Assessing the Impact of Translation Errors on Machine Translation Quality with Mixed-effects Models

Marcello Federico, Matteo Negri, Luisa Bentivogli, Marco Turchi
FBK - Fondazione Bruno Kessler
Via Sommarive 18, 38123 Trento, Italy
{federico,negri,bentivogli,turchi}@fbk.eu

Abstract

Learning from errors is a crucial aspect of improving expertise. Based on this notion, we discuss a robust statistical framework for analysing the impact of different error types on machine translation (MT) output quality. Our approach is based on linear mixed-effects models, which allow the analysis of error-annotated MT output taking into account the variability inherent to the specific experimental setting from which the empirical observations are drawn. Our experiments are carried out on different language pairs involving Chinese, Arabic and Russian as target languages. Interesting findings are reported, concerning the impact of different error types both at the level of human perception of quality and with respect to performance results measured with automatic metrics.

1 Introduction

The dominant statistical approach to machine translation (MT) is based on learning from large amounts of parallel data and tuning the resulting models on reference-based metrics that can be computed automatically, such as BLEU (Papineni et al., 2001), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), GTM (Turian et al., 2003). Despite the steady progress in the last two decades, especially for few well resourced translation directions having English as target language, this way to approach the problem is quickly reaching a performance plateau. One reason is that parallel data are a source of reliable information but, alone, limit systems knowledge to observed positive examples (*i.e.* how a sentence should be translated) without explicitly modelling any notion of error (*i.e.* how a sentence should *not* be translated). Another reason is that, as a development and evaluation criterion, automatic metrics provide a holistic view of systems' behaviour without identifying the specific issues of a translation. Indeed, the global scores returned by MT evaluation metrics depend on comparisons between translation hypotheses and reference translations, where the causes and the nature of the differences between them are not identified.

To cope with these issues and define system improvement priorities, the focus of MT evaluation research is gradually shifting towards profiling systems' behaviour with respect to various typologies of errors (Vilar et al., 2006; Popović and Ney, 2011; Farrús et al., 2012, *inter alia*). This shift has enriched the traditional MT evaluation framework with a new element, that is the actual errors done by a system. Until now, most of the research has focused on the relationship (*i.e.* the correlation) between two elements of the framework: humans and automatic evaluation metrics. As a new element of the framework, which becomes a sort of "evaluation triangle", the analysis of error annotations opens interesting research problems related to the relationships between: *i)* error types and human perception of MT quality and *ii)* error types and the sensitivity of automatic metrics.

Besides motivating further investigation on metrics featuring high correlation with human judgements (a well-established MT research sub-field, which is out of the scope of this paper), connecting the vertices of this triangle raises new challenging questions such as:

(1) Which types of MT errors have the highest impact on human perception of translation quality? Surprisingly, little prior work focused on this side of the triangle. Error annotations have been considered to highlight strengths and weaknesses of MT engines or to investigate the influence of different error types on post-editors' work. However, the direct connection between er-

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1643–1653,
October 25-29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

rors and users' preferences has been only partially understood, mainly from a descriptive standpoint and through rudimentary techniques unsuitable to draw clear-cut conclusions or reliable inferences. **(2) To which types of errors are different MT evaluation metrics more sensitive?** This side of the triangle has been even less explored. For instance, little has been done to understand which automatic metric is more suitable to assess system improvements with respect to a specific issue (*e.g.* word order or morphology) or to shed light on the joint impact of different error types on performance results calculated with different metrics.

To answer these questions, **we propose a robust statistical framework to analyse the impact of different error types**, alone and in combination, both on human perception of quality and on MT evaluation metrics' results. Our analysis is carried out **by employing linear mixed-effects models**, a generalization of linear regression models suited to model responses with fixed and random effects. Experiments are performed on data covering three translation directions (English to Chinese, Arabic and Russian). For each direction, two automatic translations were collected for around 400 sentences and were manually evaluated by expert translators through absolute quality judgements and error annotation.

Building on the advantages offered by linear mixed-effects models, our main contributions include:

- A rigorous method, novel to MT error analysis research, to relate MT issues to human preferences and MT metrics' results;

- The application of such method to three translation directions having English as source and different languages as target;

- A number of findings, specific to each language direction, which are out of the reach of the few simpler methods proposed so far.

Overall, our study has clear practical implications for MT systems' development and evaluation. Indeed, the proposed statistical analysis framework represents an ideal instrument to: *i)* identify translation issues having the highest impact on human perception of quality and *ii)* choose the most appropriate evaluation metric to measure progress towards their solution.

