

Is MT Fit For Purpose?

**A Comparative Evaluation of Two Machine Translation
Systems**

**Bogdan Babych, Debbie Elliott, Tony Hartley
Centre for Translation Studies, University of Leeds
www.leeds.ac.uk/cts**

Contents

Is MT Fit For Purpose?.....	1
A Comparative Evaluation of Two Machine Translation Systems.....	1
Contents	1
1 Executive Summary.....	2
2 Objectives of the evaluation.....	3
3 Design of the two MT systems.....	3
4 Source texts	4
4.1 Linguistic features of the whitepaper.....	4
4.2 Linguistic features of the emails.....	4
5 Gold standard human translations	5
6 Dictionary updating	5
7 Evaluation procedure	6
7.1 Evaluators.....	6
7.2 Evaluations and materials.....	6
7.2.1 Emails: usability.....	6
7.2.2 Whitepaper: fluency	7
7.2.3 Whitepaper: adequacy	8
8 Results.....	9
8.1 Human evaluation of fluency, adequacy and usability	9
8.2 Impact of dictionary update.....	10
8.3 Human evaluation – frequency polygons.....	11
8.4 Human evaluation – inter-evaluator agreement.....	13
8.5 Automated scoring.....	14
8.5.1 Automatic evaluation – BLEU method.....	15
8.5.2 Automatic evaluation – LTV method.....	16
8.6 Correlation between human evaluation scores	17
9 Conclusions.....	17
10 Recommendations.....	18
11 References	19

1 Executive Summary

We evaluated the French-to-English versions of two rules based machine translation (MT) systems (referred to as s01 and s02 in this document) in order to assess the quality of their output and to determine whether updating the system dictionaries brought about an improvement in performance.

The input for the evaluation were a whitepaper from the European Commission and a collection of business and personal emails, both provided by Translution as typifying the anticipated translation and information needs of their target users.

The evaluations were performed by 70 judges –42 business people recruited by Translution and 28 postgraduate students from the University of Leeds who knew very little or no French.

Judgments were solicited on three attributes of text quality: Adequacy (meaning preservation), Fluency and Usability.

There is very high agreement among evaluators that MT preserves intelligible meaning (Adequacy), even if naturalness (Fluency) is impaired by translation. For Adequacy the most frequent score is always ‘five’ (the best).

Dictionary update increases the number of segments with this top score. The gain in performance varies from 25% for Usability to around 10% for Fluency and 3% for Adequacy.

We found that s01 performs better than s02 on both types of text – whitepaper and emails – and in terms of all three attributes of text quality evaluated. s01 gives a more stable quality across genres, while the quality for s02 is more dependent on the genre of the translated text: dictionary update greatly improves its quality for ‘easier’ texts, such as the emails, as compared to ‘hard’ texts.

However, the difference is not great and s02 after dictionary update exceeds the quality of translation achieved by s01 before dictionary update.

The commercial niche for MT is ‘assimilation’, where acceptable quality has been achieved (for Adequacy), but where using human translation is prohibitively expensive, so there would otherwise be no translation at all.

The human evaluations can be closely mimicked by two automated methods – BLEU and LTV – of which LTV proves superior. Both automated methods reliably reproduce the rankings of the two MT systems on the two text types and both before and after dictionary update.

Both s01 and s02 are capable of producing output consistently rated highly intelligible by human judges from the business community and by postgraduate students who do not know the source language (here French). The choice between the systems is likely to hinge less on the relatively small difference in output text quality than on the quality of the commercial relationship that Translution believes it can forge with one or other of the MT vendors.

2 Objectives of the evaluation

The evaluation focused on translation from French into English by two machine translation (MT) systems: s01 and s02. It was conducted with the following aims:

- assess whether the ‘raw’, unedited translations produced by each system are of sufficient quality for business people to use them as working documents
- establish whether the quality of the translations is improved by updating the dictionaries of each system in line with a benchmark provided by a human translation
- quantify any such improvement.

The evaluation targeted the following parameters of quality, both before and after updating of the systems’ dictionaries:

- adequacy – the extent to which the content of the original is preserved in translation
- fluency – the extent to which a translation reads naturally, as if written by a native speaker of English
- usability – the extent to which a translated document is judged to be usable in the workplace.

As such, the evaluation represents what is known as a ‘black box evaluation’: it looked only at the inputs to the MT systems (the source texts) and the output translations. It did not seek to identify causes of errors or to locate the point in the processing where the error occurred: this would be the objective of a so-called ‘glass box evaluation’.

The user-friendliness of the systems and the quality of the user documentation were also beyond the scope of the evaluation.

3 Design of the two MT systems

Both systems are ‘linguistic knowledge-based systems’, relying on rules which attempt to represent the grammars of the source and target languages (here French and English, respectively). This is in contrast to statistics-based systems that rely on statistical models.

In addition to their grammar rules, both systems are supplied with large dictionaries and offer the user the possibility of creating new, customised dictionaries to fill gaps or to override existing translations of words or expressions with a preferred variant.

Architecturally, the two systems are again similar. Both are based on the so-called ‘transfer’ model, which decomposes the translation process into three main stages:

- analysis – This phase is concerned with establishing an unambiguous ‘understanding’ of the source text; it is independent of any considerations of the target language.
- transfer – This phase is concerned with mapping the more or less abstract representation of the source text into a corresponding representation of the target text; it is inherently bilingual and contrastive.

- synthesis – This phase is concerned with transforming the abstract representation of the target text into a string of grammatically and semantically correct words and sentences; it is independent of any considerations of the source language.

This modular architecture is intended to ensure the re-usability of the analysis and synthesis modules for a particular language, whatever other language it is paired with.

