# Quantitative fine-grained human evaluation of machine translation systems

Klubička, Filip; Toral , Antonio; Sánchez-Cartagena, Víctor M.

*Published in:*
Machine Translation

*DOI:*
10.1007/s10590-018-9214-x



*Publication date:*
2018

Mach Translat (2018) 32:195–215
https://doi.org/10.1007/s10590-018-9214-x

# Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian

**Filip Klubička**[1] · **Antonio Toral**[2] · **Víctor M. Sánchez-Cartagena**[3]

Received: 24 August 2017 / Accepted: 27 January 2018 / Published online: 10 February 2018


**Abstract** This paper presents a quantitative fine-grained manual evaluation approach to comparing the performance of different machine translation (MT) systems. We build upon the well-established multidimensional quality metrics (MQM) error taxonomy and implement a novel method that assesses whether the differences in performance for MQM error types between different MT systems are statistically significant. We conduct a case study for English-to-Croatian, a language direction that involves translating into a morphologically rich language, for which we compare three MT systems belonging to different paradigms: pure phrase-based, factored phrase-based and neural. First, we design an MQM-compliant error taxonomy tailored to the relevant linguistic phenomena of Slavic languages, which made the annotation process feasible and accurate. Errors in MT outputs were then annotated by two annotators following this taxonomy. Subsequently, we carried out a statistical analysis which showed that the best-performing system (neural) reduces the errors produced by the worst system (pure phrase-based) by more than half (54%). Moreover, we conducted an additional analysis of agreement errors in which we distinguished between short (phrase-level) and long distance (sentence-level) errors. We discovered that phrase-based MT approaches

✉ Filip Klubička
filip.klubicka@mydit.ie

Antonio Toral
a.toral.ruiz@rug.nl

Víctor M. Sánchez-Cartagena
vmsanchez@prompsit.com

1 Dublin Institute of Technology, Dublin, Ireland

2 University of Groningen, Groningen, The Netherlands

3 Prompsit Language Engineering, Alacant, Spain

are of limited use for long distance agreement phenomena, for which neural MT was found to be especially effective.

**Keywords** Neural machine translation · Statistical machine translation · Phrase-based machine translation · Factored models · Human evaluation · Error annotation · Multidimensional quality metrics (MQM)

## 1 Introduction

A machine translation (MT) paradigm based on deep neural networks, usually referred to as neural MT (NMT) (Bahdanau et al. 2015), has emerged in the past few years. This has disrupted the MT field since NMT, despite its infancy, has already surpassed the performance of phrase-based MT (PBMT) (Koehn et al. 2003), the mainstream approach to date.

The vast potential of NMT in terms of overall performance scores, be those automatic (e.g. BLEU) or human (e.g. system rankings) was, for example, showcased in 2016 news translation shared task at WMT,[1] where NMT systems significantly outperformed PBMT in 8 of the 9 language directions submitted where NMT systems were submitted, according to human evaluations (system rankings). In these evaluations, users (mainly MT researchers) were presented with a source-language sentence, its reference translation and a set of machine translations produced by the different systems submitted to the shared task. They had to rank the machine translations.

Additionally, monolingual direct assessment adequacy and fluency evaluations were also carried out in WMT 2016 for translations directions into English. In these evaluations, users had only to give an adequacy and fluency score to individual translations. Whereas the language pairs for which NMT outperformed PBMT according to the adequacy evaluation completely matched those in the system ranking (the only language pair in which NMT did not outperform PBMT was Russian-to-English), the fluency direct assessment showed that NMT output is more fluent than PBMT output for all the language pairs evaluated (including Russian-to-English).

In 2017 edition of the same shared task,[2] the trend has gained strength and, for all language directions, the best-performing submitted system either follows the NMT architecture or is a hybrid system that includes an NMT component.

The fine-grained human evaluation presented in this paper greatly differs from WMT evaluation: instead of just ranking translations, the annotators had to classify the errors contained in each translation produced by the MT systems being evaluated according to a complete error hierarchy and choose the particular tokens that contains the error.

Considering the high overall performance of NMT, researchers have in the past year attempted to analyse the potential of NMT in more detail. While overall scores, such as those obtained in WMT evaluation, give an indication of the general performance of a system, they do not shed light on the strengths and weaknesses of this new paradigm to

[1] http://www.statmt.org/wmt16/translation-task.html.

[2] http://www.statmt.org/wmt17/translation-task.html.

MT. Hence, two recent papers have looked at automatically conducting multifaceted evaluations:

- Bentivogli et al. (2016) performed a detailed analysis of the English-to-German language direction, comparing state-of-the-art PBMT and NMT systems on transcribed speeches. Their findings show that NMT (i) decreases post-editing effort, (ii) degrades faster than PBMT with sentence length and (iii) improves notably on reordering and inflection.
- Toral and Sánchez-Cartagena (2017) carried out a series of analysis and evaluations for NMT and PBMT systems on the news domain for 9 language pairs. Their research corroborated the findings of Bentivogli et al. (2016) regarding NMT's excellent performance on reordering and inflection and its degradation with sentence length. In addition to that, Toral and Sánchez-Cartagena's findings show that NMT systems (i) exhibit higher inter-system variability, (ii) lead to more fluent outputs and (iii) perform more reordering than PBMT, but less than hierarchical PBMT.

A limitation of these analyses lies in the fact that all of them were performed automatically (e.g. reordering and inflection errors were detected based on automatic evaluation metrics). More recently, other authors have performed human analyses of NMT's strengths and weaknesses in comparison with PBMT and rule-based paradigms. Such human evaluations do not suffer from the potential biases introduced by automatic tools employed in the above papers.

- Burchardt et al. (2017) presented a study based on an error categorization specifically tailored to the English–German language pair (in both directions) and a test set carefully designed in order to cover the most relevant linguistic phenomena. They conclude that NMT systems are able to produce translations that resemble those produced by rule-based MT without using explicit linguistic information.
- Popović (2017) also targeted the English–German language pair and identified language-related issues in the outputs of NMT and PBMT systems. She concluded that NMT systems are better than PBMT ones in handling verbs, English noun collocations, German compound words, phrase structure and articles, while PBMT systems perform better when dealing with prepositions, translation of English (source) ambiguous words and generation of English (target) continuous tenses. As the issues are complementary between the two MT paradigms analysed, results suggest that hybridisation between them could be a promising way forward.
- Castilho et al. (2017) evaluated the performance of NMT versus PBMT for three different translation domains: e-commerce product listings, patents and massive open online courses. They performed error analysis with an error taxonomy consisting of seven categories for patent translation from Chinese to English. The analysis showed that NMT made more omission errors than PBMT, while PBMT systems made more errors related to sentence structure than NMT. Overall, they concluded that, according to human evaluation, NMT has not fully reached the quality of PBMT.