2 Related Work

Error analysis, as a way to identify systems' weaknesses and define priorities for their improvement, is gaining increasing interest in the MT community (Popović and Ney, 2011; Popovic et al., 2013). Along this direction, the initial efforts to develop error taxonomies covering different levels of granularity (Flanagan, 1994; Vilar et al., 2006; Farrús Cabeceran et al., 2010; Stymne and Ahrenberg, 2012; Lommel et al., 2014) have been recently complemented by investigations on how to exploit error annotations for diagnostic purposes. Error annotations of sentences produced by different MT systems, in different target languages and domains, have been used to determine the quality of translations according to the amount of errors encountered (Popovic et al., 2013), to design new automatic metrics that take into consideration human annotations (Popovic, 2012; Bojar et al., 2013), and to train classifiers that can automatic identify fine-grained errors in the MT output (Popović and Ney, 2011). The impact of edit operations on post-editors' productivity, which implicitly connects the severity of different errors to human activity, has also been studied (Temnikova, 2010; O'Brien, 2011; Blain et al., 2011), but few attempts have been made to explicitly model how fine-grained errors impact on human quality judgements and automatic metrics.

Recently, the relation between different error types, their frequency, and human quality judgements has been investigated from a descriptive standpoint in (Lommel et al., 2014; Popović et al., 2014). In both works, however, the underlying assumption that the most frequent error has also the largest impact on quality perception is not verified (in general and, least of all, across language pairs, domains, MT systems and post-editors). Another limitation of the proposed (univariate) analysis lies in the fact that it exclusively focuses on error types taken in isolation. This simplification excludes the possibility that humans, when assigning a global quality score to a translation, may be influenced not only by the error types but also by their interaction. The implications of such possibility call for a multivariate analysis capable to model also error interactions.

In (Kirchhoff et al., 2013), a statistically-grounded approach based on conjoint analysis has been used to investigate users' reactions to different types of translation errors. According to their results, word order is the most dispreferred error type, and the count of the errors in a sentence is not a good predictor of users' preferences. Though more sophisticated than methods based on rough error counts, the conjoint model is bound to several constraints that limit its usability. In particular, the application of conjoint analysis in this context requires to: *i)* operate with semi-automatically created (hence artificial) data instead of real MT output, *ii)* manually define different levels of severity for each error type (*e.g.* high/medium/low), and *iii)* limit the number of error types considered to avoid the explosion of all possible combinations. Finally, the conjoint analysis framework is not able to explicitly model variance in the translated sentences, the human annotators, and the SMT systems used to translate the source sentences. Our claim is that avoiding any possible bias introduced by these factors should be a priority in the analysis of empirical observations in a given experimental setting.

So far, the relation between errors and automatic metrics has been analysed by measuring the correlation between single or total error frequencies and automatic scores (Popović and Ney, 2011; Farrús et al., 2012). Using two different error taxonomies, both works show that the sum of the errors has a high correlation with BLEU and TER scores. Similar to the aforementioned works addressing the impact of MT errors on human perception, these studies disregard error interactions, and their possible impact on automatic scores.

To overcome these issues, we propose a robust statistic analysis framework based on mixed-effects models, which have been successfully applied to several NLP problems such as sentiment analysis (Greene and Resnik, 2009), automatic speech recognition (Goldwater et al., 2010), and spoken language translation (Ruiz and Federico, 2014). Despite their effectiveness, the use of mixed-effects models in the MT field is rather recent and limited to the analysis of human post-editions (Green et al., 2013; Läubli et al., 2013). In both studies, the goal was to evaluate the impact of post-editing on the quality and productivity of human translation assuming an ANOVA mixed model for a between-subject design, in which human translators either post-edited or translated the same texts. Our scenario is rather different as we employ mixed models to measure the influence of different MT error types - expressed as continuous fixed effects - on quality judgements and automatic quality metrics. Mixed models, having the capability to absorb random variability due to the specific experimental set-up, provide a robust multivariate method to efficiently analyse the importance of error types.

Finally, differently from all previous works, our analysis is run on language pairs having English as source and languages distant from English (in term of morphology and word-order) as target.

3 Mixed-effects Models

Mixed-effects models - or simply mixed models - like any regression model, express the relationship between a *response variable* and some *covariates* and/or *contrast factors*. They enhance conventional models by complementing *fixed effects* with so-called *random effects*. Random effects are introduced to absorb random variability inherent to the specific experimental setting from which the observations are drawn. In general, random effects correspond to covariates that are not - or cannot be - exhaustively observed in an experiment, *e.g.* the human annotators and the evaluated systems. Hence, mixed models permit to elegantly cope with experimental design aspects that hinder the applicability of conventional regression models. These are, in particular, the use of repeated and/or clustered observations that introduce correlations in the response variable that clearly violate the independence and homoscedasticity assumptions of conventional linear, ANOVA, and logistic regression models. Significance testing with mixed models is in general more powerful, *i.e.* less prone to Type II Errors, and also permits to reduce the chance of Type I Errors in within-subject designs, which are prone to the "fallacy of language-as-a-fixed-effect" (Clark, 1973).

Random effects can be directly associated to the regression model parameters, as *random intercepts* and *random slopes*, and have the same form of the generic error component of the model, *i.e.* normally distributed with zero mean and unknown variance. As random effects introduce hidden variables, mixed models are trained with Expectation Maximization, while significance testing is performed via likelihood-ratio (LR) tests.

