Applying the Cognitive Machine Translation Evaluation Approach to Arabic

Irina Temnikova, Wajdi Zaghouani[†], Stephan Vogel, Nizar Habash[‡]
Qatar Computing Research Institute, HBKU
[†]Carnegie Mellon University in Qatar, [‡]New York University Abu Dhabi
itemnikova@qf.org.qa, wajdiz@cmu.edu, svogel@qf.org.qa, nizar.habash@nyu.edu

Abstract

The goal of the cognitive machine translation (MT) evaluation approach is to build classifiers which assign post-editing effort scores to new texts. The approach helps estimate fair compensation for post-editors in the translation industry by evaluating the cognitive difficulty of post-editing MT output. The approach counts the number of errors classified in different categories on the basis of how much cognitive effort they require in order to be corrected. In this paper, we present the results of applying an existing cognitive evaluation approach to Modern Standard Arabic (MSA). We provide a comparison of the number of errors and categories of errors in three MSA texts of different MT quality (without any language-specific adaptation), as well as a comparison between MSA texts and texts from three Indo-European languages (Russian, Spanish, and Bulgarian), taken from a previous experiment. The results show how the error distributions change passing from the MSA texts of worse MT quality to MSA texts of better MT quality, as well as a similarity in distinguishing the texts of better MT quality for all four languages.

Keywords: Arabic, machine translation evaluation, post-editing

1. Introduction

Machine Translation (MT) today is used more and more by professional translators, including freelancers, companies, and official organisations, such as, for example, the European Parliament. MT output, especially of publicly available MT engines, such as Google Translate,[1] is, however, well known to contain errors and lack fluency from human expectations' point of view. For this reason, the MT translated texts often need manual (or automatic) corrections, known as "Post-Editing" (PE) (Allen, 2001; Somers, 2003). Although there are fast and simple measures of post-editing cost, such as time to post-edit, or edit-distance, these measures do not reflect the cognitive difficulty involved in correcting the specific errors in the MT output text. As the MT output texts can be of different quality and thus contain errors of different difficulty to be corrected, fair compensation of post-editing should take into account the difficulty of the task, which should thus be measured in the most reliable way. The best solution for this would be to build an automatic classifier which (a) assigns each MT error into a specific correction class, (b) assigns an effort value which reflects the cognitive effort a post-editor needs to make in order to make such a correction, and (c) gives a post-editing effort score to a text. On our way of building such a classifier, we investigate whether an existing cognitive effort model (Temnikova, 2010) could provide a fairer compensation for the post-editor, by testing it on a new language which strongly differs from the previous languages on which this methodology was tested.

The model made use of the Statistical Machine Translation (SMT) error classification schema proposed in Vilar et al. (2006), from which the error classes were subsequently re-grouped and ranked in an increasing order, so as to reflect the cognitive load post-editors experience while correcting the MT output. Error re-grouping and ranking was done on the basis of relevant psycholinguistic error correction literature (Harley, 2013; Larigauderie et al., 1998; Baddeley and Hitch, 1974). The aim of proposing such an approach was to create a better metric for the effort a post-editor faces while correcting MT texts, instead of relying on a non-transparent MT evaluation score such as BLEU (Papineni et al., 2002). Figure 1. shows the previous error ranking. The easiest errors to correct were considered those which required only a small change inside the word (*CInF*), followed by errors requiring replacing or adding a word (*Styl*, *InW*, etc.), while the hardest errors were considered those which required understanding of the whole sentence (e.g. *InP*, *MissP*, *WoW* and *WoPh*).

Text Level	Error Category
Morpho	1. Correct word, incorrect form (CInF)
Lexical	2. Incorrect style synonym (Styl)
	3. Incorrect word (InW)
	4. Extra word (ExW)
	5. Missing word (MissW)
	6. Idiomatic expression (Idiom)
Syntactic	7. Wrong punctuation (InP)
	8. Missing punctuation (MissP)
	9. Word order at Word level (WoW)
	10. Word order at Phrase level (WoPh)

Figure 1: Temnikova (2010)'s Error Ranking.