This paper adds to the body of research dealing with manual analysis of NMT systems by conducting a detailed human analysis of the outputs produced by NMT and

PBMT systems when translating news texts in the English-to-Croatian language direction. We manually annotate the errors found according to a detailed error taxonomy that is compliant with the hierarchical listing of issue types defined as part of the multidimensional quality metrics (MQM) (Lommel et al. 2014a). First, we define an error taxonomy that is relevant to the problematic linguistic phenomena of this language pair. Subsequently, we annotate the errors produced by 3 state-of-the-art translation systems that belong to the following paradigms: PBMT, factored PBMT (Koehn and Hoang 2007) and NMT. Finally, we analyse the annotations and draw conclusions.

This paper's main contribution can thus be summarised as follows:

1. We conduct one of the first human fine-grained error analyses of NMT in the literature and, to the best of our knowledge, the first one in which a Slavic language is involved.
2. We analyse NMT in comparison not only to pure PBMT and hierarchical PBMT, as in other previous work, but also with respect to factored models.
3. We develop an MQM-compliant error taxonomy for Slavic languages. It is much more detailed in terms of error categories than that followed by Castilho et al. (2017) in their Chinese-to-English human evaluation, to account for the grammatical features of Slavic languages. Additionally, unlike the taxonomies used by Burchardt et al. (2017) and Popović (2017), ours is not restricted to a single language pair, and is at the same time based on a well-known error categorization framework (MQM).
4. Unlike Burchardt et al. (2017) and Popović (2017), we included two annotators in our evaluation so that each sentence is annotated twice. This allows us to compute inter-annotator agreement, which increases the reliability of our results.
5. We also employ a statistically grounded approach to analyzing and interpreting the results of MQM error annotation that goes beyond simple counting of errors.

This paper builds upon our recent work on this topic (Klubička et al. 2017), which is here extended in a number of directions:

1. We have performed additional categorisation and analysis of agreement errors, in order to investigate whether there is a difference in the number of agreement errors produced in regards to their scope, i.e. we looked at whether the reduction in agreement errors equally affect phrase (or short distance) agreement and sentence (or long distance) agreement.
2. We have included some examples of sentences from the dataset used in the experiments to better illustrate the different MQM error types.
3. We have included a more detailed discussion, expanded some points and added an explanation of the statistics calculated from the MQM annotation.

The rest of the paper is organized as follows. Section 2 describes the MT systems and the datasets used in our experiments. Section 3 includes the definition of the error taxonomy and explains the annotation setup and guidelines given to annotators. Next, Sect. 4 presents the results obtained and their discussion. Section 5 describes the additional annotation focused on agreement errors and analysis thereof. Finally, Sect. 6 outlines the conclusions and lines of future work.

## 2 MT systems and datasets

This section describes the MT systems and the datasets used in our experiments. We built PBMT, factored PBMT and NMT systems.

The three systems were trained on the same parallel data. We considered a set of publicly available English–Croatian parallel corpora, comprising the DGT Translation Memory,[3] HrEnWaC,[4] JRC Acquis,[5] OpenSubtitles 2013,[6] SETIMES[7] and TED talks[8], many of which can be obtained from OPUS[9] (Tiedemann 2009, 2012). We concatenated all of these corpora and performed cross-entropy based data selection (Moore and Lewis 2010) using the development set. Once the data is ranked we keep the 25% highest-ranked sentence pairs (4,786,516). Data selection was carried out in order to speed up training and discard the training parallel sentences that are too different from the domain of the development and test sets (news) and hence could have a negative impact on the results.

PBMT systems also require monolingual data for language modeling. To this end we concatenated the hrWaC corpus (Ljubešić and Klubička 2014) with the target side of the aforementioned parallel corpora.

As our development set we used the first 1000 sentences of the English test set used at the WMT12 news translation task,[10] translated by a professional translator into Croatian. Similarly, our test set is comprised of the first 1000 sentences of the English test set of the WMT13 translation task,[11] again manually translated into Croatian.

The PBMT system was built with Moses v3.0[12] (Koehn et al. 2007). In addition to the default models we also used hierarchical reordering (Galley and Manning 2008), an operation sequence model (Durrani et al. 2011) and a bilingual neural language model (Devlin et al. 2014).

The factored PBMT system maps one factor in the source language (surface form) to two factors in the target (surface form and morphosyntactic description). This system is described in detail by Sánchez-Cartagena et al. (2016).

The NMT system is based on the sequence-to-sequence architecture with attention (Bahdanau et al. 2015) and it was built with Nematus (Sennrich et al. 2017). We applied sub-word segmentation with byte pair encoding (Sennrich et al. 2016) jointly on the source and target languages. We performed 85,000 join operations. We defined a hidden layer size of 1000 and an embedding layer size of 620. We used Adadelta (Zeiler 2012) with a minibatch size of 80, and reshuffled the training set

[3] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory.

[4] https://www.clarin.si/repository/xmlui/handle/11356/1058.

[5] http://tinyurl.com/CroatianAcquis.

[6] http://www.opensubtitles.org/.

[7] http://opus.nlpl.eu/SETIMES2.php.

[8] http://opus.nlpl.eu/TedTalks.php.

[9] http://opus.nlpl.eu/.

[10] http://www.statmt.org/wmt12/translation-task.html.

[11] http://www.statmt.org/wmt13/translation-task.html.

[12] https://github.com/moses-smt/mosesdecoder/tree/RELEASE-3.0.

**Table 1** Automatic evaluation (BLEU and TER scores) of the 3 MT systems

| System | BLEU | TER |
|---|---|---|
| PBMT | 0.2544 | 0.6081 |
| Factored PBMT | 0.2700 | 0.5963 |
| NMT | 0.3085 | 0.5552 |

between epochs. We applied gradient clipping (Pascanu et al. 2013) with a cutoff of 1.0. Training was run for 10 days and a model was saved every 4.5 h. We decoded the test set using an ensemble of four models. These were the four models with the highest BLEU scores on the development set.

Table 1 reports the scores obtained in terms of the BLEU (Papineni et al. 2002) and TER (Snover et al. 2006) automatic evaluation metrics on the three systems previously described. It can be observed from the table that the use of factored models leads to a substantial improvement upon pure PBMT (6% relative in terms of BLEU). NMT allows us to obtain a further notable improvement; 14% relative in terms of BLEU compared to the factored PBMT system and 21% compared to the initial PBMT system. All the differences are statistically significant according to paired bootstrap resampling (Koehn 2004) ($p \leq 0.05$, 1000 iterations).

