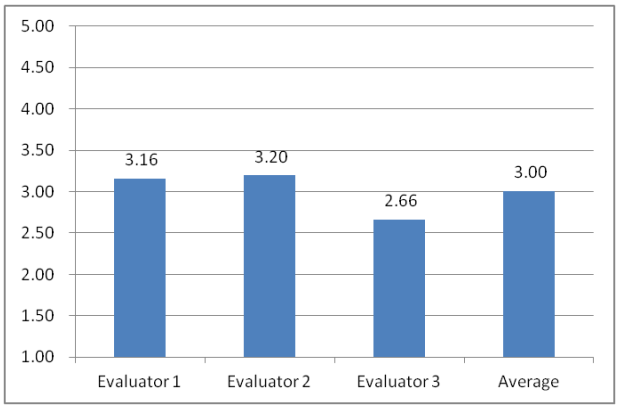

Figure 19: Human adequacy evaluation scores English → Chinese

Observations:

- The scores of the different evaluators vary between 3.20 (Evaluator 2) and 2.66 (Evaluator 3). This means a scores difference of 0.54.
- Average score for the English into Chinese language pair is 3.00, which is about average.

Figure 20 below shows the adequacy scores obtained for translations from Chinese into English:

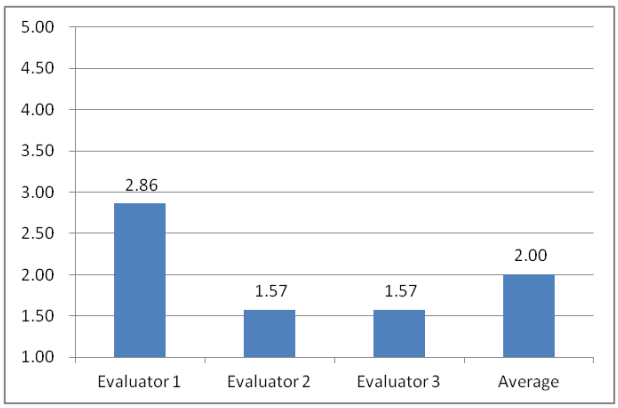

Figure 20: Human adequacy evaluation scores Chinese → English

Observations:

- The scores of the different evaluators vary between 2.86 (Evaluator 1) and 1.57 (Evaluators 2 and 3). This means a scores difference of as much as 1.29.
- Average score for the Chinese into English language pair is 2.00, which is fairly low.

Benchmarking

Figure 21 below shows how evaluators have ranked the PLuTO English into Chinese 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.

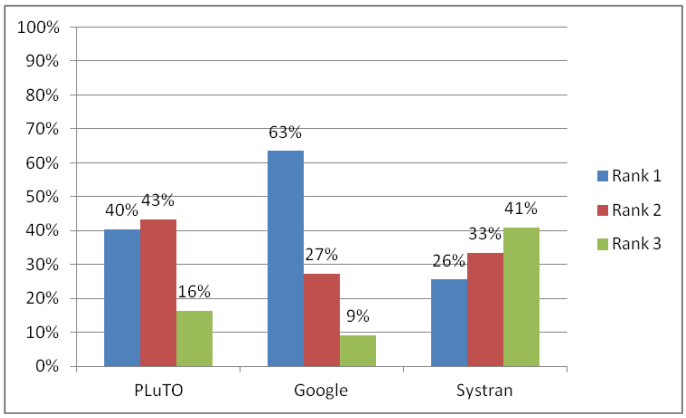

Figure 21: Human benchmarking evaluation English → Chinese

Observations:

- Evaluators seem to have a clear preference for the Google output. It was selected as the best performing engine in 63% percent of the cases.
- PLuTO output is the second best and still distinctly favoured compared to the Systran output.
- These results confirm the findings of the automatic evaluation.

