Automated Language Translation: Opportunities and Impact on the Society

S. T. Nandasara
University of Colombo School of Computing
Sri Lanka

Yoshiki Mikami
Nagaoka University of Technology
Niigata, Japan

AIC. Mohideen
University of Aberdeen
School of Engineering
Aberdeen, UK

K. G. D. Tharangie
Gradhat, UK

ABSTRACT
Language is a system of communication that links different multilingual societies. Machine language translation techniques have improved productivity and quality in translation, online communication, online business and trade. It also demonstrates the integral need for innovative technological solutions to the age-old difficulties of the digital divide due to the language barriers. The language translation plays an essential role in crossing through different cultures and communication channels. This paper presents the leading roles of automated translation in propagating important social ideas between two or more languages, and examines the difficulties and opportunities that translation techniques face in the process.

Further, this paper also aims to present critical and unintended social issues triggered by the process of automated translation. An acceptable automated translator should be aware of the cultural factors, simultaneously, customs and traditions, and consider the chronological orders, specific meaning, development of related disciplines, and historical and religious sensitivity of the content. Further, it is essential to evoke the same response as the source text attempted to and avoid inserting irrelevant new words or essence into the language used by people. This paper emphasises the need for the consideration of all these factors into account in the translating process.

General Terms
Language Translation

Keywords
Translation quality, consistency, translation tools, Google Translate, Facebook Translator

1. INTRODUCTION
In the constant development of humanity, language translation has always played a crucial role, especially in digital communication, by allowing for the sharing of knowledge and culture between different languages. An ample of the wealth of knowledge and richness of experience that is constructed and documented exist in our societies, however, confined within language silos, to which access is restricted for most of us, even with our favourite Internet search engines.

In profiling characteristic of Internet users versus world population of that language in 2019, the available data reveal that the number of English-speaking users, at 1,105 million (25.2%), followed by Chinese-speakers, at 863 million (19.3%) and then drops to 344 million (7.9%) for Spanish users—in a total user base of 4.3 billion—see Figure 1 and Figure 2 (Internet World Stats, April, 2019; W3Techs, March, 2018). Further, according to the Internet World Stats-2019, out of the estimated 97 million individuals in the world that speak German, 95.1% are internet users, out of the estimated 126 million individuals speak Japanese, 93.5% are internet users, out of the estimated 143 million persons speak Russian, 76.1% are internet users and out of the estimated 1,485 million persons speak English, 74.5% are internet users. The content available on the internet to these users, English leads at 54%, with an immediate plunge to Russian at 6%, German at 5.9%, Spanish at 4.9%, French at4% and Chinese at 1.7% which is much less compared to the 3% in 2015. Similarly, English content is at 2% less compared to the 2015 statistics (Internet World Stats, 2015).

The statistics on the growth rates of the use of languages also provide more insights into the internet users. The growth rate in the number of English-speaking users has continued steadily at a rate of about 685.7% from 2000 to 2019, and it is overshadowed greatly by other global languages. Arabic grew by 8,917.3%, and Russian and Chinese grew by 3,434.0% and 2,572.3%, respectively, with other languages showing considerable growth in the same period—for example, Portuguese at 2,164.8%, and Indonesian/Malaysian at 2,861.4% and Chinese at 2,572.3% (Internet World Stats, 2019; W3Techs, 2018). This trend mirrors the composition of the automatic language translation domain during the same period.

Analysts from the translation industry report [3] indicate that only a tiny amount of digital content, less than 0.1%, is currently being translated. New technologies have been looked at to provide solutions to translate this explosion of content that traditional human translation processes cannot manage.

This paper demonstrates how language translation techniques are driven by two primary automated machine translation tools., have fundamentally changed how humen communicate today. The development of translation tools and the associated positive and negative consequences, situated within the context of a fast-changing internet field are also discussed. Thus, this paper critically reviews how translators now translate (process); what is being translated (product); and how the role of the translation technique has diversified to include various professional specialisations and technical competencies as well as everyday users (society).