Creating a new dictionary entry is, then, a matter of associating a word or expression in the analysis dictionary (or ‘source lexicon’) with a corresponding word or expression in the synthesis dictionary (or ‘target lexicon’). This association is stored in the ‘transfer lexicon’, which is thus specific to translation from a particular source language into a particular target language.

4 Source texts

Two types of document were used to evaluate the quality of English machine translations of French texts:

- The first 3,334 words (in the source text) of a European Commission Whitepaper on youth policy. This was divided into 120 text segments. Each segment comprised a complete sentence or heading, with the exception of very long or short sentences, which were split or merged as appropriate.
- A set of 36 French emails (24 business-related and 12 personal). The emails varied in length between 31 and 210 words, the average being 107 words.

These texts were supplied by Translution as being representative of the anticipated use of MT by the target business community.

All the original French texts were carefully checked and errors (often involving accents) were corrected before the source texts were used as input for machine translation.

4.1 Linguistic features of the whitepaper

This document presented a number of challenges to the MT systems.

- This contained many strings of noun and names of organisations, policies and treaties, which were not likely to be found in the system dictionaries, e.g. *Convention des Nations unies relative aux droits de l’enfant* – UN Convention on the Rights of the Child.
- Some sentences were very long and complex.

4.2 Linguistic features of the emails

The emails were expected to pose a different set of problems for the MT systems:

- abbreviated words: *ordi* - *ordinateur* (computer), *t’as* - *tu as* (‘you have’) and many others
- sound effects: *ouf*, etc.
- colloquial or new words: *biper* (‘to beep’), *tchatcher* (‘to yak’)

- colloquial phrases akin to speech, which would probably not be found in system dictionaries
- words foreign to the source language: eg. ‘main challenge’ which is already in English and should not be translated. (French *main* in means ‘hand’, which would trigger a mis-translation.)
- occasional long sentences, and instances of email writers simply forgetting to use full stops
- acronyms used in business, unknown by system dictionaries
- names of places, companies, etc.

5 Gold standard human translations

Translution supplied us with translations of all the texts produced by a professional translator. We used these as a gold standard reference for creating new dictionary entries. These human translations also figured in the evaluation exercise.

For the emails, we also had translations produced by a non-professional, French-speaking translator. This was intended to simulate a situation where, in the absence of MT, the author of the email would have to write in a foreign language (here English). We anticipated that the quality would be judged lower than the professional, native speaker translations.

6 Dictionary updating

We first generated translations of all source texts using the dictionaries supplied with the MT systems. We did not use any of the specialised dictionaries supplied with s01, because tests showed that none was appropriate for this set of texts.

We then created a new user dictionary for each system. Two native speakers of English, fluent in French and experienced in translation, each analysed the raw output from one MT system, along with the corresponding source text and gold standard translation. Candidate French terms for dictionary update were marked and their English equivalents were taken from the gold standard. To maintain consistency, the two judges discussed candidate entries and agreed on which terms to update or discard.

For both the whitepaper and the emails, the vast majority of items chosen for dictionary update were nouns. In the whitepaper, many of these were noun strings and named entities, the longest containing 10 words in the source text (*Politique en faveur de l'enfance et de la jeunesse* – Child and youth policy). Some or all of the components of French noun strings were often translated correctly by the MT systems (e.g. ‘the policy of the youth’). However, many of these terms appeared frequently in the text and adding them to the user dictionary would improve fluency and reduce the annoyance factor for the reader. Many nouns in the emails were updated because the item was incorrect in context (e.g. the preferred translation of *colocataire* was ‘flatmate’ rather than ‘joint tenant’, which appeared in the MT output).

Legitimate variants were not updated (e.g. *thème* was always machine translated as ‘topic’ even though the expert translation varied between ‘theme,’ ‘topic’ and ‘area’).

Words with more than one meaning were carefully considered. For example, *cours* appeared seven times in the emails and was translated by s01 as ‘course’ on each occasion, even though it should have been translated as ‘class’ or ‘classes’ some of the time. A decision was taken not to update this term.

The user-defined dictionaries were then set to take precedence over the default system dictionaries and a second batch of translation was generated.

7 Evaluation procedure

We used a combination of evaluations by human judges and two automated metrics developed at IBM and at Leeds.

7.1 Evaluators

Seventy evaluators were required to judge the emails for usability and the whitepaper for both fluency and adequacy.

Thirty business professionals, all volunteers, each evaluated a set of twenty-four emails. Business people were considered to be representative users of emails (two-thirds of which were business-related), and were deemed suitable judges for the usability of information contained within them.

Two sets of judges were required to evaluate fluency and adequacy respectively. Eight post-graduate students at the University of Leeds and twelve business professionals rated the whitepaper for fluency. Twenty post-graduate students rated adequacy. The majority of the students were in their mid-twenties and had very little or no knowledge of French. This was by design, as knowledge of the source language can influence judgements when words left untranslated by an MT system are understood by the evaluator.

7.2 Evaluations and materials

For all tasks, the evaluators entered their scores electronically, which enabled us to automate their collation and avoid all possibilities of transcription errors.

7.2.1 Emails: usability

Five translations of each email were evaluated: the four machine translations (s01 and s02 before and after dictionary update) and one translation by a French native speaker. The judges were not told that the texts were translations. The evaluation was designed to provide ten judgements for each machine translation of each email in order to be able to draw statistically valid generalisations about this essentially subjective task. The design of the evaluation meant that twenty scores were available for each human translation.

Initially, forty-eight emails were machine translated. The thirty-six emails containing the most items for dictionary update (across both MT systems) were selected for evaluation. This yielded a greater variation in quality between MT output before and after dictionary update.

