Vilnius University Press

Contents lists available at Vilnius University Press

Vertimo studijos
2021, vol. 14, pp. 56–70

eISSN 2029-7033
DOI: https://doi.org/10.15388/VertStud.2021.4

Machine Translation Quality in Mobile Apps for Text-based Image Translation

Eglė Miltakienė

Kaunas University of Technology
Faculty of Social Sciences, Arts and Humanities
egle.jasenaite@ktu.edu
https://orcid.org/0000-0002-6890-3885

Abstract. With the advancement of mobile applications, now it is possible to perform instant text translation using a smartphone's camera. Because text translation within images is still a relatively new field of research, it is not surprising that the translation quality of these mobile applications is under-researched. This study aims to determine the image-to-text translation quality in the English to Lithuanian language direction using popular machine translation apps. To classify errors and evaluate the quality of translation, the present study adopts and customizes the Multidimensional Quality Metrics (MQM) framework (Lommel 2014). The obtained results indicate that image-to-text machine translation apps produce exceptionally low-quality translations for the English-Lithuanian language pair. Therefore, the quality of machine translation for low-resource languages such as Lithuanian remains an issue.
Keywords: machine translation, image-to-text applications, translation quality assessment, translation errors, multidimensional quality metrics

Mašininio vertimo kokybė vertimo programėlėse su integruotu vaizdo atpažinimu

Santrauka. Šiandien naujausiomis technologijomis grįstos vertimo programėlės su integruotu vaizdo atpažinimu suteikia galimybę išmaniuoju telefonu aptikti tekstą vaizde ir jį greitai išversti į norimą užsienio kalbą. Teksto vertimas vaizde yra dar visai nauja mokslinių tyrimų kryptis, tad šių mobiliųjų programėlių vertimo kokybė yra nepakankamai ištirta. Šio darbo objektas yra tekstų, išverstų pasitelkiant populiariąsias programėles su integruotu vaizdo atpažinimu, vertimo kokybė. Vertimo atlikto iš anglų kalbos į lietuvių kalbą su vaizdo atpažinimą integruojančiomis mašininio vertimo programėlėmis klaidų analizei pasirinkta adaptuota daugiamatė kokybės vertinimo sistema (angl. Multidimensional Quality Metrics) klasifikacija. Apibendrinus rezultatus, galima teigti, kad ištirtų vaizdo atpažinimą integruojančių programėlių vertimo iš anglų kalbos į lietuvių kalbą kokybė buvo itin prasta.
Pagrindiniai žodžiai: mašininis vertimas, vaizdo atpažinimą integruojančios vertimo programėlės, vertimo kokybės vertinimas, vertimo klaidos, daugiamatė kokybės vertinimo sistema

Introduction

Nowadays, with just a simple app, users can snap a picture of a sign, a newspaper or a menu, and it will instantly provide a translation into a selected language. The application identifies the text from the scene image and displays back the translation onto the phone's screen. The benefits of image-to-text mobile translation applications are obvious, as it may provide an aid to travellers, language learners and may even assist visually impaired navigate their surroundings. (Ramiah and Jayabalan 2015). The most widely used freely available machine translation (MT) applications, e.g., Google Translate, Microsoft Translator or Yandex, already incorporate an instant camera translation feature in their applications. Despite the substantial progress in technologies, text translation within the image is a relatively new field of research. Research on its translation quality is scarce in both international and native context. It is worth a mention that so far there are only a few machine translation quality studies for the English-Lithuanian language pair (Petkevičiūtė and Tamulynas 2011; Stankevičiūtė et al. 2017; Kasperavičienė et al. 2020, among others). The present study aims to determine the image-to-text translation quality in the English to Lithuanian language direction using best all-round machine translation systems.

The objectives set for this study are to overview the recent studies on mobile applications for text-based image translation and translation quality assessment; to identify and classify errors present in the book cover translations made by image-to-text translation apps in English to Lithuanian language direction using a customized error typology based on the Multidimensional Quality Metrics (MQM) framework (Lommel 2014); to identify the factors that contribute to the quality of translation produced by image-to-text mobile applications; to determine the most common translation errors and compare the quality of translation of the selected image-to-text applications output. This study adopts quantitative and qualitative descriptive research design.