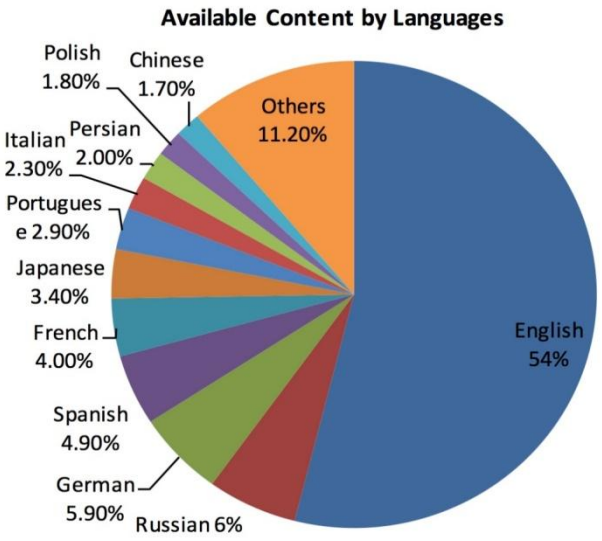

Fig 1. Internet content available by Language

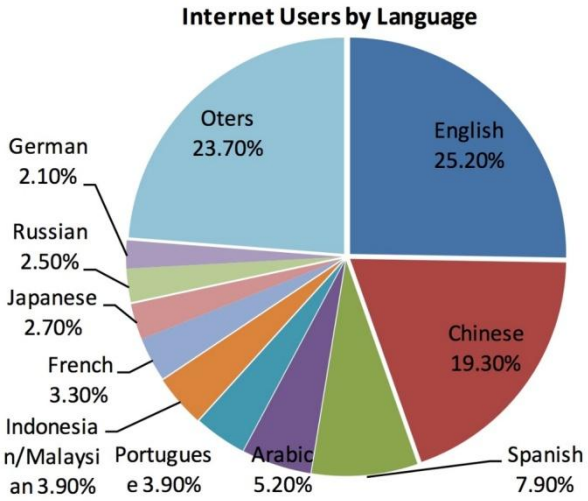

Fig 2. Internet users by Language

2. RESEARCH METHODOLOGY

The process of this study consists of two major phases. The first phase is a translation experiment between Russian to English through fourteen different languages (intermediate). One phrase of Russian statement was selected and translated into fourteen other languages using Google Translate (GT) [1] and Facebook Translator (FT) [2]. Then the result of translation is translated back to English. The intermediate translation and English translation are assessed with a manual evaluation across morphological errors of words and sense relationships. This study decided to choose intermediate translation to be culturally different languages as they are all available on Google and Facebook. The second part is a semantic and lexical study of Google and Facebook translation, with a set of comprehensive articles that required for a good machine translation, and each translation services examined.

This study is divided into three stages:

Stage-1: A study on real-world examples of Sinhala and English posts translation on Facebook is considered and its impact on the society is investigated.

Stage-2: Compare the morphological error testing with Google Translate and Facebook Translators with a selected short phrase.

Stage-3: The results generated by Google and Facebook Translators on thirty (30) articles of UDHR (The Universal Declaration of Human Rights) analysed with manual evaluation method will be used to evaluate lexical and semantic errors.

Under manual evaluation analysis results generated by Google and Facebook contain errors such as Missing Words, Quality of Translation and Word Order.

2.1 Why Language Translation is Important to Sri Lanka?

Sri Lanka is home to many ethnicities – Sinhalese, Tamil, Moors/Muslims, Burghers/Eurasians, Veddahs, and Malays - which add to the dynamics of the languages spoken in the country. The colonial powers of Portuguese, Dutch and English, have had significant influences in the development of Sri Lanka's languages.

Sri Lanka is a multiracial society comprising approximately 75% of Sinhala speaking population and around 15% of the Tamil-speaking population. Tamil and Sinhala languages are Sri Lanka's two official languages, and English is termed as the link language. All these languages are widely spoken throughout the country by 21.6 million population. English is preferred in governmental policies and practices, and all these documents are available in all three languages, including information available in most of the government websites. In July 1996, Sri Lanka launched its first National Website (http://www.lk), initially consisting of information entirely in the English language and from 1997 (http;//www.gov.lk) it supports all three languages, Sinhala, Tamil and English [4].

2.2 Languages used in Sri Lanka

Sinhala is the official language of administration of Sri Lanka throughout most of its history under the Sinhala kingdom. The orders were communicated to the public in Sinhala even in the north, and then it was part of the ancient Rajarata. P. E. Pieris says [5] the pact signed by the Jaffna ruler, Cankili I (King of the Jaffna Kingdom, 1519-61) with the Portuguese, in 1560, was in Portuguese and Sinhala, not Tamil. During the Dutch period, Dutch have administered Sri Lanka in the Sinhala language. At independence in 1948, Sinhala was spoken by over two-thirds of the population.

Tamil language, in Sri Lanka, there are approximately 4.7 million Tamil speakers. Tamil is Sri Lanka's second official language since 1972, which is about 15% of the population.

3. MACHINE TRANSLATION TOOLS

The field of machine translation (MT) has a long and turbulent history. The Georgetown-IBM experiment of January 1954 [6], the first time in effecting machine translation from Russian into English on a limited basis gave, much encouragement to research in the field. As a result, in 1956, the Institute of Precision Mechanics and Computer Technology of the U.S.S.R. announced the successful performance of translation of English into Russian on their BESM ("Bolshaya Elektronno-Schetnaya Mashina") computer and acknowledged the relationship between their undertaking and the Georgetown-IBM experiment.

The task of MT defined [7] the way computer must be able to obtain an input as a text in one language (SL, source

language) and produce as an output a text in another language (TL, target language) so that the meaning of the TL text is the same as that of the SL text. However, the real story behind the genesis of machine translation is that the transference of meaning from one patterned set of signs occurring in a given culture into another set of patterned signs occurring in another related culture [8].

3.1 Error Typology Employed in this Study

As discussed by Carbonell, J. G., Cullingford, R, E. and Gershman A. G [9], it is clear that finding a way of maintaining invariance of meaning is the crucial problem in MT research. There are multiple dimensions of 'quality' in the translation process, to wit:

- **Semantic invariance:** Preserving invariant the meaning of the source text as it transformed into a target text.
- **Pragmatic invariance:** Preserving the implicit intent or illocutionary force of an utterance.
- **Structural invariance:** Preserving as far as possible the syntactic structure of the text under translation.
- **Lexical invariance:** Preserving a one-to-one mapping of words or phrases from source to the target text.
- **Spatial invariance:** Preserving the external characteristics of the text, such as its length and location on the page.

