Beyond Linguistic Equivalence. An Empirical Study of Translation Evaluation in a Translation Learner Corpus

Mihaela Vela **Anne-Kathrin Schumann** **Andrea Wurm**
Department of Applied Linguistics, Translation and Interpreting
Saarland University, Saarbrücken, Germany
{m.vela, anne.schumann, a.wurm}@mx.uni-saarland.de

Abstract

The realisation that fully automatic translation in many settings is still far from producing output that is equal or superior to human translation has lead to an intense interest in translation evaluation in the MT community. However, research in this field, by now, has not only largely ignored the tremendous amount of relevant knowledge available in a closely related discipline, namely translation studies, but also failed to provide a deeper understanding of the nature of "translation errors" and "translation quality". This paper presents an empirical take on the latter concept, translation quality, by comparing human and automatic evaluations of learner translations in the KOPTE corpus. We will show that translation studies provide sophisticated concepts for translation quality estimation and error annotation. Moreover, by applying well-established MT evaluation scores, namely BLEU and Meteor, to KOPTE learner translations that were graded by a human expert, we hope to shed light on properties (and potential shortcomings) of these scores.

1 Translation quality assessment

In recent years, researchers in the field of MT evaluation have proposed a large variety of methods for assessing the quality of automatically produced translations. Approaches range from fully automatic quality scoring to efforts aimed at the development of "human" evaluation scores that try to exploit the (often tacit) linguistic knowledge of human evaluators. The criteria according to which quality is estimated often include *adequacy*, the degree of meaning preservation, and *fluency*, target language correctness (Callison-Burch et al., 2007). The goals of both "human" evaluation and fully automatic quality scoring are manifold and cover system optimisation as well as benchmarking and comparison.

In translation studies, the scientific (and prescientific) discussion on how to assess the quality of human translations has been going on for centuries. In recent years, the development of appropriate concepts and tools has become even more vital to the discipline due to the pressing needs of the language industry. However, different from the belief, typical to MT, that the "goodness" of a translation can be scored on the basis of linguistic criteria alone, the notion of "translation quality", in translation studies, has assumed a multi-faceted shape, distancing itself from a simple strive for equivalence and embracing concepts such as functional, stylistic and pragmatic appropriateness as well as textual coherence. In this section, we provide an overview over approaches to translation quality assessment developed in MT and translation studies to specify how "quality" is being defined in both fields and which methods and features are used. Due to the amount of available literature, this overview is necessarily incomplete, but still insightful with respect to differences and commonalities between MT and human translation evaluation.

1.1 Automatic MT quality scores

MT output is usually evaluated by automatic language-independent metrics which can be applied to any language produced by an MT system. The use of automatic metrics for MT evaluation is legitimate, since MT systems deal with large amounts of data, on which manual evaluation would be very time-consuming and expensive.

Automatic metrics typically compute the closeness (adequacy) of a "hypothesis" to a "reference" translation and differ from each other by how this closeness is measured. The most popular MT eval-

Workshop on Humans and Computer-assisted Translation, pages 47–56,
Gothenburg, Sweden, 26 April 2014.

uation metrics are IBM BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) which are used not only for tuning MT systems, but also as evaluation metrics for shared tasks, such as the Workshop on Statistical Machine Translation (Bojar et al., 2013).

IBM BLEU uses n-gram precision by matching machine translation output against one or more reference translations. It accounts for adequacy and fluency by calculating word precision, respectively the n-gram precision. In order to deal with the over generation of common words, precision counts are clipped, meaning that a reference word is exhausted after it is matched. This is then the modified n-gram precision. For N=4 the modified n-gram precision is calculated and the results are combined by using the geometric mean. Instead of recall, the brevity penalty (BP) is used. It penalizes candidate translations which are shorter than the reference translations.

The NIST metric is derived from IBM BLEU. The NIST score is the arithmetic mean of modified n-gram precision for N=5 scaled by BP. Additionally, NIST also considers the information gain of each n-gram, giving more weight to more informative (less frequent) n-grams and less weight to less informative (more frequent) n-grams.

Another often used machine translation evaluation metric is Meteor (Denkowski and Lavie, 2011). Different from IBM BLEU and NIST, Meteor evaluates a candidate translation by calculating precision and recall on the unigram level and combining them into a parametrized harmonic mean. The result from the harmonic mean is then scaled by a fragmentation penalty which penalizes gaps and differences in word order.

Besides these evaluation metrics, several other metrics are sometimes used for the evaluation of MT output. Some of these are the WER (word error-rate) metric based on the Levensthein distance (Levenshtein, 1966), the position-independent error rate metric PER (Tillmann et al., 1997) and the translation edit rate metric TER (Snover et al., 2006) with its newer version TERp (Snover et al., 2009).