Mobile applications for text-based image translation

By employing artificial neural networks, natural language processing and neural machine translation, the language translation is becoming a rather simple task for various translation applications. Translation applications like text-to-text, text-to-speech, speech-to-speech and other translation apps on mobile devices are reshaping interlingual communication, and at the same time hint at the shift to a more innovative future of technology. In addition to the fact that the implementation of translation technologies appears to have formed a completely different approach to translation and its quality, this topic has captured the entire attention of major technology companies, scholars, industry professionals and users.

Numerous mobile applications with text extraction and translation tools have been proposed to facilitate communication among speakers of different languages. Real-time image (of words) to text translation mobile applications have been actively researched over the last decade e.g. Canedo-Rodríguez et al. (2009), Fragoso et al. (2011), Petter et al. (2011). Researchers also underline the immeasurable value of text extraction and translation technologies for the blind and visually impaired. Mobile applications that detect and read a text in natural scenes are an indispensable aid when navigating in both indoor and outdoor environments, retrieving a text and providing contextual clues for numerous vision tasks, e.g. Epshtein et al. 2010, among other authors.

Thus, as for the performance of mobile applications for text-based image translation, the studies indicate certain patterns in text detection, text extraction and text translation quality. Hamad and Mehmet (2016) report on the most prevalent challenges faced by optical character recognition (OCR) technologies in terms of good quality and high accuracy character recognition, because multiple mistakes while detecting and extracting text from images may occur. The authors provide the major determining factors for the accuracy and success of text detection and extraction: scene complexity, conditions of uneven lighting, skewness (rotation), blurring and degradation, tilting (perspective distortion), fonts, multilingual environments, warping (Mehmet 2016).

Above all, the pace of technological development has gained tremendous momentum, and it is clear that the most recent translation functions of mobile applications are far advanced. Modern mobile translation tools such as mobile applications for text-based image translation break down language barriers, accelerate cultural integration and enable easy access to the world's ideas and knowledge to anyone.

Approaches to machine translation quality assessment

The evaluation of MT quality is a fundamental field of research. MT error analysis is a way to identify weaknesses of translation systems, analyse and improve their performance. MT system quality can be measured in different ways; however, the best practice for the evaluation of MT quality is considered either human or automatic assessment.

The most reliable method so far for judging and measuring MT quality is human expert evaluation, where the quality of translation output is assessed by bilingual professionals in linguistics and translation. Though human MT evaluation has many benefits and is even considered the golden standard (Läubli et al. 2020), human evaluation is time-consuming, expensive and rather subjective. Due to these limitations, today, in MT research, automated evaluation methods are the fastest, cheapest and effortless way to measure the effectiveness of an MT systems. For this reason, auto-

mated metrics are a particularly common option for MT quality assessment. The most prominent metrics in the field are BLEU (Papineni et al. 2002) and METEOR (Banerjee and Lavie 2005). Even though these metrics are extensively used in MT research, scientists identify potential limitations of automated metrics (Babych 2014). All automated evaluation methods provide an overall translation performance assessment and quantitative scores; however, they lack detail on translation errors types. That being the case, to identify the MT systems strengths and weaknesses in more detail, various error typologies have been proposed for the comprehensive evaluation of MT quality and the classification of translation errors (Flanagan 1994; Vilar et al. 2006; Farrús et al. 2010, etc.).

Given the importance of MT quality assessment on the global level, in Lithuania, however, only a few studies exist for English to Lithuanian MT quality evaluation. One of the earliest attempts to discuss the importance of machine translation efficiency for the Lithuanian language was made by Labutis (2005). Daudaravičius (2006) briefly described features of MT systems and the evaluation criteria for quality assessment. Rimkutė et al. (2007) emphasised the importance of MT quality assessment and discussed linguistic peculiarities that arise in the machine translation process. Among the later approaches for English to Lithuanian, machine translation quality evaluation for neural translation systems was conducted by Stankevičiūtė et al. (2017) and Kasperavičienė et al. (2020). However, one of the most valuable studies on the MT for English-Lithuanian quality evaluation was published by Petkevičiūtė and Tamulynas (2011), who determined the main indicators for translation quality as well as identified many practical translation problems faced by MT systems.

