Date of publication MAR-29, 2021, date of current version FEB-05, 2021.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.20.1.2094

Google/Yandex Translation Detection in the Patterns Identifying System of Multilingual Texts

VLADIMIR KULIKOV, VALENTINA KULIKOVA, GULNUR YERKEBULAN
North Kazakhstan State University M. Kosybaev, 150000, 86 Pushkina St., Petropavlovsk, Kazakhstan Republic
(e-mail: qwertyra@mail.ru, v4lentina@mail.ru, erkgulnur@mail.ru)

Corresponding author: Gulnur Yerkebulan (e-mail: erkgulnur@mail.ru).

ABSTRACT The object of this work is to develop a script for evaluating the ability of online translators to translate text from one language to another. For this purpose, we used Google Translate and Yandex.Translate. Examples from English, Kazakh and Russian languages were used for the analysis of 147 news items and about 1800 sentences. The texts are taken from an Internet resource astana.gov.kz. A corpus of parallel texts for three languages has been created. We used development for the “sentence” pattern with the prospect of further development for the “text” pattern. We analyzed errors in the following categories: untranslated/omitted words, extra words, incorrect word endings, incorrect word order, punctuation errors, mutilate translation and incorrect translation. Based on the analysis of the obtained data we have concluded that it is better to do the translation of the Russian text into Kazakh or English in the YandexTranslate than in Google Translate. The developed comparison script and error analysis script are available on the Internet in open access.

KEYWORDS Google Translate, Yandex.translate, English, Russian, Kazakh, FuzzyWuzzy.

I. INTRODUCTION

IN the list of languages by number of native speakers, English is ranked 3rd (379 million of people), Russian – 7th (154 million of people), and Kazakh – 76th (12.9 million of people) [1]. According to the analysis of multilingualism, the number of people who speak three languages fluently is 13% of the world's population, 42% speak two languages [2]. In the Republic of Kazakhstan knowledge of three languages - Russian, Kazakh and English - is almost a prerequisite for career growth and higher pay. The state language of the Republic of Kazakhstan is Kazakh. Kazakh and Russian are used on equal grounds. Since 2014 all schools have started teaching English from grade 1.

The language industry is a big business. TechNavio analysts in their “Global Language Services Market 2020-2024” research forecast a market growth by 9.72 billion USD in 2020-2024, with a CAGR of 4% [3].

Thus, all these facts indicate that the issue of quality of translators is an urgent problem.

The object of this work is to develop a script for evaluating the ability of online translators to translate text from one language to another. For this purpose, we used Google Translate [4] and Yandex.Translate [5], examples from English, Kazakh and Russian languages were used for the analysis. Research papers on comparing data from online translators are available on the Internet, but they explore other language combinations and algorithms for comparing translations, and there are no errors that we found [6-9]. For our research, we chose only two online translators, since they have the ability to translate into all three languages we need. Google Trends confirms the demand for these translators [10].

Translation analysis was performed by the similar_text function [11] and token_set_ratio from the FuzzyWuzzy library [12] for the “sentence” pattern. The script works for the “sentence” pattern. It is planned to Refine the system for the “text” pattern and connect the analysis to other common online translators. The similar_text function is a PHP string function and is widely used in cases where fuzzy string comparison is necessary. At the moment, the token_set_ratio function from the FuzzyWuzzy library is also a popular PHP-solution for fuzzy string comparison [13, 14].

To perform the experiments, a corpus of parallel texts was collected in the amount of 1773 sentences in Kazakh, Russian and English languages. We used news published on the official Internet resource of the akimat of Nur-Sultan: http://astana.gov.kz/en.

Creation of a script for evaluating the quality of translation using the coefficients similar_text/token_set_ratio for different language pairs can be considered as the research contribution.

This script is required in the multilingual text pattern identification system (hereinafter - the system), which generates a hybrid translation from sentences with the highest similar_text/token_set_ratio coefficient based on Yandex and Google translations. How will the system work?