The approach does not rely on using specific software, in contrast to PE cognitive evaluation approaches which are based on keystroke logging (Carl et al., 2011; Krings and Koby, 2001; Koponen et al., 2012) or eye-tracking (Carl et al., 2011; Vieira, 2014; Stymne et al., 2012; Doherty and O'Brien, 2009; O'Brien, 2011). Furthermore, the approach is more objective than the approaches which rely on human scores for perceived post-editing effort (Specia, 2011; De Sousa et al., 2011; Koponen, 2012; Vieira, 2014). In its essence, it is similar to other error classification approaches, such as (Flanagan, 1994; Font-Llitjós et al., 2005;

[1] https://translate.google.com

Vilar et al., 2006; Farrús Cabeceran et al., 2010; Blain et al., 2011; Stymne, 2011; Koponen, 2012; Koponen et al., 2012; Fishel et al., 2012; Stymne et al., 2012; Vieira, 2014; Avramidis et al., 2014). It is enriched, however by error ranking, based on information specifying which errors require more cognitive effort to be corrected, and which less. In this way, the approach only requires counting the number of errors of each type in the MT output. And thus it allows the comparison of the post-editing cost of different output texts of the same MT engine, the same text as an output of different MT engines, or for different language pairs.

Temnikova (2010) tested her approach on two emergency instructions texts, one original (called "Complex") and one manually simplified (called "Simplified"), according to Controlled Language (CL) text simplification rules (Temnikova et al., 2012). Both texts were translated using the web version of Google Translate into three languages: Russian, Spanish, and Bulgarian. The MT output was manually post-edited by 3-5 human translators per language and then the number of errors per category was manually counted by one annotator per language.

Several researchers based their work on Temnikova (2010)'s cognitive evaluation approach. Among them, Koponen et al. (2012) have modified the error classification by adding one additional class: "Typographical, upper/lowercase or similar orthographical edits", and splitting the "Incorrect Word" (*InW*) class into three sub-classes: 1) Different word but same POS, 2) Different POS, and 3) Untranslated source word in MT. Lacruz and Munoz (2014) enriched our original error ranking/classification with numerical weights from 1 to 9, which showed a good correlation with another metric they used (Pause to Word Ratio), but did not normalise the scores per text length. The weights were added to form a unique score for each text called Mental Load (ML).

The current work presented in this paper makes the following contributions, compared to our previous work:

1. We separate the Controlled Language (CL) evaluation as it was in Temnikova (2010) from the MT evaluation and applies it only as MT evaluation.
2. We test the error classification and ranking method on a new (Non-Indo-European) language (Modern Standard Arabic, MSA).
3. We increase the number of annotators and textual data.
4. We test the approach on new text genres (news articles).

2. Experimental Settings

In this section we present the settings of our experiment.

2.1. Texts Used

To test the MT Cognitive Evaluation Approach on Arabic, we used three texts (Text1, Text2, and Text3), which were Wikinews[2] articles translated with Google Translate[3] into Modern Standard Arabic (MSA). The three texts are part of a post-edited corpus, which we plan to distribute, together with the post-editing guidelines in a future shared task on automatic error correction.[4]

The texts were selected in a way to contain different amounts of post-editing corrections, in order to represent different MT output quality: Text1 had 46% corrections changes, Text2 – 37% correction changes, and Text3 - 14% correction changes. The percentage of correction changes was calculated by dividing *the number of actions, recorded by the post-editing tool* by *the number of the MT output text tokens* (stated above). Each action corresponded to changes related to 1 word. In this way, we can consider Text1 as the hardest to correct, and Text3 as the easiest to correct. For the rest of this paper, we refer to the three texts by the percentage of errors they contain, namely Text1 (with 46% changes) is **Ar46**, Text2 (with 37% changes) is **Ar37**, and Text3 (with 14% changes) is **Ar14**.

The length of the MT output versions of the three texts were respectively 193 tokens (words and punctuation included) for Ar46, 194 tokens for Ar37, and 134 tokens for Ar14.

The specific texts used were created as a part of the Qatar Arabic Language Bank (QALB) project, a large-scale manually annotated annotation project (Zaghouani et al., 2014b; Zaghouani et al., 2015; Mohit et al., 2014; Rozovskaya et al., 2015; Zaghouani et al., 2016). The project goal was to create an error corrected 2M-words corpus for online user comments on news websites, native speaker essays, non-native speaker essays and machine translation output.

The MT output texts were first automatically corrected with the spelling errors correction tool MADAMIRA (Pasha et al., 2014) for common errors such as the Ya/Alif-Maqsura, Ha/Ta-Marbuta and the Hamzated Alif forms. Although usually MT systems do not produce spelling errors, unless trained on very noisy data, this is not the case of Arabic. In many cases we noticed that the MT system produced spelling errors such as the spelling of the Arabic letter Hamza, which is due to common mistakes related to that particular letter in Arabic.[5]

Next, the texts were post-edited by the QALB team of human post-editors (native speakers of Arabic), who were specially trained beforehand following a specific set of guidelines (Zaghouani et al., 2014a; Zaghouani et al., 2016). The post-editors used QAWI (QALB Annotation Web Interface), a web post-editing tool (Obeid et al., 2013). The QAWI tool records the post-editing actions such as correcting, modifying, moving, deleting, merging, splitting and inserting words in an XML file.