As the widespread adoption of artificial intelligence in MT engines continues to improve the quality of translation, repeated assessment of MT output is utterly important. Major tech companies and MT system developers are constantly improving the quality based on weaknesses of translations MT systems produce. As a consequence, the output must be measurable in terms of quality and this is especially crucial for low-resource languages lacking large monolingual or parallel corpora.

Multidimensional Quality Metrics

In regards to the limitations of previous MT quality assessment methods, Lommel et al. (2014) have developed the most detailed and exhaustive evaluation scheme called Multidimensional Quality Metrics (MQM). This framework is based on the best existing translation quality assessment practices and, on that account, is a highly systematic and unified method to evaluate MT quality (Lommel 2018). MQM can be easily adapted to manual, semi-automatic, and automatic evaluation environments. Lommel

et al. (2014) indicate that users can customize their metrics, and for this reason, MQM can be used in many different quality evaluation environments and tasks. Another advantage of MQM is that it is a language-neutral metric and applicable to any existing language. For its many benefits, MQM has already been employed by a number of researchers in the field (Klubička et al. 2017; Vardaro et al. 2019, etc.).

The hierarchical list of quality issue types is a fundamental component of a systematic MQM framework. The hierarchy contains over 100 error types of different levels, which cover all major existing translation quality evaluation metrics (Lommel 2018). The core of MQM is organized into eight primary branches or dimensions: accuracy, fluency, terminology, locale-convention, style, verity, design, internationalization (Lommel et al. 2018).

The idea of this research is to evaluate the machine-translated text present in the text-based image produced by mobile applications. The main reasons for choosing the MQM framework are the following. Since the environment of the research material, e.g. text on images is quite uncommon, hence it is not clear what types of errors will be required to perform error analysis and what problems will arise in the MT output. For this reason and those mentioned above, it can be assumed that this unified systematic framework metric is best suited for the purposes of this research.

Methodology

Non-fiction book cover titles in English were selected as research material for this study. The rationale for choosing book titles is that the translation of the book title using image-to-text mobile application could perfectly indicate the effectiveness of such an application when rendering the meaning from one language to another language, as well as demonstrate whether translating with these applications can truly help the user to understand any textual content provided in the foreign language. The choice of selecting Google Translate and Microsoft Translator mobile applications has been inspired by the fact that they are the most widespread among users and support the largest number of language options; in this way, they increasingly expand their usability. Both Google and Microsoft applications combine image processing, optical character recognition (OCR) and language identification technologies. The applications recognize the written text from the image and overwrite it with a translation.

Following Lithuanian language peculiarities and the specifics of the evaluation task, it was decided to partially rearrange and supplement the MQM hierarchy. The custom error classification is arranged in hierarchical levels. Two main branches of the taxonomy, accuracy and fluency, descend into the following levels: category, subcategory, children, sub children. The following modifications were made to the proposed MQM tag set. On the accuracy level, which tackles the relationship between the source

and the target text, the following changes were applied. To distinguish terminology errors from mistranslation errors, a supplementary category of terminology errors has been added to the accuracy branch. Additionally, to get a view of how much content is being left untranslated in the target text by image-to-text applications, the untranslated error category was extended in three additional levels: word, phrase, sentence-level issue types. Furthermore, since one of the objectives of this study was to identify factors contributing to the quality of translation produced by image-to-text mobile applications and taking into account that image-to-text applications face text detection and extraction issues, a custom text detection/extraction subcategory was added to the tag set. Considerably, this custom error type will give insights into how text extraction and detection issues affect the overall quality assessment score. In a similar case, the fluency branch, which addresses the linguistic form or content of a text, was marginally extended. Taking into consideration the grammatical categories and morphological complexity of the Lithuanian language, custom issue types of person, number, gender and case were added to the grammar child category agreement. This possibly will be useful when assessing the quality of MT systems adapted to the Lithuanian language. Furthermore, to distinguish between common spelling errors and capitalization errors present in the translation, the subcategory capitalization was added to the spelling category. Figure 1 shows the custom MQM-compliant taxonomy used for the manual error annotation in this study.