## 3 Error analysis

The fact that Croatian is rich in inflection, has rather free word order and other similar phenomena not present in English gives rise to specific translation issues. For example, grammatical categories that do not exist in English, like gender or case inflections in nouns, may be particularly hard to generate reliably in a Croatian translation. We built our factored PBMT system (cf. Sect. 2) aiming to directly address such issues. Similarly motivated was our goal to find out how an NMT system would grapple with the same issues. Existing research on this tells us that both systems should lead to improvements on such linguistic aspects. However, this would happen for different reasons: factored SMT deals with explicit linguistic knowledge about grammatical categories, while NMT combined with sub-word representation (e.g. byte pair encoding) solves the problem implicitly in an unsupervised manner, without actually knowing what the grammatical categories are.

Indeed, as shown in Sect. 2, both systems lead to significant improvements compared to the pure PBMT system in terms of automatic evaluation metrics. However, as is the nature of automatic scoring methods, these provide solely an overall score for each system, but do not indicate whether any of the linguistic problems mentioned earlier have been addressed by the systems. Hence, the question of whether the linguistic quality (or rather, grammaticality) of the output is improved has not been answered by automatic evaluation. Are cases and gender handled better? Has agreement been improved?

In order to provide answers to these research questions, we decided to thoroughly compare these systems by systematically analyzing their outputs via manual error analysis. In this way we can obtain a more complete picture of what is happening in the

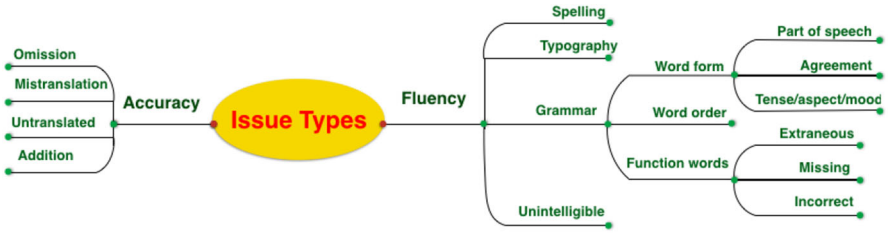


**Fig. 1** The core set of error categories proposed by the MQM guidelines

translation, which can provide pointers on where to act to obtain further improvements in the future. In the remainder of this section, we describe the annotation framework, overall annotation process and show the level of agreement between the annotators who took part in the process.

### 3.1 Multidimensional quality metrics and the Slavic tagset

We decided to make use of the MQM framework, developed in the QTLaunchpad project,[13] for performing the task of manual evaluation via error analysis. It is a framework for describing and defining custom translation quality metrics. It provides a flexible vocabulary of quality issue types and a mechanism for applying them to generate quality scores. It does not impose a single metric for all uses, but rather provides a comprehensive catalogue of quality issue types, with standardized names and definitions, that can be used to describe particular metrics for specific tasks.

The main reason we chose the MQM framework was the flexibility of the issue types and their granularity; it gave us a reliable methodology for quality assessment, that still allowed us to choose which error tags we wanted to use.

The MQM guidelines propose a great variety of tags on several annotation layers.[14] However, the full tagset is too comprehensive to be viable for any annotation task, so the process begins with choosing the tags to use in accordance with our research questions. It is good practice to start with the so-called core tagset, a default set of evaluation metrics (i.e. error categories) proposed by the MQM guidelines, shown in Fig. 1.

However, given the morphological complexity of Croatian and the way our MT systems were constructed, we found that these core categories were not detailed enough, or rather, did not allow us to conduct an analysis of the specific phenomena we were interested in. Some categories that were of interest to us, like specific *Agreement* types, were not present in the tagset, while some errors, such as *Typography*, were irrelevant to our research questions.

For these reasons, we defined our own set of tags by modifying the core set, rearranging the hierarchy, adding new tags and removing those that were of little relevance. We call this new tagset "the Slavic tagset", as its expansion allows for the identification

[13] http://www.qt21.eu/mqm-definition/definition-2015-06-16.html.

[14] http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html.

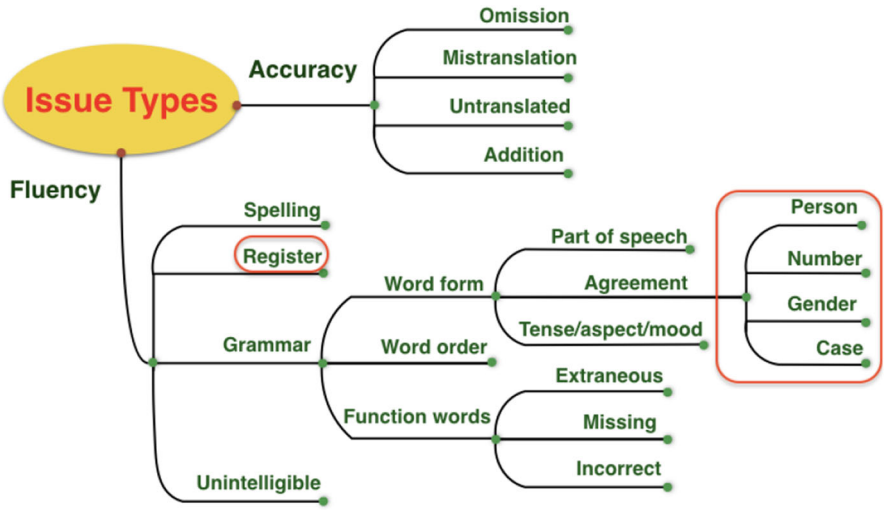


**Fig. 2** The Slavic tagset, a modified version of the MQM core tagset. The additional categories are highlighted with a red rectangle

of grammatical errors which are commonly shared by Slavic languages. This tagset is outlined in Fig. 2.

As evidenced by a comparison of the two figures, we did not change anything about the *Accuracy* branch, but rather modified *Fluency*. As mentioned earlier, we removed *Typography*, but added *Register* in its place. *Register* was included because preliminary insights into the data showed a potential usefulness for annotating a breach of standardness, which has indeed cropped up a couple of times in the systems' outputs. For example, sometimes a synonym for a word can be used, one that is a correct translation in a very general sense, but is actually sub-standard and would not normally be found in that sentence or that particular context [e.g. "She was the first **woman** in space." should be translated as "Bila je prva **žena** u svemiru.", but is instead translated as "Bila je prva **ženska** u svemiru.", roughly corresponding to "She was the first **broad** in space." (broad, n. = woman, informal)]

In addition to this change, and much more importantly, we added another level to the hierarchy, specifically to the *Agreement* error tag, which we expanded to cover the specific grammatical categories that need to agree in Croatian (nominal categories such as *Gender*, *Number* and *Case*, and the verbal category of *Person*). For example, if the sentence "The cats walk.", which should be translated as "Mačk**e** hodaju." is instead translated as "Mačk**a** hodaju." [The cat walk.], this is to be marked as an error in *Agreement_Number*.

