Original scientific paper

# A comparative error analysis of English and German MT from and into Croatian

Marija Brkic Bakaric
Department of Informatics, University of Rijeka
Radmile Matejčić 2, 51000 Rijeka
mbrkic@inf.uniri.hr

Nikola Babic
Faculty of Humanities and Social Sciences
Sveučilišna avenija 4, 51000 Rijeka
nbabic@ffri.hr

Luka Dajak
Faculty of Humanities and Social Sciences
Sveučilišna avenija 4, 51000 Rijeka
ldajak@ffri.hr

Maja Manojlovic
Faculty of Humanities and Social Sciences
Sveučilišna avenija 4, 51000 Rijeka
mmanojlovic@ffri.hr

**Summary**

*This paper first gives an insight into the problem of evaluating translations and lists existing approaches. After introducing error taxonomies, a comparative error analysis of Google Translate results is conducted based on the selected taxonomy. The analysis covers translations from English and German into Croatian, and from Croatian into English and German. The intra and inter-annotator agreements, which are rarely accounted for, are reported where applicable. The aim of the paper is better understanding of different notions of error analysis and detecting weak points of such analysis. A preliminary list of guidelines for error analysis is suggested.*

**Key words:** error analysis, machine translation, intra-annotator, inter-annotator agreement, guidelines

## Introduction

Quality assessment is one of the most debated topics in translation. There is no single standard for translation quality assessment because quality is context dependent (Secară, 2005). Furthermore, one sentence can be translated in multiple

ways. All this makes machine translation (MT) evaluation inherently subjective. Error analysis is a means to assess translations in qualitative terms. It refers to the identification and classification of individual errors in a translated text. However, like other subjective approaches, it is susceptible to low inter-annotator agreement (Stymne & Ahrenberg, 2012).

Related work, presented in the next section, is divided into research on error analysis in general and research with a focus on Croatian. The research presented in this paper aims at gaining a better understanding of different notions of error analysis and detecting weak points of such analysis. Besides compiling a list of guidelines for error analysis, the research aims at answering what we can conclude about Google Translate (GT) engine for the selected language pairs and what types of errors the system makes most often. Section three gives details on the error analysis conducted on GT engine from English and German into Croatian and vice versa. It is followed by results and discussion. Conclusion summarizes main findings and gives directions for future work.

## Related work

### Error analysis

Translation error can be defined as a semantic component not shared by source and target texts (Koponen, 2010). This component can be larger than individual words (e.g. compound nouns, names, idioms, etc.). Error analysis gives a qualitative view on the MT system and should be an integral part of MT development (Stymne S., 2011). It can point to strengths and problem areas for a certain machine translation system, which is not possible using automatic evaluation metrics (Stymne & Ahrenberg, 2012). Automatic metrics, as well as some forms of human evaluation, such as fluency and adequacy scoring or system ranking, provide quantitative system evaluation (Stymne S., 2011). However, research community would like to get answers to what kind of errors the system makes most often, whether a particular modification improves some aspect of translation although the overall score is intact, whether one system is superior to another in all aspects of translation or just in some, etc. (Popovic & Burchardt, 2011). Context and extra-linguistic knowledge often subconsciously guide us into correcting certain errors and this alone proves that some errors are less severe than others. The author in (Secară, 2005) emphasizes that in a post-editing scenario precedence should be given to error categories over numerical scores, since the effort put into correcting each error type is as important as the final score. MQM resulted from a thorough investigation of major human and machine translation assessment metrics. It is a general mechanism for declaring specific metrics for general quality assessment and error annotation tasks, since various error taxonomies have been suggested for the task of error analysis (Flanagan, 1994; Font-Llitjós, Carbonell, & Lavie, 2005; Elliott, Hartley, & Atwell, 2004; Vilar, Xu, d'Haro, & Ney, 2006; Farrús, Costa-Jussa, Marino, Poch, Hernández, Henríquez, Fonollosa, 2011; Costa, Ling, Luís, Correia, &

Coheur, 2015). Although error analysis is subjective, Stymne and Ahrenberg (2012) argue it is possible to get a reasonable agreement either when using a simple error taxonomy or when using a more detailed taxonomy and a set of guidelines. The study in (Elliott, Hartley, & Atwell, 2004) emphasizes the importance of assigning weights to different error categories to make them correlate with intuitive human judgements of translation quality. Moreover, the focus of MT evaluation research is gradually shifting towards profiling systems with respect to various error taxonomies (Federico, Negri, Bentivogli, Turchi, & Kessler, 2014). One reason is that parallel data limits system knowledge to the observed positive examples. Another is that majority of automatic metrics provide only a holistic view of system performance. The authors in (Federico, Negri, Bentivogli, Turchi, & Kessler, 2014), therefore, propose a robust statistical framework to analyse the impact of different error types on human perception of quality and on automatic metrics.

