Noname manuscript No.
(will be inserted by the editor)

Involving Language Professionals in the Evaluation of Machine Translation

Maja Popović[1] · **Eleftherios Avramidis**[1] · **Aljoscha Burchardt**[1] · **Sabine Hunsicker**[2] · **Sven Schmeier**[1] · **Cindy Tscherwinka**[2] · **David Vilar**[1] · **Hans Uszkoreit**[1]

Received: date / Accepted: date

Abstract Significant breakthroughs in machine translation only seem possible if human translators are taken into the loop. While automatic evaluation and scoring mechanisms such as BLEU have enabled the fast development of systems, it is not clear how systems can meet real-world (quality) requirements in industrial translation scenarios today. The TARAXŰ project has paved the way for wide usage of multiple machine translation outputs through various feedback loops in system development. The project has integrated human translators into the development process thus collecting feedback for possible improvements. This paper describes results from detailed human evaluation. Performance of different types of translation systems has been compared and analysed via ranking, error analysis and post-editing.

Keywords machine translation · human evaluation · error analysis

1 Introduction

Translation is a difficult task – even for humans. Machine translation (MT) quality has improved greatly over the last years, nevertheless the evaluation of machine translation output is intrinsically difficult as well. A widely used practice for MT evaluation is to let human annotators do ranking of outputs by different machine translation systems. While this is an important step towards an understanding of their quality, it does not provide enough scientific insights.

[1] DFKI – Language Technology Lab
Berlin, Germany
firstname.lastname@dfki.de

[2] euroscript Deutschland
Berlin, Germany
firstname.lastname@euroscript.de

This paper describes the results of two large-scale human evaluation rounds carried out in the framework of the TARAXŰ[1] project. The approach arises from the need to detach MT evaluation from a pure research-oriented development scenario and to bring it closer to the potential end users in translation industry. Therefore, evaluation rounds have been performed in close co-operation with several Language Service Providers (LSP). The evaluation process has been designed in order to answer particular questions closely related to the applicability of MT within a real-time professional translation environment. All evaluation tasks have been performed by qualified professional translators.

To our best knowledge, no previous work has combined the aspects of professional human assessment of machine translation outputs including ranking, post-editing and error analysis. The shared tasks of the workshops of machine translation WMT (e.g. [Callison-Burch et al (2010)]) included human ranking of MT outputs done by volunteers rather than professional translators. In [Vilar et al (2006)], a framework for human error analysis and error classification has been proposed and has become widely used in the machine translation community. [Specia and Farzindar (2010)] estimate post-editing effort based on professionally post-edited data. [He et al (2010)] compare the preferences and behavior of professional post-editors when offered TM and MT outputs. [Farzindar and Lapalme (2009)] investigated statistical machine translations of legal texts and presented results of a human evaluation in the form of edit distances and numbers of post-edit operations.

However, no study has been carried out yet which puts all these aspects together.

2 Human evaluation design

Several large-scale human evaluation rounds have been performed by the TARAXŰ project focussing on different aspects such as translation quality, analysis of translation errors and post-editing effort. The first two rounds and the results are presented in this paper. The involved languages were German, English, Spanish, French and Czech. The evaluation tasks were performed by external Language Service Providers, as they offer human translation services and act as experts.

2.1 Translation systems used

The translation outputs evaluated during the rounds presented in this work are produced by German-English, Czech-English, German-French and German-Spanish machine translation systems in both directions as well as one Translation Memory System (TMS) widely used in translation industry. While a TMS does not generate translations per se, it is a valuable instrument for

[1] http://taraxu.dfki.de/

storing translations that have been manually created and validated during a translation job, and their use a de-facto standard in translation and localisation industry. New source language text to be translated is compared to the source language segments in the memory and a match value defining the degree of similarity is computed. If the sentence to be translated is identical to a source language segment in the Translation Memory (TM), this is a 100% match and in the best case also a perfect translation in the given translation context and no editing effort is needed. Differences in wording and formatting lower the match rate according to their severity, which depends partly on user's settings and the algorithm itself. In case of so called fuzzy matches (matches below 100%), the user is presented with translations of the respective similar source language sentences and the translator edits the target language segment to reach a good translation. The source language sentence and the translation are stored as a new translation unit in the memory. We decided to set the threshold as low as possible to be able to define the boundary below which a translation suggestion from the memory would not be of use for the translator anymore - meaning that editing a translation suggestion would require more time than translate from scratch. In the end there were not enough Trados matches to define that but experience shows that matches below 75-85% are not useful for the translator with respect to editing time. Although that differs from memory to memory depending on quality and maintenance effort taken.

The corpora used for the MT systems consist of two domains: news texts taken from WMT tasks [Callison-Burch et al (2010)] and technical documentation extracted from the freely available OpenOffice project [Tiedemann (2009)].

A TMS comes with no content in its memory and is filled manually during translation. As such the TMS employed in the TARAXŰ project would be of little value for the evaluation. To mirror a professional translation environment the TMS has therefore been filled with the same parallel data that was used to train the Moses system.

In the first round, four different translation systems were considered:

Moses [Koehn et al(2007)]: a phrase-based statistical machine translation (SMT) system.

Google Translate[2]: a web-based machine translation engine also based on statistical approach. Since this system is known as one of the best general purpose MT engines, it has been included in order to allow us to assess the performance level of our Moses system and also to compare it directly with other MT approaches.