Figure 22 below shows how evaluators have ranked the PLuTO Chinese into English output in comparison with the Google Translate and Systran output:

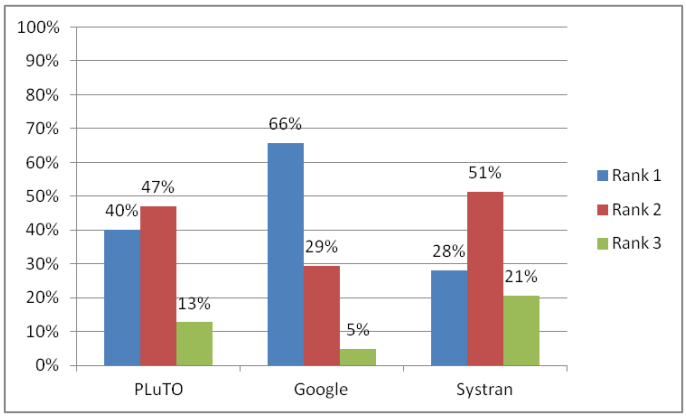

Figure 22: Human benchmarking evaluation Chinese → English

Observations:

- The picture is almost the same as for the reverse language pair: evaluators have a clear preference for the Google output. PLuTO output comes second and Systran third.
- Again, these results seem to confirm the results of the automatic evaluation.

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 23 below shows how errors were classified for the English into Chinese language pair:

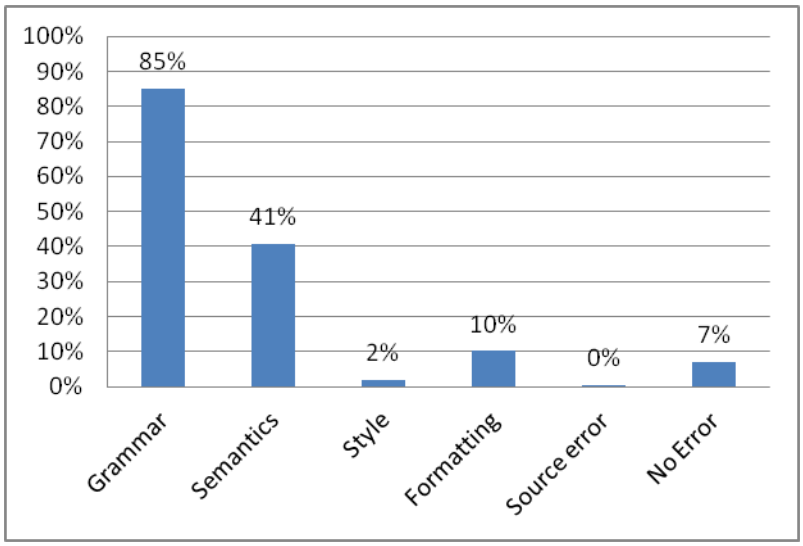

Figure 23: Error classification English → Chinese

Observations:

- The rate of semantic issues (41%) and especially grammatical errors (85%) is very high. See the examples below for more details.
- In spite of the low automatic and human scores and the high rate of grammatical and semantic errors, still 7% of all segments in the evaluation showed no errors at all.

Table 15 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
The present invention has a simple production process, quick gelation, and the gel pads made therefrom are smooth.	**本**发明的生产过程简单，快速胶凝，以及凝胶垫由其制成平滑的。	Grammatical error
A food processor with a safety protection device comprises a base (3) provided with a power unit.	**食品加工机具有安全保**护装置，其包括基座(3)设置有电源单元。	Grammatical error
STEEL CABLE PRE-TENSION DEVICE	预应力钢索装置	Semantic error
Disclosed are urea compounds represented by formula I or pharmaceutically acceptable salts, polymorphic forms, solvates or stereoisomers thereof; as well as preparation methods, intermediates and uses thereof.	**本**发明公开了脲式I**表示的化合物或其**药用盐、溶剂合物、多晶型或立体异构体；以及制备方法、中间体及其用途。	Semantic error

Table 15: Error type examples for English → Chinese

Figure 24 below shows how errors were classified for the Chinese into English language pair:

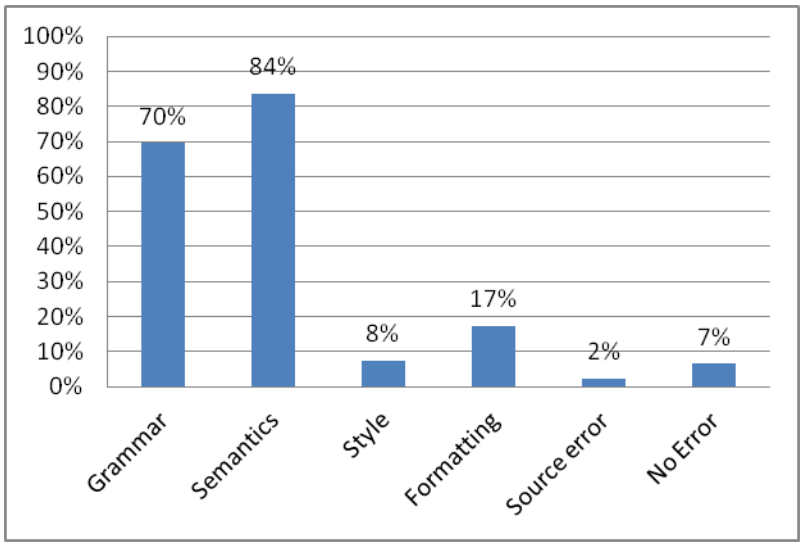

Figure 24: Error classification Chinese → English

Observations:

- Errors found in the translation output are predominantly of a semantic (84%) and grammatical (70%) nature. See the examples below for more details.
- As for the reverse language pair, a surprising 7% of the segments in the evaluation set showed no errors at all.

Table 16 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
人造咬胶皮、其制备方法及用途	Artificial biting rubber, its preparation method and use	Semantic error
本发明的化合物具有很好的水溶性和稳定性，在本类药物中本发明的化合物选择性好，具有很高的治疗指数。	The compounds of the invention have excellent water solubility and stability, within this class medicament compounds of the present invention good selectivity, with very high therapeutic index.	Semantic error
结构更加简单,**提高了**产品生产效率，降低了生产成本。	Even simpler structure, which increases the product production efficiency, reducing the production costs.	Grammatical error
整个空腔阻隔件结构简单、通用性强、减振降噪效果明显。	The entire cavity barrier arrangement simple, general more powerful, a vibration damping noise reduction are significant.	Grammatical error
本发明还提供了一种上下文重用的系统。	The present invention also provides A context reuse system.	Formatting error
一种串口波特率配置方法及相关设备、系统	A serial port baud rate configuration method, and associated apparatus, system	Formatting error

Table 16: Error type examples for Chinese → English

Discussion

As far as system preferences go, the automatic and human scores seem to correlate fairly well: both show that Google is scoring better than PLuTO and PLuTO better than Systran. They do not correspond so well in terms of absolute scoring: whereas the automatic scores suggest that the Chinese into English system provides better translations, the human adequacy evaluation indicates the opposite.

We think this might be explained by the type of errors that are found in the output. Both outputs show a high degree of grammatical errors, but there is a significant difference in the amount of semantic errors. Even though the amount of semantic errors is quite high in the English into Chinese output (41%), too, it is still a lot lower than the number of errors of this nature that is found in the Chinese into English output (84%). Since the adequacy of a translation first and foremost relates to the transfer of meaning, and since especially errors in the semantic category indicate problems having to do with meaning, the difference in prevalence of semantic issues in both outputs might explain why the adequacy of the Chinese into English translations is valued lower than that of the English into Chinese translations, although the automatic scores suggest otherwise. It should also be noted that the difference in the automatic scores that are reported for the different engines is not that big (less than 3 BLEU points for each of the engines).

Overall, the evaluations suggest that the output of neither of the language directions or systems is very usable. For example, even though the benchmarking evaluations showed that evaluators had a clear preference for the Google output, the small difference between the Google scores and the PLuTO scores in the automatic evaluations seem to suggest that chances are small that the Google output will be usable.

The scores that were obtained here for the English—Chinese language pair are not exceptional. Scores reported on the NIST website[2] for this language pair might be slightly higher (in the range of 25-30 BLEU), but they are still very much in the same order of magnitude. Taking into account that the patent domain is a particularly difficult domain for MT, it is not surprising that the scores we observed are a little bit lower than those reported by NIST. The bottom line is that the language pair English—Chinese continues to be a challenging one for statistical machine translation (and by extension for machine translation in general).