The user enters the text (for example Russian text), and selects the language to translate the text into (for example, English), then clicks “Translate” button. The system performs direct and double translation of the received text through Google Translate and Yandex.Translate. A double translation is a translation through an interim language, for example, Russian text translated into English and back from English to Russian. The similar_text/token_set_ratio coefficients show how much the original sentence changed after the double translation. The higher the coefficients, the fewer changes occurred in the text, which means that the direct translation of this sentence into English provided by online translator is of high quality.

The system calculates the similar_text/token_set_ratio coefficients for each sentence using the created script, which is described in this article. Then, based on these coefficients decides: to display the Google Translate or Yandex.Translate translations. The system displays the translation with the highest coefficient, this means that the displayed translation is more stable for double translations.

Thus, the user gets the most reliable translation, which is compiled with the help of two of the most popular online translators in Kazakhstan.

II. RESEARCH

A. TOOLS

A php-script “Text Comparison” was developed for the analysis [15].

The program is executed in the php programming language with the “utf-8 without bom” encoding. The code can be easily adjusted to perform similar analysis for many other language pairs.

In the program code we used information cleansing to text normalization.

To text normalization the lowercase line feed function mb_strtolower [16] and regular expression replacement preg_replace [17] was used. In preg_replace, $pattern='/[^ a-za-яё\d]/ui' was substituted.

Short description of the regular expression: the “u” flag indicates that the searched expression and the text use utf-8 encoding, not just Latin letters. The “I” flag does not require uppercase characters. In the spectrum “а-я” there is no “ё” symbol, so we specify it additionally, and “\d” – any number. In addition, $replacement contains a space, so we changed all characters other than letters, numbers, and spaces to a space.

- Porter stemmer.

Porter stemmer is a stemming algorithm published by Martin porter in 1980 [18]. The original version of stemmer was intended for English and was written in BCPL. Martin later created the Snowball project and, using the basic idea of the algorithm, wrote stemmers for common Indo-European languages, including Russian [19]. The algorithm does not use word bases, but removes word endings and suffixes based on the features of the language by applying a range of rules.

– similar_text php-function.

Similar_text determines the similarity of two lines using the Oliver algorithm [11]. The Function returns the percentage of two lines matching in $percent.

– token_set_ratio function from the FuzzyWuzzy library [20]. Out of the four available functions in the FuzzyWuzzy library we selected token_set_ratio. This function does not depend on the word order and their repetition, and produces the best result based on the matching of lines [12]. “Token_set_ratio=100” means 100% match of the compared lines.

For the translation of manually compiled corps of parallel texts in Kazakh, Russian and English free online translators Google Translate and Yandex.Translate were used. We also intended to use a comparison with Bing [21], but, unfortunately, this translator does not support the Kazakh language.

B. RESEARCH METHODOLOGIES

For the research we created a corpus of parallel texts in three languages, made by professional interpreter: English (En), Russian (Ru) and Kazakh (Kz). For the case we used news published on the official Internet resource of the akimat of Nur-Sultan http://astana.gov.kz/en from 26.12.2018 to 28.11.2019. Note that sometimes news was first created in Russian or Kazakh, and then translated by a professional interpreter into two other languages.

When creating the corpus, it was essential to have all three language versions of texts and the same number of sentences in each. If the number of sentences in translations did not match, the sentence in the text was omitted. There were 48 news and 591 sentences for each language, and in total 144 news and 1773 sentences compiled by a professional interpreter. Each news item in the corpus files is separated from the other by a double line. The corpus and all the research data can be downloaded here [22].

After forming the corpus, we started creating translations of sentences in accordance with Fig. 1.

Both online translators provided translations through interim languages (RuEnRu, RuKzRu) and direct translations (KzRu, EnRu).

Table 1 shows the average number of words in a sentence, the longest and the shortest by word count.