1.2 Human MT quality evaluation

Human evaluation of MT output is performed in different ways. The most frequently used evaluation method seems to be a simple ranking of translated sentences by a "reasonable number of evaluators" (Farrús et al., 2010). According to Birch et al. (2013), this form of evaluation was used, among others, during the last STATMT workshops and can thus be considered rather popular. APPRAISE (Federmann, 2012) is a tool that can be used for such as task, since it allows for the manual ranking of sentences, quality estimation, error annotation and post-editing.

Other forms of evaluation, however, exist. For example, Birch et al. (2013) propose HMEANT, an evaluation score based on MEANT (Lo and Wu, 2011), a semi-automatic MT quality score that measures the degree of meaning preservation by comparing verb frames and semantic roles of hypothesis translations to their respective counterparts in the reference translation(s). Unfortunately, Birch et al. (2013) report difficulty in producing coherent role alignments between hypotheses and translations, a problem that affects the final HMEANT score calculation. This, however, seems hardly surprising given the difficulty of the annotation task (although, following the authors' description, some familiarity of the annotators with the linguistic key concepts can be assumed) and the fact that guidelines and training are meant to be minimal.

Another (indirect) human evaluation method for MT that is also employed for error analysis are reading comprehension tests (e.g. Maney et al. (2012), Weiss and Ahrenberg (2012)). Moreover, HTER (Snover et al., 2006) is a TER-based repair-oriented metric which uses human annotators (the only apparent qualificational requirement being fluency in the target language) to generate "targeted" reference translations by post-editing the MT output or the existing reference translations, following the goal to find the shortest path between the hypothesis and a "correct" reference. Snover et al. (2006) report a high correlation between evaluation with HTER and traditional human adequacy and fluency judgements. Last but not least, Somers (2011) mentions other repair-oriented measures such as post-editing effort measured by the amount of key-strokes or time spent on producing a "correct" translation on the basis of MT output.

1.3 The notion of quality in translation studies

Discussions of translation "quality", in translation studies, for a long time focused on *equivalence*

which, in its oldest and simplest form, used to echo *adequacy* as understood by today's MT researchers: "good" translation was viewed as an optimal compromise between meaning preservation and target language correctness, which was especially relevant to the translation of religious texts. For example, Kußmaul (2000) emphatically cites Martin Luther's famous Bible translation into German as an example of "good" translation because Luther, according to his own testimony and following his reformative ambition, focused on producing fluent, easily understandable text rather than mimicking the linguistic structures of the Hebrew, Aramaic and Greek originals (see also Windle and Pym (2011) for a further discussion).

More recent work in translation studies has abandoned one-dimensional views of the relation between source and target text and postulates that, depending on the communicative context within and for which a translation is produced, this relation can vary greatly. That is, the degree of linguistic or semantic "fidelity" of a good translation towards the source text depends on functional criteria. This view is echoed in the concepts of "primary vs. secondary", "documentary vs. instrumental" and "covert vs. overt" translation (Hönig, 2003). The consequence of this shift in paradigms is that, since different *translation strategies* may be appropriately adopted in different situations, evaluation criteria become essentially dependent on the *function* that the translation is going to play in the target language and culture. This view is most prominently advocated by the so-called *skopos theory* (cf. Dizdar (2003)). *Translation errors*, then, are not just simple violations of the target language system or outright failures to translate words or segments, but violations of the *translation task* that can manifest themselves on all levels of text production (Nord, 2003). It is important to point out that, in this framework, *linguistic errors* are just one type of error covering not only one of the favourite MT error categories, namely un- and mistranslated words (compare, for example, Stymne and Ahrenberg (2012), Weiss and Ahrenberg (2012), Popović et al. (2013)), but also phraseological, idiomatic, syntactic, grammatical, modal, temporal, stylistic, cohesion and other kinds of errors. Moreover, *translation-specific errors* occur when the translation does not fulfill its function because of pragmatic (e.g. text-type specific forms of address), cultural (e.g. text conventions, proper names, or other conventions) or formal (e. g. layout) defects (Nord, 2003). Depending on the appropriate translation strategy for a given translation task, these error types may be weighted differently. Furthermore, the communicative and functional view on translation also dictates a change in the concept of equivalence which is no longer considered to be adequately described by the notions of "meaning preservation" or "fidelity", but becomes dependent on aesthetic, connotational, textual, communicative, situational, functional and cognitive aspects (for a detailed discussion see Horn-Helf (1999)). In MT evaluation, most of these aspects have not yet or only in part been considered.