Lucy MT [Alonso and Thurmair (2003)]: a commercial rule-based machine translation (RBMT) system with sophisticated hand-written transfer and generation rules, which has shown good performance on previous shared tasks.

Trados[3]: a professional Translation Memory System (TMS); Trados comes with a sophisticated user interface in which, depending on the user's set-

[2] http://translate.google.com/

[3] http://www.trados.com/en/

tings, translation suggestions are displayed. According to the users's settings translations up to a certain percentage are replaced automatically or source language content is kept and a correct translation is filled in manually by the translator. In case there exist user-specific dictionaries, the user can decide to replace matching entries automatically in the content, too, which then also applies to matches below the defined match rate.

In the second round, two more systems were investigated:

Jane [Vilar et al (2010)]: a hierarchical phrase-based SMT system.
Rbmt: Another widely used commercial rule-based machine translation system whose name is not mentioned here.[4]

The obtained outputs were then given to the professional human annotators in order to perform several sentence-level evaluation tasks using the browser-based evaluation tool Appraise [Federmann (2010)].

3 First evaluation round

In the first evaluation round, the systems were prepared as follows:

Lucy MT: No special adaptation.
Moses: Trained on the standard Europarl and News corpora of WMT 2010 containing about 1.7 million parallel sentences. 5-gram language models were trained on 1.9 million target language sentences from these corpora.
Google Translate: not available
Trados: Automatically filled with the same data as Moses. rados performs at a disadvantage when the texts to be translated vary a lot from the data in the Translation Memory as is the case for the news texts used in our evaluation. To level the field, 100 sentences of these test data had been manually edited to reach a high matching degree for at least this part of the data. [5]

As test sets, 1000 sentences from the News domain and 400 sentences from the OpenOffice domain were translated by the described systems. The following three tasks were carried out on produced German-to-English, Spanish-to-German and English-to-German translation outputs:

Ranking: *rank the four different MT outputs* relatively to each other, from best (rank 1) to worst (rank 4). Ties were not allowed, i.e. the ranks had to be always 1,2,3,4 although perhaps the outputs with ranks 2 and 3 are of the same quality (or are even identical).

[4] For reasons of required anonymisation.
[5] Note that this must be seen as an experiment. This was done in order to simulate the use of TMs, although it does not mirror the exact use of them in the translation industry.

Table 1 Human ranking results in the first evaluation round as the average position of each system in each task.

human ranking	Lucy	Moses	Google	Trados
Overall	2.00	2.38	**1.86**	3.81
News	**2.52**	2.59	2.69	3.78
OpenOffice	1.72	2.77	**1.56**	3.95
de-en	2.01	2.46	**1.73**	3.80
es-de	**1.85**	2.42	1.99	3.84
en-de	2.12	2.28	**1.89**	3.81

Error classification: select the best ranked translation output and *define the two main types of errors* (if any) in it. A following subset of the error types suggested by [Vilar et al (2006)] is used: missing content word(s), wrong content word(s), wrong functional word(s), incorrect word form(s), incorrect word order, incorrect punctuation and other error.

Post-editing: select the translation output *which is easiest to post-edit* (which is not necessarily the best ranked) and perform the editing. The translators were asked to perform only the minimal post-editing necessary to achieve acceptable translation quality.

It should be noted that the Google Translate system was not considered as an option for error classification and post-editing. We took this decision in order to avoid futile efforts because we have no way to influence on improving this system. In case Google was the best ranked system, the translators were offered the second ranked system for the classification task, whereas for the post-editing task they could choose among ranks 2, 3 and 4. In order to speed up the evaluation process, the tasks were carried out separately. Each task was split into multiple parts to allow for parallel processing. The results of the first round are presented in the following sections.

3.1 Ranking

The results for the ranking task are shown in Table 1. The first row presents the overall average ranks for the four listed systems, **bold face** indicating the best system. Furthermore, the results are presented separately for each domain, namely news and technical domain, as well as for each translation direction, namely German-to-English, Spanish-to-German and English-to-German. The domain results are the average over all language pairs and the language pair results are the average over the two domains.

The average ranks of the machine translation systems are comparably close. There is not one single best system. This is an indication that a hybrid system could be a good option. A noticeable result is that Google performs worst on the WMT corpus although the data should – in principle – have been available online for training. This might, however, explain the good perfomance of this web-based system on the OpenOffice corpus. On the other hand, for the OpenOffice task Moses showed the worst performance – the reason is that

Table 2 Average BLEU scores (%) for each system and each task.

BLEU (%)	Lucy	Moses	Google	Trados
Overall	15.6	15.0	**20.2**	2.9
News	14.1	16.3	**17.7**	2.1
OpenOffice	19.5	12.0	**26.6**	4.9
de-en	22.7	18.8	**29.8**	5.0
es-de	12.3	12.9	**15.9**	1.8
en-de	13.6	14.6	**17.2**	2.5

Table 3 Human error classification in the first evaluation round: overall percentage of most severe errors for each translation system.