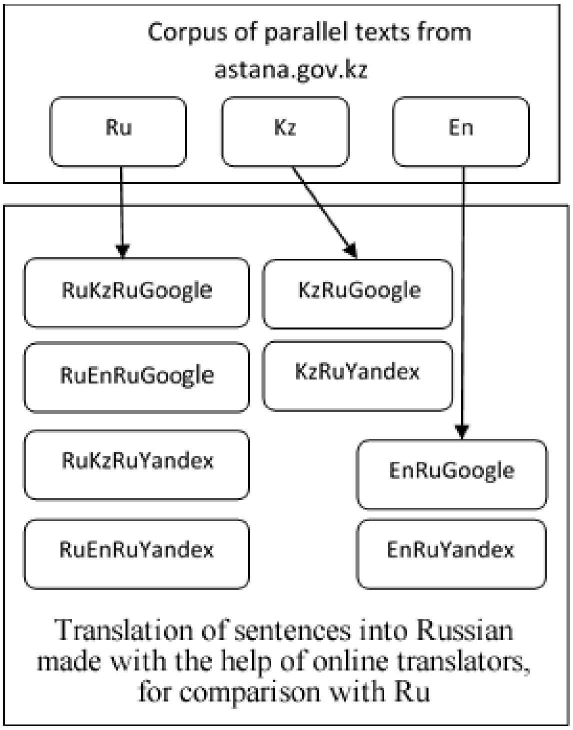

Corpus of parallel texts from
astana.gov.kz
Ru
Kz
En
RuKzRuGoogle
KzRuGoogle
RuEnRuGoogle
KzRuYandex
RuKzRuYandex
EnRuGoogle
RuEnRuYandex
EnRuYandex
Translation of sentences into Russian made with the help of online translators, for comparison with Ru

Figure 1. Translation structure for analysis

Table 1. Word Statistics

Texts	Average number of words in a sentence	Maxi-mum number of words in a sentence	Mini-mum number of words in a sentence
Ru	14.4	53	1
RuEnRuGoogle	14.7	72	1
RuKzRuGoogle	14.0	91	2
RuEnRuYandex	14.6	49	2
RuKzRuYandex	13.7	52	2
Kz	13.3	41	2
KzRuGoogle	13.6	43	2
KzRuYandex	13.5	45	2
En	18.7	67	2
EnRuGoogle	14.6	49	1
EnRuYandex	14.6	48	2

In the original corpus of parallel texts, English sentences (En) were the longest, with an average of 18.7 words per sentence, while Russian (Ru) and Kazakh (Kz) sentences were 14.4 and 13.3 words per sentence.

Double translation of Russian text shows that the average number of words in sentences in Russian Ru=4.4 increased when translated through English in both online translators (RuEnRuGoogle = 14.7 and RuEnRuYandex = 14.6) and decreased when translated through Kazakh (RuKzRuGoogle = 14.0 and RuKzRuYandex = 13.7).

To compare texts for similarities we performed text normalization. All lines were converted to lowercase. Replaced all characters other than letters, numbers, and spaces with spaces, since the following cases occurred in the case: “start/finish”, “10 km(from 16 years old)”, etc.

Then we applied the Porter stemmer for the Russian language.

Next, we performed a fuzzy line comparison using the php (similar_text) and FuzzyWuzzy functions before and after the Porter stemmer.

As a result, the php-script returned a table with the results of similar_text in % and token_set_ratio from FuzzyWuzzy for comparison pairs Ru – RuKzRuGoogle, Ru – RuKzRuYandex, Ru – RuEnRuGoogle, Ru – RuEnRuYandex, Ru – KzRuGoogle, Ru – KzRuYandex, Ru – EnRuGoogle, Ru – EnRuYandex. All data can be viewed at the link [22].

III. RESULTS

A. COMPARISON ANALYSIS OF SIMILAR_TEXT AND TOKEN_SET_RATIO

As a result of the received data, it was decided not to use Porter stemmer in the system for identifying patterns of multilingual texts. First, the script execution time using the stemmer increases it almost 2 times (from 1.98 seconds to 4.19 seconds on average). Second, Porter stemmers are not designed for all languages, which will be an obstacle when connecting other languages to the system.

To evaluate offers for translation quality, we selected three parameters: “correct translation”, “mutilate translation”, and “incorrect translation”.