Last but not least, the translation industry has developed normative standards and proofreading schemes. For example, the DIN EN 15038:2006-08 (Deutsches Institut für Normung, 2006) discusses translation errors, quality management and qualificational requirements for translators and proofreaders, while the SAE J2450 standard (Society of Automotive Engineers, 2005) presents a weighted "translation quality metric". An application perspective is given by Mertin (2006) who discusses translation quality management procedures in a big automotive company and, among other things, develops a weighted translation error scheme for proofreading.

1.4 Discussion

The above discussion shows that, while the object of evaluation is the same for both MT and translation studies, namely translation, the differences between evaluation approaches developed in both fields are considerable. Most importantly, in translation studies, translation evaluation is considered an *expert task* for which fluency in one or several languages is certainly not enough, but for which *translation-specific expert knowledge* is required. Another important distinction is that evaluation, again in translation studies, is normally not carried out on the sentence level, since sentences are usually split up into several "units of translation" and can certainly contain more than one "translation problem". Consequently, the popular MT practice of ranking whole sentences according to some automatic score, by anonymous evaluators or even users of Amazon Turk (e.g. in the introduction to Bojar et al. (2013)), from a translation studies point of view, is unlikely to provide reason-

able evaluations. Last but not least, the MT community's strive for adequacy or meaning preservation does not match the notions of weighting translation errors, of adopting different translation strategies and, consequently, does not fit the complicated source/target text relations that have been acknowledged by translation studies. Evaluation methods that are based on simple measures of linguistic equality such as n-gram overlap (BLEU) or, just slightly more complicated, the preservation of syntactic frames and semantic roles (MEANT) fail to provide straightforward criteria for distinguishing between *legitimate* and *illegitimate* variation. Moreover, semantic and pragmatic criteria as well as the notion of "reference translation" remain, at best, rather unclear.

On the other hand, the MT community has recognised translation evaluation as an unresolved research problem. For example, Birch et al. (2013) state that ranking judgements are difficult to generalise, while Callison-Burch et al. (2007) carry out extensive correlation tests of a whole range of automatic MT evaluation metrics in comparison to human judgements, showing that BLEU does not rank highest, but still remains in the top segment. It still needs to be shown how MT research can benefit from more sophisticated evaluation measures and whether all the parameters that are considered relevant to the evaluation of human translations are relevant for MT usage scenarios, too. In the remainder of this paper, we present a study on how much and possibly for which reasons automatic MT evaluation scores (namely BLEU and Meteor) differ from translation expert quality judgements on extracts of a French-German translation learner corpus.

2 The KOPTE corpus

2.1 General corpus design

The KOPTE project (Wurm, 2013) was designed to enable research on translation evaluation in a university training course (master's level) for translators and to enlighten students' translation problems as well as their problem solving strategies. To achieve this goal, a corpus of student translations was compiled. The corpus consists of several translations of the same source texts produced by student translators in a classroom setting. As a whole, it covers 985 translations of 77 source texts amounting to a total of 318,467 tokens. Source texts were taken from French newspapers and translated into German in class over a span of several years, the translation brief calling for a ready-to-publish text to be printed in a German national newspaper. Consequently, all translation tasks include the use of idiomatic language, explanations of culture-specific items, changes in the explicitness of macrotextual cohesive elements, etc.[1]

2.2 Annotation of translation features and translation evaluation in KOPTE

Student translations were evaluated by one of the authors, an experienced translation teacher, with the aim of giving feedback to students. All translations were graded and *errors* as well as *good solutions* were marked in the text according to a fine-grained evaluation scheme. In this scheme, the weight of evaluated items is indicated through numbers ranging from plus/minus 1 (minor) to plus/minus 8 (major). Based on these evaluations, each translation was assigned a final grade according to the German grading system on a scale ranging from 1 ("very good") to 6 ("highly erroneous") with in-between intervals at the levels of .0, .3 and .7. To calculate this grade, positive and negative evaluations were summed up separately, before the negative score was subtracted from the positive one. A score of around zero corresponds to the grade "good" (=2), to achieve "very good" (=1) the student needs a surplus of positive evaluations.

The evaluation scheme based on which student translations are graded is divided into external and internal factors. *External* characteristics describe the communicative situation given by the source text and the translation brief (author, recipient, medium, location, time). *Internal* factors, on the other hand, comprise eight categories: form, structure, cohesion, stylistics/register, grammar, lexis/semantics, translation-specific problems, function. These categories are containers for more fine-grained criteria which can be applied to segments of the (source or target) text or even to the whole text, depending on the nature of the criterion. Some internal subcriteria of the scheme are summarised in Table 1. A quantitative analysis of error types in KOPTE shows that semantic/lexical errors are by far the most common error in the student translations (Wurm, 2013).