2.2. Error Category Annotation

As the aim of the current experiment was to test the previous evaluation approach on Arabic, we decided to keep the original error categories and ranking from Temnikova (2010) without making any changes, with a future aim to adapt those to Arabic, as needed. Since in this annotation effort there was no phrase-level word move, the only

[2]https://en.wikinews.org

[3]https://translate.google.com. English-Arabic language pair, web version of the engine used in October 2015.

[4]This data can be obtained from the authors after the signature of a free license agreement.

[5]For more information on challenges to Arabic natural language processing, see (Habash, 2010).

change made was to merge the following two categories into a new word-order category *WoE (Word Order Error)*:

1. *WoW - Word Order error at Word level* category, i.e., error correction requiring moving single words.

2. *WoPh - Word Order error at Phrase level* category, i.e., error correction requiring moving whole phrases.

As the original error categories did not reflect two changes in the same word, e.g., moving a word and then correcting it, the annotators were also asked to annotate those new cases.
For our annotation, we had four annotators (all Native speakers of Arabic), two of which are graduate students in translation and interpretation studies. All the annotators were also post-editors of these texts. Ar46 was annotated by three annotators, Ar37 and Ar14 by two annotators each. The three texts were annotated by different combinations of annotators, depending on the availability of the annotators. Table 1 shows the distribution of annotators per text.

Text	Annotators
Ar46	Ann1, Ann2, Ann3
Ar37	Ann3, Ann4
Ar14	Ann1, Ann2

Table 1: Assignment of Annotators per Text.

Google Spreadsheets[6] were used for annotation, with error categories provided in drop down lists. The annotation guidelines were translated into Arabic, coupled with examples for each category (see Figure 2.2.).[7] The annotation spreadsheet contained the MT output tokens, the post-edited tokens, a column for selecting an error category from a drop-down list, a column for marking the presence of two changes in the same word, and a screenshot (taken from the web post-editing tool) of the post-edited text with changes highlighted (see Figure 2.2.). The annotators were asked to assign only one category and put a flag if there were two changes in the same word.
The Inter-Annotator Agreement (IAA) was 88.49% for Ar46, 88.50% for Ar37, and 99% for Ar14 (the text with the highest MT quality). The average IAA for all three texts was 92%.

2.3. Research Hypotheses

Our research hypotheses are:

1. The distributions of categories of errors in the three Arabic texts will differ.

2. The Arabic language will show different error distributions from the languages of the previous experiment.

We test Hypothesis 1 by comparing the error categories distributions for the 3 Arabic texts. We test Hypothesis 2 by comparing the error categories distributions of the three Arabic texts with the previous Russian, Spanish, and Bulgarian pairs of Complex and Simplified texts. Specifically Hypothesis 2 is motivated by the fact that Arabic is an Afro-Asiatic, Semitic language (differently from Bulgarian, Spanish and Russian, which are all Indo-European), and thus should have substantial differences from these three languages.

3. Results and Discussion

3.1. Results

Figures 4, 5, 6, and 7 show the results of the experiment for Arabic (the distributions of error categories in the three Arabic texts and the percentage of errors in each Arabic text). In order to make the results more comparable, the mean number of errors in Figure 4 has been normalized by the number of tokens (including words and punctuation signs) in each text.
The total count of the errors of the category *Correct Word Incorrect Form* (*CInF*) includes the spelling errors automatically corrected by MADAMIRA (Pasha et al., 2014), which are, respectively 12 for Ar46, 7 for Ar37, and 8 for Ar14. Also, the number of not corrected tokens per text are, respectively: 135.7 (on average per annotator) for Ar46; 144.5 in Ar37; and 120.7 for Ar14. It is clear that the number of not corrected tokens is the highest in Ar14 (the text with the best MT output quality).
Respectively, Figure 8 and Figure 9 show the results of the comparison between the error categories present in the MT-translated Arabic texts and the error categories present in the Russian, Spanish, and Bulgarian "Complex" and "Simplified" pairs of texts from the previous experiment (Temnikova, 2010). Table 2 gives the acronyms we use to refer to the Russian, Spanish, and Bulgarian texts. We compare separately between them the more difficult-to-correct texts (Ar46, Ar37, RuC, EsC, and BgC) in Figure 8 and separately the easier-to-correct texts (Ar14, RuS, EsS, and BgS) in Figure 9.