Since the task of error analysis is labour-intensive and time-consuming, and requires either professionals or native speakers of a language in question, a lot of effort has been put into automatic error classification (Fishel, Bojar, Zeman, & Berka, 2011; Popovic & Burchardt, 2011). Benefits of coupling automatic and manual error classifications are shown in (Popovic & Arcan, 2016). Furthermore, the authors show that conducting manual error annotation on pre-annotated texts, where reference translations are post-edited translation outputs, can give much better and reliable insights into particular flaws of an automatic error classification tool.

Since different language combinations exhibit different error distributions in the translation output which often relates to the linguistic characteristics of the languages involved and divergences between them (Popovic & Arcan, 2016), the rest of the paper is tied to the research involving Croatian as either source or target language.

**Croatian language error analysis**

There has been abundant work on error analysis involving Croatian. Tree texts of different types are translated from Croatian into English by GT and errors are analysed on *lexical*, *syntactic*, *morphological*, *semantic* and *punctuation* level in a descriptive manner and corroborated by examples and by *fluency* and *adequacy* judgements (Brkic, Vicic, & Seljan, 2009). Short texts from four different domains and genres are translated from Croatian into English by four translation services, including GT, and evaluated by 48 evaluators on a 1-5 scale according to *fluency* and *adequacy* (Brkic, Seljan, & Matetic, 2011). The evaluation for the opposite direction in (Seljan, Brkic, & Kucis, 2011) included only GT and 50 human evaluations in total. The following four categories are covered: *morphological errors*, *untranslated words*, *lexical errors* (which also comprise semantic errors), and *syntactic errors* (which also comprise punctuation errors). The criterion of *adequacy* is mostly affected by *untranslated words*, while the

criterion of *fluency* is more affected by *lexical* and *syntactic errors*. The research in (Brkic, Seljan, & Matetic, 2011) extends the methodology by including three automatic metrics in the evaluation. GT from Croatian into English is compared to an in-house system and human translations in (Brkic, Basic Mikulic, & Matetic, 2012) by six case-sensitive metrics in the legislative and mixed domains (religion, psychology, computer games). BLEU scores on multiple reference translations and human *fluency* and *adequacy* judgements of English-Croatian GT in the domain of legislation are enriched by error analysis in (Seljan, Brkic, & Vicic, 2012), with a special focus on sentence length. The MT error taxonomy used resembles the one in (Vilar, Xu, d'Haro, & Ney, 2006) when first-level categories are taken into account, with a major difference that *extra words* and *incorrect form* are separated out as categories of the highest level. The criterion of *adequacy* is mostly affected by *semantic* and *lexical errors*, while the criterion of *fluency* is mostly affected by *morphological errors*, but also *missing words*, i.e. *lexical errors*. Similarly, English-Croatian GT is evaluated in legislative and general domains by including three additional automatic metrics and investing the impact of lowercasing, tokenization and punctuation in (Brkic, Seljan, & Vicic, 2013). A new language-pair, i.e. Russian-Croatian, is introduced into the analysis in (Seljan, Tucaković, & Dunđer, 2015) with an additional online translation service, i.e. Yandex.Translate. A comparison with four automatic metrics can be found in (Seljan & Dunđer, 2015b). GT is also evaluated by four automatic metrics for Croatian-English and English-Croatian translations in sociological-philosophical-spiritual domain in (Seljan & Dunđer, 2015a). Better results are obtained for the Croatian-English translation direction.