Given the notoriously low agreement on similar annotation tasks (cf. Subsect. 3.4), it stands to reason that even the development of such a taxonomy is already prone to human error or disagreement. This is why we made sure that the categories we added were in line with the MQM guidelines; they were already present in the expanded tagset (e.g. *Register*), and those that were not (e.g. the different agreement types) are analogous to tags that are. Still, in order to make sure that we did not taking any missteps

**Table 2** Example of a *Mistranslation* error that also causes an *Agreement* error

| | |
|---|---|
| Source: | For example, **websites** provide... |
| Correct: | Na primjer, **internetske stranice** pružaju... |
| Translation: | Na primjer, **internetska stranica** pružaju... |
| Gloss: | For example, **website** provide... |

in the construction of the taxonomy, we additionally discussed our changes with other researches and colleagues not directly involved in this particular piece of research. Consequently, the taxonomy was verified by both a traditional and computational linguist who respectively specialise in both English and Croatian linguistics.

### 3.2 Accuracy versus fluency

Unrelated to our interventions in the taxonomy, one important thing to note about the annotation process, as stated in the MQM usage guidelines, is that

> *Accuracy* addresses the extent to which the target text accurately renders the meaning of the source text, whereas *Fluency*, on the other hand, relates to the monolingual qualities of the source or target text, relative to agreed-upon specifications, but independent of relationship between source and target. [15]

In other words, fluency issues can be assessed without regard to whether the text is a translation or not. So for example, if a translated text tells the user to push a button when the source tells the user not to push it, there is an accuracy issue, while a spelling error or a problem with register remain issues regardless of whether the text is translated.

It has to be said that at first look this distinction might seem obvious and clear-cut, but in practice it is anything but. Very often examples can seem like they belong to either category, and so it is up to the annotators' judgement to decide which level is a better fit, and then being consistent in following through on the decisions made regarding dubious examples.

An example of an error category that might cause trouble for annotators is *Mistranslation*, which describes issues that arise when the content on the target side of the translation does not accurately represent the content on the source side. The issue is that it can seemingly overlap with the *Fluency* branch; according to the guidelines, only one error should be tagged, and *Accuracy* trumps *Fluency* if the required information is present in the source text.

An example of this is shown in Table 2, where the only actual error is the translation of 'website' in the singular rather than the plural, which is explicitly encoded via the -s morpheme in the source text. However, this error then causes a subject-verb agreement error, where the translated subject is singular, but the verb has been correctly translated as plural. This example should, according to the guidelines, be classified only as *Mistranslation*, even though it also shows problems with agreement. If the subject had been translated properly (as the plural), the subject-verb agreement prob-

[15] http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf.

lem would be resolved, so in this case only 'internetska stranica' should be tagged as a *Mistranslation*.

### 3.3 Annotation setup

In order to carry out the annotations we used `translate5`,[16] a web-based tool that implements annotations of MT outputs using hierarchical taxonomies, as is the case of MQM.

We had two annotators with very similar backgrounds at our disposal. Both are native speakers of Croatian, and both have prior experience with MQM as well as the same academic background; an MA in English linguistics and information science. All of these aspects of the annotators' backgrounds are relevant: their language and linguistics background is necessary given that English is the source language, and Croatian is the target language of our systems, while the information science background promises, at the very least, a basic understanding of what MT is and how it works. Thus, both annotators are well-equipped to handle the task.

Prior to annotation, they were thoroughly familiarized with the `translate5` system and the official MQM annotation guidelines, which offer detailed instructions for annotation within the MQM framework.[17]

The annotators annotated 100 randomly selected sentences from the test set introduced in Sect. 2, while presented with the English source text, a Croatian reference translation and the three unannotated system outputs at the same time. They could choose in which order to annotate, but did not know which translations belonged to which system, thus performing blind annotation. The two annotators did not operate completely independently of each other; they occasionally discussed particularly difficult or ambiguous sentences and how to approach them.

All three translations were annotated by both annotators, meaning that each system translated the same 100 sentences, each annotator annotated the resulting 300 translated sentences (100 source sentences for 3 MT systems), producing a total of 600 annotated sentences (300 translated sentences for 2 annotators). We have made the annotated dataset publicly available on GitHub.[18]

Once the sentences were annotated and the annotation data was extracted, we calculated inter-annotator agreement (reported in Sect. 3.4) and analyzed the output to determine the performance of each system for each error category (cf. Sect. 4).

### 3.4 Inter-annotator agreement

Though carefully thought out and developed, the MQM metrics (and manual MT evaluation in general) are notorious for resulting in low inter-annotator agreement (IAA) scores. This is attested by the body of work that has addressed this issue, most

[16] http://www.translate5.net/.

[17] The instructions include a handy decision tree to aid in the annotation process. It can be found at the following URL: http://www.qt21.eu/downloads/annotatorsGuidelines-2014-06-11.pdf.

[18] https://github.com/GreenParachute/mqm-eng-cro/.

**Table 3** Inter-annotator agreement (Cohen's $\kappa$ values) for the MQM evaluation task

| Error type | PBMT | Factored | NMT | Concat |
|---|---|---|---|---|
| Accuracy | 0.66 | 0.62 | 0.56 | 0.61 |
| Mistranslation | 0.51 | 0.48 | **0.58** | 0.53 |
| Omission | 0.34 | 0.39 | 0.37 | 0.37 |
| Addition | 0.50 | 0.54 | 0.33 | 0.47 |
| Untranslated | **0.86** | **0.86** | − 0.02 | **0.72** |
| Fluency | 0.50 | 0.41 | 0.29 | 0.43 |
| Unintelligible | 0.39 | 0.32 | 0.00 | 0.35 |
| Register | 0.37 | 0.20 | 0.22 | 0.27 |
| Spelling | 0.00 | 0.00 | 0.00 | 0.00 |
| Grammar | 0.50 | 0.43 | 0.33 | 0.45 |
| Word order | 0.56 | 0.33 | 0.21 | 0.40 |
| Function words | 0.43 | 0.27 | 0.36 | 0.35 |
| Extraneous | 0.56 | 0.32 | 0.49 | 0.46 |
| Incorrect | 0.37 | 0.18 | 0.34 | 0.29 |
| Missing | 0.00 | 0.49 | 0.00 | 0.33 |
| Word form | 0.48 | 0.46 | 0.36 | 0.47 |
| Part of speech | − 0.03 | 0.10 | 0.00 | 0.04 |
| Tense... | 0.44 | 0.36 | 0.15 | 0.38 |
| Agreement | 0.52 | 0.52 | 0.49 | 0.53 |
| Number | 0.53 | 0.55 | 0.52 | 0.54 |
| Gender | 0.46 | 0.59 | 0.48 | 0.53 |
| Case | 0.53 | 0.49 | 0.52 | 0.56 |
| All errors | **0.56** | **0.49** | **0.44** | **0.51** |
| Any errors | **0.80** | **0.67** | **0.51** | **0.64** |

The highest scores for any individual system and the concatenation, as well as the overall scores, are shown in bold

notably Lommel et al. (2014b), who worked specifically on MQM, and Callison-Burch et al. (2007), who investigated several tasks. This is why it is important that we check how well our annotators agree on the task at hand, and whether this is consistent with prior work done with MQM.