Acronym	Text
RuC	Russian Complex
EsC	Spanish Complex
BgC	Bulgarian Complex
RuS	Russian Simplified
EsS	Spanish Simplified
BgS	Bulgarian Simplified

Table 2: Acronyms of Previous Texts.

3.2. Discussion

3.2.1. Distribution of Error Categories in Arabic

As can be seen from Figure 4, the three texts show different error distributions (which confirms our first hypothesis), with the major differences occurring in *CInF*, *Styl*, *InW*, *MissW*, and *WoE* error categories. We can also observe that the number of *CInF*, *InW*, and *WoE* errors decreases when the quality of the MT text improves. We also see that in none of the texts there are any idiom errors, and only Ar46

[6] https://www.google.com/sheets/about/

[7] The annotation guidelines are available upon request from the authors.

	Please read the original and the corrected texts and fill in the spreadsheet. Assign an error category to each corrected piece of the text (underlined in light blue). If a word has not been corrected, leave the NO_Correct category. Fill in the Excel spreadsheet cells in order of encountering the editing corrections. E.g.: The categories with their definitions and examples are given below:	
Code	Definition	Examples
NO_Correct	A word that has not been correct or changed. كلمة غير مصححة	
ClnF	The post-editor changed an ending of the word, or added a diacritic. ...خطأ إعراب صرف إملاء	تصحيح كلمة المدينه - المدينة
Styl	The post-editor changed a word with a synonym (= another word with the same or similar meaning). تصحيح بمرادف	تغيير كلمة أسد بكلمة ليث
InW	The post-editor changed a word with a word with a totally different meaning. تغيير كلمة بمعنى مغاير	تغيير كلمة أبيض بكلمة أسود
ExW	The post-editor removed a word. حذف كلمة	
MissW	The post-editor added a missing word. إضافة كلمة	
Idiom	The post-editor corrected a wrongly translated idiomatic expression. تصحيح ترجمة مثل شعبي	
InP	The post-editor changed a punctuation sign. تغيير في علامات الترقيم	تغيير النقطة بنقطة استفهام
MissP	The post-editor added a missing punctuation sign. إضافة علامة ترقيم	
WoE	The post-editor moved a word or several words from one part of the text to another. تغيير وضع الكلمة في الجملة	تغيير مثل الولد ذهب - ذهب الولد

Figure 2: Error Category Annotation Guidelines.

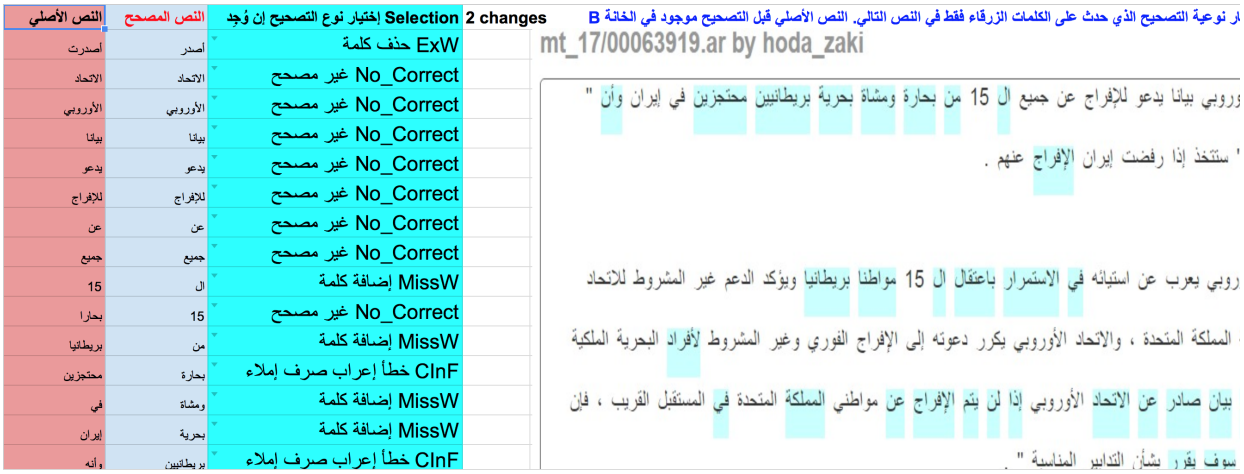

Figure 3: Annotators' screen.