“Correct translation” means that the whole sentence is translated correctly. In this sentence, you can use synonyms.

“Mutilate translation” means that the meaning of the translated sentence is slightly distorted, but generally remains the same.

“Incorrect translation” means that the meaning of the translated sentence is significantly distorted.

As the result of the error analysis we formed a table of intervals for the system for identifying patterns of multilingual texts (Table 2).

Unfortunately, the analysis shows a strong intersection of intervals and we cannot be sure that the sentence translated from one language into another using Yandex.Translate or Google Translate will relate to a specific translation quality parameter. Therefore, it was decided not to include the table in the system.

What is the most reliable indicator of translation quality? Similar_text (ST) or token_set_ratio (TSR)? To do this, the following Ru – RuEnRuGoogle comparison graphs are created for the “Incorrect translation” (Fig. 2) and “Distorted translation” (Fig. 3) parameters. A more accurate result is given by Similar_text, showing lower coefficients for these parameters.

Table 2. Table of language pair intervals

Name of the translation quality Parameter	Ru – RuKzRuGoogle		Ru – RuKzRuYandex		Ru – RuEnRuGoogle		Ru – RuEnRuYandex	
	ST	**TSR**	**ST**	**TSR**	**ST**	**TSR**	**ST**	**TSR**
correct translation	36-100	35-100	41-100	66-100	40-100	79-100	30-100	23-100
mutilate translation	31-99	35-100	44-98	55-100	28-98	35-100	35-97	63-100
Incorrect translation	13-93	5-100	19-94	3-100	35-99	54-100	24-98	47-100

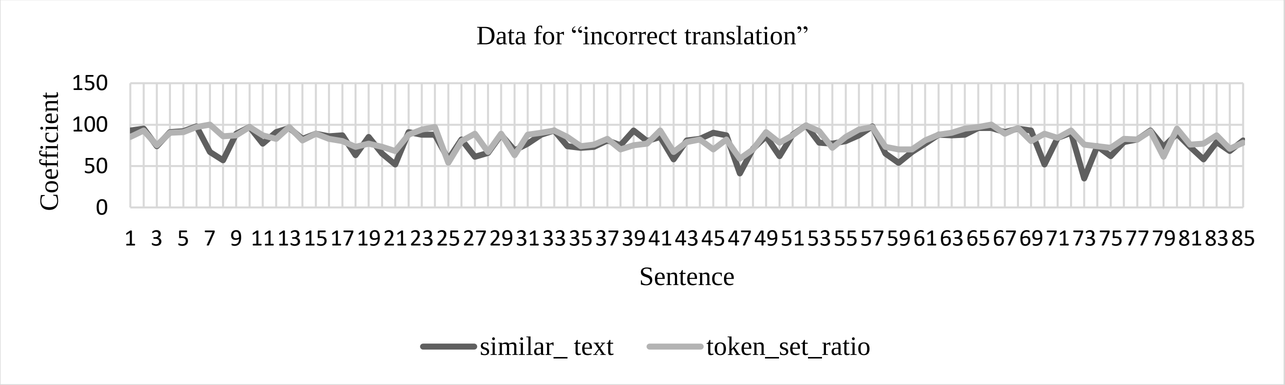


Figure 2. ST and TSR chart for “Incorrect Translation”