Evaluations in KOPTE were carried out by just

[1]More information about KOPTE is available from http://fr46.uni-saarland.de/index.php?id=3702&L=%2524L.

one evaluator for the reason that, in a classroom setting, multiple evaluations are not feasible. Although multiple evaluations would have been considered highly valuable, the data available from KOPTE was evaluated by an experienced translation scholar with long-standing experience in teaching translation. Moreover, the evaluation scheme is much more detailed than error annotation schemes that are normally described in the literature and it is theoretically well-motivated. An analysis of the median grades in our data sample (compare Tables 2–4) shows that grading varies only slightly between different texts, considering the maximum variation potential ranging from 1 to 6, and thus can be considered consistent.

Criteria	**Examples of subcriteria**
author, recipients, medium, topic, location, time	—
form	paragraphs, formatting
structure	thematic, progression, macrostructure, illustrations
cohesion	reference, connections
stylistics	style, genre
grammar	determiners, modality, syntax
semantics	textual semantics, idioms, numbers, terminology
translation problems	erroneous source text, proper names, culture-specific items, ideology, math. units, pragmatics, allusions
function	goal dependence

Table 1: Internal evaluation criteria in the KOPTE annotation scheme.

3 Experiments

The goal of our experiments was to study whether the human translation expert judgements in KOPTE can be mimicked using simple automatic quality metrics as used in MT, namely BLEU and Meteor. More specifically, we aim at:

- studying how automatic evaluation scores relate to fine-grained human expert evaluations,
- investigating whether a higher number of references improves the automatic scores and why (or why not),
- examining whether a higher number of references provides more reliable evaluation scores as measured by an improved correlation with the human expert judgments.

In order to study the behaviour of automatic MT evaluation scores, we conducted three experiments by applying IBM BLEU (Papineni et al., 2002) and Meteor 1.4 (Denkowski and Lavie, 2011) to a sample of KOPTE translations that were produced by translation students preparing for their final master's exams. Scores were calculated on the complete texts. To evaluate the overall performance of the automatic evaluation scores on these texts, we calculated Kendall's rank correlation coefficient for each text following the procedure described in Sachs and Hedderich (2009). Correlations were calculated for:

- the human expert grades and BLEU scores for each translation,
- the human expert grades and Meteor scores for each translation,
- BLEU and Meteor scores for each translation.

3.1 Experimental setup and results

In a first experiment, we applied the automatic evaluation scores to the source texts given in Table 2, choosing, for each text, the student translation with the best human grade as reference translation. The median human grades as well as mean BLEU and Meteor and correlation scores obtained for each text (excluding the reference translation) are included in Table 2. In a second experiment, we repeated this procedure, however, using a set of three reference translations. Results are given in Table 3. Finally, in a last experiment we used five reference translations selected according to their human expert grade (Table 4). In both steps, source texts for which less than four hypotheses were available were excluded from the data sets.

3.2 Discussion

The tables show that in the first experiment a set of 152 translations was evaluated, whereas in the second and third experiment these numbers were reduced to 108 and 68 respectively due to the selection of more references. The human expert evaluations rated most of these translations at least as acceptable, as can be seen from the median grade for each experiment which was 2.3 in the first experiment and consecutively decreased to 3.0 for the third experiment, again due to the selection of more "good" translations as references. The

Source text	Human trans./ source text	Median grades	Mean BLEU	Mean Meteor	Correlation Human-BLEU	Correlation Human-Meteor	Correlation BLEU-Meteor
AT001	7	2. 7	0. 15	0. 33	−0. 39	−0. 73	0. 24
AT002	12	2. 3	0. 15	0. 35	−0. 20	−0. 43	0. 49
AT004	12	2. 7	0. 19	0. 37	0. 14	0. 11	0. 63
AT005	12	2. 3	0. 20	0. 36	0. 32	0. 45	0. 45
AT008	10	2. 15	0. 23	0. 38	−0. 43	−0. 29	0. 78
AT010	11	2. 7	0. 25	0. 41	0. 06	−0. 10	0. 56
AT012	9	2. 0	0. 22	0. 40	−0. 30	−0. 36	0. 50
AT015	5	2. 0	0. 11	0. 28	0. 36	0. 12	0. 60
AT017	7	2. 3	0. 22	0. 38	−0. 20	0. 06	0. 71
AT021	4	3. 0	0. 18	0. 39	−0. 55	−0. 55	1. 00
AT023	6	2. 3	0. 22	0. 38	0. 50	−0. 07	−0. 20
AT025	4	2. 15	0. 13	0. 36	0. 33	0. 0	0. 00
AT026	21	3. 0	0. 12	0. 26	−0. 19	−0. 35	0. 67
AT039	13	3. 0	0. 10	0. 29	−0. 08	0. 03	0. 49
AT052	7	2. 0	0. 17	0. 31	−0. 32	0. 05	0. 00
AT053	7	2. 3	0. 18	0. 32	0. 62	0. 39	0. 33
AT059	5	2. 0	0. 24	0. 36	0. 00	0. 22	0. 80

Table 2: Source texts, number of human translations per source text, median of the obtained grade per source text, mean of the BLEU and Meteor scores per source text and Kendall's rank correlation coefficients for the first experiment.