## Experimental setup

### Tool and error taxonomy

For the error analysis conducted within this research, the tool BLAST [4] is chosen. The Vilar's taxonomy [7] is used as a starting point with a goal to check its suitability for language directions covered by the study. GT is used as the translation engine. The following abbreviations are introduced and used hereinafter: errs (errors), avg sen len (average sentence length), wo (*word order*), unk (*unknown*), punct (*punctuation*), form (*incorrect form*) E1 (1st annotator), E2 (2nd annotator), de (German), en (English), hr (Croatian).

### Test set

There are 24 sentences in our evaluation. The text is constructed for the purpose of this research. It is originally written in Croatian and then manually translated into German and English. The text is in the form of a magazine article which is a reflection on major events in the last year. Croatian original has 513 words, while translations in English and German have 601 and 576 words, respectively. German original has 579 words while its translation in Croatian has 504 words.

English original has 566 words while its translation in Croatian has 478 words. These three texts are then translated by GT into English and German for the Croatian source, and into Croatian for the other two languages. The average translation sentence length is 25 for English, 24 for German, while it ranges from 20 to 21 for Croatian. Croatian as a morphologically rich and pronoun-dropping language has the shortest average sentence length, while English has the longest, due to its poor morphology.

**Annotators**

The error analysis is performed by four annotators in total who are either native speakers or have a formal university-level education of a language in question finished or nearly finished. Croatian translations are assessed by one final-year student at the graduate study of Croatian who is also native in Croatian and by one native speaker of Croatian. They are given access only to the reference sentences, and not to the source sentences, as we do not want to presume the knowledge of English and/or German. Similar methodology is applied in (Elliott, Hartley, & Atwell, 2004). German translation is assessed by one final-year student at the graduate study of German. English translations are assessed by one final-year student at the graduate study of English and one professional translator who graduated from the same study. Additionally, quality judgements are collected by asking the annotators to rate each translation on 1-5 Likert scale, where 1 means incomprehensible translation, and 5 means perfect translation.

**Intra and interannotater agreement**

The inter-annotator agreement per each language direction and error category and/or subcategory on a sentence basis is calculated in the first phase of the research. The calculation is performed by the equation in (1), where *all* stands for the total number of annotations by each annotator, and *agree* stands for the number of annotations on which agreement is reached. The practice is taken from (Stymne & Ahrenberg, 2012). Since annotators might agree on the label but not on the position merely because there are no guidelines on how to conduct annotation, only the agreement on the categories is reported, with the presumption that detecting errors is what matters after all, and not their precise position in sentences. Two out of four annotators asked questions prior to annotation. The questions concerned the appropriate positions for annotations of *missing word* and *word order* categories.

$$Agreement = \frac{2A^{agree}}{A1^{all}+A2^{all}}. \quad (1)$$

## Results

Table 1 shows that Croatian-English is the best scoring translation direction according to human annotators. The worst scoring direction is German-Croatian. Figure 1 presents the distribution of quality scores per each annotator and language direction. Reluctance in assigning scores 1 and 5 has been observed. Interestingly, out of three translation directions which contain sentences scored 5, two of them also contain sentences scored 1. Both annotators evaluating the German-Croatian translation direction agree that there are some extremely bad sentences, while both annotators evaluating the Croatian-English direction agree that some sentence translations are perfect.

Pearson correlation coefficients are calculated between error frequencies and human sentence scores, and between error frequencies of selected categories and human sentence scores. The results are presented in Table 2. The coefficients between the total number of errors and human scores, as well as between the number of *incorrect words* and human scores are significant at $p<0.05$.

Numbers of errors per each top-level category are presented in Figure 2. The most represented category per all language directions is *incorrect word* category, followed by *missing word* and *word order*. According to the number of errors at the intermediate level of detail, which is not included due to space considerations, the most frequent subcategory is *extra word* for translations into English and German, while the biggest issue for translations from English into Croatian is detected with the *incorrect form* category.