	Lucy	Moses	Trados
Missing content word(s)	3.2	16.8	12.6
Wrong content word(s)	34.6	24.6	33.2
Wrong functional word(s)	18.6	11.8	11.0
Incorrect word form(s)	13.1	14.6	9.1
Incorrect word order	16.1	22.0	13.4
Incorrect punctuation	3.7	3.4	2.1
Other error	10.8	6.7	18.6

it has been trained only on the WMT data, which is out of domain in this case.

Table 2 shows the average BLEU scores [Papineni et al (2001)] for illustration. The main difference is that the Google Translate system is the best one for each language pair and each task. This can be expected, since in several WMT evaluation tasks it is shown that the correlation between the BLEU score and the human rankings is not particularly high, mainly because the BLEU score is biased towards statistical systems thus underestimating rule-based systems (such as Lucy).

3.2 Error classification

The results of the shallow human error classification are presented in Table 3. The most frequent errors that the judges considered as most important in all systems are *wrong lexical choices* (wrong content and functional words), and the next frequent error type is *incorrect word order*. This indicates the need for improvement of reordering and lexical choice techniques for all translation approaches. Another interesting observation is the very low number of missing content words for the Lucy system.

3.3 Post-editing

The human post-edits were used a) to study the difference between selecting sentences for post-editing and ranking and b) to get further insight into the errors the system made.

Table 4 Example of discrepancy between ranking and post-editing in the first evaluation round: the worst ranked sentence is chosen for post-editing.

Rank	Translation output
1	Our experience shows that the majority of the customers in the three department stores views not more at all on the prices.
2	Our experience shows that the majority of the customers doesn't look on the prices in the three department stores any more.
3	Our experience shows that the majority of the customers does not look at the prices anymore at all in the three department stores.
4	Our experience shows that the majority of customers in the three Warenhäusern do not look more on prices.
Edited	Our experience shows that the majority of customers in the three department stores no longer look at the prices.

3.3.1 Ranking for translation quality vs. selection for post-editing

A central question that is to be answered by the evaluation round is whether there is a difference between those sentences that are ranked best (i.e. that are the best MT result) and those sentences that are chosen by human professionals as the easiest for post-editing. It has been shown that 74% of those hypotheses selected for post-editing were ranked as the best or the second best in the ranking task. 20% were ranked third and 6% had the worst rank. An example of discrepancy between the "best ranked" and "easiest to post-edit" sentence is presented in Table 4. The chosen sentence contains untranslated words (Warenhäusern) and therefore got a bad ranking – on the other hand, such a lexical error is very easy to post-edit. Another example is a missing or extra negation particle ("not") – this is a very severe error in terms of translation quality, i.e. conveying the meaning of the source sentence, but very easy to post-edit.

3.3.2 Automatic classification of edits

In order to obtain more insight into the nature of errors corrected by post-editing thus learning more about differences between the systems and possibilities for improvement, we performed an automatic error analysis with the Hjerson tool [Popović (2011)] using the post-edited translations as references. The original translation hypotheses were compared with the post-edited ones in order to estimate which type of editing are most frequent for each of the systems. The following five types of edits were taken into account: correcting word form (morphology), correcting word order, adding missing word, deleting extra word and correcting lexical choice. Table 5 presents the edit rates for each of the five correction types for the three systems. Edit rate is defined as the percentage of performed edits, i.e. number of edit operations normalised over the total number of words generated by the translation system.

The main observation from the overall results is that the most frequent correction for all systems (not including translation memory Trados) is the

Table 5 Five types of edit rates for three translation systems in the first evaluation round: values are percentages of performed edits, i.e. number of edits normalised over the total number of words generated by the corresponding system.

edit rates (%)	correcting word form	correcting word order	adding missing word	deleting extra word	correcting lexical choice
Lucy	4.3	7.0	4.4	6.2	23.7
Moses	4.9	9.0	7.5	4.9	21.8
Trados	2.6	4.9	8.1	6.5	47.7

Table 6 Five types of edit rates separately for each domain and each language pair in the first evaluation round: values are percentages of perfomed edits, i.e. number of edits normalised over the total number of words generated by the corresponding system. Trados is not taken into account (see the main text).

edit rates (%) Lucy/Moses	correcting word form	correcting word order	adding missing word	deleting extra word	correcting lexical choice
News	4.3 / 5.7	6.8 / 8.6	3.7 / 6.6	5.3 / 4.7	19.2 / 20.7
OpenOffice	2.9 / 4.1	6.8 / 11.2	2.8 / 7.0	6.3 / 8.0	16.6 / 26.9
de-en	2.4 / 2.6	7.8 / 9.7	4.3 / 7.3	6.3 / 5.3	20.6 / 21.6
en-de	5.8 / 6.4	7.4 / 8.8	5.8 / 6.8	5.0 / 3.8	26.3 / 20.8
es-de	5.9 / 5.9	5.9 / 7.3	3.2 / 8.3	7.2 / 5.4	26.3 / 22.6

lexical choice and the next frequent correction is the word order, which suggests the same as the human error classification: the main weak points of all systems are incorrect lexical choice and incorrect word order. Furthermore, it can be seen that the rule-based Lucy system is better handling morphology and word ordering, whereas the statistical-based Moses system produces less lexical errors.