Source text	Human trans./ source text	Median grades	Mean BLEU	Mean Meteor	Correlation Human−BLEU	Correlation Human-Meteor	Correlation BLEU-Meteor
AT001	5	3. 0	0. 17	0. 36	−0. 12	0. 36	0. 60
AT002	10	2. 3	0. 17	0. 36	−0. 14	0. 05	0. 38
AT004	10	2. 85	0. 20	0. 37	0. 39	0. 16	0. 51
AT005	10	2. 3	0. 20	0. 40	−0. 10	0. 05	0. 47
AT008	8	2. 5	0. 25	0. 45	−0. 67	−0. 15	0. 00
AT010	9	2. 7	0. 23	0. 41	−0. 10	−0. 50	0. 28
AT012	7	2. 3	0. 23	0. 43	0. 00	0. 11	0. 52
AT017	5	2. 3	0. 21	0. 43	0. 12	0. 36	0. 60
AT023	4	2. 5	0. 21	0. 38	0. 41	0. 81	0. 67
AT026	19	3. 3	0. 10	0. 26	−0. 31	−0. 41	0. 77
AT039	11	3. 0	0. 11	0. 34	0. 06	0. 14	0. 74
AT052	5	2. 0	0. 18	0. 40	0. 12	0. 36	0. 20
AT053	5	2. 3	0. 17	0. 35	0. 36	−0. 12	0. 40

Table 3: Source texts, number of human translations per source text, median of the obtained grade per source text, mean of the BLEU and Meteor scores per source text and Kendall's rank correlation coefficients for the second experiment.

grades for the best translations selected as references range for the first and second experiment between 1.0 and 2.3, whereas for the third experiment the selected references were evaluated with grades between 1.0 and 2.7. Nevertheless, the median grade for the references in all three experiments is always 1.7. From the overall median grade and the median grade of the selected translations as reference we can notice, that the translations selected as references were indeed "better" than the remaining ones.

The BLEU and Meteor scores given in the tables are mean values over the individual translations' scores for each source text. These scores are very low, reaching a maximum of 0.25 over all three experiments for BLEU and 0.45 for Meteor. However, given the human expert grades the translations cannot be considered unreadable. In fact, the correlation coefficients show that neither BLEU nor Meteor (except a few exceptional cases) correlate with the human quality judgements, however, they show a (weak) tendency to correlate with each other. Moreover, the data shows that the addition of reference translations results neither in significantly higher BLEU or Meteor scores nor in improved correlation.

3.3 Qualitative analysis

Our finding that human quality judgements do not correlate with automatic scores if the object of evaluation is a translation produced by a human (as opposed to a machine) matches earlier results presented by Doddington (2002) within the context of evaluating NIST. Doddington (2002) proposes the explanation that "differences between professional translators are far more subtle [than differences between machine-produced translations, the authors] and thus less well characterized

Source text	Human trans./ source text	Median grades	Mean BLEU	Mean Meteor	Correlation Human-BLEU	Correlation Human-Meteor	Correlation BLEU-Meteor
AT002	8	2. 5	0. 17	0. 36	−0. 08	0. 00	0. 43
AT004	8	3. 0	0. 20	0. 36	0. 00	0. 23	0. 71
AT005	8	2. 3	0. 20	0. 42	0. 00	0. 08	0. 43
AT008	6	2. 85	0. 26	0. 45	−0. 55	−0. 14	0. 33
AT010	7	2. 7	0. 23	0. 41	0. 00	−0. 12	0. 05
AT012	5	2. 3	0. 23	0. 43	0. 22	0. 22	0. 40
AT026	17	3. 3	0. 11	0. 31	−0. 24	−0. 34	0. 62
AT039	9	3. 0	0. 10	0. 37	0. 22	0. 55	0. 22

Table 4: Source texts, number of human translations per source text, median of the obtained grade per source text, mean of the BLEU and Meteor scores per source text and Kendall's rank correlation coefficients for the third experiment.

by N-gram statistics." We conducted a qualitative analysis of some KOPTE translations in order to check whether the differences between individual translations are indeed as subtle as suggested by Doddington and to come up at least with hypotheses that could explain the poor performance of the automatic scores. We selected three source texts used in the second experiment, namely AT008, AT023 and AT053 and compared their respective reference translations to selected hypothesis translations. This analysis was conducted on the lexical level alone, that is, most of the features of KOPTE's elaborated evaluation scheme were not even considered. The analysis, however, shows that the amount of variation that can be found just on the lexical level is almost overwhelming. Some examples are listed in Appendix A.