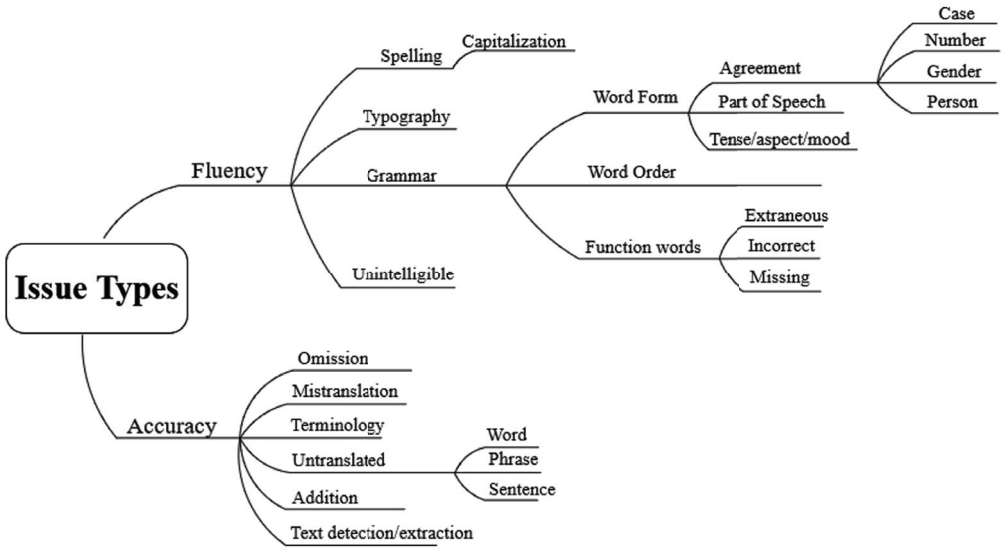


Fig. 1. The custom MQM-compliant taxonomy

Accordingly, errors present in translations were manually annotated and analysed in compliance with a developed custom metric. Given the fact that the Lithuanian language is morphologically, lexically and semantically complex, the granularity of this metric will help to grasp the language-specific nuances in the translated texts. Furthermore, it provides a detailed insight into particular errors produced by image-to text applications.

Results

This section presents the results of Google Translate and Microsoft Translator image-to-text applications translation error analysis. Each application translated a total of 355 book covers, making it a total of 2786 (17862 characters, no spaces) input words per system. In the form of raw translation error count for each application, Google Translate made 1105 translation errors and Microsoft Translator made 1227 translation errors. The quantitative data collected from both applications output indicate the exceptionally low translation quality. Google image-to-text application translated 4.79% of all book covers correctly, while Microsoft application translated 3.66% of book covers correctly. Table 1 illustrates the total number and percentages of correctly and incorrectly translated book titles for each system.

Table 1. Count of correctly and incorrectly translated book titles

App	Error		No error		Total	
Google	338	95.21%	17	4.79%	355	100%
Microsoft	342	96.34%	13	3.66%	355	100%

Google application produced a total of 338 incorrect translations, of which 48.96% were accuracy errors and 51.04% were fluency errors. Meanwhile, Microsoft application made 342 incorrect translations, of which 48.49% were accuracy errors and 51.51% were fluency errors. It was indicated that the majority of translation errors were found within the grammar, mistranslation, spelling, terminology, and untranslated categories. The overall study result implies that the Google application achieves slightly better results than the Microsoft application. Table 2 illustrates translation error counts and error distribution of major categories in both image-to-text applications after the manual annotation.

Table 2. Translation error counts and distribution of major categories

	Google App		Microsoft App	
	Error counts (n, %)		Error counts (n, %)	
Accuracy				
Addition	15	1.36%	19	1.55%
Mistranslation	218	19.73%	209	17.03%
Omission	35	3.17%	67	5.46%
Terminology	172	15.57%	123	10.02%
Untranslated	58	5.25%	99	8.07%
Text/detection extraction	43	3.89%	78	6.36%
Accuracy total	541	48.96%	595	48.49%
Fluency				
Grammar	373	33.76%	394	32.11%
Spelling	171	15.48%	205	16.71%
Typography	13	1.18%	20	1.63%
Unintelligible	7	0.63%	13	1.06%
Fluency total	564	51.04%	632	51.51%