Once the data was annotated, agreement was observed at the sentence level, and inter-annotator agreement was calculated using Cohen's Kappa ($\kappa$) (Cohen 1960). Agreement was calculated on the annotations of each system separately, as well as on the concatenation of the annotations for the 3 systems together. This way we can (i) investigate whether there are differences in agreement across systems, and also (ii) gain insight into the overall agreement between the two annotators. In addition, Cohen's $\kappa$ was also calculated for every error type separately. Results can be found in Table 3.

The 'Any errors' IAA value presented at the bottom of the table is the most general agreement measure—it represents agreement on there being any sort of error in a given

**Table 4** Example of annotator disagreement on error span on the example of a *Word order* error

| | |
|---|---|
| Source: | Trakhtenberg was the presenter of many programs before Hali-Gali times |
| Annotator_1: | Bio **je** voditelj **Trakhtenberg** brojnih programa Hali-Gali prije puta |
| Annotator_2: | **Bio je voditelj Trakhtenberg** brojnih programa Hali-Gali prije puta |

sentence. These values will logically be higher than the IAA values of the 'All errors' measure (which looks at the total of error agreement, but of specific error categories in a given sentence), and much higher than the agreement calculated for each of the individual, specific error categories.

Examining the table reveals that our annotators agree most on evaluations of the PBMT system, less so on evaluations of the Factored SMT system, and least on evaluations of the NMT system. The drop in agreement scores for the NMT system is a bit striking. Our intuition is that, because the outputs of the NMT system are much more fluent and grammatically correct (cf. Sect. 4), errors become less clear cut, and more difficult for our annotators to detect. Or rather, any errors produced by the system are more debatable and the tags are subject to the annotators' interpretation, rather than grounded in some sort of objective truth.

Still, the comparison of IAA between the different systems is likely not that meaningful, as it involves a slightly different sample size due to the different lengths of the outputs. Besides, even disregarding this discrepancy, agreement scores are relatively low overall, with the average total $\kappa$ being 0.51. Indeed, the $\kappa$ scores are relatively consistent across all error types for each system, mostly ranging between 0.35 and 0.55. According to Cohen, such scores constitute moderate agreement. As already stated, this is to be expected, given the complexity of the problem and annotation schema. In fact, the IAA scores in this work are notably higher than those that have been reported in similar work, e.g. Lommel et al. (2014b), who achieved $\kappa$ scores ranging between 0.25 and 0.34.

That said, this comparison should be taken with a grain of salt, given that in our setup we looked at sentence-level agreement, while they calculated agreement on the token level. The calculations are approached differently here in order to attempt to account for some of the problems that come with span-level annotation. As Lommel et al. (2014b) point out, a "fundamental issue that the QTLaunchPad annotation encountered was disagreement about the precise scope of errors". In other words, though annotators can agree that a sentence contains the same issue, they might disagree on the span that the issue covers. An example is shown in Table 4 (annotations marked in bold).

This case shows that annotators can agree on the nature and categorization of issues, yet still disagree on their precise span-level location. Even though they are instructed to mark minimal spans, i.e. spans that cover only the issue in question, they frequently disagree as to what the scope of these issues is. Lommel et al (2014b, 4) hypothesize that this may be due to the fact that the two reviewers perceive the issue differently, and so see different spans as cognitively relevant. In some instances this disagreement may reflect differing ideas about optimal solutions, while in others the problem may have more to do with perceptual units in the text.

**Table 5** Total errors per system and annotator, ass annotated in MQM

| System | Annotator 1 | | | Annotator 2 | | |
|---|---|---|---|---|---|---|
| | PBMT | Factored | NMT | PBMT | Factored | NMT |
| Total errors | 317 | 276 | 178 | 264 | 199 | 132 |

In cases where annotators disagree on the span of the annotation, even Lommel et al are uncertain as to how best to assess IAA. Thus, building on their work and exploring a sentence-level approach is a direction we deemed worth pursuing, as there seems to be no optimal solution, given that both the sentence- and token-level approach come with certain drawbacks. However, to dispel any doubt regarding the reliability of the annotators' judgements on the task at hand, further analysis of the results shows that both annotators' annotations point to comparable conclusions, both when considered separately and together. This is elaborated on in Sect. 4.

## 4 Results

Directly extracting raw annotation data from the `translate5` system provides a sum of error tags annotated for each error type by each annotator and system. The total values are presented in Table 5.

Looking at the aggregate data alone, one can easily detect that both annotators have judged that the PBMT system contains the most errors, and that the NMT system contains the smallest number of errors. This trend is consistent across most fine-grained error categories too, as we will see later on in this section.

However, even though simply counting the errors can provide insight into which system performs better, it does not allow us to draw statistically meaningful conclusions from the results. Error counts cannot be directly compared because different MT systems may output sentences of different lengths, which is indeed the case in the data explored here: in the 100 annotated sentences, the phrase-based system produced an average of 18.99 tokens per sentence, the factored system averaged on 18.89, while the neural system produced 18.36 tokens per sentence. Hence, we need to normalize the scores.

There seems to be no related work on how to approach normalization of MQM results. In all the work published so far, authors simply count the number of MQM tags and stop there. Our normalization approach is rather straightforward: instead of counting just error tags produced by each annotator, we count the tokens that these errors are assigned to.

Once these counts are divided by the total number of tokens in the system's output, they provide a ratio of tokens with errors, as shown in Eq. (1):

$$error\ ratio = \frac{output\ tokens\ \ with\ \ errors}{total\ \ output\ \ tokens} \tag{1}$$

Given that, according to this equation, the numerator counts words in the output that contain an error, the ratio is biased in favour of systems that produce shorter output. However, this is not a problem in our setup, as our taxonomy includes an *Omission* error category. So if a word, segment, or phrase (or whatever the annotators deem as the basic unit) was not translated from the source sentence, the target sentence is tagged with an *Omission* error. While counting error tokens for our error ratio, we assume that 1 token was omitted for every omission error in the output, and so every omission error was given one phantom token to latch on to. This allows us to perform the calculations and prevents translations that lack some of the information of the source language sentence from having a low error rate.