Each email was treated as one unit of text to be scored. Six evaluator packs were compiled and five identical copies of each pack were produced to provide material for the thirty evaluators. Each pack contained twenty-four different emails. For each email, a judge would see and evaluate two of the machine translations and the translation by the French speaker. Materials were so designed that across the six different packs, each machine translation of each email would be rated by two different evaluators.

Given the number of texts per pack, each evaluation would take an estimated two hours, the same time as for the whitepaper evaluations described below.

For each email, translations by the four MT systems and one human were arranged across the packs so that the maximum number of orderings and combinations were presented. Furthermore, the twenty-four emails were placed in a different order in each of the six packs. This was done to neutralise the possible impact of tiredness, hunger, boredom, halo effects and other factors.

This is the usability task set to the evaluators:

Using each reference email on the left, rate the three alternative versions on the right according to **how usable you consider them to be for getting business done**.

Please **DO NOT** go back to a segment once you have made a judgement.

Give each alternative version a score of 5,4,3,2 or 1 where:

5 = The email is **as usable as** the reference email

1 = The email is **totally unusable**

7.2.2 Whitepaper: fluency

The four machine translations were evaluated for fluency by twenty judges. The judges were unaware that the candidate texts were translations. The method described below was selected in order to provide five fluency judgements for each segment of each machine translation.

Each of the four translations of the Whitepaper was divided into 120 segments (often sentences or headings). One quarter of the segments was then taken from each version to form an evaluator pack. This meant that each evaluator would judge all 120 segments in order, unaware that they came from four different sources. In this way, each evaluator would see (and intuitively compare) segments of varying quality instead of output from one system alone. Twenty different packs were made, so that each evaluator would see a different combination of segments.

Each evaluator pack was stored as a separate Word document. The segments were presented in the form of a table containing 120 rows with a scoring box adjacent to each segment. Student judges worked in a computer cluster and keyed their scores directly into the document. Business people followed the same procedure, working at their own convenience. The time taken for students to complete the evaluation varied between one and two hours.

This is the fluency task set to the evaluators:

Look carefully at each segment of text and give each one a score according to how much you think the text reads like fluent English written by a native speaker. Enter your scores in the appropriate boxes in the right hand column. Please DO NOT go back to a segment once you have made a judgement.

Give each segment of text a score of 1, 2, 3, 4, or 5 where:

5 = All of this segment reads like fluent English written by a native speaker.

1 = None of this segment reads like fluent English written by a native speaker.

NB Please bear in mind that this is a running piece of text and that it has been segmented in this way only for the purposes of this experiment.

7.2.3 Whitepaper: adequacy

Twenty students rated the four machine translations for adequacy. Again, judges were unaware that the candidate texts were translations. Packs were compiled in exactly the same way as the above fluency evaluations, resulting in five adequacy scores per segment. However, this time each segment was paired with the 'gold standard' human translation, referred to as a 'reference text'.

This is the adequacy task set to the evaluators:

For each segment, read carefully the reference text on the left. Then judge how much of the same content you can find in the candidate text. Enter your scores in the appropriate boxes in the right hand column. Please DO NOT go back to a segment once you have made a judgement.

Give each segment of text a score of 1, 2, 3, 4, or 5 where:

5 = All of the content is present

1 = None of the content is present (OR the text completely contradicts the information given on the left hand side).

NB Please bear in mind that this is a running piece of text and that it has been segmented in this way only for the purposes of this experiment.

At the beginning of the session, some evaluators asked whether they should take grammatical errors into consideration when scoring. Consequently, all judges were told to ignore ungrammaticality or disfluency and focus on content alone. Students again worked in a computer cluster. The time taken for this evaluation (bearing in mind that the reference translation doubled the amount of reading material) varied between one and a half and two and a quarter hours.

8 Results

8.1 Human evaluation of fluency, adequacy and usability

Figure 1 and Table 1 summarise the results of human evaluation for the three different evaluation tasks:

1. FLU – Fluency of the whitepaper translations (the 2 MT systems before and after dictionary update), judged by students (40%) and business users (60%)
2. ADE – Adequacy of the White Paper translation (the 2 MT systems before and after dictionary update), judged by students
3. USL – Usability of the E-mail translations (the 2 MT systems before and after dictionary update and a non-native speaker translation), judged by business users