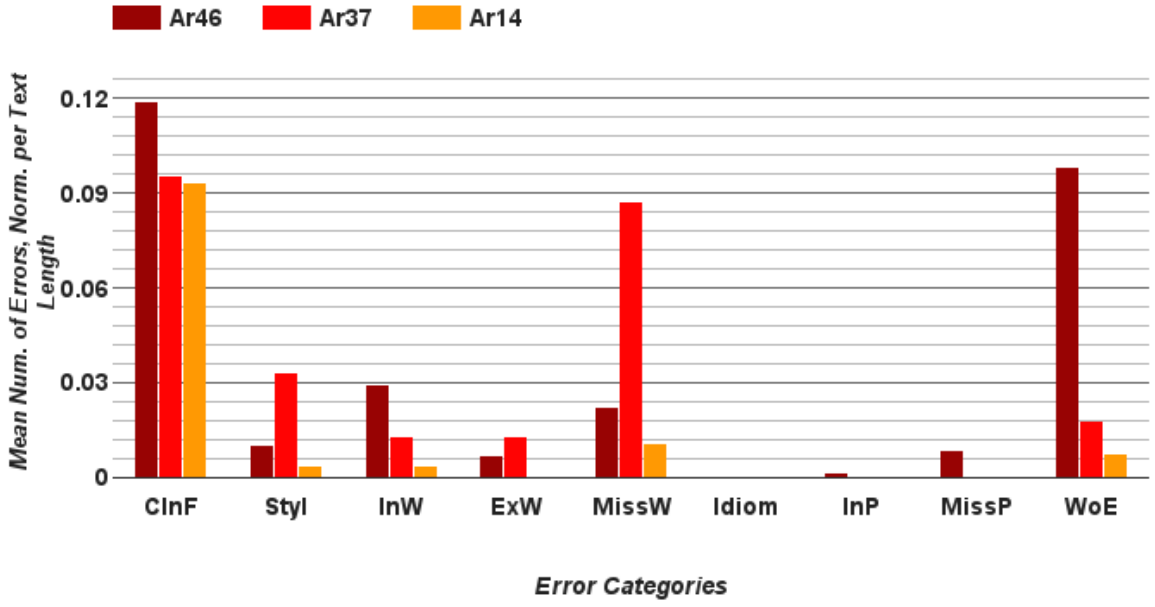

Figure 4: Comparison of Error Distributions for the Three Arabic Texts.

exhibits some punctuation errors. We hypothesize that the high number of *Styl* and *MissW* errors in Ar37 can be text-specific, which requires additional analysis of Ar37.

In terms of relative percentages of error categories in each text, we observe that text with different MT qualities have different distributions of errors (which again confirms our Hypothesis 1). For example, the highest number of errors in Ar14 (the one with the best MT quality) are *ClnF* er-

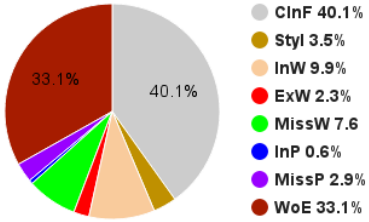

Figure 5: Percentages of Errors in Ar46.

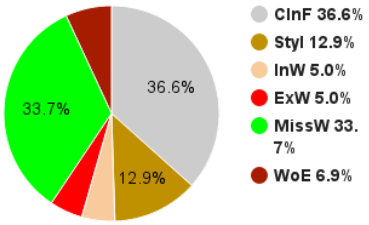

Figure 6: Percentages of Errors in Ar37.

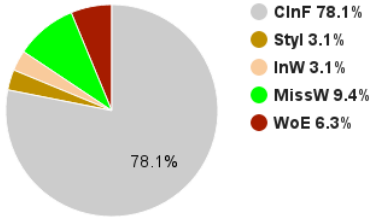

Figure 7: Percentages of Errors in Ar14.

rors, and it also has less categories of errors (5 categories only) than the other two texts (6 categories for Ar37 and 8 categories for Ar46).

We also observe that the worst quality text (Ar46) is characterized by a high number of *WoE* and *CInF* errors, which can be seen also from Figure 4.

In terms of cognitive effort, from Figure 4 we observe that Ar37 and Ar14 have a lower number of *InP*, *MissP*, and *WoE* errors, which are *cognitively difficult to correct*. The results also show decrease in the number of all errors for Ar37 and Ar14, except for the categories *Styl*, *ExW*, and *MissW* in Ar37, which is again something related to this specific text. In particular, the much lower number of all categories of errors in Ar14 shows that Ar14 requires the least cognitive effort to be post-edited.