Nevertheless, early MT systems sought to preserve lexical invariance in the hope that all other invariance would follow; modern approaches take a somewhat more realistic view.

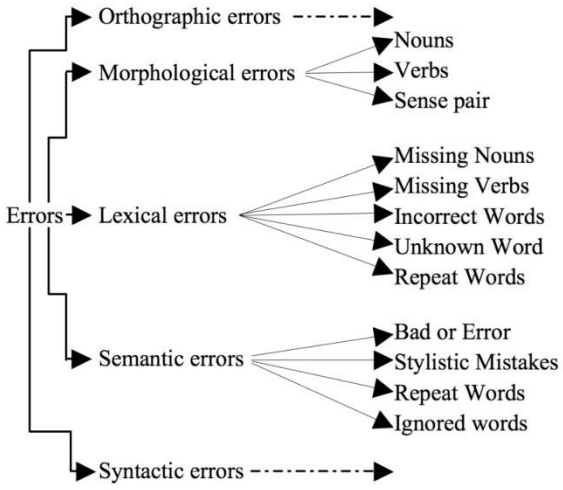

Fig 3. Classification of translation errors

The other classification scheme employed by Farrús Cabeceran et al. [10] in 2010, linguistic-based typologies tend to offer more information about the types of errors found. In this classification, at the first level, errors are split into five major categories: orthographic errors, morphological errors, lexical errors, semantic errors, and syntactic errors. The current schemes borrow with a few improvements, and this paper considered adjustments the linguistic-based categories, mainly, morphological errors, lexical errors and semantic errors at the first level while having subcategories that are more suitable for the English to Sinhala language pair. It has a hierarchical structure, as shown in Fig. 3.

3.2 Machine Translation Techniques

On the practical side, of course, available computing hardware is much more powerful than one could reasonably have expected in the 1960s, and improvements in memory size and processing speed continue to make. In the 1990s, recognising the need to translate marketable products to be successful in international markets, software companies, and several other technology-related industries, sought a way to increase productivity in translation. As a result of this, computer-assisted translation (CAT) tools provided the major technological shift in the present-day translation industry [11].

Statistical machine translation (SMT) learns how to translate by analysing existing human translations (known as bilingual text corpora). In contrast to the CAT approach that is usually word based, most modern SMT systems are phrased based and assemble translations using overlap phrases. In phrase-based translation, the aim is to reduce the restrictions of word-based translation by translating whole sequences of words, where the lengths may differ [12]. Although useful, SMT methods suffered from a narrow focus on the phrases translated, losing the broader nature of the target text. The hard focus on data-driven approaches also meant that methods might have ignored important syntax distinctions known by linguists. Finally, the statistical approaches required careful tuning of each module in the translation pipeline. Translation tools, such as Google Translate, have traditionally been built around phrase-based statistical machine translation. However, its effectiveness depends much on the quality of the original language samples, and it is prone to mistakes. For these reasons, in 2016, Alan Packer, director of engineering language technology at Facebook, said on BBC, that "statistical machine translation was reaching -the end of its natural life-" [13]. Instead, translation technology is now moving towards Neural Machine Translation (NMT).

Neural Machine Translation (NMT) is an end-to-end learning approach for automated translation, with the potential to overcome many of the weaknesses of conventional phrase-based translation systems. These are structured similarly to the human brain and use complex algorithms to select and use the appropriate translation. However, rather than translate the words, a neural network can learn metaphors and the meaning behind the language, allowing it to select a translation that means the same thing to a different culture, rather than a direct literal translation which may in some cases offend. Unfortunately, NMT systems are known to be computationally expensive both in training and in translation inference, not every researcher or institute can afford.

Although, the key benefit to the approach is that a single system can be trained directly on the source and target text — the pipeline of specialised systems used in statistical machine translation. However, the traditional phrase-based translation system which consists of many small sub-components that are tuned separately, NMT attempts to build and train a single, extensive neural network that reads a sentence and outputs a correct translation. Furthermore, a neural network can learn the meaning behind the language and allowing it to translate to their own culture, rather than a direct literal translation.

3.3 Online Language Translation Tools

According to Seljan [14], at present, the online machine translation systems and tools underwent accepting new information and communication knowledge and skills, as well as adopting the usage of modern multilingual technologies,

Over the last ten years, the United Nations Educational, Scientific and Cultural Organization (UNESCO), the United Nation (UN) and other international organization such as moreover, the European Union (EU) have been intensively thinking with the inherent problems of a multilingual environment, due to the digital divide, which is a demanding and ambitious project. Language translations have to be unambiguous and terminologically consistent. Such unambiguity can only be achieved through the consistent and synchronized use of language terminology databases and other translation tools. In the next section, this paper will present the freely and widely available language translation tools today using the most advanced techniques.

3.3.1 Google Translate

Google Translate (GT) is a free text translation service developed by Google. Google Translate can translate to and from over 100 languages, including Sinhala. It has included a "detect language" feature, which means, language identification generally refers to a process that attempts to classify a text in a language to one in a pre-defined set of known languages. It is an essential technique for Natural Language Processing (NLP), especially in manipulating and classifying text according to language [15] [16] [17]. Google Translate was launched in 2006 and gathered linguistic data from UN and EU official documents. Neural machine translation engine was adopted in 2016, and the service uses Artificial Intelligence (AI) to translate sentences at a time. GT can pronounce some translated words, highlight similar words in the source text and translated text, and act as a single-word dictionary.