Table 1. Human scores and total number of errors per direction and annotator

| **Translation direction** | | **HR → DE** | **HR → EN** | **DE → HR** | **EN → HR** |
|---|---|---|---|---|---|
| Avg score | E1 | 3.33 | 3.04 | 2.54 | 3.08 |
| | E2_1/E2_2 | -/- | 3.71/3.63 | 2.58/- | 3.17/3.13 |
| | Avg | 3.33 | 3.46 | 2.54 | 3.13 |
| # of errs | E1 | 148 | 120 | 319 | 161 |
| | E2_1/E2_2 | -/- | 79/96 | 216/- | 133/131 |
| | Avg | 148 | 98.33 | 267.5 | 141.66 |

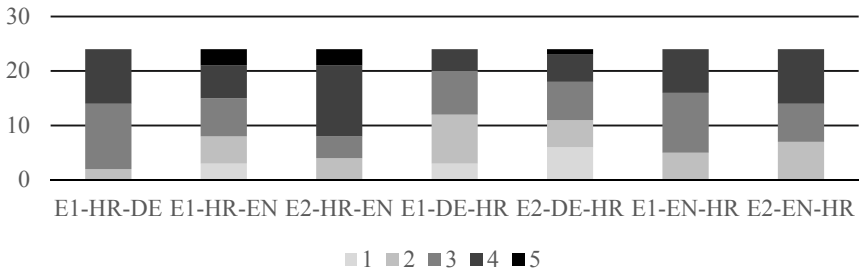


Figure 1. Distribution of quality scores per each annotator and language direction

Table 2. Pearson correlation coefficients between selected categories and human scores

| Pearson | errs | missing | wo | incorrect | unk | punct | form | sense |
|---|---|---|---|---|---|---|---|---|
| Avg score | **-0.97** | -0.78 | -0.1 | **-0.97** | 0.02 | 0.39 | -0.90 | -0.65 |

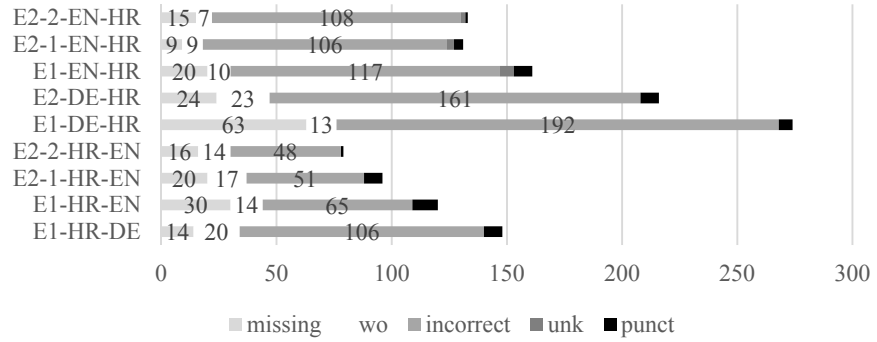


Figure 2. Number of errors per each annotator and translation direction at lowest level of detail

The intra and inter-annotator agreements for the top-level categories are given in Figure 3. Inter-annotator agreement could not be calculated for the Croatian-German language direction since there was no second annotator at disposal. Intra-annotator agreement is calculated only for translation directions involving English and Croatian due to time considerations. Intra annotator agreement is consistently over 60%, except for the *word order* and *punctuation* categories. This is due to low representativeness of these categories. As far as *incorrect words* are concerned, the agreement in translating into morphologically rich Croatian is somewhere between 65 and 75%.

## Discussion

The Pearson correlation coefficients indicate that there is a statistically significant relationship between the total number of errors and human scores, as well as between the number of *incorrect words* and human scores.

In the study presented no guidelines are given to annotators. This is done purposefully to better understand different notions of error analysis. Since annotators might agree on the label but not on the position merely due to the lack of guidelines, positions are excluded from the study. If included, the agreement on different categories would be differently affected. A short reflection on the task follows. The annotators had trouble deciding on the number of errors in a phrase, i.e. should a phrase be treated as a unique unit or not (e.g. "Am schlimsten Jahre" can be annotated either as three errors of i*ncorrect form* subcategory or as one error of the same type). Furthermore, they lack determination on deciding whether a *missing word* is *content* or *filter*. One of the annotators showed the tendency to follow references too strictly and mark differences automatically, machine-like. Although at first glance it might seem that human annotators are rather forgiving as far as style is concerned, this can be attributed to the genre of the text. Journalistic texts are broadly represented on the Internet so they make an important part of GT training data. Therefore, the *style* category could be excluded from our further studies, except in highly specialized domains with a pronounced style, e.g. the domain of law. *Punctuation* category is pretty straightforward, i.e. annotators who are not language experts may fail to detect such error, but they will not mistakenly take it for another error type. It could be abandoned from further studies as well, in order to reduce the dimensionality of calculations.