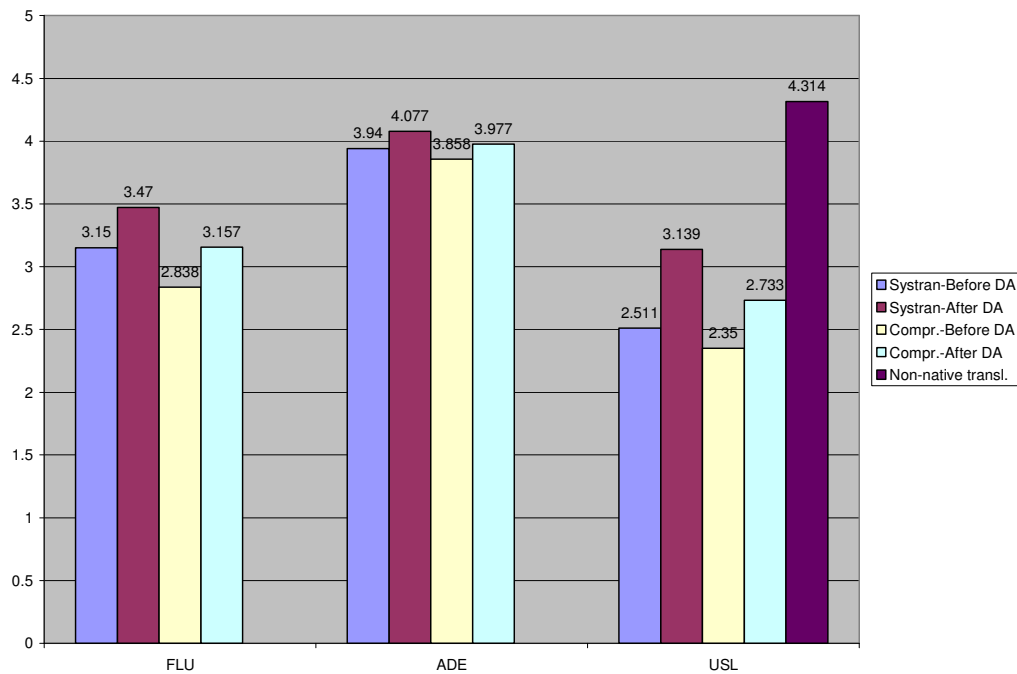


Figure 1. Human evaluation results

	FLU	ADE	USL
s01-Before DA	3.15	3.94	2.51
s01-After DA	3.47	4.08	3.14
s02-Before DA	2.84	3.86	2.35
s02-After DA	3.16	3.98	2.73
Non-native translation			4.31

Table 1. Human evaluation results

It can be seen from the figures that the results for Adequacy are very high: on average the MT systems scored ‘four’ on the five-point scale. The results for Fluency are less good: ‘three’ on the five-point scale is the most likely score for MT systems.

This shows that MT is useful primarily for the purposes of ‘assimilation’, i.e. where the users try to grasp the meaning and are less interested in having grammatically and lexically impeccable and stylistically natural sentences. Such qualities are important for the purposes of ‘dissemination’ (e.g. publication), for which MT is still not so good.

On the other hand, the Usability score has most probably integrated both ‘Fluency’ and ‘Adequacy’ aspects of text quality (and has perhaps been negatively influenced by the presence of the non-native human translation). It is natural to suggest that a text which is easier to read requires less effort on the part of the user to reconstruct the meaning. From the point of view of Usability, the Fluency and Adequacy errors aggravate each other, so the scores for Usability are lower than for the other two attributes.

8.2 Impact of dictionary update

All human scores for texts after dictionary update are consistently higher both for s02 and for s01, but the degree of improvement is different: it is the biggest for Usability of the e-mail translations (25% for s01 and 16% for s02), and the smallest for Adequacy of the White Paper translation (3.5% for s01 and 3.1% for s02). Table 2 summarises the ratios of improvement across systems after dictionary update for each task:

	FLU	ADE	USL
s01	0.101587	0.034772	0.250100
s02	0.112403	0.030845	0.162979

Table 2. Ratios of improvement after dictionary update

All corresponding scores for s01 are somewhat higher than for s02. s01’s advantage is in the range of 2.5% - 14.9%. The ratio of s01’s advantage as compared to s02 is summarised in Table 3:

	FLU	ADE	USL
Before DA: s01 ahead by	0.109937	0.021255	0.068511
After DA: s01 ahead by	0.099145	0.025145	0.148555

Table 3. Ratios of advantage of s01 over s02

However, the scores for s02 after dictionary update are as good as s01 before update (or slightly better – up to about 8.8% for Usability), which indicates that s02 is capable of reaching s01’s baseline quality. Table 4 summarises this difference:

FLU	ADE	USL
0.002222	0.009391	0.088411

Table 4. s02–after DA: advantage over s01–before DA

8.3 Human evaluation – frequency polygons

This section summarises counts of the scores received by each systems for each of the evaluated attributes of MT quality. Such information is important if we want to know the ratios of real scores given by individual evaluators to each of the compared MT systems. The following figures show how many times each system received the score ‘one’ (‘very bad’), ‘two’, ‘tree’ (‘average’), ‘four’, ‘five’ (‘very good’). Frequency polygons allow us to see most frequent scores given for each system in each case.

Figure 5 and Table 5 represent the frequency polygon of scores for Fluency; Figure 6 and Table 6 – for Adequacy and Figure 7 and Table 7 – for Usability.