A very low number of double changes per word (e.g. a word is moved, and then corrected, see Section 2.2.) has been observed. The normalized means per text length (number of tokens, including punctuation), are: Ar46 - 0.035; Ar37 - 0.013; Ar14 - 0.007. As can be seen, the number of double changes decreases with the increasing MT quality of the texts.

3.2.2. Comparison of Arabic with the Three Previous Languages

The comparison between Ar46, Ar37, RuC, EsC, and BgC (see Figure 8) shows higher numbers of *CInF*, *MissW*, *MissP*, and *WoE* for Arabic, compared to the other languages. On the other hand Russian, Spanish, and Bulgarian have higher number of errors for Incorrect Word (*InW*), than Arabic. These two findings confirm our Hypothesis 2, and namely, that there are some language-specific differences for Arabic. The lowest number of errors for all these languages in the difficult-to-correct texts can be observed at *Idiom*, *InP*, and *MissP*.

The comparison between the easier-to-correct-texts (Ar14, RuS, EsS, and BgS) show similar tendencies for all the languages, namely the number of errors decreases from the easier-to-correct categories of errors (e.g. Correct Word, Incorrect Form *CInF*) to the more difficult-to-correct-errors (e.g. Word order Errors *WoE*). Moreover, there are no errors of categories *Idiom*, *InP*, and *MissP* for all these languages. These findings show that although there are some language-specific characteristics, our evaluation approach can be used for distinguishing the easier-to-correct texts from the more difficult-to-correct ones for all these languages without modification.

4. Conclusions and Future Work

On our way of building a classifier which would assign post-editing effort scores to new texts, we have conducted a new experiment, aiming to test whether a previously introduced approach (Temnikova, 2010) applies also to Arabic, a language different from those for which the cognitive evaluation model was initially developed.

The results of the experiment confirmed once again that Machine Translation (MT) texts of different translation quality exhibit different distributions of error categories, with the texts with lower MT quality containing more errors, and error categories which are more difficult to correct (e.g. word order errors). The results also showed some variation in the presence of certain categories of errors, which we deem being typical for Arabic. The comparison of texts of better MT quality showed similar results across all four languages (Modern Standard Arabic, Russian, Spanish, and Bulgarian), which shows that the approach can be applied without modification also to non-Indo-European languages in order to distinguish the texts of better MT quality from those of worse MT quality.

In future work, we plan to adapt the error categories to Arabic (e.g., add the category "merge tokens"), in order to test if such language-specific adaptation would lead to better results for Arabic. We plan to use a much bigger dataset and extract most of the categories automatically. We also plan to assign weights and develop a unique post-editing cognitive difficulty score for MT output texts. We are confident that this will provide a fair estimation of the cognitive effort required for post-editors to edit such texts, and will help

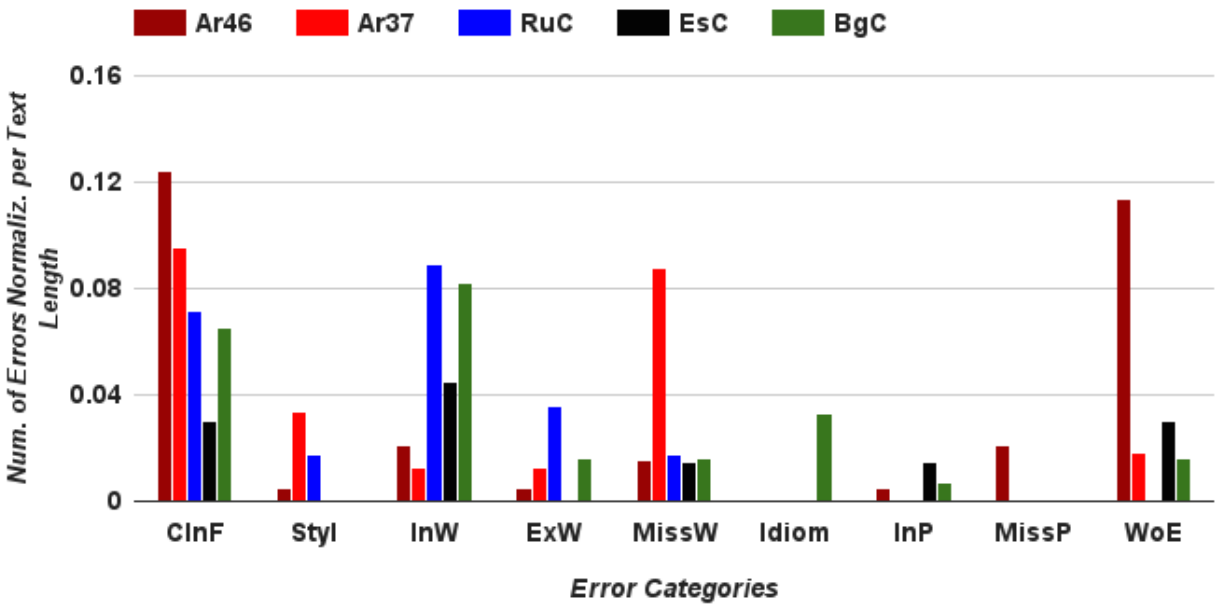

Figure 8: Comparison of the Error Distributions for the Arabic, Russian, Spanish, and Bulgarian Difficult Texts.