The results for Trados should be interpreted as follows. A large portion of the evaluation data did not reach a high degree of similarity for the content of the Trados Memory. Therefore many sentences remained untranslated which accounts for the high number of lexical errors. In this case the translator was presented with the source language sentence. The low number of morphological and reordering errors is easily explained by the fact that the content of the memory stems from human translations in the first place. The fact that morphological and reordering errors occur at all indicates that the training material that has been used to enrich the memory contained some impure translations – as we cannot say that the translations of the training data with which the memory was filled were done by professional translators or even native speakers, there might be lexical or other errors in the translation suggestions that were corrected by the experts. Also, language quality is subjective, changes might resemble personal preferences in style that the annotators have before the suggested translations.

More detailed results can be seen in Table 6. The edit rates are presented separately for each domain and each language pair. However, these detailed

results are not reported for Trados for the reasons explained above, but only for Lucy and Moses. The following can be observed:

- Word forms: Lucy performs better than Moses for the English-to-German task. For the other tasks, results are comparable. The reason is the rich morphology of the German language which can be better dealt with a rule-based system. Nevertheless, Spanish-to-German is "easier" for statistical systems than English-to-German in terms of word forms since the Spanish morphology is richer than English. These results indicate possibilities for improving the English-to-German Moses system.
- Word order: Lucy performs better than Moses for all language pairs and domains. Possible improvements could be achieved by better reorderings for Moses.
- Missing words: Again significantly lower numbers for Lucy outputs. One of the reason for both reorderings and for missing words could be the special positions of German verbs which are hard to deal with by statistical translation systems. This indicates a possibility for improvement using more linguistic knowledge.
- Extra words: for this error type, Moses performs better than Lucy. This can be attributed to word and phrase penalties in statistical translation systems.
- Lexical choice: for German-to-English and for the News domain, both systems have similar performance. However, for translation into German, Moses performs significantly better than Lucy. The probable reason is that whereas rule-based systems deal better with linguistic characteristics, statistical ones better handle lexical variations if trained in-domain. Further illustration of this can be seen from the results of the OpenOffice domain: Lucy performs significantly better, since Moses was trained on the out-of-domain WMT data. Possible directions for improvements are including appropriate terminologies into Lucy, as well as in-domain training or domain adaptation for Moses.

3.4 Lessons learnt

The first evaluation round can be seen both as a pilot and as a baseline. Some of the design decisions of the first evaluation round were guided by the wish to keep this pilot simple, e.g. allowing the human translators to annotate at most two errors per sentence. In addition, translation systems were not particularly adapted to the domains.

The design of the second evaluation round was based on the lessons learnt from the first round. The task specifications were more sophisticated and the engines were adapted to the respective domains and generally improved. The following requirements for the second round were inspired by the experience in the first round:

Ranking: Ties need to be allowed. For example, if four systems are involved, it should be possible to assign ranks 1,2,2,3 in order to avoid artificial

ranks. In the second evaluation round, about 28% of ranked sentences were actually tied.

Google translation outputs: In both evaluation rounds, Google translation outputs are taken into account only for the ranking task. However, in the first round, the process of avoiding Google in the post-editing and error classification task was carried out in the following way: if the Google was the best ranked system, the translators were offered the second best system for error classification and ranks 2, 3 and 4 for selection for post-editing. This way is nevertheless complicated and not optimal – it does not allow exact comparison between the first ranked outputs and outputs selected for post-editing for obvious reasons. Therefore, for the error classification and both post-editing tasks in the second evaluation round, all Google outputs were removed.

Error analysis: Word based error analysis should be carried out instead of sentence based for obvious reasons. Analysis should be carried out on all translation outputs of a given source sentence instead of the best ranked one only, thus enabling better comparison between translation systems.

Terminology error: One should distinguish between terminology error and wrong lexical choice. While *automobile* might be an adequate translation in certain context, in another context the customer might require that the term *car* is used.

Translate From Scratch: This option needs to be offered for the post-editing task. Otherwise, translators will simply delete the whole machine output and the post-editing distance will look artificially high.

Post-Edit All: Ask for post-edits of all systems' output for a subset of the data – this makes it possible to better compare the differences between translation engines. In addition, as we receive the number of edits performed for every output sentence, we can check if the sentence selected in the task *select-and-post-edit* was valid, e.g. if the translator choose one of the sentences with the minimal number of required edits.

Translation Memory: Translation memories are the most widely used translation tool used by human translators. Their performance depends on their content, which is usually extended and maintained over years. Within the project, it was not possible to design a fair and meaningful comparison between this technology and MT systems in general. Therefore, Trados was only to be used where applicable.

4 Second evaluation round

In the second evaluation round, the systems were prepared as follows:

Moses: Designed to support both domains, i.e. trained on in-domain data for each task. The News domain systems were trained on the Europarl and News corpora of WMT 2011 summing up to 1.9 million parallel sentences. The OpenOffice domain systems were trained on Openoffice3 corpus [Tiedemann (2009)] containing about 71000 parallel sentences. The

Table 7 Test sets for the second evaluation round – number of source sentences per language pair and domain.

number of source sentences	News	OpenOffice	Total
de-en	1788	418	2206
de-es	514	414	928
de-fr	912	412	1324
en-de	1744	414	2158
es-de	101	413	514
fr-de	1852	412	2264
Total	6911	2483	9394

monolingual side of the MULTIUN corpus [Eisele and Chen (2010)] was also included into the interpolated 5-gram language model.