3.3.2 Microsoft Bing Translator

Microsoft Bing Translator (BT) is a cloud service that translates between more than 60 languages and uses an automatic translation engine that employs machine learning to generate statistical translation models. In addition to the powering Bing translation for Search, it powers translations in Microsoft products such as Microsoft Office, Yammer, Skype Translator, Internet Explorer, and many others. Bing Translator requires a significant amount of high-quality translation text, typically over one million words, to build a translation system for the language. Microsoft Bing translator is yet to support the Sinhala Language.

3.3.3 Facebook Translator

Facebook Translator (FT) introduced a new translation tool and methodology that allowed its users to perform translation of the site into users' native languages. The Facebook translation tool works by asking the users to submit possible translations of phrases and then soliciting their votes on the most accurate translation. Facebook human-powered approach juxtaposes quite sharply with Google's service, which uses technology to automatically translate Web sites and text — with occasional unintentionally comical results. The Facebook tool, of course, has had to handle a relatively small number of phrases, and Facebook currently translates over 100 unique languages, including Sinhala.

3.4 Real-World Examples: Social Media

Present days, Facebook has become one of the most prominent social media application; it has opened up the translation application to everyone to translate Facebook into their particular native language. Consequently, Facebook has also been translated into Sinhala, which is helpful to users who are not competent in English. However, Fig. 4 below shows that the first line of Sinhala text translated into English as "Where is the picture of the underwear ?". However, there is no related word for "underwear" in Sinhala phrase, but the word "යට" ('*yata'* meaning 'below') followed by "පිංතූරය" ('*pinthuraya'* meaning 'picture') translated into 'underwear' by Facebook. Similarly, the next two lines of Sinhala phrase translated as "The Muslim of the top." and "Sinhalese in the underwear.". However, the meaning of these Sinhala phrases is to be like "Muslims are in the top picture" and "Sinhalese are in the below picture" respectively.

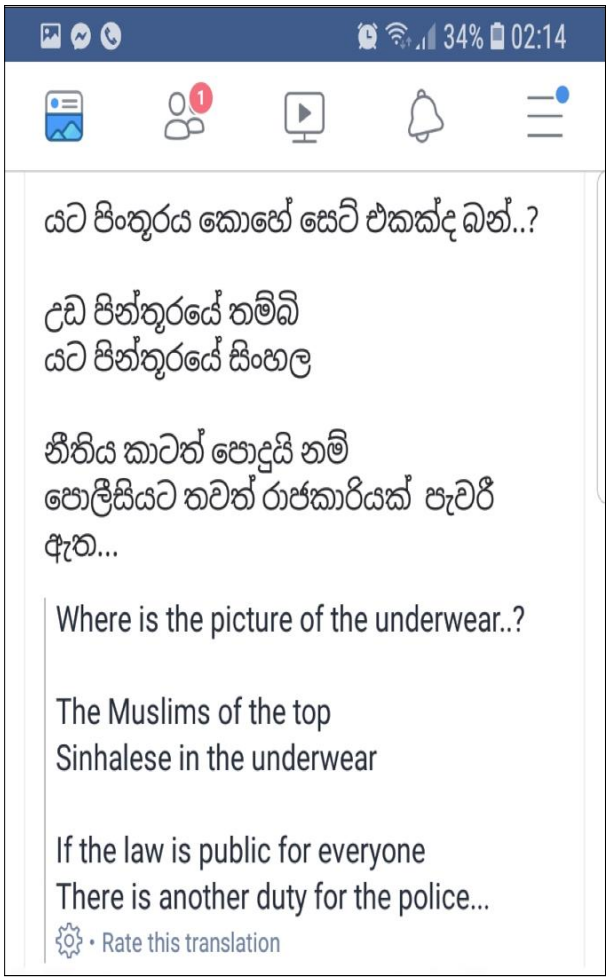

Fig. 4: Contovesale translation on Facebook (Example 1)

Fig. 5 and Fig. 6, below, are other controversial posts that appeared on Facebook recently, and moreover, the meaning of the translated English phrase in Fig. 5 renders a completely opposite meaning. In Fig. 5, a comment by the users, it says, "Facebook users should seriously consider switching off automatic English-Sinhala translation".

Three weeks after, on 13th May 2019, the Easter Sunday terrorist attack on Roman Catholic churches and top level hotels in Sri Lanka, the English post appeared on Facebook posted by a Muslim shopkeeper, had written "Don't laugh more, 1 day u will cry" and that was translated to Sinhala as " සිනාසෙන්න එපා, දවසක් අඬන්න" (see Fig. 6). Sri Lanka's authorities blocked the social media, including Facebook, after a post, local Christians took the post as a warning of a next attack. A little while later, a violent mob smashed his shop and vandalised a mosque nearby, and this situation spread quickly to two other provinces — a curfew was imposed until dawn on that Monday.