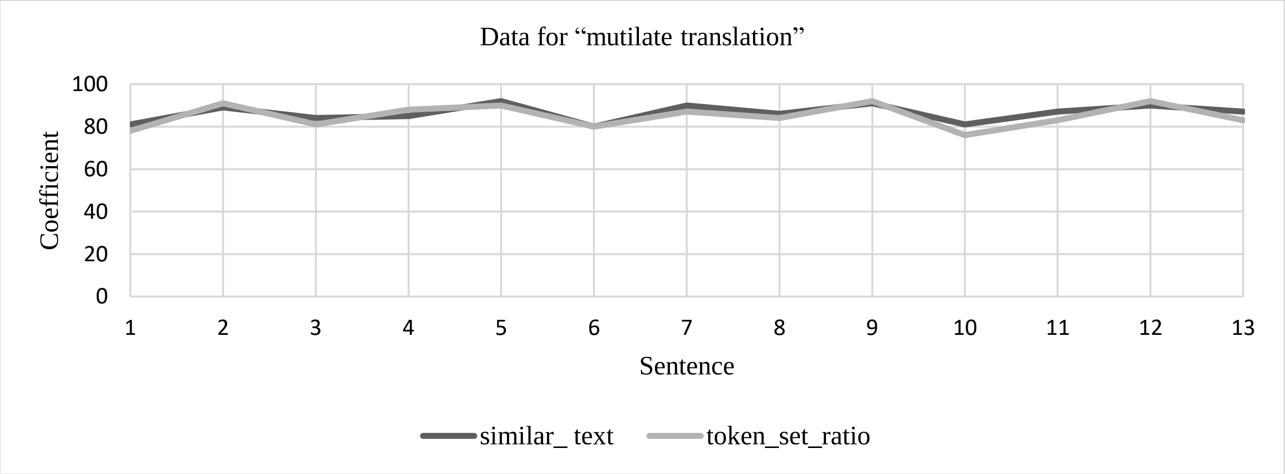


Figure 3. ST and TSR chart for “Mutilate translation”

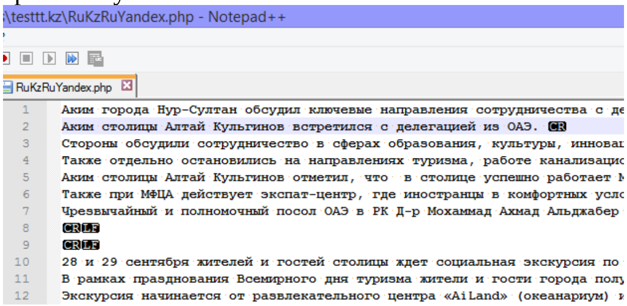
B. *ERROR ANALYSIS*

When developing a script for the system for identifying patterns of multilingual texts, the following errors were noticed.

Google Chrome browser (Version 81.0.4044.138) was used to create double translations. At the same time, it was discovered that the translation to the new line is not always saved. Also, Yandex.Translate does not take into account the line breake “cr”, but only the combination “cr” “lf” (Fig. 4). In Google Translate, all types of line breaks are taken into account [23].

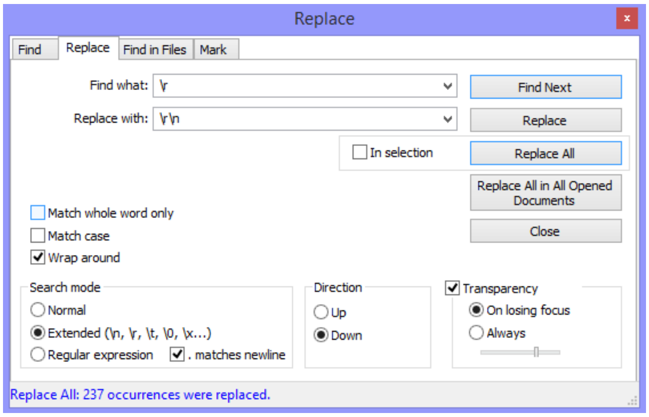
To save all line breaks after translation in Yandex.Translate following steps were performed. Replacing “\r” with “\r\n” in the Notepad++ text editor in “Advanced” search mode (Fig. 5).

Replacing “\n\n” with “\n” (Fig. 6). All news is still separated by two line breaks.



```
Аким города Нур-Султан обсудил ключевые направления сотрудничества с де
Аким столицы Алтай Кульгинов встретился с делегацией из ОАЭ. CR
Стороны обсудили сотрудничество в сферах образования, культуры, инновац
Также отдельно остановились на направлениях туризма, работе канализацио
Аким столицы Алтай Кульгинов отметил, что  в столице успешно работает М
Также при МФЦА действует экспат-центр, где иностранцы в комфортных усло
Чрезвычайный и полномочный посол ОАЭ в РК Д-р Мохаммад Ахмад Альджабер
CRLF
CRLF
28 и 29 сентября жителей и гостей столицы ждет социальная экскурсия по
В рамках празднования Всемирного дня туризма жители и гости города полу
Экскурсия начинается от развлекательного центра «AiLand» (океанариум) и
```

Figure 4. Screenshot of the Notepad++ text editor



Find what: \r

Replace with: \r\n

☐ In selection

☐ Match whole word only

☐ Match case

☑ Wrap around

Search mode

○ Normal

● Extended (\n, \r, \t, \0, \x...)