In this work we employ mixed *linear* models to measure the influence of different MT error types, expressed as continuous fixed effects, on quality judgements or on automatic quality metrics.[1]

We illustrate mixed linear models (Baayen et al., 2008) by referring to our analysis, which addresses the relationships between a quality metric (y) and different types of errors (*e.g.* A, B, and C)[2] observed at the sentence level. For the sake of simplicity, we assume to have balanced repeated observations for one single crossed effect. That is, we have $i \in \{1, \ldots, I\}$ MT systems (our groups) each of which translated the same $j \in \{1, \ldots, J\}$ test sentences. Our response variable y_{ij} - a numeric quality score - is computed on each (sentence, system) pair, and we aim to investigate its relationship with error statistics available for each MT output, namely A_{ij}, B_{ij} and C_{ij}. A (possible) linear mixed model for our study would be:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1\, A_{ij} + \beta_2\, B_{ij} + \beta_3\, C_{ij} + \\ &\quad b_{0,i} + b_{1,i} A_{ij} + b_{2,i} B_{ij} + b_{3,i} C_i + \epsilon_{ij} \end{aligned} \quad (1)$$

The model is split into two lines on purpose. The first line shows the fixed effect component, that is intercept (β_0) and slopes ($\beta_1, \beta_2, \beta_3$) for each error type. The second line specifies the random structure of the model, which includes random intercept and slopes for each MT system and the residual error. Borrowing the notation from (Green et al., 2013), we conveniently rewrite (1) in the group-wise arranged matrix notation:

$$y_i = x_i^T \beta + z_i^T b_i + \epsilon_i \quad (2)$$

where y_i is the $J \times 1$ vector of responses, x_i is the $J \times p$ design matrix of covariates (including the intercept) with fixed coefficients $\beta \in \mathcal{R}^{p \times 1}$, z is the random structure matrix defined by $J \times q$ covariates with random coefficients $b_i \in \mathcal{R}^{q \times 1}$, and ϵ_i is the vector of residuals (in our example, $p = 4$ and $q = 4$). By packing together vectors and matrices indexed over groups i, we can rewrite the model in a general form (Baayen et al., 2008), which can represent any possible crossed-effects and random structures defined over them allowing, at the same time, for a compact model specification:

$$\begin{aligned} y &= X^T \beta + Z^T b + \epsilon \\ &\quad \epsilon \sim \mathcal{N}(0, \sigma^2 I),\ b \sim \mathcal{N}(0, \sigma^2 \Sigma), b \perp \epsilon \end{aligned} \quad (3)$$

where Σ is the relative variance-covariance $q \times q$ matrix of the random effects (now $q = 4I$), σ^2 is the variance of the per-observation term ϵ, the symbol $\perp$ denotes independence of random variables, and $\mathcal{N}$ indicates the multivariate normal distribution. While b, σ, and Σ are estimated via maximum likelihood, the single random intercept and slope values for each group are calculated subsequently. They are referred to as Best Linear Unbiased Predictors (BLUPS) and, formally, are *not* parameters of the model.

The significance of the contribution of each single parameter (*e.g.* single entries of Σ) to the goodness of fit can be tested via likelihood ratio. In this way, both the fixed and random effect structure of the model can be investigated with respect to its actual necessity to the model.

4 Dataset

For our analysis we used a dataset that covers three translation directions, corresponding to English to Chinese, Arabic, and Russian. An international organization provided us a set of English sentences together with their translation produced by two anonymous MT systems. For each evaluation item (source sentence and two MT outputs) three experts were asked to assign quality scores to the MT outputs, and a fourth expert was asked to annotate translation errors. The four experts, who were all professional translators native in the examined target languages, were carefully trained to get acquainted with the evaluation guidelines and the annotation tool specifically developed for these evaluation tasks (Girardi et al., 2014). The annotation process was carried out in parallel by all annotators over one week, resulting in a final dataset composed of 312 evaluation items for the ENZH direction, 393 for ENAR, and 437 for ENRU.

4.1 Quality Judgements

Quality judgements were collected by asking the three experts to rate each automatic translation according to a 1-5 Likert scale, where 1 means “incomprehensible translation” and 5 means “perfect translation”. The distribution of the collected annotations with respect to each quality score is shown in Figure 1. As we can see, this distribution reflects different levels of perceived quality across languages. ENZH, for instance, has the highest number of low quality scores (1 and 2), while ENRU has the highest number of high qual-

[1] Although mixed *ordinal* models (Tutz and Hennevogl, 1996) are in principle more appropriate to target quality judgements, in our preliminary investigations mixed linear models showed a significantly higher predictive power.

[2] Here, A, B and C represent three generic error classes. Their actual number in a given experimental setting will depend on the granularity of the reference error taxonomy.

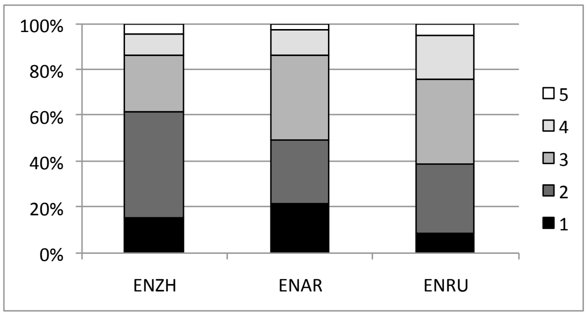

Figure 1: Distribution of quality scores.