Fig. 5: Contovesale translation on Facebook (Example 2)

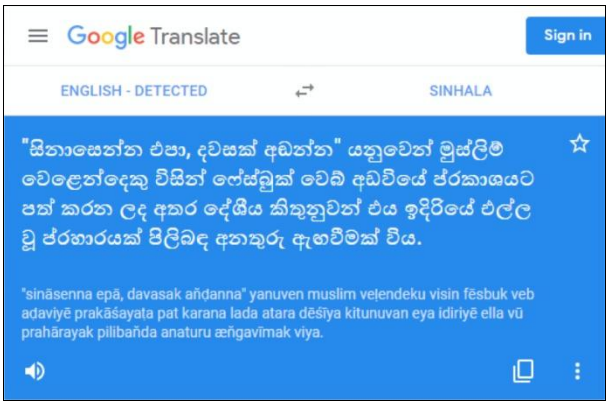

Fig 6: Sinhala translation of "Don't laugh more, 1 day u will cry"

4. TRANSLATION EVALUATION

Over the last fifteen years or so, information and communication technology has developed rapidly in the cultural and social sector, mainly in areas such as globalisation, internationalisation, localisation, and translation. Due to EU encouragement and the use of the English language as lingua franca on one side, and the interest for the protection of national cultures and identities on the other side, the development of multilingual tools and services play a crucial role in communication. At present day, language translation tools work not only translating from English to another language; it also translates from any to any language. The following tables, Table 1 and Table 2, demonstrate the translated phrase from Russian as a source language to the individual fourteen target languages using GT and FT. The fourteen target languages covered by Roman & Latin languages, Indic languages, Arabic languages, and the Ideographic languages. The 'phrase' is the name of a series of Soviet mainframe computers built in 1950–60s. The name BESM is an acronym for "Большая Электронно-Счётная Машина" ("Bolshaya Elektronno-Schetnaya Mashina"), literally "Large Electronic Computing Machine."

4.1 Morphological Error Testing

As the first step of comparison between GT and FT, this section discusses the semantic problems of the machine translation program has faced while translating the texts. The meaning of a word is for the most part based on its sense relationships towards other words surrounding it in a semantic field or by the 'role' it fulfils within the action described within a sentence. These sense relationships, which are a source of translation problems as there will be hardly similar sense relationships between words of different languages even though they have the same meaning [18][19]. Some semantic-translation problems are related to ambiguity when SL item has a mainly restricted range of meaning that it may not be possible to match this restriction in the TL, while others are related to collocations. The process, as shown in Fig. 7, Russian phrase (as SL) translated to fourteen languages (three letter language codes based on the ISO 639-3 standards) as TL, and then those translated phrases of related language phrase use as SL and translated to English. Table 1 and Table 2 illustrates the translated phrases to relevant languages. The next column of Table 1, illustrates the acceptability of the translated phrase as SL to English as TL using GT and FT, and marked as "acceptable or not acceptable" in the column with the tick marks.

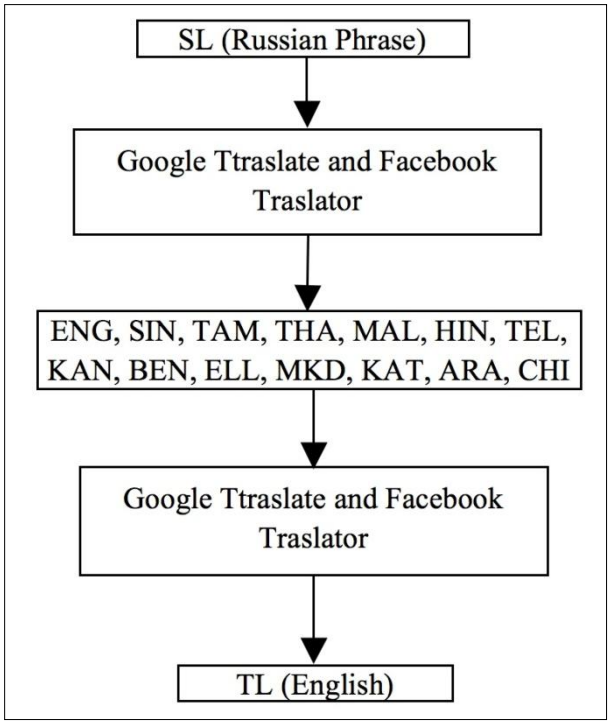

Fig. 7 Translation Process of Russian to English through an intermediate Languages

According to Table 1 below, the Russian phrase "Большая Электронно Счётная Машина" has been translated into several languages, and consider by comparing this with the human translated English phrase "Large Electronic Computing Machine." The word "Computing" has been translated to a sense relations word of the "Counting" by GT

for the fourteen languages and translated English phrases read as "Large Electronic Counting Machine" for all the languages.

Table 1. Google translation from Russian phrase to the respective languages

Language	Translated Phrase by GT (1st June 2019)	
Russian	Большая Электронно Счётная Машина	✓
English	Large Electronic Counting Machine	-
Sinhala	විශාල ඉලෙක්ට්‍රෝනික ගණන් කිරීමේ යන්ත්‍රය	✓
Tamil	பெரிய மின்னணு எண்ணும் இயந்திரம்	✓
Thai	เครื่องนับอิเล็กทรอนิกส์ขนาดใหญ่	✓
Malayalam	വലിയ ഇലക്ട്രോണിക് കൗണ്ടിംഗ് മെഷീൻ	✓
Hindi	बड़ी इलेक्ट्रॉनिक गिनती मशीन	✓
Telugu	పెద్ద ఎలక్ట్రానిక్ కౌంటింగ్ మెషిన్	✓
Kannada	ದೊಡ್ಡ ಎಲೆಕ್ಟ್ರಾನಿಕ್ ಲೆಕ್ಕಪರಿಶೋಧಕ ಯಂತ್ರ	✓
Bengali	বড় বৈদ্যুতনি গণনা মশেনি	✓
Greek	Μεγάλο ηλεκτρονικό μηχάνημα καταμέτρησης	✓
Macedonian	Голема електронска машина за броење	✓
Georgian	დიდი ელექტრონული დათვლის მანქანა	✓
Arabic	آلة عد إلكترونية كبيرة	✓
Chinese	大型電子計數機	✓