○ Regular expression ☑ . matches newline

Direction

○ Up

● Down

☑ Transparency

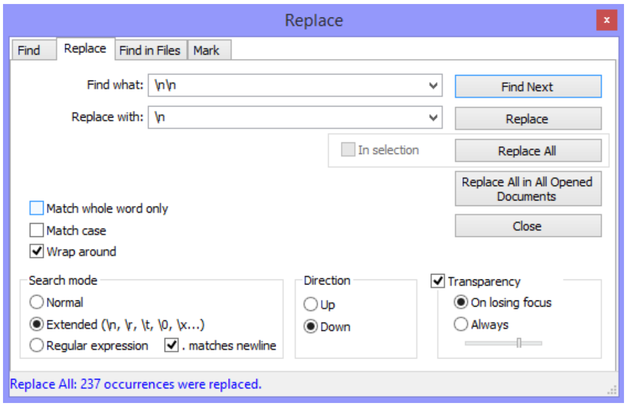
● On losing focus

○ Always

Replace All: 237 occurrences were replaced.

Figure 5. Replacing “\r” with “\r\n” in the Notepad++

Yandex.Translate does not use quotation marks correctly after the translation. For more information, see the files with the name containing the word “yandex” at the link [22].



Find what: \n\n

Replace with: \n

☐ In selection

☐ Match whole word only

☐ Match case

☑ Wrap around

Search mode

○ Normal

● Extended (\n, \r, \t, \0, \x...)

○ Regular expression ☑ . matches newline

Direction

○ Up

● Down

☑ Transparency

● On losing focus

○ Always

Replace All: 237 occurrences were replaced.

Figure 6. Replacing “\n\n” with “\n” in the Notepad++

As a result of comparing the text in Russian with double and direct translations, the following table was formed. To create the table, we used a modified script of the main comparison script, which helps to create a table for Excel [15].

Table 3. Error statistics in % to the total number of sentences

Texts	**Untrans-lated/ omitted words**	**Extra words**	**Incor-rect word endings**	**Incor-rect word order**	**Punc-tuation errors**	**Mutilate trans-lation**	**Incorrect trans-lation**	**Mutilate + Incorrect translation**
RuEnRuGoogle	17.6	14.9	19.6	19.0	25.9	**60.4**	**14.7**	**75.1**
RuKzRuGoogle	49.4	27.9	13.4	30.5	12.9	**14.4**	**63.1**	**77.5**
RuEnRuYandex	5.2	10.8	4.2	11.3	3.0	**27.4**	**12.2**	**39.6**
RuKzRuYandex	8.8	6.3	6.9	4.7	0.5	**7.6**	**7.6**	**15.2**
KzRuGoogle	53.1	43.1	17.4	52.3	6.4	**53.5**	**42.0**	**95.5**
KzRuYandex	27.4	23.5	6.3	2.5	15.1	**18.1**	**20.3**	**38.4**
EnRuGoogle	26.6	16.6	12.9	18.8	5.4	**39.1**	**12.2**	**51.3**
EnRuYandex	20.0	20.3	7.1	15.1	2.9	**38.7**	**6.8**	**45.5**

Comparing the Ru text with the received direct translations from Kz and from En (KzRuGoogle, KzRuYandex, EnRuGoogle, EnRuYandex), we can say that the ability to determine the identity of multilingual texts created by professional interpreters is low. If you add up the percentages “Mutilate translation” and “Incorrect translation”, then the similarity of the meaning of the Ru and Kz texts is easier to detect using Yandex.Translate (38.4% vs. 95.5%). At the same time, the proximity of Ru and En is almost the same in both online translators (45.5% vs. 51.3%).