The results of our error ratio calculations again show that the PBMT system has the largest error/token ratio (0.2633), while the factored system has a smaller ratio (0.212), and the NMT system has the smallest one (0.1277). This is further backed up by a pairwise chi-squared ($\chi^2$) statistical significance test (Plackett 1983); we calculate statistical significance from $2 \times 2$ contingency tables for every system pair (PBMT $\times$ Factored, PBMT $\times$ NMT and Factored $\times$ NMT). In one such contingency table, the rows contain token counts for each of the systems, while the columns contain counts of tokens with and without errors. The null-hypothesis in this setting states that there is no link between the MT system and the number of tokens with or without errors that it produces (i.e. that no matter which system is employed, the number of errors is relatively similar). With the $p$ value lower than 0.0001 in all three comparisons, we can safely dismiss the null hypothesis, showing that the difference in the total counts of tokens with errors is statistically significant for all three system pairs.

These error/token ratios provide an overall score for each system. At this point we would like to delve deeper and discover the performance of each system for each error type. To this end, we repeated these same measurements, but instead of performing them on all error types concatenated, they were performed separately for each specific error category. The combined results of the aforementioned calculations and transformations are presented in Table 6.

We can derive several findings from this table. As mentioned earlier, looking at the grand total of tokens with and without errors, the difference between the systems is statistically significant by a wide margin. When looking at PBMT and factored PBMT, the factored system has significantly fewer errors than the pure PBMT system. The overall error rate is in this case reduced by 20% (809 vs. 1010 errors, cf. last row in Table 6). In addition, a separate analysis of specific error types that contribute to this score reveals that only some of the error categories are significantly different between the two systems. In the table, those categories are filled in with a bold background. One can see that, when it comes to agreement errors, the only agreement error type that results in a significantly smaller number of errors with the factored PBMT system compared to the pure PBMT system is agreement in case.

However, taking a look at NMT shows that, not only does it result in a 42% overall error reduction compared to the factored system (469 vs. 809 errors), and 54% with respect to pure PBMT (469 vs. 1010 errors), but it also produces even less agreement errors—overall, as well as at the level of number, gender and case—while not using any kind of explicit linguistic information. This might in part be due to the use of sub-word segmentation, as inflections in Croatian are relatively regular. In addition to

**Table 6** Processed annotation data from both annotators concatenated: each system's total number of tokens with and without errors

| Error type | PBMT | | Factored | | NMT | |
|---|---|---|---|---|---|---|
| | OK | Error | OK | Error | OK | Error |
| Accuracy | 3467 | 369 | **3525** | **291*** | 3402 | 266 |
| Mistranslation | 3547 | 289 | **3586** | **230*** | 3471 | 197 |
| Omission | 3801 | 35 | 3793 | 23 | *3619* | *49** |
| Addition | 3814 | 22 | 3797 | 19 | 3655 | 13 |
| Untranslated | 3813 | 23 | 3797 | 19 | **3662** | **6*** |
| Fluency | 3195 | 641 | **3298** | **518*** | **3465** | **188**** |
| Unintelligible | 3790 | 46 | 3769 | 47 | **3668** | **0**** |
| Register | 3810 | 26 | 3794 | 22 | 3646 | 22 |
| Spelling | 3833 | 3 | 3812 | 4 | 3659 | 9 |
| Grammar | 3270 | 566 | **3371** | **445**** | **3497** | **156**** |
| Word order | 3752 | 84 | 3752 | 64 | **3646** | **22**** |
| Function words | 3801 | 35 | 3780 | 36 | **3650** | **18*** |
| Extraneous | 3829 | 7 | 3810 | 6 | 3664 | 4 |
| Incorrect | 3810 | 26 | 3790 | 26 | **3655** | **13*** |
| Missing | 3834 | 2 | 3812 | 4 | 3667 | 1 |
| Word form | 3389 | 447 | **3471** | **345*** | **3538** | **102**** |
| Part of speech | 3822 | 14 | 3800 | 16 | **3663** | **5*** |
| Tense... | 3775 | 61 | 3765 | 51 | **3648** | **20*** |
| Agreement | 3466 | 370 | **3540** | **276*** | **3566** | **102**** |
| Number | 3778 | 58 | 3772 | 44 | **3646** | **22*** |
| Gender | 3788 | 48 | 3756 | 60 | **3644** | **24*** |
| Case | 3614 | 222 | **3694** | **122*** | **3622** | **46**** |
| Person | 3836 | 0 | 3816 | 0 | 3664 | 4 |
| Total errors | 2826 | 1010 | **3007** | **809**** | **3199** | **469**** |

Statistical significance for a system, when compared to the system on its left, is marked with * where $p$ value is $< 0.05$ and ** where $p$ value is $< 0.0001$. Cells with a bold background indicate that the system has fewer errors than the one on its left, while those in italic indicate that it has more. In both cases, the bold/italic background is only displayed when the difference between the error ratios is statistically significant

improving in the Agreement category, NMT also produces significantly fewer errors in many more categories than the factored model does. Interestingly, it produces more Omission errors than either of the other two systems. It seems that NMT tends to sacrifice completeness of translation in order to increase overall fluency. This result is compatible with the average token per sentence ratio mentioned above: the NMT system has the lowest one (18.36; while PBMT has 18.99 and factored PBMT has 18.89).

## 5 Additional agreement annotation

In this section we look at the agreement error category in more detail. Our motivation for picking this error type is twofold: (i) significant gains have been obtained in this

**Table 7** Example sentences showcasing the two different spans an agreement error can take. The first sentence features disagreement in case, whereas the second one features disagreement in gender

| | |
|---|---|
| Phrase disagreement: | Veliki broj ljudi radi **u palijativne skrbi** |
| Sentence disagreement: | Stalna antikorupcijska **jedinica**, koja se bori protiv svakog oblika korupcije, **nastao** je 2011. godine |

error category (cf. Table 6) by NMT compared to the two PBMT systems, and (ii) this error category constitutes the main branch that we added to the core MQM tagset (to be able to evaluate the performance of MT on relevant linguistic phenomena present in Slavic languages, cf. Fig. 2).

Agreement is also worth exploring further because two syntactically different types of agreement are subsumed under the MQM Agreement tags, namely:

- Local, short-distance agreement (or phrase agreement), which concerns agreement of elements within a phrase.[19]
- Long-distance agreement (or sentence agreement), which concerns agreement of elements at the sentence level, outside phrase boundaries. These elements have wider spans and can be much further apart.

For example, local agreement would be agreement between an adjective and a noun, or between a preposition and the following noun, while sentence agreement would be agreement between a noun and a verb. Table 7 contains an example of agreement errors at these two levels. The phrase bolded in the first sentence contains disagreement in case: the preposition "u" should introduce a phrase in the dative case ("palijativn**oj** skrbi"), but the translation is in the accusative case ("palijativn**e** skrbi"), which is morphologically marked. The phrase bolded in the second sentence contains disagreement in gender: the noun "jedinica" ("unit", feminine) is the subject of the sentence and as such should agree with the verb "nastati" ("was created") that follows it in gender, number, case and person; however, in the translation, the verb is marked for masculine gender ("nasta**o**") instead of the required feminine ("nasta**la**").