Table 2. Facebook translation from Russian phrase to the respective languages

Language	Translated Phrase by FB (1st June 2019)	
Russian	Большая Электронно Счётная Машина	Large
English	Large electronic counter machine	Large
Sinhala	විශාල ඉලෙක්ට්‍රොනික ප්‍රති යන්ත්‍රය	Big
Tamil	பெரிய மின்னணு கவுண்டர் இயந்திரம்	A Large
Thai	เครื่องคอมพิวเตอร์อิเล็กทรอนิกส์ขนาดใหญ่	Big
Malayalam	വലിയ ഇലക്ട്രോണിക് കൗണ്ട മെഷീൻ	Big
Hindi	बड़ी इलेक्ट्रॉनिक काउंटर मशीन	The Big
Telugu	పెద్ద ఎలక్ట్రానిక్ కౌంటర్ యంత్రం	Big
Kannada	ದೊಡ್ಡ ಪದ್ಮನ್ಮಾನ ಕೌಂಟರ್ ಯಂತ್ರ ದೊಡ್ಡ ಎಲೆಕ್ಟ್ರಾನಿಕ್ ಕೌಂಟರ್ ಯಂತ್ರ	Big
Bengali	বড় ইলকেট্রনকি কাউন্টার মশেনি	Great
Greek	Μεγάλη ηλεκτρονική μηχανή μετρητή	Big
Macedonian	голема електронска контра машина	Big
Georgian	Not Available	-
Arabic	الة مضادة الكترونية كبيرة	Big
Chinese	大型電子櫃台	Large

The Table 2 above, shows the Russian phrase "Большая Электронно Счётная Машина", has been translated to individual languages by FB, and the word "Счётная" (literal meaning "Computing") translated to the word "Counter" in English, except for Bengali. The Bengali translated the phrase to English and read as "Great electronic meter machine". The two words "Great" and "meter" are the words translated by FB for Bengali through the process shown in Fig. 7. In the case relevant to the Sinhala language, "ප්‍රති යන්ත්‍රය" (*prathi yantraya*) has no meaning with the English translation, i.e., "Computing Machine". The Russian phrase "Большая Электронно Счётная Машина" has been translated by FB to Chinese as "大型電子櫃台", and then the Chinese phrase translation to English by FB read as "Large E-Counter", which has no real and relevant meaning. As far as the words "Электронно" and "Машина" in the Russian language is concerned, it has translated to relevant languages and those words translated to English and read as "electronic" and "machine". However, the first word "Большая" (Lateral meaning is "Large") of the phrase in Russian has been translated to English using FB through the process as described in Fig. 7. The last column of Table 2 has given five different words as a result. Although the meaning is similar for all five; they are not the most suitable word to get the quality and correct translations.

4.2 Classification Error Testing

The previous stage translation is used to check Russian phrase translation into fourteen languages, and then those language phrases translated into English and evaluated the morphological errors and sense translation quality. After that, the study examined the real world scenario with facebook translation with few samples. In this stage, this study considered the broader evaluation based on the lexical and semantic invariance through real-world samples.

4.2.1 Lexical Errors

The error classification scheme used in this accuracy testing derives from the model proposed by Farrús Cabeceran [10]. He suggests a classification scheme with a precise linguistic categorisation at the first level: orthographic, morphological, lexical, semantic, and syntactic errors. At the lexical invariance level, the study have split the errors into five big classes: "Missing Nouns", "Missing verbs", "Incorrect Words", "Unknown Words" and "Repeat words". Next, study examined the meaning of the translation in semantic invariance level. Therefore, phrase or sentence errors have grouped according to the number of errors in morphological and lexical level is "Bad" if the number of errors >= 3, is considered as "Moderate" if the number of errors> 0 and < 3, and is considered "Good" if the number of errors equal to zero.

For our evaluation, sample sentences are collected from the historical document known as "The Universal Declaration of Human Rights (UDHR)" [20]. This document consists of thirty articles and by now, this document is translated into more than 500 languages by human translators. The articles include 64 sentences and 1371 words, and human translated UDHR Sinhala document [21] too consists of 64 sentences and 1234 words. This human translated Sinhala document used as a source for error checking.

Each English sentence is translated into Sinhala using Google Translate and Facebook Translator, and the translated document has 1167 and 1078 words, respectively. Every 64 sets of sentences in the 30 articles of the UDHR is checked carefully with translated sentences from both Google and Facebook translator and scored for each sentence with the criteria discussed above, and the results given in the following Tables 3, 4 and 5.

There were a total of 39 lexical errors identified in GT out of 1167 words and 191 lexical errors identified out of 1078 words in FT translation during this study and result shown in Table 3. Table 4 shows the semantic errors, and they consist of 10 and 80 semantic errors for all four categories for GT and FT, respectively.