Comparing Ru with double translations (RuEnRuGoogle, RuKzRuGoogle, RuEnRuYandex, RuKzRuYandex), we find the following results.

Multiple distortions of the text in Russian were detected when double-translated through Kazakh and English using Google Translate (RuKzRuGoogle=77.5% and RuEnRuGoogle=77.5%). Yandex.Translate works much better in this language groups (RuKzRuYandex=15.2% and RuEnRuYandex=39.6%).

Thus, although it is better to translate Ru to Kz or En in Yandex.Translate, when using the system, the translation subtleties can be taken into account using Similar_text calculations, which will help generate text based on more accurate translations of sentences in online translators.

Detected errors can be used to further improve the program for comparing two texts.

IV. CONCLUSION

The collected corpus of parallel texts in English, Kazakh and Russian is available for download and further use for research in the field of multilingual texts [22]. At the time of writing, we have not found a similar case on the Internet. We think this will save time for other researchers.

The program code can be used for research of many other language pairs based on the Latin or Cyrillic alphabet. If you modify the code based on the observed errors specified in clause 3.3 Error Analysis of this article, you can get more accurate results.

The main comparison script “Text Comparison” and the script for creating a table for comparing sentences in Excel are available on the Internet in open access [15]. The developed scripts are intuitively simple, so they can be scaled up by further development and used for comparison with other online translators as well.

Highlighting with a certain color depending on the similar_text coefficient will allow interpreters to quickly create translations that do not require high accuracy. They will be able to focus on sentences with a low coefficient of similar_text, while skipping sentences with a high coefficient that characterizes a more accurate translation.

References

[1] Wikipedia, *List of languages by number of native speakers*, 2020, [Online]. Available at: https://en.wikipedia.org/w/index.php?title=List_of_languages_by_number_of_native_speakers&oldid=957968997.

[2] Language Learning, *Multilingual People*, 2018, [Online]. Available at: http://ilanguages.org/bilingual.php.

[3] Research and Markets, *Global Language Services Market 2020-2024*, 2020. [Online]. Available at: https://www.researchandmarkets.com/reports/4894434/global-language-services-market-2020-2024.

[4] Google Translator, 2020. [Online]. Available at: https://translate.google.com/.

[5] Yandex.Translator, 2020. [Online]. Available at: https://translate.yandex.com/.

[6] S. Seljan, M. Tucaković, I. Dunđer, "Human evaluation of online machine translation services for English/Russian-Croatian," in: A. Rocha, A.M. Correia, S. Costanzo, L.P. Reis (Eds.), New Contributions in Information Systems and Technologies, Springer International Publishing, Cham, 2015, pp. 1089–1098. https://doi.org/10.1007/978-3-319-16486-1_108.

[7] A. Sukhoverkhov, D. DeWitt, I. Manasidi, K. Nitta, V. Krstic, "Lost in machine translation: Contextual linguistic uncertainty," *Science Journal of VolSU. Linguistics*, vol. 18, pp. 129–144, 2019. https://doi.org/10.15688/jvolsu2.2019.4.10.

[8] Z. Bülbül, A. Çetinkaya, and F. Arıcı, *Google Translate and Yandex Translate's Differences in Naturalness, Clarity, and Accuracy: A Comparison Study on Machine Translation*, 2020, [Online]. Available at: https://www.researchgate.net/publication/339029502_Google_Translate_and_Yandex_Translate's_Differences_in_Naturalness_Clarity_and_Accuracy_A_Comparison_Study_on_Machine_Translation.

[9] O. Mohammed, S. Samad, "Machine translation strategies of translating death euphemistic expressions from Arabic into English and vice versa," *An International Peer-Reviewed Open Access Journal*, pp. 114-121, 2020.