A common phenomenon is simple variation due to synonyms or the use of phrasal variants or paraphrases. Moreover, the listed examples show that lexical variation can be triggered by different source text elements. The phenomena shown in the tables are well-known translation problems, e.g. proper names, colloquial or figurative speech or numbers. The other categories in the table are less clear-cut, that is, they can overlap. In our analysis, source text elements that cannot be translated literally, but instead call for a creative solution were classified as translation problems. Different translation strategies can be applied to different kinds of problems, most importantly to the translation of culture-specific items, proper names, underspecified source text elements or culture-specific arguments. The respective table and other examples that we analysed show that for this category some translators chose to add additional information, to adapt the perspective to the German target audience (for example, by adapting pronouns or deictic elements) or to adapt the formatting choices to the variant preferred by the target culture (e.g. commas instead of fullstops, different types of quotation marks), whereas other translators chose to translate literally. Both strategies are legitimate under certain circumstances, however, it can be assumed that adaptations require a greater cognitive effort. Source ambiguities, according to our preliminary typology, are source text features that can be interpreted in different ways - at least for a translator translating from a foreign language (as opposed to a native speaker). Obviously, the line between this category and outright translation errors is not very clear.

However, it needs to be stated that also for the other categories - while many variants are correct and legitimate - not all are equally good. Best solutions for given problems are distributed unequally across the translations studied. Beyond the purely lexical level, extensive variation can be witnessed on the syntactic, but also the grammatical level. For example, some translators chose to break the rather complicated syntax of the French original into simpler, easily readable sentences, producing, in some cases, considerable shifts in the information structure of the text - often a legitimate strategy.

With respect to the performance of the automatic scores, our preliminary study - that still calls for larger-scale and in-depth verification - suggests that neither BLEU nor Meteor are able to cope with the amount of variation found in the data. More specifically, they cannot distinguish between legitimate and illegitimate variation or grave and slight errors respectively, but seem to fail to match acceptable variants because of lexical and phrasal variation or divergent grammatical structures resulting in different verb frames, word sequences and text lengths, not to talk even about acceptable variation on higher linguistic levels. Therefore, automatic scores seem to overrate surface differ-

ences and thus assign very low scores to many translations that were found to be at least acceptable by a human expert.

Considering the impact of these findings for MT evaluation purposes, it is not straightforward to assume that the differences that we have observed between the human translations are more "subtle" (in the sense of being unimportant) than the ones produced by machine translation systems. On the contrary, our analysis suggests that "good" translations are characterised by creative solutions that are not easily reproducible but that help to achieve target language readability and comprehensibility. This is a fundamental quality aspect of translation independently of its production mode. Moreover, it is difficult to see why some of the variants that we observed in the human translations selected from KOPTE, once the context shifts from human to machine translation, should be found valid in one situation and invalid in another, depending on the training and test data used for developing an MT system: A high amount of the variation found in the human translations goes back to the legitimate use of the creative and constructive powers of natural language, and it is, among others, these powers that should be mimicked by MT output.

4 Conclusion and future work

In this paper, we have studied the performance of two fully automatic MT evaluation metrics, namely BLEU and Meteor, in comparison to human translation expert evaluations on a sample of learner translations from the KOPTE corpus. The automatic scores were tested in three experiments with a varying number of reference translations and their performance was compared to the human evaluations by means of Kendall's rank correlation coefficient. The experiments suggest that both BLEU and Meteor systematically underestimate the quality of the translations tested, that is, they assign scores that, given the human expert evaluations, seem to be by far too low. Moreover, they do not consistently correlate with the human expert evaluations. Coming up with explanations for this failure is not straightforward, however, the results of our qualitative and explorative analysis suggest that lexical similarity scores are not able to cope satisfactorily neither with standard lexical variation (paraphrases etc.) nor with dissimilarities that can be traced back to the specific nature of the translation process, leave alone linguistic levels beyond the lexicon. For Meteor, this shortcoming may partly be alleviated by the provision of richer sets of synonyms and paraphrases, however, the amount of uncovered variation is still immense. In fact, it seems that many more reference translations would be needed in order to cover the whole range of legitimate variants that can be used to translate a given source text - a scenario that seems hardly feasible! So how can BLEU or Meteor scores be interpreted when they are given in MT papers? Based on our analyses, it seems clear that these scores are based on a data-driven notion of translation quality, that is, they measure the degree of compliance of a hypothesis translation with some reference set. This is insofar problematic as studies based on different reference sets cannot be compared, neither can BLEU or Meteor scores be generalised to other domains. Even more importantly, BLEU or Meteor scores cannot be used to measure a data-independent concept of quality or even the usability of a translation for a target audience which, as we have shown, depends on many more factors than just lexical surface overlap.