Jane: Designed to support both domains, i.e. trained on in-domain data for each task. The systems are trained on the same data as Moses, except that a separate language model is used for each domain, without interpolation.

Google Translate: n.a.

Lucy: Adapted to domains by importing domain-specific terminology. Lucy generates a list of words that are not in the systems' dictionaries (unknown words). For system improvement the translations used by the translators in the first round were imported into the system to improve translation performance in the second round. In the second round terminology lists were imported. Those lists were generated manually by the translators during previous translation jobs for the customers.

Rbmt: Adapted to domains by importing the same terminology as used for Lucy.

Trados: Automatically filled with the same data as Moses.

More language pairs were taken into account, and the tasks were modified according to experiences from the first round. The number of source sentences in the test sets for each language pair and each domain is shown in Table 7. The translation outputs were produced by systems for the German-English, German-Spanish and German-French language pair in both translation directions, and the evaluation tasks were defined as follows:

Ranking: The concept of this task is basically the same as for the first round: *rank the outputs* of five different MT systems according to how well these preserve the meaning of the source sentence. The only difference is that ties were allowed.

Error classification: The error classification was performed on the word level, and all of the errors were marked (not only the main errors). For the translation outputs of particular low quality, a special category "too many errors" was offered. Apart from that, the subset of the error types was also slightly changed so that following error categories were taken into account: incorrect lexical choice, terminology error, morphological error, syntax error, misspelling, insertion, omission, punctuation error and other

Table 8 Human ranking results in the second evaluation round as the average position of each system in each task.

human ranking	Google	Jane	Moses	Rbmt1	Rbmt2
Overall	**2.0**	3.4	3.3	2.9	3.3
de-en	**2.0**	3.4	3.4	2.9	3.2
de-es	**2.0**	3.2	3.3	**2.2**	4.2
de-fr	**2.1**	3.2	3.3	2.9	3.4
en-de	**1.9**	3.5	3.4	3.0	3.2
es-de	**1.9**	3.2	3.4	2.7	3.7
fr-de	**2.1**	3.6	3.4	3.0	2.9
News	**2.1**	3.5	3.4	2.8	3.2
OpenOffice	**1.8**	3.2	3.3	3.0	3.6

error. For each error type except for missing words, two grades were defined: severe and minor. As for missing words, the evaluators should only decide if there are missing words in the sentence or not.

Post-editing: In this evaluation round, this task was divided into two sub-tasks:

Select and post edit: This part is basically the same as in the first round: select the translation output *which is easiest to post-edit* (which is not necessarily the best ranked) and perform the editing.

Post-edit all: In order to be able to compare better the translation systems, for each source sentence in the selected subset, post-editing was performed on all produced translation outputs.

For both post-editing sub-tasks, the translators were asked to perform only the minimal post-editing necessary to achieve acceptable translation quality. An option "Translate from scratch" was added: the translators were instructed to use it when they think that a completely new translation is faster than post-editing.

The results of the second round are presented in the following sections. The Google Translate system again was not considered as an option for error classification and post-editing. The difference from the first round is that the Google outputs were not only skipped when the Google was ranked as the best system, but were simply not offered at all.

4.1 Ranking

The results for the ranking task are shown in Table 8. Again, the first row presents the overall average ranks for the five systems, and **bold face** indicates the best system. Separate results are also presented for each translation direction as well as for each domain.

Similarly to the first round, Google performs best most often. However, for the German-to-Spanish translation Rbmt1 performs almost equally. In general, it can be observed that the two rule-based systems perform comparably well except for the language pair Spanish-German where the Rbmt1 system

performs significantly better than Rbmt2. This strengthens the observation that rule-based systems heavily rely on the amount of effort that the provider puts into the development of certain language pairs.

Apart from that, all systems perform comparably close, even closer as in the first round. In other words, system improvement makes the choice of a best system even harder.

4.2 Error classification

The error classification results are presented in Table 9 – for each translation system, raw error counts in its output were normalised over total number of sentences generated by this system. Thus, the percentage "12.9" in the column "Jane" and row "lexical choice (minor)" can be interpreted as follows: in 100 sentences translated by Jane there is a total of 12.9 minor lexical choice errors. The exact error distribution among the sentences is not indicated by these numbers. It can be seen that the most frequent errors in all systems are again *wrong lexical choices, terminology errors* and *syntax errors*. One observation is that the rule-based systems have more problems with terminology errors than statistical systems. The statistics referring to terminology errors should be considered with care. During analysis of the evaluation data it became obvious that the assessment of terminology error versus lexical choice was not consistently done by the annotators. Most terminology errors turned out to be errors of lexical choice. An example can be seen in the second sentence in table 10. Translating "Einstellung" as "attitude" is not merely using the wrong technical term in a specific customer context but semantically wrong, therefore it is not a terminological error but the wrong lexical choice. On the other hand, the rule based systems produce less severe morphological errors and significantly less omissions.

Another interesting observation is that the number of severe errors for all error types is higher than the number of minor errors. Exceptions include incorrect lexical choice for the rule-based systems, where both are nearly the same.

From this table, one can now generate recommendations for the most effective system improvements. It seems that better syntactical modelling and improvement of terminology use should have the biggest effect. Lexical choice in general is an issue, but the solution would probably require modelling of meaning, context and world knowledge, which may be even more demanding.

Table 10 presents three examples of human error classification performed on German-to-English translation outputs.