[10] Google Trends, *Comparison*, 2020. [Online]. Available at: https://trends.google.com/trends/explore?date=all&geo=KZ&q=%2Fm%2F025sndk,%2Fg%2F11x1nzgtw,%2Fm%2F02z9kkt.

[11] PHP, *similar_text – Manual*, 2020. [Online]. Available at: https://www.php.net/manual/en/function.similar-text.php.

[12] ChairNerd, *FuzzyWuzzy: Fuzzy String Matching in Python*, 2011. [Online]. Available at: https://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/.

[13] R.S. Sandhu, J. Shin, K.C. Wang, and G. Shih, "Single-center experience implementing the LOINC-RSNA radiology playbook for adult Abdomen/Pelvis CT and MR procedures using a semi-automated method," *Journal of Digital Imaging*, vol. 31, pp. 124–132, 2018. https://doi.org/10.1007/s10278-017-0016-0.

[14] P. Kanani and Dr. M. Padole, "Improving pattern matching performance in genome sequences using run length encoding in distributed Raspberry Pi clustering environment," *Procedia Computer Science*, vol. 171, pp. 1670–1679, 2020. https://doi.org/10.1016/j.procs.2020.04.179.

[15] G. Yerkebulan, *Scripts developed to compare Google translate and Yandex.Translator*, 2020. [Online]. Available at: http://102030.kz/works.php.

[16] PHP, *mb_strtolower – Manual*, 2020. [Online]. Available at: https://www.php.net/manual/en/function.mb-strtolower.php.

[17] PHP, *preg_replace – Manual*, 2020. [Online]. Available at: https://www.php.net/manual/en/function.preg-replace.php.

[18] M. Porter, *Porter Stemming Algorithm*, 2006. [Online]. Available at: https://tartarus.org/martin/PorterStemmer/

[19] M. Porter, *Russian Stemming Algorithm*, 2020. [Online]. Available at: http://snowball.tartarus.org/algorithms/russian/stemmer.html.

[20] Wyndow, *fuzzywuzzy*, 2017. [Online]. Available at: https://github.com/wyndow/fuzzywuzzy.

[21] Bing Microsoft Translator, 2020. [Online]. Available at: https://www.bing.com/translator.

[22] G. Yerkebulan, *Yandex and Google Translate Compare – Google Disk*, 2020. [Online]. Available at: https://drive.google.com/drive/folders/1tPI42nCbaNZvlggnxclkf1vQoQS0ecgk?usp=sharing.

[23] Wikipedia, *Newline*, 2020. [Online]. Available at: https://en.wikipedia.org/w/index.php?title=Newline&oldid=957966639.

***VLADIMIR KULIKOV,** Candidate of Physical Mathematical Sciences. Associate Professor of "Information and Communication Technologies" Department at the Manash Kozybayev North Kazakhstan State University. Graduated from the Lomonosov Moscow State University in 1982, applied mechanics. Vladimir Kulikov received his degree in 1985. Main areas of research interests: analysis and modeling of socio-information processes of organizations; methods and technologies of informatization of organizational and managerial decisions in business process optimization; mathematical models of management processes, recommendations on the use of information and communication technologies in institutions in order to improve the efficiency of their main activities.*

***VALENTINA KULIKOVA,** Candidate of Engineering Sciences. Associate Professor of "Information and Communication Technologies" Department at the Manash Kozybayev North Kazakhstan State University. Graduated from the Lomonosov Moscow State University in 1982, mathematician (discrete mathematics). Received her Ph.D. in engineering from the L. Gumilyov Eurasian National University in 2003. Main areas of research interests: scientific and innovative potential of the education system in the development of the intellectual nation: problems of methodology and assessment; mathematical support for information systems that support organizational and managerial decision-making, resource allocation and optimization, impact forecasting.*

***GULNUR YERKEBULAN,** doctoral candidate at the Manash Kozybayev North Kazakhstan State University majored in Computer Science, Computer Technology and Management. Graduated Kanysh Satpayev Kazakh National Technical University in 2008 with a degree in "Organization and Technology of Data Protection". Main areas of research interests: modern online technologies for task automation, web development, chatbots.*