A first draft of the guidelines which we either adopt from related work or compile based on the results obtained in this study is given as follows: (1) only after reading and comprehending translation, check your understanding by consulting source or reference sentence; (2) register all possible errors on a word; (3) mark as few errors as possible to make the sentence grammatically correct and semantically equivalent to the source; (4) if the meaning is affected, wrong preposition should be annotated as a *disambiguation* error; (5) use higher-level categories when it is not possible to use deeper-level categories; (6) mark *filter* word if unsure whether the *missing word* is *content* or *filter*; (7) if unsure whether an error is a *disambiguation* or a *lexical choice* error, consult the corresponding source word and confirm whether it can be used in both senses in order to opt for *disambiguation* error. Agreement per sentence reveals that sentences with many errors are too hard to annotate unanimously and should be excluded from assessment. The authors in (Lommel, Burchardt, Popovic, Harris, Avramidis, & Uszkoreit, 2014) use more than three errors as a threshold.

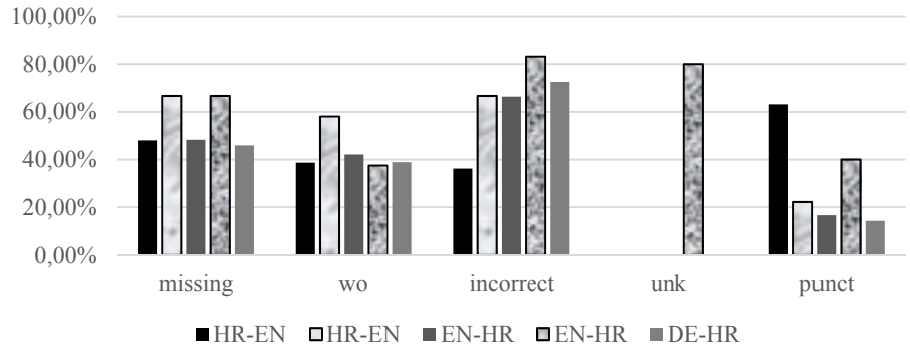


Figure 3. Intra and inter-annotator agreement per translation direction at lowest level of detail

## Conclusion

The annotations obtained in this study enable us to detect the system and language-specific distributions of errors. The paper confirms that it is possible to get a reasonable agreement even without any guidelines. However, a first step towards compiling a complete list of guidelines is made. The analysis confirms the intuitive notion that the system best handles translations from a morphologically rich into a morphologically poor language. The opposite direction generates many *incorrect word* errors. A discrepancy detected in the number of *missing word* errors opens up new questions. Should it be attributed merely to the lack of guidelines since annotators might annotate one *incorrect form* error for two errors of types *missing word* and *extra word*? Presenting fine-grained agreement analysis where the kappa values are given for each error category is left for our future work. It would be interesting to show confusion matrices at each level of Vilar's taxonomy within the super-category. Such presentation would be informative enough, i.e. it would suffice to know that *sense,* no matter of what subcategory, might be confused with *incorrect form*. By analysing errors the annotators do not agree on, categories which are most easily confused may be pinpointed, and a list of guidelines may be expanded. Furthermore, it would be interesting to see how automatic error classifications proposed by (Popovic & Burchardt, 2011) or (Fishel, Bojar, Zeman, & Berka, 2011) correlate with the results obtained in this study.

## Acknowledgement

This work has been fully supported by the University of Rijeka under the project number 16.13.2.2.01.

## References

Brkic, M., Basic Mikulic, B., Matetic, M. "Can we beat GT?" Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces (ITI). 2012, 381--386.

Brkic, M., Seljan, S., Matetic, M. "Machine Translation Evaluation for Croatian-English and English-Croatian Language Pairs." NLPCS Workshop: Human-Machine Interaction in Translation. Copenhagen: Copenhagen Business School. 2011, 93--104.

Brkic, M., Seljan, S.; Vicic, T. "Automatic and Human Evaluation on English-Croatian Legislative Test Set." International Conference on Intelligent Text Processing and Computational Linguistics. 2013, 311--317.