4.3 Post-editing

The human post-edits were again used to study the difference between selecting sentences for post-editing and ranking as well as to get details about the errors the system made.

Table 9 Human error classification in the second evaluation round: for each system, raw error counts are normalised over the total number of evaluated sentences generated by this system.

		Jane	Moses	Rbmt1	Rbmt2
lexical choice	minor	12.9	15.1	23.1	18.4
	severe	22.1	21.8	23.0	18.7
terminology	minor	5.9	6.6	11.6	10.4
	severe	27.0	29.3	44.2	37.2
morphology	minor	17.8	8.0	9.2	7.8
	severe	14.3	16.2	11.7	11.9
syntax	minor	7.8	7.6	9.8	9.7
	severe	27.5	36.6	28.3	28.4
misspelling	minor	5.8	2.5	3.4	3.8
	severe	5.6	1.7	1.8	1.3
insertion	minor	3.6	3.1	6.7	4.1
	severe	7.0	6.7	7.1	7.1
missing words		19.7	20.5	10.1	8.7
punctuation	minor	5.4	5.4	9.8	8.1
	severe	2.6	2.9	1.7	4.0
other	minor	1.3	1.5	1.6	2.8
	severe	3.4	2.3	2.7	3.3
too many errors		41.2	42.4	33.3	41.2

4.3.1 Ranking for translation quality vs. selection for post-editing

Comparison between sentence ranking and selection for post-editing in the second round confirmed the results obtained in the first round: the results can be seen in Table 11. An interesting detail is that percentage of selected sentences which are also best ranked is substantially higher for rule-based systems than for statistical systems.

4.3.2 Post-edit all

As in the first round, for this task automatic error analysis was performed using the post-edited translations as references in order to obtain more insight into the nature of post-edit corrections and to learn more about differences between the systems. The same five types of edits were taken into account: correcting word form (morphology), correcting word order, adding missing word, deleting extra word and correcting lexical choice. The results for each of the five edit types for the five systems are shown in Table 12 in the form of edit rates. Edit rate is defined as number of edited words normalised over the total number of words in the translation output.

The most frequent correction for all systems is again the lexical choice followed by the word order. Furthermore, it can be seen that the rule-based systems better handle morphology and induce less missing words. The same tendencies were also observed in the human error classification results.

The distribution of edit types for Trados is same as in the first round, i.e. the majority of edits are lexical due to untranslated portions which are not

Table 10 Examples of human error classification in German-to-English translation outputs.

source:	Der Gelatiere, auf deutsch etwas altertümlich der Eiskonditor, stellt die Creme nach allen Regeln der italienischen Eismacherkunst her.
errors:	severe lexical choice (SLC), severe terminology (ST), severe morphology (SM), severe syntax (SS), minor punctuation (MP)
translation:	The Gelatiere in German old-fashioned Eiskonditor **something**$_{SLC,SS}$ **,**$_{MP}$ **creating**$_{SM}$ the cream **after**$_{SLC}$ all the rules of the Italian **Eismacherkunst**$_{SLC}$.
reference:	The Gelatiere – in German the slightly more old-fashioned Eiskonditor – creates the cream according to all the rules of the Italian ice cream making art.
source:	Einstellung durch Drücken des roten Knopfes bestätigen.
errors:	missing words, severe terminology (ST), severe lexical choice (SLC)
translation:	**Attitude**$_{ST}$ by pressing the red **Knopfes**$_{SLC}$.
reference:	Confirm setting by pressing the red button.
source:	Das Schalten der Pumpe über Triacs / Halbleiterrelais ist im Einzelfall zu prüfen.
errors:	severe terminology (ST), minor morphology (MM)
translation:	The switching of the **credits**$_{ST}$ **above**$_{ST}$ **bidirectional**$_{ST}$ **triode**$_{ST}$ **thyristor**$_{ST}$ / semiconductor **relay**$_{MM}$ is to be checked on an individual basis.
reference:	Pump operation via triacs / semi-conductor relays must be checked in individual cases.
source:	Überraschenderweise zeigte sich, dass die neuen Räte in Bezug auf diese neuen Begriffe etwas im Dunkeln tappen.
errors:	missing words, minor lexical choice (MLC), severe syntax (SS)
translation:	Surprisingly , the new **councils**$_{MLC}$ in relation to which the new **terms**$_{SS}$ **somewhat**$_{SS}$ **in**$_{SS}$ **the**$_{SS}$ **dark**$_{SS}$.
reference:	Surprisingly , it turned out that the new council members do not understand the well-known concepts .
source:	Um das zu wissen, muss man nicht nach Anzola dell'Emilia in der Provinz Bologna fahren.
errors:	missing words, severe insertion (SI), minor misspelling (MMS)
translation:	In order to know , **at**$_{SI}$ **the**$_{SI}$ **end**$_{SI}$ **of**$_{SI}$ Anzola **Dell**$_{MMS}$ Emilia in the province of Bologna.
reference:	You do not need to go to Anzola dell'Emilia in the province of Bologna to know that .

Table 11 Percentage of sentences with a given rank selected as the best for post-editing; no Google for selection, no Trados for ranking.

