Umanistica Digitale – ISSN: 2532-8816 – n. 10, 2021
G. M. Di Nunzio, G. Cambedda, V. Nosilia – *A Study on Machine Translation Tools: A Comparative Error Analysis Between DeepL and Yandex for Russian-Italian Medical Translation*
DOI: http://doi.org/10.6092/issn.2532-8816/12631

A Study on Machine Translation Tools: A Comparative Error Analysis Between DeepL and Yandex for Russian-Italian Medical Translation

[1]Giulia Cambedda, [2]Giorgio Maria Di Nunzio, [3]Viviana Nosilia

[1,2,3]Università di Padova, Italy
[1]giulia.cambedda@studenti.unipd.it
[2]giorgiomaria.dinunzio@unipd.it
[3]viviana.nosilia@unipd.it

Abstract

The present research is aimed at conducting a study on Russian-Italian medical translation with regard to the current development of two Machine Translation tools that feature prominently in today's Neural Machine Translation framework, namely DeepL and Yandex. For the purpose of our research, we have selected three highly specialized and three popular-science articles concerning coronavirus pandemic. Such a choice is justified by the willingness not only to analyse recently published scientific documents but also to investigate the particular linguistic implications of 2020's coronavirus pandemic outbreak, which has introduced in every-day communication a whole set of terms whose use was previously limited to the language of science, as well as coined a group of new terms, which entered the boundaries of scientific terminology. We have considered this existing linguistic phenomenon as a proper condition to test the performances of Machine Translation tools. In particular, we discuss the most relevant features of the comparative error analysis as well as the BLEU metric for both DeepL and Yandex.

La presente ricerca ha lo scopo di condurre uno studio nell'ambito della traduzione medica russo-italiana sull'attuale sviluppo di due strumenti di traduzione automatica che hanno un ruolo di primo piano nell'attuale quadro della traduzione automatica "neurale", DeepL e Yandex. Ai fini della nostra ricerca, abbiamo selezionato articoli medici scritti in russo, in particolare: tre articoli medici altamente specializzati e tre di scienza popolare riguardanti la pandemia di coronavirus. Tale scelta è giustificata dalla volontà non solo di analizzare documenti scientifici di recente pubblicazione, ma anche di indagare le particolari implicazioni linguistiche dell'epidemia di coronavirus del 2020. Infatti, durante il periodo di pandemia sono stati introdotti e coniati nella comunicazione quotidiana tutta una serie di termini il cui uso era precedentemente limitato al linguaggio della scienza e che improvvisamente sono entrati nei confini della terminologia scientifica. Abbiamo considerato questo fenomeno linguistico esistente come una condizione

adeguata per testare le prestazioni degli strumenti di traduzione automatica. In particolare, discuteremo la valutazione della traduzione automatica analizzando le caratteristiche più rilevanti dell'analisi comparativa degli errori e la metrica BLEU per DeepL e Yandex.

1. Introduction

Over the years, Machine Translation (MT) has undergone rapid growth, mainly due to extensive research in the field and an increasing worldwide interest in computer science. By the middle of the previous century, several approaches have been implemented within the field of Machine Translation, in order to obtain machine translation systems combining high-quality results and low costs of implementation ([10]).

In this paper, we present a qualitative and quantitative analysis of two neural MT tools that feature prominently in today's MT panorama: DeepL and Yandex. The former was launched in 2017 by DeepL GmbHI as Neural Machine Translation system, whereas the latter, launched in 2011 by the Russian Internet company Yandex as a Statistical Machine Translation tool, shifted from a purely statistical translation system to a hybrid translation system, consisting of both a statistical and a neural approach, in 2017 (www.Yandex.com). In order to compare the effectiveness and reliability of DeepL and Yandex, we perform a manual comparative error analysis of the translations provided by the two tools. We detect the errors committed by the two translation tools and provide a linguistic analysis. Each error is marked as belonging to one or more of categories specifically selected according to the linguistic features of the texts and the language pair under examination. After completing the manual error analysis, the two tools' translation performances are evaluated by means of the BLEU metric, which calculates the percentage of textual similarity between the machine translations under examination and a set of reference translations ([24]). The choice of the specific language pair under examination, namely Russian-Italian, is certainly due to the major role as vehicular language of scientific literature that Russian played for a considerable time. Therefore, since a significant amount of Russian relevant scientific documents still persist in the original language, an improvement in the translation performances of Machine Translation tools would ensure the availability of a relevant scientific research heritage outside Russian borders. Of primary interest for our research is the study of DeepL's and Yandex's translation performances with regard to medical language. In particular, we decided to analyse how the grammatical, syntactical, and lexical disparities between medical highly specialized and popular-science texts affect the two translation tools' behaviour and overall translation performances. Therefore, for the purposes of our study, we have selected three Russian highly specialized medical texts and three Russian popular-science medical texts concerning the Coronavirus pandemic outbreak. The analysis and translation of the whole documents lay outside the scope of the present research. Nonetheless, several fragments of each document are translated into Italian and analysed. More specifically, special attention is devoted to the specialized texts' titles and abstracts, and the popular-science texts' titles and first paragraphs.