This distinction is important not only linguistically, but can also be informative from a technical perspective. Thus, we conducted an additional layer of annotation outside the framework of MQM: each agreement error was categorized as corresponding to either phrase or sentence level. Additionally, the type of elements participating in the error was marked as well, in order to obtain more fine-grained insights.

For phrase agreement, the phrases in question can be prepositional phrases (PP) that contain a noun phrase (NP), noun phrases that contain an adjective (ADJ) and a noun (N), noun phrases comprised of two nouns (N + N) and noun phrases containing numerals (NUM + NP). In sentence agreement, elements that often need to agree are subjects and verbs (S + V, usually noun and verb), verbs and objects (V + O, usually verb and noun), two or more noun phrases coordinated with a conjunction (NP + C + NP, usually "i" ["and"]), and a noun phrase followed by a subordinating

[19] Unlike in SMT jargon, here a phrase refers to a grammatical unit, not just a string of contiguous words.

**Table 8** Breakdown and categorization of agreement errors found in the annotated data

| Elements | Phrase agreement | | | | Sentence agreement | | |
|---|---|---|---|---|---|---|---|
| | PBMT | Factored | NMT | Elements | PBMT | Factored | NMT |
| PP + NP | 24 | 14 | 3 | S + V | 20 | 19 | 7 |
| ADJ + N | 15 | 5 | 2 | V + O | 5 | 3 | 3 |
| N + N | 4 | 7 | 1 | NP + C + NP | 6 | 7 | 1 |
| NUM + NP | 1 | 1 | 0 | NP + CSUB | 1 | 2 | 0 |
| Total | 44 | 27 | 6 | Total | 32 | 31 | 11 |

conjunction (NP + CSUB, usually "koji/koja/koje" ["which" or "that"]). The results of applying this categorisation to our dataset are presented in Table 8.

As the table shows, the factored PBMT model leads to quite a large improvement upon pure PBMT when it comes to phrase agreement, but the improvement is almost negligible when it comes to sentence agreement (phrase agreement sees a $\sim$ 38% relative reduction in errors, while the number of sentence agreement errors is reduced by $\sim$ 4% relative). Meanwhile, the NMT model produces substantially less agreement errors of both agreement types ($\sim$ 86% relative reduction in phrase agreement errors and $\sim$ 66% relative reduction in sentence agreement errors, when compared to pure PBMT).

Knowing that both the factored model and NMT model produce less agreement errors overall when compared to PBMT (cf. Table 6), it is no surprise that they produce overall less of either level (phrase and sentence) of agreement errors. However, just as in the MQM analysis conducted in the previous section, simply counting errors is not enough to know whether the difference in the number of errors between two MT paradigms is statistically significant. Thus, to determine whether these differences are statistically significant overall, we once again normalized the errors to the token level and employed a chi-squared ($\chi^2$) test. We calculate statistical significance from $2 \times 2$ contingency tables for every system pair (PBMT $\times$ Factored, PBMT $\times$ NMT and Factored $\times$ NMT), for each type of error (overall phrase agreement and overall sentence agreement), as well as for the elements that make up these errors. In these contingency tables, rows contain token counts for each system, while columns contain counts of tokens with and without agreement errors. The null-hypothesis states that there is no link between the MT system and the frequency of a given agreement error that it produces.

As shown in Table 9, the total counts show that when looking at phrase agreement, there is steady improvement between the systems: the factored system has significantly less tokens with a phrase-agreement error than the PBMT system ($p = 0.004$), while the NMT system has significantly less than the factored system does ($p < 0.0001$). On the other hand, looking at sentence agreement and comparing pure PBMT to the factored PBMT model yields a $p$ value of 0.8799, revealing no statistical significance, while comparing the factored model to the neural model yields a $p$ value of 0.00002, indicating a statistically significant difference in the number of tokens with errors. In other words, when compared to PBMT, both the factored model and the NMT model

**Table 9** Normalized agreement annotation data: each system's total number of tokens with and without agreement errors, also including data with regards to which elements contained errors

| Error type | PBMT | | Factored | | NMT | |
|---|---|---|---|---|---|---|
| | OK | Error | OK | Error | OK | Error |
| Total | | | | | | |
| Phrase | 1811 | 88 | **1835** | **54*** | **1824** | **12**** |
| Sentence | 1835 | 64 | 1827 | 62 | **1814** | **22**** |
| Phrase agreement | | | | | | |
| PP + NP | 1851 | 48 | **1861** | **28*** | **1830** | **6*** |
| ADJ + N | 1869 | 30 | **1879** | **10*** | 1832 | 4 |
| N + N | 1891 | 8 | 1875 | 14 | **1834** | **2*** |
| NUM + NP | 1897 | 2 | 1887 | 2 | 1836 | 0 |
| Sentence agreement | | | | | | |
| S + V | 1859 | 40 | 1851 | 38 | **1822** | **14*** |
| V + O | 1889 | 10 | 1883 | 6 | 1830 | 6 |
| NP + C + NP | 1887 | 12 | 1875 | 14 | **1834** | **2*** |
| NP + CSUB | 1897 | 2 | 1885 | 4 | 1836 | 0 |

Statistical significance for a system, when compared to the the one on its left, is marked in bold. If $p$ value is $< 0.05$, it is marked with *, and ** where $p$ value is $< 0.0001$

significantly reduce the number of phrase-agreement errors, whereas the factored model does not significantly reduce the number of sentence-agreement errors, but the neural system does.

These results are in line with previous research that showed how, for the English-to-Croatian language pair, factored PBMT struggles with sentence agreement due to the limitations of n-gram language models: Sánchez-Cartagena et al. (2016) showed that using high-order language models (with order higher than 3) for morphosyntactic tags leads to a degradation in translation quality because of the free word order of Croatian. On the contrary, the power of recurrent neural network units to model long-distance phenomena allows the NMT system to improve on both phrase and sentence agreement.

## 6 Conclusion

This paper describes a fine-grained human evaluation of three approaches to MT (pure PBMT, factored PBMT and NMT). Our analysis has provided answers to several questions, one of which was the main drive behind the development of a factored system for English-to-Croatian: is there a way to better handle agreement when translating to a morphologically rich language? We can now confidently claim that factored models result in significantly less agreement errors overall compared to pure PBMT, when translating from English to Croatian.

We can also confidently conclude that NMT handles all types of agreement better than both pure PBMT and factored PBMT, which corroborates the findings of other

researchers' NMT evaluations conducted for other language pairs. Our NMT system produces sentences with far fewer errors, and output that is more fluent and more grammatical, which should be of help when it comes to the task of post-editing.