However, our study also leads to some open research questions. One of these questions is whether automatic evaluation scores can still be used for more coarse-grained distinctions, that is, to distinguish "really bad" translations from "really good" ones. The fine-grained distinctions made by the evaluator of KOPTE on generally rather good translations do not allow us to answer this question. Future work will also deal with a comparison of mistakes made by MT systems as opposed to human translators as well as with the question how (and which) translation-specific aspects can be applied to the evaluation of MT systems.

References

Alexandra Birch, Barry Haddow, Ulrich Germann, Maria Nadejde, Christian Buck, and Philipp Koehn. 2013. The feasibility of HMEANT as a human MT evaluation metric. In *Proceedings of the 8th Workshop on SMT*, pages 52–61.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia, editors. 2013. *Proceedings of the 8th Workshop on SMT*. ACL.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007.

(Meta-) evaluation of machine translation. In *Proceedings of the 2nd Workshop on SMT*, pages 136–158.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the 6th Workshop on SMT*, pages 85–91.

Deutsches Institut für Normung. 2006. *DIN EN 15038:2006-08: Übersetzungsdienstleistungen-Dienstleistungsanforderungen*. Beuth.

Dilek Dizdar. 2003. Skopostheorie. In *Handbuch Translation*, pages 104–107. Stauffenburg.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd International Conference on HLT*, pages 138–145.

Mireia Farrús, Marta R. Costa-Jussà, José B. Mariño, and José A. R. Fonollosa. 2010. Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proceedings of the 14th Annual Conference of the EAMT*, pages 167–173.

Christian Federmann. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. *PBML*, 98:25–35, 9.

Hans Hönig. 2003. Humanübersetzung (therapeutisch vs. diagnostisch). In *Handbuch Translation*, pages 378–381. Stauffenburg.

Brigitte Horn-Helf. 1999. *Technisches Übersetzen in Theorie und Praxis*. Franke.

Paul Kußmaul. 2000. *Kreatives Übersetzen*. Stauffenburg.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Chi-Kiu Lo and Dekai Wu. 2011. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the ACL*, pages 220–229.

Tucker Maney, Linda Sibert, Dennis Perzanowski, Kalyan Gupta, and Astrid Schmidt-Nielsen. 2012. Toward determining the comprehensibility of machine translations. In *Proceedings of the 1st PITR*, pages 1–7.

Elvira Mertin. 2006. *Prozessorientiertes Qualitätsmanagement im Dienstleistungsbereich Übersetzen*. Peter Lang.

Christiane Nord. 2003. Transparenz der Korrektur. In *Handbuch Translation*, pages 384–387. Stauffenburg.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318.

Maja Popović, Eleftherios Avramidis, Aljoscha Burchardt, Sabine Hunsicker, Sven Schmeier, Cindy Tscherwinka, David Vilar, and Hans Uszkoreit. 2013. Learning from human judgements of machine translation output. In *MT Summit*, pages 231–238.

Lothar Sachs and Jürgen Hedderich. 2009. *Angewandte Statistik. Methodensammlung mit R*. Springer.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223–231.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the 4th Workshop on SMT*, pages 259–268.

Society of Automotive Engineers. 2005. *SAE J2450:2005-08: Translation Quality Metric*. SAE.

Harold Somers. 2011. Machine translation: History, development, and limitations. In *The Oxford Handbook of Translation Studies*, pages 427–440. Oxford University Press.

Sara Stymne and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. In *Proceedings of the 8th LREC*, pages 1785–1790.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Alexander Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In *Proceedings of the EUROSPEECH*, pages 2667–2670.

Sandra Weiss and Lars Ahrenberg. 2012. Error profiling for evaluation of machine-translated text: a polish-english case study. In *Proceedings of the Eighth LREC*, pages 1764–1770.

Kevin Windle and Anthony Pym. 2011. European thinking on secular translation. In *The Oxford Handbook of Translation Studies*, pages 7–22. Oxford University Press.