The paper is organized as follows: in Section 2, we provide literature review of Machine Translation Evaluation approaches; in Section 3, we present our experimental setting and our qualitative and quantitative analysis. In Section 4, we draw our conclusions and provide final remarks.

1.1 Background in Neural MT

A turning point in Machine Translation development is undoubtedly represented by the emergence, in 2014, of the first Neural Machine Translation systems. Neural Machine Translation (NMT) was initially used as a supplement of Statistical Machine Translation systems and subsequently developed its own techniques and systems. NMT systems' most innovative aspect, which undoubtedly separates them from the other MT systems, is the central role of distributional semantics and word embeddings in the translation process. The most significant contribution of the distributional semantics to the linguistic theory in general and the research in the field of machine translation is the idea that the meaning of words needs to be investigated in a contextual framework ([13]). One of the most popular NMT architectures consists of three core elements, namely: an encoder, a decoder, and an attention model. The encoder is a Bidirectional Recurrent Neural Network (BRNN) ([35]), which, using an n-gram model, analyzes the input sentence from right to left and vice versa and extracts a fixed-length vector representation of the source sentence. Afterward, the decoder, starting from the vector representation, creates a variable-length sequence, i.e., the target sentence ([34]). The attention model is a Multilayer Perception Neural Network and that aligns the source sentence words with the corresponding target sentence words. As mentioned above, an RNN may be particularly effective in performing translation tasks. Nonetheless, this kind of neural network is not able to link pieces of information that are distributed over long distances. Therefore, when contextual information is scattered over a long source text, NMT systems based on RNN cannot by their nature provide reliable translations. In order to overcome RNN's limitations, the more sophisticated Long Short Term Memory (LSTM) neural networks have been implemented. Unlike RNN, which consists of one only layer, LSTM neural networks have four layers and can solve long-distance problems and properly convey contextual meaning ([37]).

2. Machine Translation Evaluation: An Open Question

Translation Quality Assessment (TQA) has always constituted a key issue in Translation Studies and its relevance has significantly increased with the emergence of Machine Translation and the growth of the translation industry. Reaching an adequate understanding of how to properly evaluate Machine Translation systems is vital to the development of Machine Translation research. A good evaluation method can indeed shed light on a specific MT system's strengths and weaknesses and consequently suggest the necessary modifications and the appropriate line to be taken in the future. Moreover, it represents an essential tool for Machine Translation professionals to monitor the increasingly rapid progress of their systems and for users to sensibly choose the MT programs that best suit their needs ([25]). Nonetheless, since the beginning of