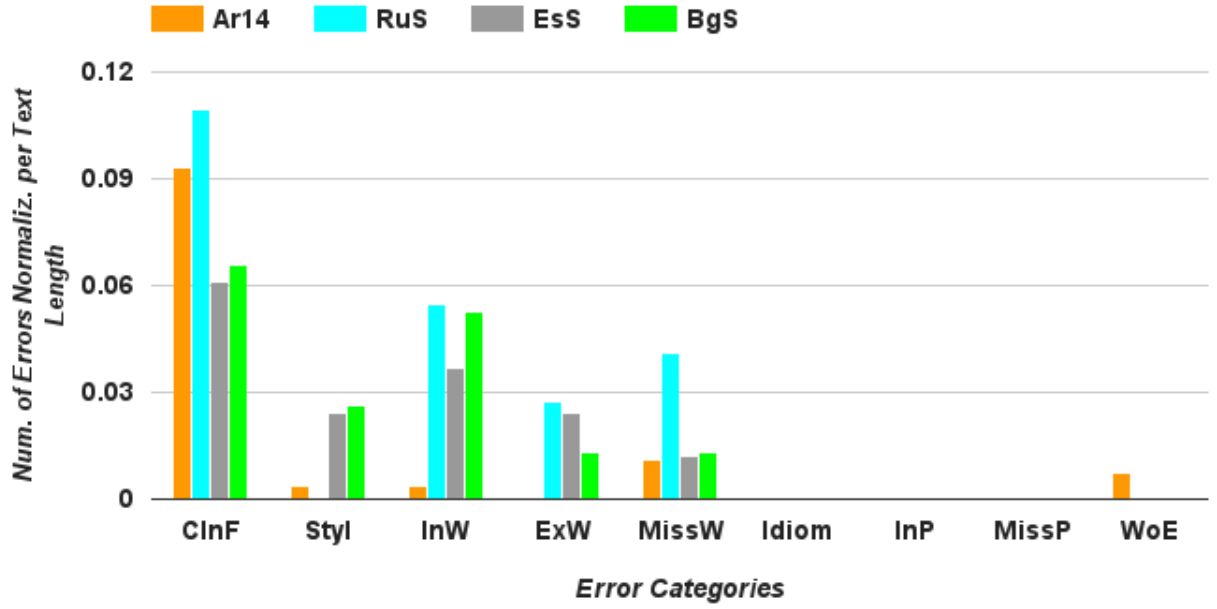

Figure 9: Comparison of the Error Distributions for the Arabic, Russian, Spanish, and Bulgarian Easy Texts.

translators to receive a fair compensation for their work.

5. Acknowledgments

We thank the anonymous reviewers for their valuable comments and suggestions. The second and fourth authors, Zaghouani and Habash, as well as the QALB post-editors, were funded by grant NPRP-4-1058-1-168 from the Qatar National Research Fund (a member of the Qatar Foundation).

6. Bibliographical References

Allen, J. (2001). Postediting: An integrated part of a translation software program. *Language International*, 13(2):26–29.

Avramidis, E., Burchardt, A., Hunsicker, S., Popovic, M., Tscherwinka, C., Vilar, D., and Uszkoreit, H. (2014). The taraXU corpus of human-annotated machine translations. *LREC*.

Baddeley, A. D. and Hitch, G. J. (1974). Working memory. *The psychology of learning and motivation*, 8:47–89.

Blain, F., Senellart, J., Schwenk, H., Plitt, M., and Roturier, J. (2011). Qualitative analysis of post-editing for high quality machine translation. *MT Summit XIII: the Thirteenth Machine Translation Summit [organized by the] Asia-Pacific Association for Machine Translation (AAMT)*, pages 164–171.

Carl, M., Dragsted, B., Elming, J., Hardt, D., and Jakobsen, A. L. (2011). The process of post-editing: A pilot study. In *Proceedings of the 8th international NLPSC workshop. Special theme: Human-machine interaction in translation*, pages 131–142.