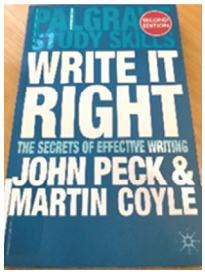
Distribution of accuracy errors

Errors within the accuracy branch address the relationship between the source text and the target text. The following categories were selected for accuracy branch error identification and analysis: addition, mistranslation, terminology, omission, untranslated (word, phrase, sentence), text detection/extraction issues. In total, under the accuracy branch, 541 errors were annotated in Google output, and 595 errors in Microsoft output (see Table 2). Mistranslation errors, a total of 218 (19.73%), were the most frequent within Google output, followed by 172 terminology (15.57%), 58 untranslated (5.25%), 43 text detection/extraction issues (3.89%), 35 omission (3.17%), and 15 addition (1.36%) errors. Similarly, the most frequent error categories within Microsoft output were mistranslation errors, a total of 209 (17.03%), followed by 123 terminology (10.02%), 99 untranslated (8.07%), 78 text detection/extraction issues (6.36%), 67 omission (5.46%), and 19 addition (1.55%) errors. Through the results, it is clear that Microsoft yielded marginally fewer mistranslation and terminology errors. However, Google produced fewer addition and untranslated content errors and had fewer text detection/extraction issues.

The following example illustrates the mistranslation error when the text is translated directly and does not reflect the original idea of the title. It is important to note

that in the original book title the text is displayed into 3 separate lines: line 1 *Write it*, line 2 *right*, line 3 *the secrets of effective writing*; it may be argued that the app detected and extracted the text from three separate segments, which results in a literal translation and definitely impacts to the translation quality.

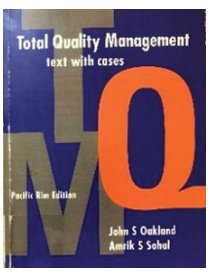
	Source (eng)	*Google App (lt)*
1)	*Write it right. The secrets of effective writing*	***Užrašyk tai teisė*** *veiksmingo rašymo* ***slaptai***

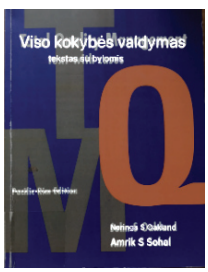


In example (1) above, the phrase *Write it right* was translated by Google Translate App as *užrašyk tai teisė*, which is a literal translation. The prefix *-už* placed at the beginning of a word *užrašyk* (*write down*) modifies its meaning. The adjective *right* is translated as a noun instead of an adjective, and for this reason, the meaning of the word becomes related to the cause of truth or justice. In addition to this, the pronoun *it/tai* should have not been translated. In the phrase *the secrets of effective writing,* the noun *secrets* was translated as an adjective *slaptai* (secretly), which results in rendering an incorrect meaning of the original title.

The second-largest category of errors within the accuracy branch was terminology errors. Terminology errors often occurred when translating book titles in economics, linguistics, or other specific fields. Example (2) below demonstrates terminology errors in both image-to-text translation applications.

	Source (eng)	*Microsoft App (lt)*
2)	*Total Quality Management: Text with Cases*	***Viso kokybės valdymas*** *tekstas su byloms*





Microsoft App translated the term Total Quality Management as *Viso kokybės valdymas*. To clarify, the application simply translates the three-word term word-for-word, which is why the output loses its meaning. The correct translation approved by the Term Bank of the Republic of Lithuania is *visuotinės kokybės vadyba*. It should be pointed out that the source of the text was laid out on one single line, so the application had full potential to detect the whole term and translate it correctly. It is noticeable that the translation quality of the terminology is dependent on a few factors. Again, terminology errors may be due to the fact that applications do not include these terms in their corpora. In addition to that, the layout of the text on the cover plays a big role in the MT output quality. In cases where the words of the multiword term are laid out in different lines, the applications do not see them as a single term and this highly affects the translation output yielded from the MT engine.

Mistranslation and terminology were the most common errors annotated in both image-to-text applications output. The main factors that resulted in incorrect translations of the content and terminology were text layout, the complexity of the book's cover design and the training corpora. Microsoft App achieved better results than Google App in terms of correct terminology; however, more errors occurred in omission, untranslated and text detection/extraction categories. Another major finding of this study was that text detection and extraction success was heavily dependent on the overall book cover's design and colour, text layout and rotation, font type and colour. Most importantly, text detection and extraction issues significantly influenced the quality of the translation.