Further, Table 3 below details the distribution of errors among all subcategories. Among the top five lexical categories, were found to have the highest error counts for FB: missing nouns (46, 4.27%), missing verbs (57, 5.29%), incorrect words (45, 4.17%), unknown words (20, 1.86%), and repeat words (17, 1.58%) errors. This work compares the error rate of FT with GT. GT's error rate is below 1.20%, which means that the highest error count is 14 for wrong words for GT. Total errors occurred for the lexical category during the translation processes Google Translate made a much lower error rate (39, 3.35%) compares with the 191 errors and 17.72% error rate of

FT (191, 17.72%). Fig. 8 below shows the distribution of errors in each lexical category.

Table 3. Numbers of errors in all sub-categories of Lexical category and their respective percentages

Main Category	Sub Category	Error Count of the Words		Percentage of the Errors (%)	
		GT	FT	GT	FT
Lexical	Missing Nouns	12	46	1.03	4.27
	Missing Verbs	12	57	1.03	5.29
	Incorrect Words	14	45	1.20	4.17
	Unknown Word	0	20	0.00	1.86
	Repeat Words	1	23	0.09	2.13
	Total Word Errors	**39**	**191**	**3.35**	**17.72**

As shown in Fig. 8 below, it is clear that two subcategories, unknown words and repeated words in lexical errors are accounted zero and one for Google Translate (0.00% and 0.09% respectively), and comparing this with Facebook Translator error rates is much higher (1.86% and 2.13%) than Google Translate.

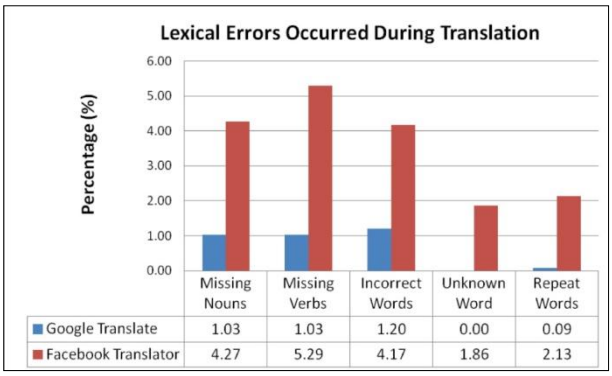

Fig. 8: Lexical error distribution in the five subcategories

4.2.2 Semantic Errors

The subcategory of Semantic Errors that accounted for errors was Bad or Error (1.11%, 20.00%), followed by Stylistic Mistakes (10.00%, 38.89%), Repeat Word (0.00%, 14.44%), Ignored Words (0.00%, 15.56%) and accordingly, Total Semantic Errors (11.11%, 88.89%) errors (see Table 4). The distribution of errors in each semantic category is shown in Fig. 9. Stylistic mistakes rate is the highest among the other errors in the subcategory for both GT and FT and the zero level error rate occurred in GT for repeat words and ignored word is notable.

Table 4. Numbers of errors in all subcategories of Semantic category and their respective percentages

Semantic Errors	Sub Category	Error Count of the Words		Percentage of the Errors (%)	
		GT	FT	GT	FT
Semantic	Bad or Error	1	18	1.11	20.00
	Stylistic Mistakes	9	35	10.00	38.89
	Repeat Words	0	13	0.00	14.44
	Ignored Words	0	14	0.00	15.56
	Total Semantic Errors	**10**	**80**	**11.11**	**88.89**

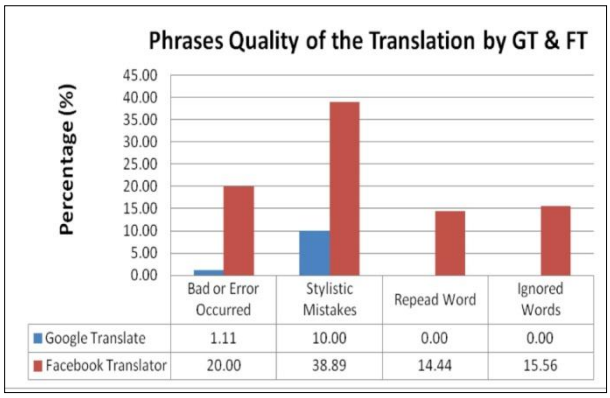

Fig. 9: Semantic error distribution in the subcategories

Table 5. Semantic Quality of the Translation

Main Category	Sub Category	Count of the Phrases		Percentage (%)	
		GT	FT	GT	FT
Semantic	Good	37	15	62.50	15.63
	Moderate	26	20	35.94	40.63
	Bad	1	29	1.56	43.75

According to Table 5, above it shows, that there are 29 'Bad' phrases out of 64 phrases has been translated by FT, which is the highest error rate (43.75%) compared with GT 'Bad' error rate is much lower and it is 1.56%. On the other hand, 62.50% of translation has identified in GT as a 'Good' translation compared with FT this rate is only 15.63%. Out of 64 phrases, 26 and 20 phrases identified as a 'Moderate' category, respectively. In this study, observation of the translation process is done, which specifically was focusing on the source phrases, and its translated result is shown in Fig. 10 below.