Productivity Evaluations

In the evaluations that have been discussed so far the focus has been on the translation quality of the systems as such, without immediately linking the results to any direct application of the systems. The productivity evaluations that we discuss in the sections below are of a more practical nature. With these evaluations we are trying to assess in how far translation suggestions generated by the machine translation systems can help translators to translate faster.

This assessment is made by having evaluators post-edit MT output on the one hand and have them translate other sentences from scratch on the other. Recording the time evaluators are spending on each of these tasks allows us to calculate what costs more time: correcting the MT output or providing translations from scratch. The productivity increase is reported as the ratio of average number of words per hour translated from scratch over average number of words per hour post-edited.

French—English

Results

Figure 25 shows the average throughput for translation and post-editing for all three evaluators.

[2] http://www.nist.gov/itl/iad/mig/openmt12results.cfm#chinese

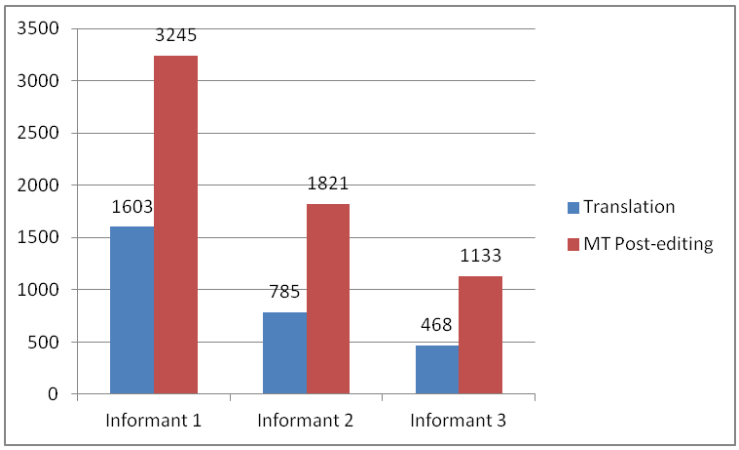

Figure 25: Translation vs. MT post-editing throughput French → English (in words/hour)

Table 17 shows the productivity increase that is obtained when the post-editing throughput is compared to the translation throughput.

	Productivity Increase
Evaluator 1	102%
Evaluator 2	132%
Evaluator 3	142%
Average	125%

Table 17: Productivity increase French → English system

Although the throughput of the three evaluators is quite different, we observe that their productivity increase is quite consistent. The fact that all three evaluators show a productivity increase of over 100% suggest that translators can translate twice as fast when they can use PLuTO MT as a starting point for their translations compared to when they would have to translate all sentences from scratch. The average productivity increase is 125%, which is very high.

Table 18 shows the similarity scores between the raw machine translation output and the post-edited version produced by the evaluators. The similarity score is a score between 0 and 100, where 0 means the outputs are completely different and 100 means they are exactly the same.

	Similarity Score
Evaluator 1	94.41
Evaluator 2	93.15
Evaluator 3	94.44
Average	94.00

Table 18: Similarity scores French → English system

The scores for the three evaluators are quite similar. They suggest that evaluators only made a minimum amount of changes, which is consistent with their high productivity increases.

Chinese—English

Results

Figure 26 shows the average throughput for translation and post-editing for all three evaluators.

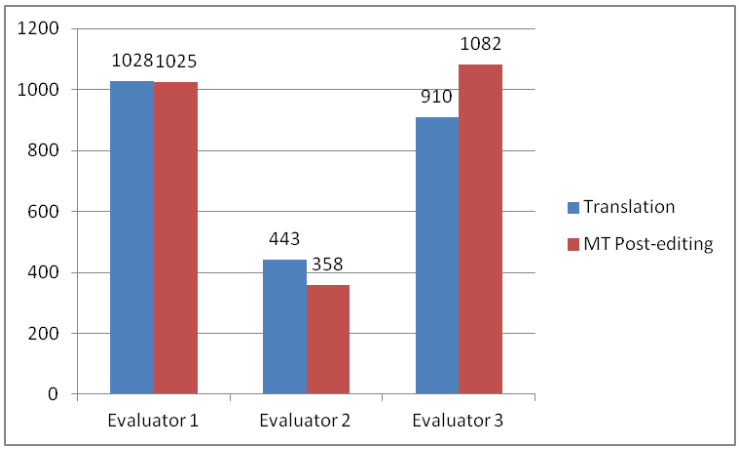

Figure 26: Translation vs. MT post-editing throughput Chinese → English (in words/hour)

Table 19 shows the productivity increase that is obtained when the post-editing throughput is compared to the translation throughput.