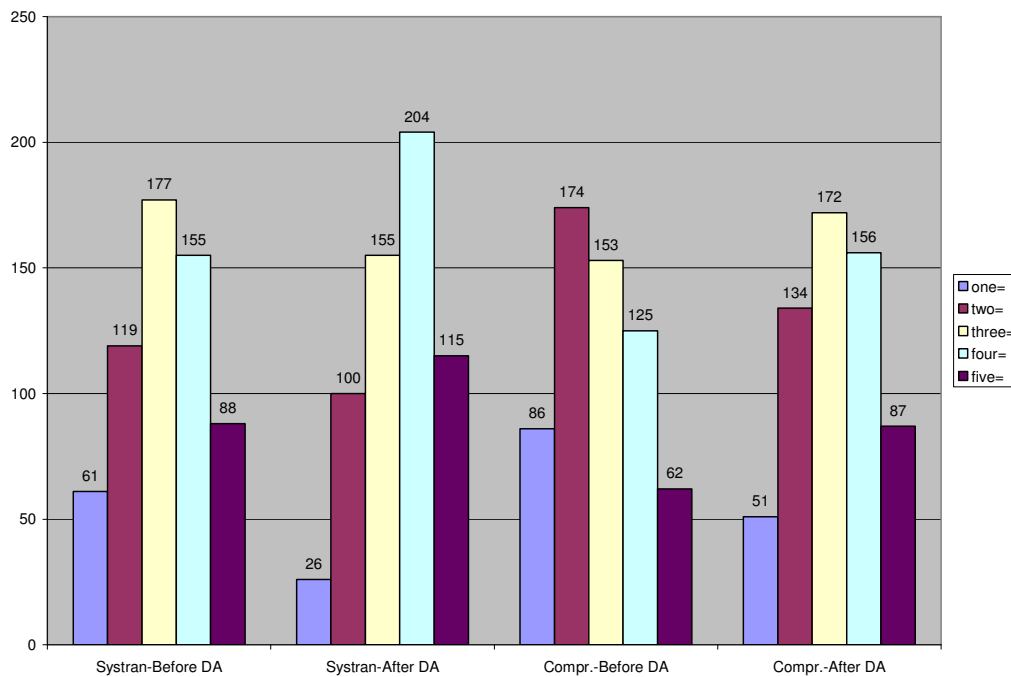


Figure 5. Frequency polygon – Fluency

	one=	two=	three=	four=	five=
s01-Before DA	61	119	177	155	88
s01-After DA	26	100	155	204	115
s02-Before DA	86	174	153	125	62
s02-After DA	51	134	172	156	87

Table 5. Frequency polygon – Fluency

For Fluency the most frequent score is different for each system. Dictionary update moves the ‘winning’ score up one point. For s01 before dictionary update ‘three’ is the most frequent score, and after update ‘four’. For s02 ‘two’ is the most frequent score before update and ‘three’ the most frequent score after. Dictionary update

clearly reduces the number of the scores ‘one’ and ‘two’ and increases the number of scores ‘five’ and ‘four’ for each system.

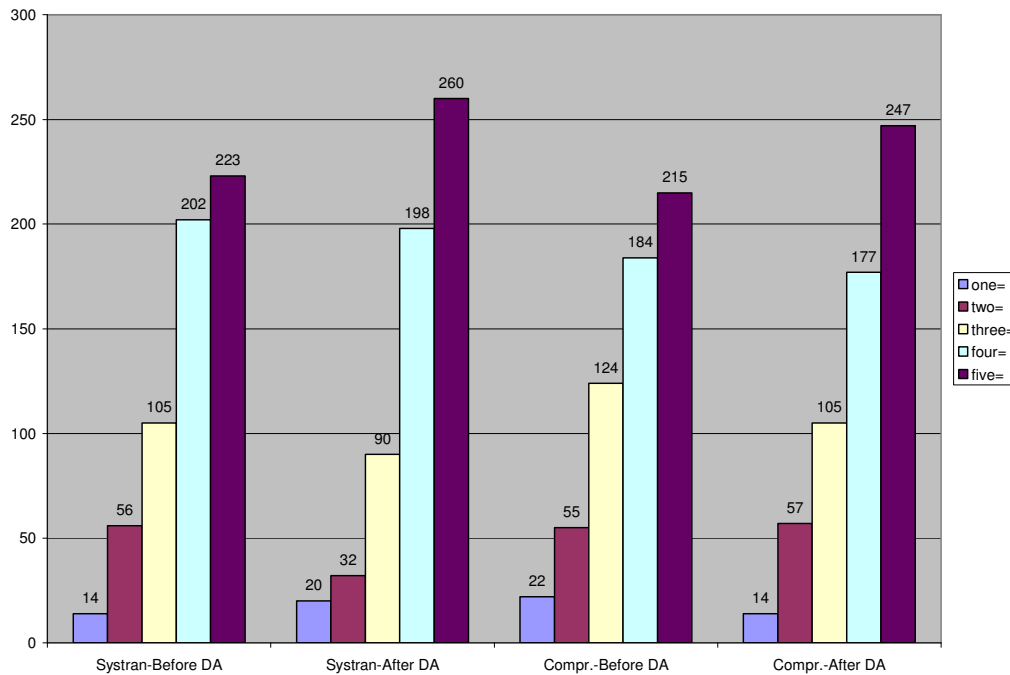


Figure 6. Frequency polygon – Adequacy

	one=	two=	three=	four=	five=
s01-Before DA	14	56	105	202	223
s01-After DA	20	32	90	198	260
s02-Before DA	22	55	124	184	215
s02-After DA	14	57	105	177	247

Table 6. Frequency polygon – Adequacy

The shape of the frequency polygon for Adequacy is different as compared to Fluency: the most frequent score is always ‘five’. Dictionary update increases the number of segments with this top score. This may be an illustration of the fact that Adequacy primarily supports ‘assimilation’. For this task MT systems are much more useful than for ‘dissemination’ tasks, where Fluency plays the most important role.

The fact that ‘one’ and ‘two’ form only a small proportion of all scores may explain why MT creates its own commercial demand: most of the scores make at least some sense for the readers of the text, so the majority of segments are indeed useful for readers who have in mind ‘assimilation’ tasks.

The commercial niche for MT is ‘assimilation’, where acceptable quality has been achieved (for Adequacy), but where using human translation is prohibitively expensive, so there would otherwise be no translation at all.