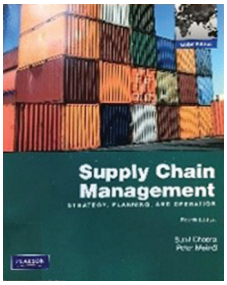
Distribution of fluency errors

Errors in this branch regard the linguistic well-formedness of the text, irrespective of whether the text is a translation or an original text. In other words, the target text has linguistic issues which prevent it from being understood. In the fluency error branch, these error categories were selected for further analysis: grammar (word form, word order, function words etc.), typography, spelling (capitalization) and unintelligible error categories. In total, under the fluency branch, 564 errors were annotated in Google output, and 632 errors in Microsoft output (see Table 2). For the Google App, the grammatical errors were the most problematic point, a total of 373 (33.76%) errors, followed by 171 spelling errors (15.48%), 13 typography errors (1.18%) and 7 unintelligible (0.63%) errors. Correspondingly, grammatical errors were the most troublesome for the Microsoft App, a total of 394 (32.11%) and were followed by 205 spelling errors (16.71%), 20 typography errors (1.63%) and 13 unintelligible (1.06%) errors. Though error distribution for Google and Microsoft output follows a somewhat similar pattern, Google image-to-text application performs better in all four categories compared with Microsoft application.

Considering that agreement errors were the most commonly detected in both systems under the word form subcategory, the given example bellow illustrates the particular case in both image-to-text apps.

Source (eng)

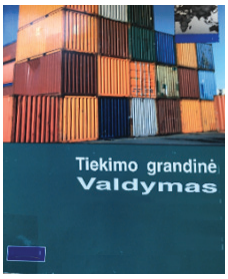
3) *Supply Chain Management: Strategy, Planning, and Operation*

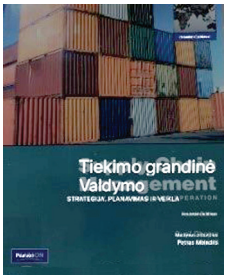


Google App (lt), Microsoft App (lt)

*Tiekimo grandin**ė** Valdymas*

*Tiekimo grandin**ė** valdym**o** strategija, planavimas ir veikla*

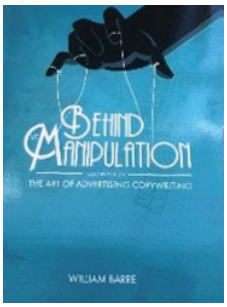




As an example, the phrase *Supply Chain Management* in both applications output text is translated disregarding the case agreement rules of the Lithuanian language. In both apps output, the noun *grandinė* (chain) which is in the nominative case should have been translated here as the genitive case *grandinės*. Also, it can be seen that Microsoft made another case agreement mistake in the same book title, i.e. the noun *valdymo* (management) which is in the genitive case should have been rendered in the nominative case. In another example, a spelling mistake is demonstrated.

Source (eng)

4) *Behind the Manipulation: The Art of Advertising Copywriting*



Google App (lt)

***Paamipulacija** reklamos **kupraktavimo** menas*

Paamipulacija is an attempt to translate the word *Manipulation*; however, the letter sequence used makes no sense and no such word exists in the Lithuanian language. In the same way, a formation *kupraktavimo* is an attempt to render the word *advertising*. Again the translation of a word is faulty and such word does not exist in Lithuanian. When examined more closely, the most likely causes of incorrectly spelt words might be related to the process of text detection and extraction; consequently, the app generates a meaningless letter combination in the output text.

The analysis showed that errors under the fluency branch made a significant impact on the translation quality. The most common issues with both image-to-text apps were grammatical and spelling errors. Taken together, the findings of the study revealed that apps had difficulties combining individual words into a coherent sense. Literal translations of words predominated, and words were rendered in incorrect word forms. Apps followed the pattern of the original text and simply replaced the words in the manner they are written on the original, disregarding the correct word form, word order, spelling or typography rules of the target language. The book cover's design, text layout, and font type all contributed to the quality of translation. Apps faced problems when translating complex text layouts or handwritten font types. Google image-to-text App surpassed Microsoft App in all fluency branch categories. However, the two apps still produced a significant number of fluency errors that resulted in extremely poor translation quality.