	% of selected sentences				
rank	Overall	Jane	Moses	Rbmt1	Rbmt2
1	71.7	65.1	63.9	78.8	73.6
2	19.1	21.4	22.9	15.3	18.7
3	6.5	9.3	9.2	4.2	5.5
4	2.7	4.2	4.0	1.7	2.1

Table 12 Five types of edit rates for five translation systems in the second evaluation round: values are percentages of performed edits, i.e. number of edited words normalised over the total number of words generated by the corresponding system.

edit rates (%)	correcting word form	correcting word order	adding missing word	deleting extra word	correcting lexical choice
Jane	5.7	9.5	6.3	7.4	25.7
Rbmt1	4.1	8.2	4.6	7.2	23.5
Moses	5.4	9.9	5.8	7.2	23.4
Rbmt2	4.9	9.9	4.5	7.5	24.7
Trados	2.5	4.0	7.8	8.7	68.4

Table 13 Examples of automatic classification of edit operations in German-to-English translation outputs (same sentences as in Table 10).

source:	Der Gelatiere, auf deutsch etwas altertümlich der Eiskonditor, stellt die Creme nach allen Regeln der italienischen Eismacherkunst her.
edits:	lexical choice (LC), word order (O), word form (F)
translation:	The Gelatiere in German $old\text{-}fashioned_O$ $Eiskonditor_O$ $something_{LC}$,$_{LC}$ $creating_F$ the cream $after_{LC}$ all the rules of the Italian $Eismacherkunst_{LC}$.
reference:	The Gelatiere – in German the slightly more old-fashioned Eiskonditor – creates the cream according to all the rules of the Italian ice cream making art.
source:	Einstellung durch Drücken des roten Knopfes bestätigen.
edits:	lexical choice (LC), word order (O), missing word (M)
translation:	$Attitude_{LC}$ by pressing the red $Knopfes_{LC}$.
reference:	$Confirm_M$ setting by pressing the red button.
source:	Das Schalten der Pumpe über Triacs / Halbleiterrelais ist im Einzelfall zu prüfen.
edits:	lexical choice (L), word form (F), extra word (E)
translation:	The_E $switching_E$ of_E the_E $credits_E$ $above_L$ $bidirectional_L$ $triode_L$ $thyristor_L$ / semiconductor $relay_F$ is_E to_L be checked on_L an_E individual $basis_L$.
reference:	Pump operation via triacs / semi-conductor relays must be checked in individual cases.

contained in the memory. This system is again not taken into account in the further analysis.

Table 13 shows three examples of automatic classification of edit operations performed in the same German-to-English translation outputs presented in Table 10.

The detailed results presented separately for each language pair and each domain can be seen in Table 14. The following can be observed:

- Word forms: Rbmt1 handles best the morphology in the majority of cases, for Spanish-to-German is however outperformed by Rbmt2, and for French-to-German all systems perform comparably.
- Word order: In majority of cases Rbmt1 performs the best, however for German-to-Spanish and German-to-French Jane is the best. The highest

Table 14 Five types of edit rates separately for each language pair and each domain: values are percentages of performed edits, i.e. number of edits normalised over the total number of words generated by the corresponding system. Trados is not taken into account (see the main text).

edit rates (%)	correcting→	form	order	missing	extra	lexical
de-en	Jane	2.9	9.3	8.0	4.5	17.2
	Moses	2.7	9.1	6.7	4.1	13.8
	Rbmt1	1.8	8.5	4.0	6.0	14.6
	Rbmt2	2.5	9.6	4.2	5.4	14.2
de-es	Jane	5.3	9.6	8.7	5.1	18.5
	Moses	5.1	10.5	7.7	4.3	19.2
	Rbmt1	4.9	10.2	5.1	5.5	19.8
	Rbmt2	7.8	20.3	7.2	3.1	23.3
de-fr	Jane	4.8	12.1	6.0	9.0	32.8
	Moses	4.9	12.5	4.6	10.4	31.3
	Rbmt1	4.5	12.7	5.2	8.8	34.0
	Rbmt2	4.6	12.7	4.3	10.2	34.5
en-de	Jane	8.0	9.7	4.5	8.9	24.2
	Moses	7.5	9.8	4.3	7.9	20.3
	Rbmt1	4.8	5.8	5.0	6.2	19.6
	Rbmt2	5.1	6.5	4.7	5.6	19.2
es-de	Jane	6.7	7.0	5.8	7.4	26.9
	Moses	6.6	8.0	6.7	6.8	25.6
	Rbmt1	4.7	5.4	4.5	7.3	23.8
	Rbmt2	3.8	9.0	4.5	8.8	28.6
fr-de	Jane	7.0	10.5	6.8	7.2	27.6
	Moses	5.7	9.4	5.6	5.0	23.0
	Rbmt1	5.7	6.9	2.7	8.6	22.5
	Rbmt2	5.7	7.5	3.2	5.9	18.1
News	Jane	5.4	10.8	6.3	7.4	25.3
	Moses	5.1	10.9	5.4	7.4	22.8
	Rbmt1	4.0	9.5	4.5	7.3	23.2
	Rbmt2	4.7	11.1	4.5	6.9	23.4
OpenOffice	Jane	6.3	6.7	6.2	7.3	26.6
	Moses	6.2	7.5	6.6	6.8	25.0
	Rbmt1	4.5	5.3	4.6	7.0	24.1
	Rbmt2	5.4	7.6	4.3	8.7	27.0

percentage of word order post-edits is present for the German-to-Spanish and German-to-French translations, and it is higher for the News than for the OpenOffice domain.
- Missing words: All systems perform comparably, Rbmt1 slightly better than others.
- Extra words: All systems perform comparably.
- Lexical choice: All systems perform comparably. The highest number of correcting lexical choice appears for the German-to-French translation and the lowest for the German-to-English and German-to-Spanish translations.