	Productivity Increase
Evaluator 1	0%
Evaluator 2	-19%
Evaluator 3	19
Average	0%

Table 19: Productivity increase Chinese → English system

Again, we note a considerable difference in absolute throughputs. Evaluators 1 and 3 translate more than double the amount of words per hour than Evaluator 2, regardless of whether they are translating from scratch or post-editing. More importantly, the results also show differences in productivity rates. Whereas the use of the PLuTO Chinese into English output seems to be beneficial to Evaluator 3, it seems to have a negative impact on the translation speed of Evaluator 2, who seems to go slower. For evaluator 1, there is no difference: he translates an equal amount of words, regardless of whether he has to translate them from scratch or can rely on a translation suggestion from MT.

Table 20 shows the similarity scores between the raw machine translation output and the post-edited version produced by the evaluators. The similarity score is a score between 0 and 100, where 0 means the outputs are completely different and 100 means they are exactly the same.

	Similarity Score
Evaluator 1	57.64
Evaluator 2	71.48
Evaluator 3	62.85
Average	63.99

Table 20: Similarity scores Chinese → English system

The similarity scores show that all three evaluators made quite some changes. This is in line with the low productivity increases we calculated: having to make a lot of changes slows the evaluators down and consequently results in low productivity increases. The amount of changes evaluators have been making, however, is not really consistent with their productivity increases. Based on the amount of changes he has been making, Evaluator 2 would seem to benefit most from the MT output.

However, the fact that his productivity increase is found to be the lowest (-19%) seems contradict this.

Discussion

Comparing MT evaluation results between different language pairs is always fraught with danger, especially, as is the case here, evaluation sets are different. We will therefore first look at the consistency of the various evaluations per language pair.

In deliverable D7.6, "First Report on Report on the Intrinsic and Extrinsic Quality of MT", we reported the evaluation scores for English—French. In that deliverable we reported good scores for language direction French into English (average of 56.26 BLEU (65.86 METEOR) across domains). The human adequacy evaluation confirmed these scores with an average score of 3.88 out of 5. Furthermore, the benchmarking evaluation showed that evaluator had a clear preference for the PLuTO output, even when compared to Google's French into English system, which is generally known to be of excellent quality.

Taking these evaluations into account, it is not surprising that the productivity evaluation yields good results. The good translation suggestions that the PLuTO systems provides allow evaluators to come up with a good translation a lot faster than if they were to translate the same sentence from scratch. Especially in the case of patent translation, where sentences tend to be longer than average, being presented with a good draft translation can be a distinct advantage for the translators. Our evaluations confirm this. The high similarity scores indicate that only relatively few changes had to be made to the machine generated translation suggestions that were provided. The low amount of changes allowed all evaluators to go at least twice as fast as they would go if they would have to translate from scratch.

The good correlation between the adequacy, benchmarking, and productivity evaluations lets us confidently state that using the French into English PLuTO systems as a translation aid for human translators will allow those translators to translate considerably faster than if they were to translate from scratch.

Along the same line of reasoning, the various results of the evaluations performed with the Chinese into English PLuTO system lead us to conclude that no productivity gains are to be had from using the output of this system as an aid for translators. Even though the outcome of the productivity evaluations as such does not seem to be decisive (with a slight productivity increase being shown with one evaluator and a slight decrease with another), we think that it is safe to say that no huge productivity benefits are to be expected with this language direction. In fact, the other evaluations, especially the low score of the human adequacy evaluation for the Chinese into English language direction (2.00), already suggested as much.

Discussion and Conclusions

In the course of the second project year, language pairs English--French and English--Portuguese were evaluated. The evaluation results for these language pairs were reported in deliverable D7.6, "First Report on Report on the Intrinsic and Extrinsic Quality of MT". For both language pairs, we were able to present good results: both automatic and human evaluations suggested that all engines for these language pairs were of good quality. When we compared the output of the PLuTO systems to that of leading providers of competing systems, such as Google and Systran, we found these good results confirmed: for all four systems that were evaluated, PLuTO output was preferred by evaluators over that of the competing systems.