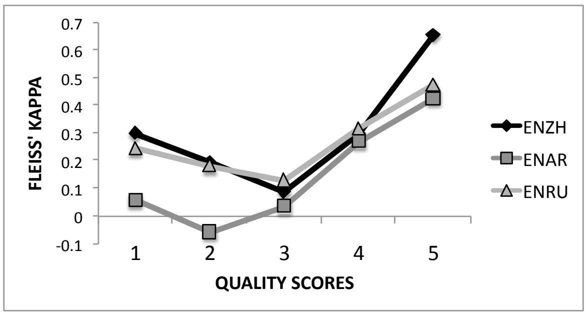

Figure 2: Class specific inter-annotator agreement.

ity scores (4 and 5).

Table 1 shows the average of all the quality scores assigned by each annotator as well as the average score obtained for each MT system. These values demonstrate the variability of annotators and systems. A particularly high variability among human judges is observed for the ENAR language direction (also reflected by the inter-annotator agreement scores discussed below), while ENZH shows the highest variability between systems. As we will see in §5.1, we successfully cope with this variability by considering systems and annotators as random effects, which allow the regression models to abstract from these differences.

	Ann1	Ann2	Ann3	Sys1	Sys2
ENZH	2.38	2.69	2.21	2.29	2.56
ENAR	2.76	2.77	1.84	2.39	2.53
ENRU	2.82	2.72	2.96	2.87	2.79

Table 1: Average quality scores per annotator and per system.

Inter-annotator agreement was computed using the *Fleiss' kappa coefficient* (Fleiss, 1971), and resulted in 22.70% for ENZH, 5.24% for ENAR, and 21.80% for ENRU. While for ENZH and ENRU the results fall in the range of "fair" agreement (Landis and Koch, 1977), for ENAR only "slight" agreement is reached, reflecting the higher annotators' variability evidenced in Table 1.

A more fine-grained agreement analysis is presented in Figure 2, where the *kappa* values are given for each score class. In general we notice a lower agreement on the intermediate quality scores, while annotators tend to agree on very bad and, even more, on good translations. In particular, we see that the agreement for ENAR is systematically lower than the values measured for the other languages on all the score classes.

4.2 Error Annotation

This evaluation task was carried out by one expert for each language direction, who was asked to identify the type of errors present in the MT output and to mark their position in the text. Since the focus of our work is the analysis method rather than the definition of an ideal error taxonomy, for the difficult language directions addressed we opted for the following general error classes, partially overlapping with (Vilar et al., 2006): *i)* reordering errors, *ii)* lexicon errors (including wrong lexical choices and extra words), *iii)* missing words, *iv)* morphology errors.

Figure 3 shows the distribution of the errors in terms of affected tokens (words) for each error type. Since token counts for Chinese are not word-based but character-based, for readability purposes the number of errors counted for Chinese translations have been divided by 2.5. Note also that morphological errors annotated for ENZH involve only 13 characters and thus are not visible in the plot. The total number of errors amounts to 16,320 characters for ENZH, 4,926 words for ENAR, and 5,965 words for ENRU.

This distribution highlights some differences between languages directions. For example, translations into Arabic and Russian present several morphology errors, while word reordering is the most frequent issue for translations into Chinese. As we will see in §5.1, error frequency does not give a direct indication of their impact on traslation quality judgements.

4.3 Automatic Metrics

In our investigation we consider three popular automatic metrics: sentence-level BLEU (Lin and Och, 2004), TER (Snover et al., 2006), and GTM (Turian et al., 2003). We compute all automatic scores by relying on a single reference and by

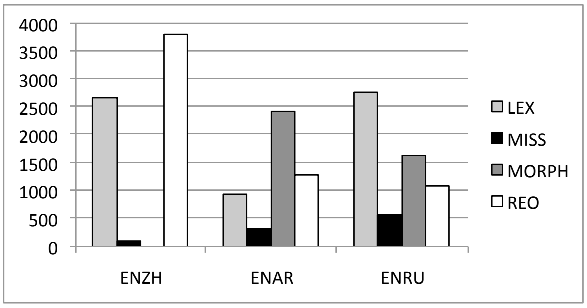

Figure 3: Distribution of error types.

means of standard packages. In particular, automatic scores on Chinese are computed at the character level. Moreover, as we use metrics as response variables for our regression models, we compute all metrics at the sentence level. The overall mean scores for all systems and languages are reported in Table 2. Differences in systems' performance can be observed for all language pairs; as we will observe in §5.2 such variability explains the effectiveness of considering the MT systems as a random effect.