De Sousa, S. C., Aziz, W., and Specia, L. (2011). Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In *RANLP*, pages 97–103.

Doherty, S. and O'Brien, S. (2009). Can MT output be evaluated through eye tracking. *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, pages 214–221.

Farrús Cabeceran, M., Ruiz Costa-Jussà, M., Mariño Acebal, J. B., Rodríguez Fonollosa, J. A., et al. (2010). Linguistic-based evaluation criteria to identify statistical machine translation errors.

Fishel, M., Bojar, O., and Popovic, M. (2012). Terra: A collection of translation error-annotated corpora. In *LREC*, pages 7–14.

Flanagan, M. (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.

Font-Llitjós, A., Carbonell, J. G., and Lavie, A. (2005). A framework for interactive and automatic refinement of transfer-based machine translation. *Computer Science Department*, page 286.

Habash, N. (2010). *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Harley, T. A. (2013). *The Psychology of Language: From data to theory*. Psychology Press.

Koponen, M., Aziz, W., Ramos, L., and Specia, L. (2012). Post-editing time as a measure of cognitive effort. *Proceedings of WPTP*.

Koponen, M. (2012). Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 181–190. Association for Computational Linguistics.

Krings, H. P. and Koby, G. S. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes*, volume 5. Kent State University Press.

Lacruz, I. and Munoz Martin, R. (2014). Pauses and objective measures of cognitive demand in post-editing. Americal Translation and Interpreting Studies Association Conference.

Larigauderie, P., Gaonac'h, D., and Lacroix, N. (1998). Working memory and error detection in texts: What are the roles of the central executive and the phonological loop? *Applied Cognitive Psychology*, 12(5):505–527.

Mohit, B., Rozovskaya, A., Habash, N., Zaghouani, W., and Obeid, O. (2014). The first QALB Shared Task on automatic text correction for Arabic. *ANLP 2014*, page 39.

Obeid, O., Zaghouani, W., Mohit, B., Habash, N., Oflazer, K., and Tomeh, N. (2013). A web-based annotation framework for large-scale text correction. In *Sixth International Joint Conference on Natural Language Processing*, page 1.

O'Brien, S. (2011). Towards predicting post-editing productivity. *Machine translation*, 25(3):197–215.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. M. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland*.

Rozovskaya, A., Bouamor, H., Habash, N., Zaghouani, W., Obeid, O., and Mohit, B. (2015). The second QALB Shared Task on automatic text correction for Arabic. *ANLP Workshop 2015*, page 26.

Somers, H. (2003). *Computers and translation: A translator's guide*, volume 35. John Benjamins Publishing.

Specia, L. (2011). Exploiting objective annotations for measuring translation post-editing effort.

Stymne, S., Danielsson, H., Bremin, S., Hu, H., Karlsson, J., Lillkull, A. P., and Wester, M. (2012). Eye tracking as a tool for machine translation error analysis. In *LREC*, pages 1121–1126.

Stymne, S. (2011). Blast: A tool for error analysis of machine translation output. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*, pages 56–61. Association for Computational Linguistics.

Temnikova, I., Orasan, C., and Mitkov, R. (2012). CLCM - A linguistic resource for effective simplification of instructions in the crisis management domain and its evaluations. In *LREC*, pages 3007–3014.

Temnikova, I. (2010). Cognitive evaluation approach for a controlled language post–editing experiment. In *LREC*.

Vieira, L. N. (2014). Indices of cognitive effort in machine translation post-editing. *Machine Translation*, 28(3-4):187–216.

Vilar, D., Xu, J., d'Haro, L. F., and Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of LREC*, pages 697–702.

Zaghouani, W., Habash, N., and Mohit, B. (2014a). The Qatar Arabic Language Bank guidelines. In *Technical Report CMU-CS-QTR-124, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, September.*

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., Farra, N., Alkuhlani, S., and Oflazer, K. (2014b). Large scale Arabic error annotation: Guidelines and framework. In *International Conference on Language Resources and Evaluation (LREC 2014)*.

Zaghouani, W., Habash, N., Bouamor, H., Rozovskaya, A., Mohit, B., Heider, A., and Oflazer, K. (2015). Correction annotation for non-native Arabic texts: Guidelines

and corpus. *Proceedings of The 9th Linguistic Annotation Workshop*, pages 129–139.

Zaghouani, W., Habash, N., Obeid, O., Mohit, B., and Oflazer, K. (2016). Building an Arabic Machine Translation Post-Edited Corpus: Guidelines and Annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.