During the third project year, one of these systems, the French into English one, was evaluated in a different application context, namely that of translation production. Again, the good quality of this system was confirmed. The productivity evaluation results suggest that translators will be able to translate at least twice as fast when they use the French into English translation output as a starting point as opposed to when they would translate from scratch.

In parallel with the existing systems being evaluated, new systems for new language combination were being built. Now, at the end of the project, eight more systems have been built and evaluated (English--German, English--Japanese, English--Spanish, and English--Chinese, all bi-directional). Unfortunately, the evaluations seem to indicate that the translation quality that the new systems are able to produce is not as high as that of the systems that were built during the first year.

Whereas our English--German systems still seemed to be performing quite well, beating Google's scores on all metrics and being ranked above Google's output by our human evaluators, the rather low adequacy score of the English into German system (2.42 out of 5) indicated that there was still room for improvement with this language pair.

The results for the remaining language pairs were rather disappointing: for all three language pairs (English--Japanese, English--Spanish, and English--Chinese), both automatic and human scores were low (English--Spanish) to very low (English--Japanese and English--Chinese), and for all three of these language pairs Google output was preferred over PLuTO output by our evaluators during the ranking evaluation.

We see a number of reasons why we did not seem to be able to keep up the same level of quality as with the initially built systems. First, there is the nature of the language pairs. As a response to the findings of the user survey (which were presented in deliverable D7.2, "First report on survey's results"), it was decided to replace a number of language pairs that we would initially build systems for (English--Dutch and English--Swedish) by language pairs that appeared to be more in demand by patent information specialists (English--Japanese and English--Chinese). This way, the Asian languages were introduced into the project. As literature and open MT competitions such as NIST show, these language pairs are not the easiest ones for MT. Automatic evaluation scores of 25 to 30 BLEU are very common for the English--Chinese language pair, with scores for English--Japanese generally being even lower. This level of automatic scores already puts into question the usability of these systems. A concern which appears to be confirmed when we look at the prevalence of semantic and grammatical errors in the output of these language pairs. And even though human evaluators seemed to express a clear preference for the Google translations, especially for the English--Chinese language pair, the automatic evaluations indicate that Google's engines for this language pair are not necessarily so much better than PLuTO's (ZH>EN: 23.0 (Google) vs. 18.8

(PLuTO) and EN>ZH: 20.6 (Google) vs. 17.4 (PLuTO)). This seems to indicate that none of the systems that are currently available is actually ready to fulfil a role in the patent search scenario.

Even though the Asian language pairs have proven to be challenging, we feel that it is still worth pursuing quality improvements. As the user survey has shown, these languages are clearly in demand, and any provider that would be able to produce usable translation output for these languages would have a clear competitive edge. In deliverable D5.1 we discuss a number of methods that have been tried to improve the English--Japanese systems.

A second reason that we see for the lower scores is the quality of the training data. This is particularly true in the case of the English--Spanish language pair. Whereas for all other language pairs, the consortium had large data set of good quality patent data at its disposal for training the MT systems, this was not the case for English--Spanish. For this language pair, only a limited section of the data was patent-specific. The majority of the data, however, was of a more general nature. It is a known fact that the performance of an MT system is directly related to quality of the training data and its degree of resemblance to the input data. Therefore, we think the rather low scores for the English--Spanish language pair might be explained by the disconnect that existed between our test set data (which was patent-specific) and our training data (which primarily consisted of general domain data).

In conclusion, we think that, at this stage, we should distinguish between language pairs that are ready for market (English--French, English--Portuguese) and language pairs that require more work (English--Chinese, English--Japanese, and English--Spanish). Results for English--German are inconclusive. In any case, comparisons with leading competing translation systems have shown that all PLuTO engines are able to produce translations of about 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. In addition, the productivity evaluation of the French into English language pair has demonstrated that for language pairs that score well on automatic and human evaluation, productivity gains are to be expected when these language pairs are used as a translation aid.

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 21: Adequacy Evaluation Guidelines