	BLEU		TER		GTM	
	Sys1	Sys2	Sys1	Sys2	Sy1	Sys2
ENZH	27.95	44.11	64.52	48.13	62.15	72.30
ENAR	19.63	25.25	68.83	63.99	47.20	52.33
ENRU	27.10	31.07	60.89	54.41	53.74	56.41

Table 2: Overall automatic scores per system.

5 Experiments

To assess the impact of translation errors on MT quality we perform two sets of experiments. The first set (§5.1) addresses the relation between errors and human quality judgements. The second set (§5.2) focuses on the relation between errors and automatic metrics. In both cases, before measuring the impact of different errors on the response variable (respectively quality judgements and metrics), we validate the effectiveness of mixed linear models by comparing their prediction capability with other methods.

In all experiments, error counts of each category were normalized into percentages with respect to the sentence length and mapped in a logarithmic scale. In this way, we basically assume that the impact of errors tends to saturate above a given threshold, hypothesis that also results in better fits by our models.[3] Notice that while the chosen log-10 base is easy to interpret, linear models can implicitly adjust it. Our analysis makes use of mixed linear models incorporating, as fixed effects, the four types of errors (*lex*, *miss*, *morph* and *reo*) and their pairwise interactions (the product of the single error log counts), while their random structure depends on each specific experiment. For the experiments we rely on the R language (R Core Team, 2013) implementation of linear mixed model in the *lme4* library (Bates et al., 2014).

We assess the quality of our mixed linear models (*MLM*) by comparing their prediction capability with a sequence of simpler linear models including only fixed effects. In particular, we built five univariate models and two multivariate models. The univariate models use as covariates, respectively, the sum of all error types (*baseline*), and each of the four types of errors (*lex*, *miss*, *morph* and *reo*). The two multivariate models include all the four error types, considering them without interactions (*FLM w/o Interact.*) and with interactions (*FLM*).

Prediction performance is computed in terms of Mean Absolute Error (MAE),[4] which we estimate by averaging over 1,000 random splits of the data in 90% training and 10% test. In particular, for the human quality classes we pick the integer between 1-5 that is closest to the predicted value.

5.1 Errors vs. Quality Judgements

The response variable we target in this experiment is the quality score produced by human annotators. Our measurements follow a typical within-subject design in which all the 3 annotators are exposed to the same conditions (levels of the independent variables), corresponding in our case to perfectly balanced observations from 2 MT systems and N sentences. This setting results in repeated or clustered observations (thus violating independence) corresponding to groups which naturally identify possible random effects,[5] namely the annotators (3 levels with 2xN observations each), the systems (2 levels and 3xN observations each), and the sen-

[3] In other words, we assume that human sensitivity to errors follows a log-scale law: *e.g.* more sensitive to variations in the interval [1-10] that in the interval [30-40].

[4] MAE is calculated as the average of the absolute errors $|f_i - y_i|$, where f_i is the prediction of the model and y_i the true value for the i^{th} instance. As it is a measure of error, lower MAE scores indicate that our predictions are closer to the true values of each test instance.

[5] In all our experiments, random effects are limited to random shifts since preliminary experiments also including random slopes did not provide consistent results.

Model	ENZH	ENAR	ENRU
baseline	0.58	0.73	0.67
lex	0.67	0.78	0.72
miss	0.72	0.89	0.74
morph	0.72	0.89	0.74
reo	0.70	0.82	0.76
FLM w/o Interact.	0.59	0.77	0.65
FLM	0.57	0.72	0.63
MLM	0.53	0.61	0.61

Table 3: Prediction capability of human judgements (MAE).

Error	ENZH	ENAR	ENRU
Intercept	4.29	$3.79_\bullet$	4.21
lex	-1.27	-0.96	-1.12
miss	-1.76	-0.90	-1.30
morph	$-0.48_\circ$	-0.83	-0.51
reo	-1.01	-0.75	-0.18
lex:miss	1.00	0.39	0.68
lex:morph	-	0.29	0.32
lex:reo	0.50	0.21	-
miss:morph	-	0.35	-
miss:reo	0.54	0.33	-
morph:reo	-	0.37	-

Table 4: Effect of translation errors on MT quality perception on all judged sentences. Reported coefficients (β) are all statistically significant with $p \leq 10^{-4}$, except those marked with $_\bullet$ ($p \leq 10^{-3}$), and $_\circ$ ($p \leq 10^{-2}$).

tences (N levels with 6 observations each). In principle, such random effects permit to remove systematic biases of individual annotators, single systems and even single sentences, which are modelled as random variables sampled from distinct populations.

Table 3 shows a comparison of the **prediction capability** of the mixed model[6] with simpler approaches. While the good performance achieved by our strong baseline cannot be outperformed by separately counting the number of errors of a single type, lower MAE results are obtained by methods based on multivariate analysis. Among them, FLM brings the first consistent improvements over the baseline by considering error interactions, while MLM leads to the lowest MAE due to the addition of random effects. The importance of random effects is particularly evidenced by ENAR (12 points below the baseline). Indeed, as discussed in §4.1, for this language combination human annotators show the lowest agreement score. This variability, which hides the smaller differences in systems' behaviour, demonstrates the importance of accounting for the erratic factors that might influence empirical observations in a given setting. The good performance achieved by MLM, combined with their high descriptive power,[7] motivates their adoption in our study.