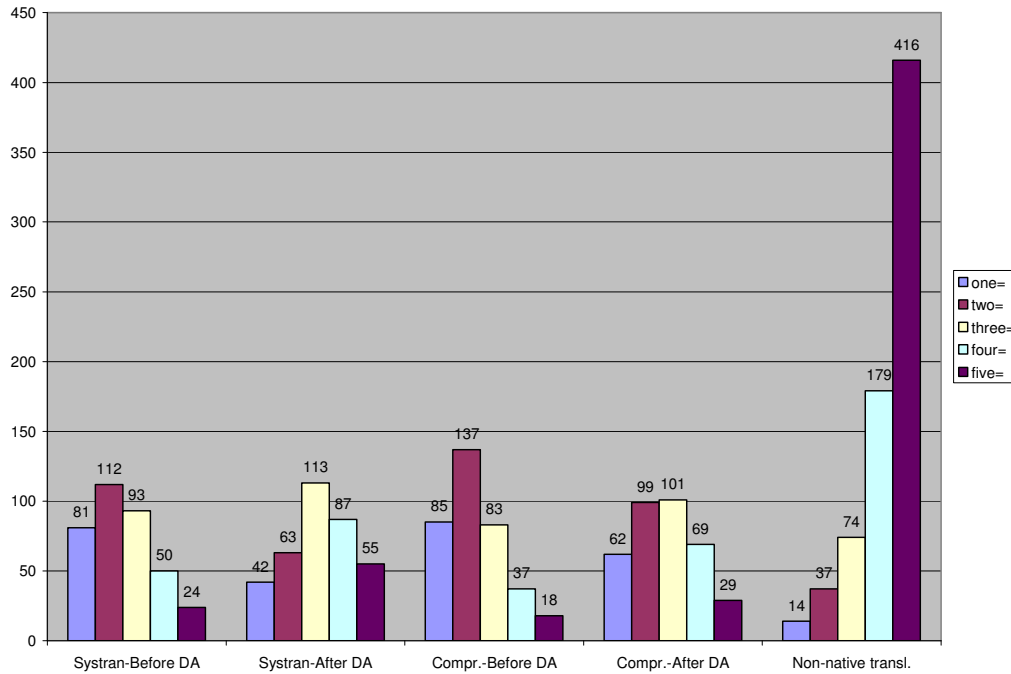


Figure 7. Frequency polygon – Usability

	one=	two=	three=	four=	five=
s01-Before DA	81	112	93	50	24
s01-After DA	42	63	113	87	55
s02-Before DA	85	137	83	37	18
s02-After DA	62	99	101	69	29
Non-native transl.	14	37	74	179	416

Table 7. Frequency polygon – Usability

The frequency polygon for Usability resembles the one for Fluency for all MT systems, and is much different for the non-native human translation. However, the most frequent scores are the same for both systems, and the dictionary update moves the most frequent score one point higher in both cases. Dictionary update improves the Usability scores most of all – much more than any other scores.

Frequency polygons for the Adequacy of MT output have a shape similar to the ‘ideal’ frequency polygon for Usability of the non-native human translation. This can be interpreted as the fact that MT Adequacy approaches the standards of human translation, but serious gaps in Fluency inhibit the Usability of MT.

8.4 Human evaluation – inter-evaluator agreement

Each segment (sentence) of the whitepaper document and each email were scored by five different people. In order to measure the degree to which human judges agreed with each other, we computed the standard deviation for each segment and took the average of these standard deviations for each evaluated system. A higher score represents a greater disagreement between human judges when evaluating that particular MT system. Figure 8 and Table 8 represent these average standard deviations for different judges of the same segment.

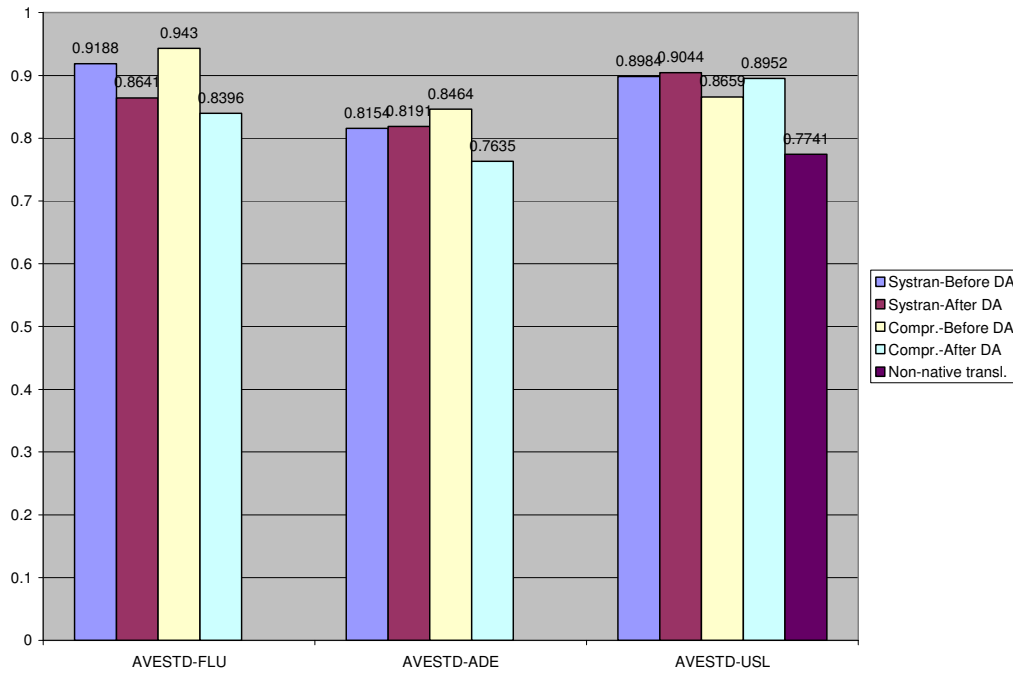


Figure 8. Inter-Annotator Disagreement: Average Standard Deviations for Evaluators

	AVESTD-FLU	AVESTD-ADE	AVESTD-USL
s01-Before DA	0.9188	0.8154	0.8984
s01-After DA	0.8641	0.8191	0.9044
s02-Before DA	0.943	0.8464	0.8659
s02-After DA	0.8396	0.7635	0.8952
Non-native transl.	-	-	0.7741

Table 8. Inter-Annotator Disagreement: Average Standard Deviations for Evaluators

The average standard deviation is smaller than one in all cases, which indicates that human evaluators usually disagree with each other by no more than one point on the five point scale.