Although the current study has demonstrated the performance level of the image-to-text apps, yet the results are hardly comparable with most other studies on the mentioned apps. Virtually nothing was found in the literature available regarding the produced translation quality of image-to-text apps, analysis of translation errors or their usability in various communication processes within society or its groups. Moreover, since the Lithuanian language is a low-resource language, the produced quality and error rate might be different from high-resource language combinations. It would be interesting to compare the results and error tendencies image-to-text apps in other high/low-resource languages.

Conclusion

The purpose of the current study was to determine the image-to-text translation quality based on the identification and human assessment of the most common translation errors in the English to Lithuanian language direction using two well-known mobile translation applications, Google Translate App and Microsoft Translator App.

Among the different translation quality assessment methods discussed, the Multidimensional Quality Metrics framework (MQM) was selected as the most fitting for

the current study as it is one of the recent, flexible and most comprehensive, language-neutral methods for translation quality evaluation. A customized taxonomy based on MQM served as a reliable instrument to identify and classify errors present in the book cover translations produced by Google and Microsoft image-to-text translation apps in English to Lithuanian language direction. This study has found that the success of translation is highly dependent on the overall book cover's design and colour, text layout, text rotation, font type and colour. Google and Microsoft image-to-text apps face text detection and extraction issues that strongly influence the quality of the translation. Also, both apps fail to translate complex text layouts, handwritten font types or rotated text. Furthermore, the findings of the study indicate that grammar, mistranslation, spelling, terminology, and untranslated are the most common errors identified in both image-to-text apps output. The Google App achieves slightly better results than Microsoft App in most of the error categories. Microsoft App surpasses Google App only in terms of correct terminology and mistranslated content; however, it fails in all other error categories. This study has found that generally image-to-text apps follow the pattern of the original text and potentially replace words in the manner they are written on the source. For this reason, image-to-text apps are lacking MT training attributes when combining individual words into a coherent meaning or following spelling and typography rules of the target language. Literal translations dominate, disregarding the correct word form, word order, or meaning. Taken together, these results show that both Google and Microsoft image-to-text apps produce exceptionally low-quality translations for English/Lithuanian language pair and that there are so many issues to be addressed to improve their performance and the quality of translation.

References

Babych, Bogdan. 2014. Automated MT Evaluation Metrics and their Limitations. *Tradumàtica* 12. 464–470. https://doi.org/10.5565/rev/tradumatica.70

Banerjee, Satanjeev, and Lavie, Alon. 2005, June. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 65–72. Accessed March 15, 2021. https://aclanthology.org/W05-0909.pdf

Canedo-Rodriguez, Adrian, Soohyung Kim, Jung H. Kim, and Yolanda Blanco-Fernandez. 2009, March. English to Spanish Translation of Signboard Images from Mobile Phone Camera. In: *IEEE Southeastcon 2009*. IEEE, 356–361. https://doi.org/10.1109/secon.2009.5174105

Daudaravičius, Vidas. 2006. Pradžia į begalybę. Mašininis vertimas ir lietuvių kalba. *Darbai ir dienos,* 45. 7–18. Accessed March 21, 2021. https://www.ceeol.com/search/article-detail?id=209872

Epshtein, Boris, Eyal Ofek, and Yonatan Wexler. 2010. Detecting Text in Natural Scenes with Stroke Width Transform. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE*, 2963–2970. https://doi.org/10.1109/cvpr.2010.5540041

Farrús Cabeceran, Mireia, Ruiz Costa-Jussà, Marta, Mariño Acebal, José Bernardo, and Rodríguez Fonollosa, José Adrián. 2010. Linguistic-based Evaluation Criteria to Identify Statistical Machine Translation Errors. In: *14th Annual Conference of the European Association for Machine Translation*. 167–173. Accessed March 5, 2021. https://upcommons.upc.edu/handle/2117/7492

Flanagan, Mary A. 1994, October. 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*. 65–72. Accessed March 8, 2021. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.467.1013&rep=rep1&type=pdf

Fragoso, Victor, Steffen Gauglitz, Shane Zamora, Jim Kleban, and Matthew Turk. 2011. TranslatAR: A Mobile Augmented Reality Translator. In: *2011 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE. 497–502. https://doi.org/10.1109/wacv.2011.5711545