Concerning the analysis of **error impact**, Table 4 shows the statistically significant coefficients for the full-fledged MLM models for each translation direction. By default, all reported coefficients have p-values $\leq 10^{-4}$, while those marked with $\bullet$ and $\circ$ have respectively p-values $\leq 10^{-3}$ and $\leq 10^{-2}$. Slope coefficients basically show the impact of different error types (alone and in combination) on human quality scores. Those that are not statistically significant are omitted as they do not increase the fitting capability of our model. As can be seen from the table, such impact varies across the different language combinations. While for ENZH and ENRU *miss* is the error having the highest impact (highest decrement with respect to the intercept), the most problematic error for ENAR is *lex*. It is interesting to observe that positive values for error combinations indicate that their combined impact is lower that the sum of the impact of the single errors. For instance, while for ENZH a one-step increment in *lex* and *miss* errors would respectively cause a reduction in the human judgement of 1.27 and 1.76, their occurrence in the same sentence would be discounted by 1.00. This would result in a global judgement of 2.26 (4.29 -1.27 -1.76 +1.00) instead of 1.26. While for ENAR this phenomenon can be observed for all error combinations, such discount effects are not always significant for the other two language pairs. The existence of discount effects of various magnitude associated to the different error combinations is a novel finding made possible by the adoption of mixed-effect models.

Another interesting observation is that, in contrast with the common belief that the most frequent errors have the highest impact on human quality judgements, our experiments do not reveal such strict correlation (at least for the examined language pairs). For instance, for ENZH and ENRU the impact of *miss* errors is higher than the impact of other more frequent issues.

[6] Note that the mixed model used in prediction does not include the random effect on sentences since the training samples do not guarantee sufficient observations for each test sentence.

[7] Note that the strong baseline used for comparison is not capable to describe the contribution of the different error types.

	BLEU score			TER			GTM		
Model	ENZH	ENAR	ENRU	ENZH	ENAR	ENRU	ENZH	ENAR	ENRU
baseline	12.4	9.8	12.2	15.7	13.4	14.4	9.8	10.6	11.5
lex	12.9	10.4	13.0	16.3	13.8	14.9	9.7	10.9	12.1
miss	13.8	10.5	14.1	17.3	14.2	16.4	10.5	11.1	13.2
morph	13.9	10.3	13.6	17.5	13.8	16.3	10.5	10.9	13.1
reo	13.7	10.5	14.0	17.4	14.1	16.3	10.4	11.1	13.1
FLM w/o Interact.	12.9	9.9	12.2	16.3	13.5	14.4	9.7	10.7	11.7
FLM	12.3	9.7	12.1	15.6	13.4	14.3	9.4	10.6	11.6
MLM	10.8	9.5	12.0	14.7	13.0	14.2	8.9	10.5	11.6

Table 5: Prediction capability of BLEU score, TER and GTM (MAE).

5.2 Errors vs. Automatic Metrics

In this experiment, the response variable is an automatic metric which is computed on a sample of MT outputs (which are again perfectly balanced over systems and sentences) and a set of reference translations. As no subjects are involved in the experiment, random variability is assumed to come from the involved systems, the tested sentences, and the unknown missing link between the covariates (error types) and the response variable which is modelled by the residual noise. Notice that, in this case, the random effect on the sentences also incorporates in some sense the randomness of the corresponding reference translations, which are themselves representatives of larger samples.

The **prediction capability** of the mixed model, in comparison with the simpler ones, is reported in Table 5. Also in this case, the low MAE achieved by the baseline is out of the reach of univariate methods. Again, small improvements are brought by FLM when considering error interactions, whereas the most visible gains are achieved by MLM due to their control of random effects. This is more evident for some language combinations and can be explained by the differences in systems' performance, a variability factor easily absorbed by random effects. Indeed, the largest MAE decrements over the baseline are always observed for ENZH (for which the overall mean results reported in Table 2 show the largest differences) and the smallest decrements relate to language/metric combinations where systems' behaviour is more similar (*e.g.* ENRU/GTM).

Concerning the analysis of **error impact**, Table 6 shows how different error types (alone and in combination) influence performance results measured with automatic metrics. To ease interpretation of the reported figures we also show Pearson and Spearman correlations of each set of coefficients (excluding intercept estimates) with their corresponding coefficients reported in Table 4. In fact, our primary interest in this experiment is to see which metrics show a sensitivity to specific error types similar to human perception. As we can see, the coefficients for each metric significantly vary depending on the language, for the simple reason that also the distribution and co-occurrence of errors vary significantly across the different languages and MT systems. Remarkably, for some translation directions, some of the metrics show a sensitivity to errors that is very similar to that of human judges. In particular, BLEU for ENZH and ENAR, and GTM for ENZH show a very high correlation with the human sensitivity to translation errors, with Pearson correlation coefficient $\geq$ 0.97. For ENRU, the best Pearson correlation is instead achieved by TER (-0.78).