8.5 Automated scoring

We applied two automated metrics to the translation data and established correlations between these results and the results of the human evaluations. In summary, we found that the automated scoring is consistent with the human judgements. This will allow us in future evaluations on similar data to dispense with the relatively expensive and time-consuming human judgments, or at least to scale them down.

8.5.1 Automatic evaluation – BLEU method

The BLEU¹ automatic evaluation metric proposed in (Papineni et al., 2001) has been shown to correlate strongly with human judgements on the Fluency of knowledge-based MT systems, which is further confirmed by the results presented here. The BLEU method is based on matches of N-grams (individual words or sequences of several words, usually up to 4) between the MT output and one or more human gold standard reference translations. More specifically, BLEU measures N-gram precision – the proportion of N-grams found both in the MT output and in any of the gold standard human reference translations.

The rationale of using BLEU is to explore objective properties of the evaluated texts as compared to a human reference translation. This gives an ‘absolute’ measure for comparison across different evaluation attributes, i.e. Fluency, Adequacy and Usability (which are not directly comparable through human scoring). The BLEU scores are in the range [0, 1].

The results of BLEU evaluation for the whitepaper document and for emails are summarised in Figure 9. BLEU used a single human reference translation and counted N-grams up to N=4.

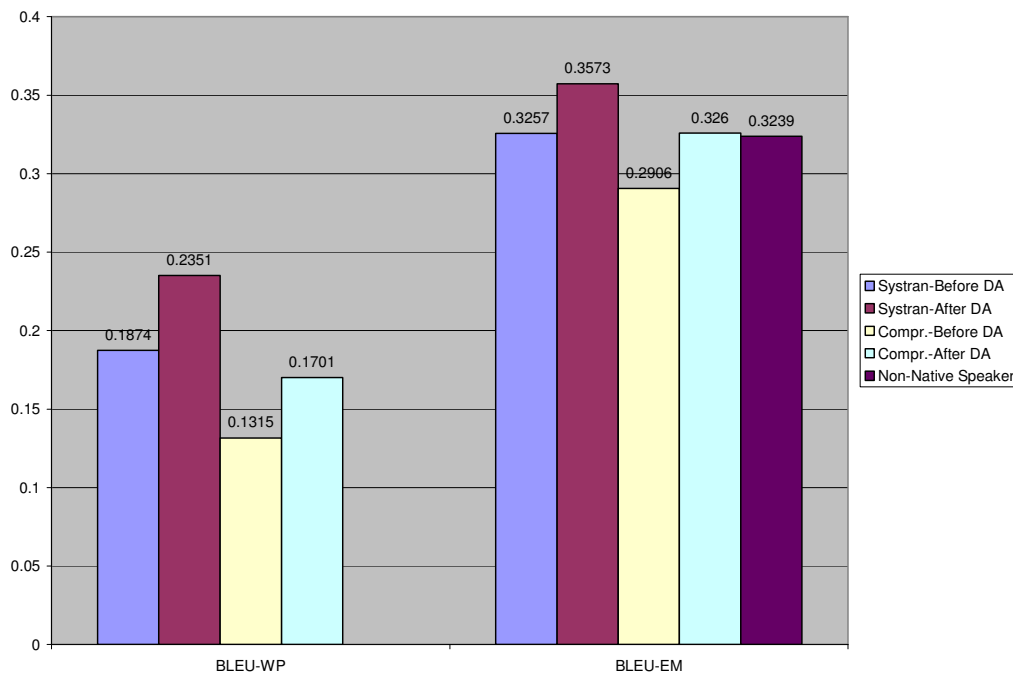


Figure 9. BLEU evaluation: whitepaper and emails

	BLEU-WP	BLEU-EM
s01-Before DA	0.1874	0.3257
s01-After DA	0.2351	0.3573
s02-Before DA	0.1315	0.2906
s02-After DA	0.1701	0.3260

¹ BLEU stands for BiLingual Evaluation Understudy

Non-Native Speaker

0.3239

Table 9. BLEU evaluation: whitepaper and emails

The BLEU evaluation results for the MT systems are consistent with the human scores: results after the dictionary update are better than before the update, and scores for s01 are somewhat higher than for s02; however, s02 is shown to be capable of reaching s01’s baseline quality (the quality before update) after its dictionary has been updated. The ratios of improvement and ratios of differences between systems are close to the ratios for human evaluation. This is an indication that human intuitive judgements about Fluency, Adequacy and Usability of MT output across systems and before and after dictionary update are confirmed by the objective criterion of precision of N-gram matches between the MT output and the ‘gold standard’ translation.

Another aspect of the BLEU evaluation is a possible comparison between the whitepaper text and in the business emails. There are much more matches of N-grams in the emails as compared to the whitepaper. Table 10 summarises the ratio of matches between these two types of documents.

s01-Before DA	0.737994
s01-After DA	0.519779
s02-Before DA	1.209886
s02 After DA	0.916520

Table 10. Ratio of N-gram matches in the E-mails over the White Paper

The table shows that translating emails is objectively easier for MT systems than translating the legal documents. However, human judges adjust the scores according to the evaluation task, so the difference becomes apparent only with automatic evaluation. In our experiment, since the human non-native translation was involved in Usability evaluation of the emails, a kind of ‘masking effect’ was introduced, so the scores for Usability were lower than for Adequacy or Fluency (where there was no comparison with the human translation). Therefore the BLEU score allows us to make comparisons between different types of texts which were not directly compared in our evaluation. It shows that translating emails is easier for MT systems and much better results are objectively achievable, in comparison with legal documents like the whitepaper.