A detailed analysis for Trados has not been made as the matching score of the system already entails the estimated effort for the translator to produce a high-quality human translation. Roughly one can say a 100% match

is considered as no effort needed, towards 85% post-editing effort is less than translating from scratch and below 85% creating a new translation is faster than post-editing. These matching scores are considered in pricing: usually LSPs negotiate rebates with translators on the basis that higher valued fuzzy matches do not require as much manual work than low fuzzies or new content and thus are paid less. The rebates are staggered, so the translator is paid more for the lower the match rate. For the LSP to have a real use from MT a comparable categorization of machine translation quality is sorely needed to estimate post-editing effort before translation as translation offers are usually created before starting to translate.

5 Summary

In this work, we have presented the results of a broad human evaluation where professional translators have judged machine translation outputs of distinct systems via three different tasks: ranking, error classification and post-editing. We have systematically analysed the obtained results in order to better understand the selection mechanisms of human evaluators as well as differences between machine translation systems. The most severe problems that machine translation systems encounter are related to terminology/lexical choice and syntax. Human annotators seem to prefer well-formed sentences over unstructured outputs, even if the latter contain the "material" needed for creating a good translation. Further work is needed to study these hypotheses in more depth.

Other results of the TARAXŰ project not covered in this article are the overall system architecture (see [Burchardt et al (2013)]) and an automatic selection mechanism (see [Avramidis et al (2011)]). The TARAXŰ system prototype allows the inclusion of hybrid MT into an everyday translation production workflow including handling of different file formats, use of translation memory and post-editing. The selection mechanism makes use of state-of-the art machine learning to select the best translation out of the different systems' output using features from language checking, system internal features or parsing probability. For more details, we refer the reader to the given publications.[6]

Acknowledgements This work has been developed within the TARAXŰ project financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development. Thanks to our colleague Christian Federmann for helping with the Appraise system.

References

[Alonso and Thurmair (2003)] Alonso JA, Thurmair G (2003) The comprendium translator system. In: Proceedings of the Ninth Machine Translation Summit

[6] More publications can be found online: `http://taraxu.dfki.de/publications`

[Avramidis et al (2011)] Avramidis E, Popović M, Vilar D, Burchardt A (2011) Evaluate with confidence estimation: Machine ranking of translation outputs using grammatical features. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Edinburgh, Scotland, pp 65–70

[Burchardt et al (2013)] Burchardt A, Tscherwinka C, Avramidis E, Uszkoreit H (2013) Machine Translation at Work, Studies in Computational Intelligence, vol 458, Springer, pp 241–261

[Callison-Burch et al (2010)] Callison-Burch C, Koehn P, Monz C, Peterson K, Przybocki M, Zaidan O (2010) Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, Association for Computational Linguistics, Uppsala, Sweden, pp 17–53, revised August 2010

[Eisele and Chen (2010)] Eisele A and Chen Y (2010) MultiUN: A Multilingual Corpus from United Nation Documents. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), La Valletta, Malta, pp 2868–2872

[Farzindar and Lapalme (2009)] Farzindar A and Lapalme G (2009) Machine Translation of Legal Information and Its Evaluation. In: Proceedings of the 22nd Canadian Conference on Artificial Intelligence (Canadian AI 09), Kelowna, BC, pp 64–73

[Federmann (2010)] Federmann C (2010) Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010), La Valletta, Malta

[He et al (2010)] He Y, Ma Y, Roturier J, Way A, and van Genabith J (2010) Improving the post-editing experience using translation recommendation: A user study. In: Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), Denver, Colorado

[Koehn et al(2007)] Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '07, pp 177–180

[Papineni et al (2001)] Papineni K, Roukos S, Ward T, Zhu WJ (2001) Bleu: a method for automatic evaluation of machine translation. IBM Research Report RC22176(W0109-022), IBM

[Popović (2011)] Popović M (2011) Hjerson: An open source tool for automatic error classification of machine translation output. The Prague Bulletin of Mathematical Linguistics pp 59–68

[Specia and Farzindar (2010)] Specia L, Farzindar A (2010) Estimating Machine Translation Post-Editing Effort with HTER. In: Proceedings of AMTA-2010 Workshop Bringing MT to the User: MT Research and the Translation Industry, Denver, Colorado

[Tiedemann (2009)] Tiedemann J (2009) News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In: Advances in Natural Language Processing, vol V, Borovets, Bulgaria, chap V, pp 237–248

[Vilar et al (2006)] Vilar D, Xu J, D'Haro LF, Ney H (2006) Error Analysis of Machine Translation Output. In: International Conference on Language Resources and Evaluation, Genoa, Italy, pp 697–702

[Vilar et al (2010)] Vilar D, Stein D, Huck M, Ney H (2010) Jane: Open source hierarchical translation, extended with reordering and lexicon models. In: ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, Uppsala, Sweden, pp 262–270