Besides these general observations, a closer look at the reported scores brings additional findings. In three cases (BLEU for ENZH, GTM for ENZH and ENAR) the analysed metrics are most sensitive to the same error type that has the highest influence on human judgements (according to Table 4, these are *miss* for ENZH and ENRU, *lex* for ENAR). On the contrary, in one case (TER for ENZH) the analysed metric is insensitive to the error type (*miss*) which has the highest impact on human quality scores. From a practical point of view, these remarks provide useful indications about the appropriateness of each metric to highlight the deficiencies of a specific system and to measure improvements targeting specific issues. As a rule of thumb, for instance, to measure improvements of an ENZH system with respect to *miss*ing words, it would be more advisable to use BLEU or GTM instead of TER.[8]

[8] Note that this conclusion holds for our data sample, in which different types of errors co-occur and only one reference translation is available. In such conditions, our regression model shows that TER is not influenced by *miss* errors in a statistically significant way. This does not mean that TER is insensitive to missing words when occurring in isolation,

	BLEU score			TER			GTM		
Error	ENZH	ENAR	ENRU	ENZH	ENAR	ENRU	ENZH	ENAR	ENRU
Intercept	60.55$_{\square}$	38.45$_{\circ}$	51.73	32.41$_{\square}$	52.25$_{\bullet}$	33.4$_{\bullet}$	83.57$_{\circ}$	60.11$_{\bullet}$	75.38
lex	-18.78	-9.25	-16.57	16.87	9.66	18.45	-13.63	-7.60	-16.13
miss	-23.20	-10.41	-6.75	-	-	8.24	-14.87	-	-5.98
morph	-	-9.97	-12.65	-	8.90	11.41	-	-6.60	-10.42
reo	-13.27	-7.62	-10.57	14.44	9.81	6.39	-7.29	-5.50	-7.03
lex:miss	14.37	4.97$_{\circ}$	-	-	-	-	8.24$_{\bullet}$	-	-
lex:morph	-	-	5.27$_{\bullet}$	-	-	-5.22$_{\circ}$	-	-	4.92
lex:reo	8.57	3.57$_{\circ}$	5.40$_{\bullet}$	-7.24$_{\circ}$	-4.35$_{\circ}$	-	5.46	3.22$_{\circ}$	3.65$_{\square}$
miss:morph	-	4.44$_{\circ}$	-	-	-	-	-	-	-
miss:reo	6.74$_{\circ}$	-	4.30	-	-	-6.38$_{\circ}$	5.07$_{\circ}$	-	4.71$_{\circ}$
morph:reo	-	3.81$_{\bullet}$	-	-	-4.97$_{\bullet}$	-	-	2.57$_{\circ}$	-
Pearson	**0.98**	**0.97**	0.70	-0.58	-0.78	**-0.78**	**0.98**	0.78	0.74
Spearman	**0.97**	**0.91**	0.73	-0.57	-0.59	**-0.80**	**0.97**	0.59	0.76

Table 6: Effect of translation errors on BLEU score, TER and GTM on all judged sentences and correlation with their corresponding effects on human quality scores (from Table 4). Reported coefficients (β) are statistically significant with $p \leq 10^{-4}$, except those marked with $_{\bullet}$ ($p \leq 10^{-3}$), $_{\circ}$ ($p \leq 10^{-2}$) and $_{\square}$($p \leq 10^{-1}$).

Similar considerations also apply to the analysis of the impact of error combinations. The same discount effects that we noticed when analysing the impact of errors' co-occurrence on human perception (§5.1) are evidenced, with different degrees of sensitivity, by the automatic metrics. While some of them substantially reflect human response (*e.g.* BLEU and GTM for ENZH), in some cases we observe either the insensitivity to specific combinations (mostly for ENAR), or a higher sensitivity compared to the values measured for human assessors (mostly for ENRU, where the impact of *miss:reo* combinations is discounted - hence underestimated - by all the metrics).

Despite such small differences, the coherence of our results with previous findings (§5.1) suggests the reliability of the applied method. Completing the picture along the side of the MT evaluation triangle which connects error annotations and automatic metrics, our findings contribute to shed light on the existing relationships between translation errors, their interaction, and the sensitivity of widely used automatic metrics.

6 Conclusion

We investigated the MT evaluation triangle (having as corners *automatic metrics*, *human quality judgements* and *error annotations*) along the two less explored sides, namely: *i)* the relation between MT errors and human quality judgements and *ii)* the relation between MT errors and automatic metrics. To this aim we employed a robust statistical analysis framework based on linear mixed-effects models (the first contribution of the paper), which have a higher descriptive power than simpler methods based on the raw count of translation errors and are less artificial compared to previous statistically-grounded approaches.