Table 10 also shows the difference between the whitepaper and the email matches for s01 is lower than for s02 (74% and 52% vs 121% and 91%). This shows that s01 gives more stable quality across genres, while the quality for s02 is more dependent on the genre of the translated text: its quality is greatly improved for ‘easier’ texts, such as the emails, as compared to ‘hard’ texts.

8.5.2 Automatic evaluation – LTV method

The LTV (Legitimate Translation Variation) method as described in (Babych, 2004) is based on BLEU, but the matched words in the tested MT output and the ‘gold standard’ translation have unequal weight when they are matched. More weight is given to statistically significant words in the evaluated text. Statistical significance

weights, suggested in (Babych, Hartley, Atwell, 2003), are computed by contrasting the word’s frequency in a text and in the rest of the corpus: the formula is similar to the TF/IDF score used in Information Retrieval, but the scores are normalised by the relative frequency of the word in the corpus.

Usually content words – such as names of events, event participants, and terminology – happen to be more statistically significant. The intuition is that such words normally have a unique translation equivalent, whereas function words and other words which are less frequent in a given text than in the rest of the corpus, are subject to greater Legitimate Translation Variation, i.e. they will vary across independently produced human translations of the same text. Therefore, matches of the ‘significant’ words should count more when the MT output is evaluated. This is captured in the LTV method by assigning greater weights to words whose statistical significance score is greater than one.

The LTV scores slightly differ from BLEU, in particular they more closely match the human scores for MT Usability and Adequacy. LTV also indicates that emails are easier for MT than the whitepaper text.

LTV also suggests that for harder texts Fluency is much more affected than Adequacy.

8.6 Correlation between human evaluation scores

There is a strong positive correlation between human evaluation scores for both of the MT systems, both before and after dictionary update. The correlation is highest between Adequacy and Fluency of the whitepaper document. Usability of emails has the strongest correlation with translation Adequacy of the whitepaper document, and the correlation with Adequacy plus Fluency is somewhat weaker, and with Fluency alone is relatively the weakest. However, for all attributes the scores give the same ranking of the MT systems, so the correlation coefficient is above 0.95 in all cases. Table 11 summarises correlation figures for human evaluation scores:

	FLU	ADE	ADE+FLU
ADE	0.986296		
USL	0.944951	0.983143	0.957404

Table 11. Correlation between human evaluation scores

9 Conclusions

We draw the following conclusions from the evaluation experiment:

- s01 performs better than s02 on both types of text – whitepaper and emails – and in terms of all three attributes of text quality evaluated: Adequacy, Fluency and Usability.
- s01 gives a more stable quality across genres, while the quality for s02 is more dependent on the genre of the translated text: dictionary update greatly improves its quality for ‘easier’ texts, such as the emails, as compared to ‘hard’ texts.

- However, the difference is not great and s02 after dictionary update exceeds the quality of translation achieved by s01 before dictionary update.
- There is very high agreement among evaluators that MT preserves intelligible meaning (Adequacy), even if naturalness (Fluency) is impaired by translation. For Adequacy the most frequent score is always ‘five’ (the best).
- Dictionary update increases the number of segments with this top score. The gain in performance varies from 25% for Usability to around 10% for Fluency and 3% for Adequacy.
- This may be an illustration of the fact that Adequacy primarily supports ‘assimilation’ (intelligence gathering). For this task MT systems are much more useful than for ‘dissemination’ (publication) tasks, where Fluency plays the most important role.
- The fact that ‘one’ and ‘two’ form only a small proportion of all scores explains why MT creates its own commercial demand: most of the scores make at least some sense for the readers of the text, so the majority of segments are indeed useful for readers who have in mind assimilation tasks.
- The commercial niche for MT is ‘assimilation’, where acceptable quality has been achieved (for Adequacy), but where using human translation is prohibitively expensive, so there would otherwise be no translation at all.
- There is a strong positive correlation between human evaluation scores for both of the MT systems, both before and after dictionary update.
- The human evaluations can be closely mimicked by two automated methods – BLEU and LTV – of which LTV proves superior. Both automated methods reliably reproduce the rankings of the two MT systems on the two text types and both before and after dictionary update.
- The automated methods show that translating emails is objectively easier for MT systems than translating legal documents.

10 Recommendations

Both systems are capable of producing output consistently rated highly intelligible by human judges from the business community and by postgraduate students who do not know the source language (here French). The choice between them is likely to hinge less on the relatively small difference in output text quality than on the quality of the commercial relationship that Translution believes it can forge with one or other of the MT vendors.

For the evaluation of MT systems combining other pairs of source and target languages – whether from s01, s02 or other vendors – we would propose to rely primarily on the two automated scoring metrics, which have proven capable of mimicking human judgments to a high degree. We would still require a human gold standard reference translation but should be able to dispense with the time-consuming process of recruiting human judges and administering evaluations.

We believe that the automated techniques will allow us to derive a good picture of Adequacy, the attribute of text quality that underpins the ability of an MT system to successfully occupy the niche of translation for the purposes of assimilation of information.

11 References

- Babych, B; Hartley, A.; Atwell, E. Statistical Modelling of MT output corpora for Information Extraction. In: *Proceedings of the Corpus Linguistics 2003 conference*, edited by Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery. Lancaster University (UK), 28-31 March 2003. pp. 62-70.
- Babych, B. Weighted N-gram model for evaluating Machine Translation output. In *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*. University of Birmingham, 6-7 January, 2004. pp. 15-22.
- Papineni K, Roukos S, Ward T, Zhu W-J 2001 *Bleu: a method for automatic evaluation of machine translation*. IBM research report RC22176 (W0109-022) September 17, 2001