Hamad, Karez, and Kaya Mehmet. 2016. A detailed analysis of optical character recognition technology. *International Journal of Applied Mathematics Electronics and Computers*, (Special Issue-1), 244–249. https://doi.org/10.18100/ijamec.270374

Kasperavičienė, Ramunė, Jurgita Motiejūnienė, and Irena Patašienė. 2020. Quality Assessment of Machine Translation Output. *Texto Livre: Linguagem e Tecnologia*, 13(2). 271–285. https://doi.org/10.35699/1983-3652.2020.24399

Klubička, Filip, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2017, June. Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation. *The Prague Bulletin of Mathematical Linguistics* 108, 1. 121–132. https://doi.org/10.1515/pralin-2017-0014

Labutis, Vitas. 2005. Išaugusi vertėjų paklausa–nauji pavojai lietuvių kalbai. *Bendrinė kalba (iki 2014 metų–Kalbos kultūra)*, 78. 205–209. Accessed March 3, 2021. https://www.ceeol.com/search/article-detail?id=93919

Läubli, Samuel, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral 2020. A Set of Recommendations for Assessing Human–Machine Parity in Language Translation. *Journal of Artificial Intelligence Research*, 67. https://doi.org/10.1613/jair.1.11371

Lavie, Alon, and Abhaya Agarwal. 2007, June. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In: *Proceedings of the Second Workshop on Statistical Machine Translation*. 228–231. https://doi.org/10.3115/1626355.1626389

Lommel, Arle. 2018. Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. *Translation Quality Assessment*. Springer, Cham. 109–127. https://doi.org/10.1007/978-3-319-91241-7_6

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, 12. 0455–463. https://doi.org/10.5565/rev/tradumatica.77

Mercader-Alarcón, Julia, and Felipe Sánchez-Matínez. 2016. Analysis of Translation Errors and Evaluation of Pre-Editing Rules for the Translation of English News Texts into Spanish with Lucy LT. *Tradumàtica: Tecnologies de La Traducció* 14. 172. https://doi.org/10.5565/rev/tradumatica.164

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of The 40th Annual Meeting of the Association for Computational Linguistics*. 311–318. Accessed March 15, 2021. https://aclanthology.org/P02-1040.pdf

Petkevičiūtė, Inga, and Bronius Tamulynas. 2011. Kompiuterinis vertimas į lietuvių kalbą: alternatyvos ir jų lingvistinis vertinimas. *Kalbų studijos* 18. 38–45. Accessed April 5, 2021. https://etalpykla.lituanistikadb.lt/object/LT-LDB-0001:J.04~2011~1367174892606/

Petter, Marc, Victor Fragoso, Matthew Turk, and Charles Baur. 2011. Automatic Text Detection for Mobile Augmented Reality Translation. In: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 48–55. https://doi.org/10.1109/iccvw.2011.6130221

Ramiah, Sathiapriya, Tan Yu Liong, and Manoj Jayabalan. 2015. Detecting Text Based Image With Optical Character Recognition for English Translation and Speech Using Android. In: *2015 IEEE Student Conference on Research and Development (SCOReD)*. IEEE. 272–277. https://doi.org/10.1109/scored.2015.7449339

Rimkutė, Erika, and Jolanta Kovalevskaitė. 2007. Mašininis vertimas–greitoji pagalba globalėjančiam pasauliui. *Gimtoji kalba,* 9. 3–10. Accessed March 3, 2021. http://donelaitis.vdu.lt/lkk/pdf/MV2.pdf

Stankevičiūtė, Gilvilė, Ramunė Kasperavičienė, and Jolita Horbačauskienė. 2017. Issues in Machine Translation. *International Journal on Language, Literature and Culture in Education* 4(1). 75–88. https://doi.org/10.1515/llce-2017-0005

Vardaro, Jennifer, Moritz Schaeffer, and Silvia Hansen-Schirra. 2019. Translation Quality and Error Recognition in Professional Neural Machine Translation Post-Editing. *Informatics*, 6(3). 41. https://doi.org/10.3390/informatics6030041

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Statistical Machine Translation Output. In: *LREC*. 697–702. Accessed April 5, 2021. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.582.1701&rep=rep1&type=pdf