Working on three translation directions having Chinese, Arabic and Russian as target (our second contribution), we analysed error-annotated translations considering the impact of specific errors (alone and in combination) and accounting for the variability of the experimental set-up that originated our empirical observations. This led us to interesting findings specific to each language pair (third contribution). Concerning the relation between MT errors and quality judgements, we have shown that: *i)* the frequency of errors of a given type does not correlate with human preferences, *ii)* errors having the highest impact can be precisely isolated and *iii)* the impact of error interactions is often subject to measurable and previously unknown "discount" effects. Concerning the relation between MT errors and automatic metrics (BLEU, TER and GTM), our analysis evidenced significant differences in the sensitivity of each metric to different error types. Such differences provide useful indications about the most appropriate metric to assess system improvements with respect to specific weaknesses. If learning from errors is a crucial aspect of improving expertise, our method and the resulting empirical findings represent a significant contribution towards a

but that TER becomes less sensitive to such errors when they co-occur with other types of errors. Overall, our experiments show that when MT outputs contain more than one error type, automatic metrics show different levels of sensitivity to each specific error type.

more informed approach to system development, improvement and evaluation.

Acknowledgements

This work has been partially supported by the EC-funded project MateCat (ICT-2011.4.2-287688).

References

Harald R. Baayen, Douglas J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker, 2014. *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-6.

Frédéric Blain, Jean Senellart, Holger Schwenk, Mirko Plitt, and Johann Roturier. 2011. Qualitative analysis of post-editing for high quality machine translation. In Asia-Pacific Association for Machine Translation (AAMT), editor, *Machine Translation Summit XIII*, Xiamen (China), 19-23 sept.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Herbert H. Clark. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior*, 12(4):335–359.

Mireia Farrús, Marta R. Costa-jussà, and Maja Popović. 2012. Study and correlation analysis of linguistic, perceptual, and automatic machine translation evaluations. *J. Am. Soc. Inf. Sci. Technol.*, 63(1):174–184, January.

Mireia Farrús Cabeceran, Marta Ruiz Costa-Jussà, José Bernardo Mariño Acebal, José Adrián Rodríguez Fonollosa, et al. 2010. Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation (EAMT)*.

Mary Flanagan. 1994. Error classification for mt evaluation. In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 65–72.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5).

Christian Girardi, Luisa Bentivogli, Mohammad Amin Farajian, and Marcello Federico. 2014. Mt-equal: a toolkit for human assessment of machine translation output. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 120–123, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Sharon Goldwater, Daniel Jurafsky, and Christopher D. Manning. 2010. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Spence Green, Jeffrey Heer, and Christopher D. Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 439–448. ACM.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 503–511, Stroudsburg, PA, USA. Association for Computational Linguistics.

Katrin Kirchhoff, Daniel Capurro, and Anne M. Turner. 2013. A conjoint analysis framework for evaluating user preferences in machine translation. *Machine Translation*, pages 1–17.

Richard J. Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 (1):159–174.

Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. 2013. Assessing Post-Editing Efficiency in a Realistic Translation Environment. In Michel Simard Sharon O'Brien and Lucia Specia (eds.), editors, *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, pages 83–91, Nice, France.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of Coling 2004*, pages 501–507, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit.

2014. Using a new analytic measure for the annotation and analysis of mt errors on real data. In *Proceedings of the 17th Conference of the European Association for Machine Translation (EAMT)*, Dubrovnik, Croatia, June.

Sharon O'Brien. 2011. *Cognitive Explorations of Translation*. Bloomsbury Studies in Translation. Bloomsbury Academic.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Research Report RC22176, IBM Research Division, Thomas J. Watson Research Center.

Maja Popović and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Comput. Linguist.*, 37(4):657–688, December.

Maja Popovic, Eleftherios Avramidis, Aljoscha Burchardt, Sabine Hunsicker, Sven Schmeier, Cindy Tscherwinka, David Vilar, and Hans Uszkoreit. 2013. Learning from human judgments of machine translation output. In *Proceedings of the MT Summit XIV*. Proceedings of MT Summit XIV.

Maja Popović, Arle Lommel, Aljoscha Burchardt, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Relations between different types of post-editing operations, cognitive effort and temporal effort. In *Proceedings of the 17th Conference of the European Association for Machine Translation (EAMT)*, Dubrovnik, Croatia, June.

Maja Popovic. 2012. Class error rates for evaluation of machine translation output. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 71–75, Montréal, Canada, June. Association for Computational Linguistics.

R Core Team, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Nick Ruiz and Marcello Federico. 2014. Assessing the Impact of Speech Recognition Errors on Machine Translation Quality. In *11th Conference of the Association for Machine Translation in the Americas (AMTA)*, Vancouver, BC, Canada.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *5th Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, Massachusetts, August.

Sara Stymne and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Irina Temnikova. 2010. Cognitive evaluation approach for a controlled language post-editing experiment. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Joseph P. Turian, I. Dan Melamed, and Luke Shen. 2003. Evaluation of machine translation and its evaluation. In *Proceedings of the MT Summit IX*.

Gerhard Tutz and Wolfgang Hennevogl. 1996. Random effects in ordinal regression models. *Computational Statistics & Data Analysis*, 22(5):537–557.

David Vilar, Jia Xu, Luis Fernando dHaro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 697–702.