Andrea Wurm. 2013. Eigennamen und Realia in einem Korpus studentischer Übersetzungen (KOPTE). *trans-kom*, 6(2):381–419.

A Examples of lexical variation in human translation

In the examples below, bold face indicates the French source.

A.1 Proper names

président gabonais
Präsidenten von Gabon
Präsidenten Gabuns
Präsidenten von Gabun
Präsident des afrikanischen Landes Gabon
gabunesischen Präsidenten

la Commission nationale de l'informatique et des libertés (CNIL)
Commission nationale de l'informatique et des libertés (CNIL)
französische Datenschutzbehörde (CNIL)
französische Datenschutzkommission CNIL
französische Datenschutzbehörde CNIL
französische Kommission für Datenschutz (CNIL)

A.2 Problematic source text elements (translation problems)

pivot de l'influence française
Stützpunkt des Einflusses Frankreichs
zentralen Figur des französischen Einfluss
Stütze für den Einfluss Frankreichs
Schlüsselfigur für den Einfluss Frankreichs
Garant für den französischen Einfluß

"doyen de l'Afrique"
obersten Würdenträgers Afrikas
"Alten Herrn von Afrika"
"Abtes von Afrika"
"Ältesten von Afrika"
"doyen de l'Afrique"

A.3 Paraphrases

sera-t-elle capable
es schaffen
fähig sein
in der Lage sein
sich als fähig erweisen

se tenir à la bonne distance
auf angemessener Distanz zu bleiben
sich nicht einzumischen
sich herauszuhalten
die gebührende Neutralität zu wahren

A.4 Culture-specific elements and underspecified source text items

la "Françafrique"
"Françafrique"
Französisch-Afrika ("Françafrique")
„Franzafrika“
"Frankafrika"
"Françafrique" d.h. der französisch beeinflussten Gebiete Afrikas

les "voitures Google", équipées de caméras à 360 degrés
mit 360-Grad-Kameras ausgestatteten "Google-Kamerawagen"
Kamera-Autos
Street-View-Wagen mit ihren 360°-Kameras
"Google-Autos", die auf dem Dach eine 360-Grad-Kamera montiert haben,
mit 360-Grad-Kameras ausgestatteten "Street View-Autos"

A.5 Source text ambiguities (syntactic and semantic)

la France a soutenu un régime autoritaire et prédateur, sans pitié pour les opposants
autoritären Systems [...], das kein Mitleid mit seinen Gegnern zeigte
hat Frankreich ohne Rücksicht auf Regimekritiker ein autoritäres Gewaltregime unterstützt
autoritäre und ausbeutende Regime [...], welches keine Gnade für seine Gegner kannte
autoritäres und angriffslustiges Regime [...], das kein Mitleid mit seinen Gegnern hatte
hat Frankreich dieses autoritäre und ausbeuterische System, ohne Mitleid mit dessen Gegnern, gestützt

justes paroles
hat die Wahrheit gesagt
hat [...] die richtigen Worte gefunden
hat die richtigen Worte gefunden
Aussage [...] war nichts als Worte
hat genau das Richtige gesagt

A.6 Numbers

une amende de 100 000 euros
Geldstrafe in Höhe von 100 000 Euro
Strafe von 100 000€
Geldstrafe von 100.000,- EUR
Geldstrafe in Höhe von 100.000 Euro
Bußgeld in Höhe von 100 000€

photographe Yann Arthus-Bertrand, 63 ans
63jährigen Fotografen Yann Arthus-Betrand
Fotographen Yann Arthus-Bertrand (63 Jahre)
Fotografen Yann Arthus-Bertrand (63)
63-jährigen Fotografen Y.A.B.
Fotografen Yann Arthus-Bertrand, 63

A.7 Colloquial or figurative speech

Je vais vite
Ich beeile mich
Ich mache es schnell
Ich bewege mich schnell
Ich hab's eilig
Ich beeile mich

résultats des petits frères
Einnahmen der Vorgänger
Verdienste zusätzlicher kleiner Artikel
Einnahmen durch andere Produkte
Erlöse von Merchandising
Einnahmen aus dem Merchandising

A.8 Source text element triggering correct and incorrect translations

65 chaînes de télévision, dont France 2 et 23 chaînes en Afrique
65 Fernsehsendern, darunter auch France 2 und 23 afrikanische Sender
65 Fernsehsendern, unter anderem France 2 und 23 Sender in Afrika
65 Fernsehsender, darunter der französische Sender France 2 und 23 afrikanische Sender
65 Fernsehkanälen, u.a. 2 in Frankreich und 23 in Afrika
65 Fernsehkanälen, darunter France 2 und 23 afrikanische Sender