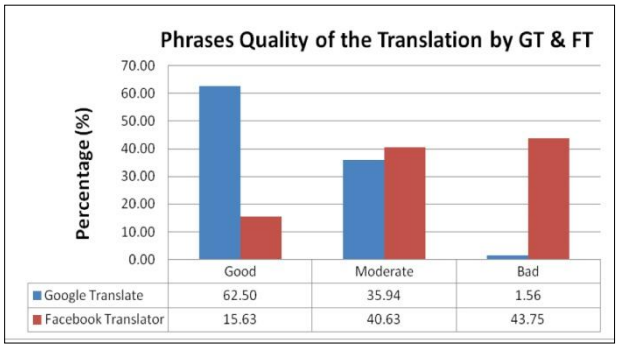

Fig. 10: Quality of the Translation by Google Translate and Facebook Translator

5. CONCLUSION

The accuracy varies greatly among languages, due to the differences in complexity and nature of language. Some languages produce better results than others. A good example is Russian-to-English. An acceptable automated translator should be aware of the cultural factors, simultaneously, customs and traditions, and consider the chronological orders, specific meaning, development of related disciplines, and historical and religious sensitivity of the content. Unintended error in the process of automated translation not only render a varied meaning but also lead to trigger social issues.

6. REFERENCES

[1] GoogleTranslate:https://translate.google.com/intl/en/about/, Google Inc, [accessed on 10th, June 2019]

[2] AutomaticTranslationonFacebook:https://code.fb.com/ml-applications/expanding-automatic-machine-translation-to-more-languages/ [accessed on 12th, June 2019]

[3] Stephen Doherty. 2016. The Impact of Translation Technologies on the Process and Product of Translation. International Journal of Communication 10 (2016). pp. 947–969.

[4] Nandasara, S. T., Leong K. Y., Samaranayake, V. K., Tan, T. W., Trilingual Sinhala-Tamil-English National Website of Sri Lanka, INET'97 Internet Society Annual Conference, Kuala Lumpur, Malaysia. (June 1997

[5] Pieris, P. E., 2019. Ceylon and the Portuguese: 1505 1658. ISBN: 1528183967 (ISBN13: 9781528183963). (Classic Reprint Published January 10th, 2019). Forgotten Books, Hardcover. 322 pages.

[6] Leon Dostert. 1957. Brief History of Machine Translation Research. Eighth Annual Round Table Meeting on Linguistics and Language Studies. Georgetown University.

[7] Mark R. Wolgemuth, John David Alberg, 2008. Proxy For Real-Time Translation of Source Objects Between A Server And A Client, Feb. 14, 2008.

[8] William N. Locke and A. Donald Booth, Written 15 July 1949. Published in: Machine translation of languages: fourteen essays, ed. by (Technology Press of the Massachusetts Institute of Technology, Cambridge, Mass., and John Wiley & Sons, Inc., New York, 1955), pp.15-23.

[9] Carbonell, J. G., Cullingford, R, E. and Gershman A. G. 1981. Steps Towards Knowledge-Based Machine Translation, IEEE Trans. PAMI, Vol. PAMI-3, No. 4.

[10] Farrús Cabeceran, M., Ruiz Costa-Jussà, M., Mariño Acebal, J. B., & Rodríguez Fonollosa, J. A., 2010. Linguistic-based evaluation criteria to identify statistical machine translation errors. 14th Annual Conference of the European Association for Machine Translation, Saint-Raphaël (pp. 167-173).

[11] Stephen Doherty, 2016. The Impact of Translation Technologies on the Process and Product of Translation, International Journal of Communication 10(2016), pp. 947–969.

[12] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin and others. 1990. A statistical approach to machine translation. Computational Linguistics. MIT Press Cambridge, MA, USA, Volume 16, Issue 2, June 1990, pp 79-85.

[13] Alan Packer, 2016. Director of Engineering Language Technology at Facebook (ETF), https://www.bbc.com/news/business-36638929, 28th, June 2016.

[14] Seljan, Sanja; Gašpar, Angelina; Pavuna, Damir, 2007. Sentence Alignment as the Basis For Translation Memory Database, The Future of Information Sciences: INFuture 2007, Digital Information and Heritage, Department of Information Science, Faculty of Philosophy Zagreb, 2007. pp. 299-311.

[15] Yoshiki Mikami, Pavol Zavarsky, Mohd Zaidi Abd Rozan, Izumi Suzuki, Masayuki Takahashi, Tomohide Maki, Irwan Nizan Ayod, Paolo Boidi, Massimo Santini, Sebastiano Vigna, 2005. The Language Observatory Project (LOP), www2005, May 2005, Chiba, Japan.

[16] Chew Choong, Yoshiki Mikami, C. A. Marasinghe and S. T. Nandasara, 2009. Optimizing n-gram Order of an n- gram Based Language Identification Algorithm for 68 Written Languages, The International Journal on Advances in ICT for Emerging Regions 2009 02 (02). pp. 21 – 28.

[17] I. Suzuki, Y. Mikami, A. Ohsato, 2002, A Language and Character Set Determination Method Based on N- gram Statistics. ACM Transaction on Asian Language Information Processing, 2002, Vol. 1. No. 3, pp. 270–279.

[18] Yule, G., 1996. Pragmatics. Oxford: Oxford University Press.

[19] D. A. Cruse, Lexical Semantics, In the series Cambridge Textbooks in Linguistics, New York: Cambridge University Press, 1996, pp. 310.

[20] Universal Declaration of Human Rights, United Nations: https://www.un.org/en/universal-declaration-human-rights/, [accessed on 10th June 2019]

[21] Universal Declaration of Human Rights, United Nations, Sinhala:https://www.ohchr.org/EN/UDHR/Pages/Language.asp x? LangID=snh, [accessed on 10th June 2019]