Brkić, M.; Vičić, T.; Seljan, S. Evaluation of the Statistical Machine Translation Service for Croatian-English. // *International Conference The Future of Information Sciences*. 2009, 319--332.

Costa, Â.; Ling, W.; Luís, T.; Correia, R.; Coheur, L. A Linguistically Motivated Taxonomy for Machine Translation Error Analysis. // *Machine Translation* (2015): 127--161.

Elliott, D.; Hartley, A.; Atwell, E. A Fluency Error Categorization Scheme to Guide Automated Machine Translation Evaluation. // *Conference of the Association for Machine Translation in the Americas*. 2004, 64--73.

Farrús, M.; Costa-Jussa, M. R.; Marino, J. B.; Poch, M.; Hernández, A.; Henríquez, C.; Fonollosa, J. A. Overcoming Statistical Machine Translation Limitations: Error Analysis and Proposed Solutions for the Catalan--Spanish Language Pair. // *Language resources and evaluation*. (2011), 181--208.

Federico, M.; Negri, M.; Bentivogli, L.; Turchi, M.; Kessler, F. F. B. Assessing the Impact of Translation Errors on Machine Translation Quality with Mixed-effects Models. // *EMNLP*. 2014, 1643--1653.

Fishel, M.; Bojar, O.; Zeman, D.; Berka, J. Automatic Translation Error Analysis. // *International Conference on Text, Speech and Dialogue*. 2011, 72--79.

Flanagan, M. Error Classification for MT Evaluation. // *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*. 1994.

Font-Llitjós, A.; Carbonell, J. G.; Lavie, A. A Framework for Interactive and Automatic Refinement of Transfer-Based Machine Translation. // *Tenth workshop of the European Association for Machine Translation (EAMT)*. 2005.

Koponen, M. Assessing Machine Translation Quality with Error Analysis. // *Electronic proceeding of the KaTu symposium on translation and interpreting studies*. 2010.

Lommel, A.; Burchardt, A.; Popovic, M.; Harris, K.; Avramidis, E.; Uszkoreit, H. Using a New Analytic Measure for the Annotation and Analysis of MT Errors on Real Data. // *Proc. of EAMT*. 2014.

Popovic, M.; Arcan, M. PE2rr Corpus: Manual Error Annotation of Automatically Pre-annotated MT Post-edits. // *LREC*. 2016.

Popovic, M.; Burchardt, A. From Human to Automatic Error Classification for Machine Translation Output. // *15th International Conference of the European Association for Machine Translation (EAMT 11)*. 2011.

Secară, A. Translation Evaluation - a State of the Art Survey. // *Proceedings of the eCoLoRe/MeLLANGE Workshop*. 2005, 39-44.

Seljan, S.; Dunđer, I. Automatic Quality Evaluation of Machine-Translated Output in Sociological-Philosophical-Spiritual Domain. // *10th Iberian Conference on Information Systems and Technologies (CISTI)*. 2015a.

Seljan, S.; Dunđer, I. Machine Translation and Automatic Evaluation of English/Russian-Croatian. // *Proceedings of Corpus Linguistics*. 2015b, 72--79.

Seljan, S.; Brkic, M.; Vicic, T. BLEU Evaluation of Machine-Translated English-Croatian Legislation. // *LREC*. 2012, 2143--2148.

Seljan, S.; Brkic, M.; Kucis, V. Evaluation of Free Online Machine Translations for Croatian-English and English-Croatian Language Pairs. // *Proceedings of the 3rd International Conference on the Future of Information Sciences: INFuture2011-Information Sciences and e-Society*. 2011, 331--344.

Seljan, S.; Tucaković, M.; Dunđer, I. Human Evaluation of Online Machine Translation Services for English/Russian-Croatian. New Contributions in Information Systems and Technologies. // *Springer International Publishing*, 2015, 1089--1098.

Stymne, S.; Ahrenberg, L. On the Practice of Error Analysis for Machine Translation Evaluation. // *LREC*. 2012, 1785--1790.

Stymne, S. Blast: A Tool for Error Analysis Of Machine Translation Output. // *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*. 2011, 56--61.

Vilar, D.; Xu, J.; d'Haro, L. F.; Ney, H. Error Analysis of Statistical Machine Translation Output. // *Proceedings of LREC*. 2006, 697--702.