Furthermore, the error taxonomy that was developed for this research, while only used for the English-to-Croatian language direction in the current work, should be applicable for the analysis of errors for any translation direction towards a Slavic language, as it takes into account specific grammatical properties shared by the members of this language family.

Among other possible lines of future work, including the application of our methodology to another language pair that involves a Slavic target language (e.g. English–Czech), performing more controlled IAA analysis or IAA adjudication, as well as comparing to an NMT model without sub-word segmentation, another direction to go in is further adapting the tagset. In its current version, it has been demonstrated to be informative when comparing PBMT to factored PBMT. However, NMT has shown itself to produce language that is so fluent that the fine-grained hierarchy in the *Fluency* branch is of little use. Meanwhile, the most common error type in the NMT output is *Mistranslation*, which, according to the MQM guidelines, covers both lexical selection and (less intuitively) translation of grammatical properties (e.g. if 'cats[pl.]' is translated into Croatian as 'mačka[sg.]', this is to be tagged as *Mistranslation*, in spite of correct lexical choice). This makes it quite a vague category, so if one would wish to perform an even more nuanced analysis of errors for NMT, adding additional layers to the *Accuracy* branch would seem a promising direction to follow.

**Acknowledgements** We would like to extend our thanks to Maja Popović, who provided invaluable advice, and Denis Kranjčić, who performed the annotation together with Filip Klubička, first author of the paper. This research was partly funded by the ADAPT Centre, which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. This research has also received funding from the European Union Seventh Framework Programme FP7/2007-2013 under Grant agreement PIAP-GA-2012-324414 (Abu-MaTran) and the Swiss National Science Foundation Grant 74Z0_160501 (ReLDI).

## References

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of international conference on learning representations, San Diego, CA, USA, 2015

Bentivogli L, Bisazza A, Cettolo M, Federico M (2016) Neural versus phrase-based machine translation quality: a case study. In: Proceedings of the 2016 conference on empirical methods in natural language processing, Austin, Texas, USA, pp 257–267

Burchardt A, Macketanz V, Dehdari J, Heigold G, Peter JT, Williams P (2017) A linguistic evaluation of rule-based, phrase-based, and neural MT engines. Prague Bull Math Linguist 108:159–170

Callison-Burch C, Fordyce C, Koehn P, Monz C, Schroeder J (2007) (Meta-) evaluation of machine translation. In: Proceedings of the second workshop on statistical machine translation, Prague, Czech Republic, pp 136–158

Castilho S, Moorkens J, Gaspari F, Calixto I, Tinsley J, Way A (2017) Is neural machine translation the new state of the art? Prague Bull Math Linguist 108:109–120

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46

Devlin J, Zbib R, Huang Z, Lamar T, Schwartz R, Makhoul J (2014) Fast and robust neural network joint models for statistical machine translation. In: Proceedings of association for computational linguistics conference, Baltimore, Maryland, USA, pp 1370–1380

Durrani N, Schmid H, Fraser A (2011) A joint sequence translation model with integrated reordering. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, vol 1. Portland, Oregon, USA, pp 1045–1054

Galley M, Manning CD (2008) A simple and effective hierarchical phrase reordering model. In: Proceedings of the conference on empirical methods in natural language processing, Waikiki, Honolulu, Hawaii, pp 848–856

Klubička F, Toral A, Sánchez-Cartagena VM (2017) Fine-grained human evaluation of neural versus phrase-based machine translation. Prague Bull Math Linguist 108:121–132

Koehn P (2004) Statistical significance tests for machine translation evaluation. In: Proceedings of the conference on empirical methods in natural language processing, Barcelona, Spain, pp 388–395

Koehn P, Hoang H (2007) Factored translation models. In: Proceedings of conference on empirical methods on natural language processing and computational natural language learning, Jeju Island, Korea, pp 868–876

Koehn P, Och FJ, Marcu D (2003) Statistical phrase-based translation. In: Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology, Edmonton, Canada, pp 48–54

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, et al (2007) Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, Prague, Czech Republic, pp 177–180

Ljubešić N, Klubička F (2014) bs,hr,sr WaC-web corpora of Bosnian, Croatian and Serbian. In: Proceedings of the 9th web as corpus workshop (WaC-9), Gothenburg, Sweden, pp 29–35

Lommel AR, Burchardt A, Uszkoreit H (2014a) Multidimensional quality metrics (MQM): a framework for declaring and describing translation quality metrics. Tradumàtica: tecnologies de la traducció pp 455–463

Lommel AR, Popovic M, Burchardt A (2014b) Assessing inter-annotator agreement for translation error annotation. In: MTE: workshop on automatic and manual metrics for operational translation evaluation, Reykjavik, Iceland

Moore RC, Lewis W (2010) Intelligent selection of language model training data. In: Proceedings of the association for computational linguistics 2010 conference short papers, Stroudsburg, PA, USA, pp 220–224

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, pp 311–318

Pascanu R, Mikolov T, Bengio Y (2013) On the difficulty of training recurrent neural networks. In: Dasgupta S, Mcallester D (eds) Proceedings of the 30th international conference on machine learning (ICML-13), JMLR workshop and conference proceedings, vol 28. Atlanta, USA, pp 1310–1318

Plackett RL (1983) Karl pearson and the chi-squared test. International statistical review/revue internationale de statistique pp 59–72

Popović M (2017) Comparing language related issues for NMT and PBMT between German and English. Prague Bull Math Linguist 108:209–220

Sánchez-Cartagena VM, Ljubešić N, Klubička F (2016) Dealing with data sparseness in SMT with factored models and morphological expansion: a case study on Croatian. Baltic J Mod Comput 4(2):354–360

Sennrich R, Haddow B, Birch A (2016) Neural machine translation of rare words with subword units. In: Proceedings of the 54th annual meeting of the association for computational linguistics, Berlin, Germany, pp 1715–1725

Sennrich R, Firat O, Cho K, Birch A, Haddow B, Hitschler J, Junczys-Dowmunt M, Lubli S, Barone AVM, Mokry J, Nadejde M (2017) Nematus: a toolkit for neural machine translation. In: Proceedings of the European association for computational linguistics 2017 software demonstrations, Valencia, Spain, pp 65–68

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of association for machine translation in the Americas, Cambridge, Massachusetts, USA, pp 223–231

Tiedemann J (2009) News from OPUS—a collection of multilingual parallel corpora with tools and interfaces. In: Recent advances in natural language processing. Borovets, Bulgaria, pp 237–248

Tiedemann J (2012) Parallel data, tools and interfaces in OPUS. In: Proceedings of the eighth international conference on language resources and evaluation, Istanbul, Turkey, pp 2214–2218

Toral A, Sánchez-Cartagena VM (2017) A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In: Proceedings of the 15th Conference of the European chapter of the association for computational linguistics, Valencia, Spain, pp 1063–1073
Zeiler MD (2012) ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701