Machine Translation growth, its evaluation has always been an extremely controversial issue, and still nowadays represents an open question. Generally speaking, when assessing a translation, be it performed by humans or computer programs, we need to start from the assumption that, given a source sentence, or text, there is not only one "perfect" translation, as well as there may be several "acceptable" translations. Moreover, with regard to Machine Translation evaluation, a core concept is introduced in the following sentence by Papineni et al. ([25]): "The closer a machine translation is to a professional human translation, the better it is". Hence, in order to undertake quality assessment for MT systems' output, a machine translation is usually judged on the basis of a numerical metric that measures its closeness to a set of reference professional human translations, and its fluency, adequacy, comprehensibility and readability are assessed ([4]). In particular, accuracy and usability constitute two major properties that are taken into consideration when assessing professional translation ([30]). Accuracy can be defined as "How much of the meaning expressed in the gold standard translation or the source is also expressed in the target translation" ([19]), while usability concerns the degree of possibility for a translation to be useful to its users in order to accomplish a set of intended tasks within a certain context of use ([4]). Moreover, translation evaluation, be it human or automatic, is strictly related to the textual type and the language pair under examination, which define some of the parameters according to which the translation itself needs to be assessed. The evaluation of specialized texts' translation, such as medical texts, implies a set of specific parameters that depend on the main characteristics of medical language and the translation's final use. This requires the translator a certain degree of expertise not only in the source and target language but also in medical language and the specific topic covered by the medical text under examination ([23]). Medical language can be defined as the "occupational register of physicians and it is largely opaque outside the medical community" ([18]). It is one of the so-called Languages for Specific Purposes (LSP), i.e. those specific registers adopted by professionals to exchange information and knowledge in professional contexts. The medical language shares some features with the other LSP and has developed over time a set of specific characteristics. A high degree of impersonality aimed at maintaining distance and objectivity undoubtedly stands out in medical texts and treatises. It usually comes along with a marked tendency towards the nominalization of verbs and adjectives, which leads to extensive use of extended nominal groups. Moreover, medical language is prone to passivize active verbs and includes highly technical phrases, which constitute the medical jargon. As for lexis, the historical origins of medical language are responsible for the persistence, within the medical vocabulary, of Greek and Latin terms in their original form, as well as the formation of new words starting from Greek and Latin suffixes and prefixes ([7]). The Russian medical language shares some features with the other technical languages and developed its own ones. Objectivity and impersonality, represented by impersonal phrases and constructions, and reflexive verbs, constitute the two main features of the Russian medical language. Moreover, nouns are more widely used than verbs and a marked tendency to formulate grammatically relatively easy sentences has been observed. However, the existence of concatenations of genitives along with widespread use of gerunds and participles certainly contributes to complicating the overall grammatical structure. The Russian language of medicine is likewise characterized by a widespread presence of synonyms, which attests that close contact between Russian and international terminology took place. Although in medical terminology, international synonyms

are usually preferred for reasons of systematization, translatability, ease of international communication, and spread of medical knowledge, the Russian medical language displays the coexistence of international terms and their Russian equivalents ([26]). As mentioned above, specialized translation evaluation is required to specifically consider the final use of the translation under examination. Since its aim is related to the transfer of knowledge, its clarity and the absence of ambiguity needs to be taken into consideration ([23]). Moreover, the correct identification of the target user is a major aspect. For this reason, a distinction has to be made when evaluating the translation of medical highly specialized texts from popular-science texts' one. Machine translation evaluation can be performed by means of automated, semi-automated, and human techniques. It must be said that a clear distinction between the mentioned categories cannot be considered obvious, as their boundaries are quite blurry, and a juxtaposition is not difficult to occur. Automated Machine Translation evaluation refers to those assessing methods that do not involve human judgment or rather in which human intellect is confined to a number of side activities, such as data collection, preparation of the reference translations, or annotation. On the contrary, non-automated or human Machine Translation evaluation includes all those approaches that directly or not involve human judgment ([5]). Both evaluation approaches suffer from substantial shortcomings. On the one hand, human evaluation of Machine Translation has been strongly criticized for depending on a specific linguistic knowledge and subjective opinion of an evaluator or group of evaluators, and it is undoubtedly slower and more costly when compared to automatic evaluation metrics. Moreover, human Machine Translation evaluation implies not only a significantly costly and time-consuming implementation (Papineni et al., 2002) but also the impossibility of reproducing the same patterns an indefinite number of times and the risk of being too subjective, as human beings may be biased by a set of different factors, related to the external environment, as well as the evaluator's previous knowledge and physical or psychological conditions ([20]). However, the scholars who support the human evaluation of Machine Translation pointedly remark the idea that a human assessment of Machine Translation is needed firstly because the output of an MT system is meant to be received, understood, and used by human beings. Secondly, only human perception of the world can efficiently detect the errors made by MT systems and assess their severity. Thirdly, professional human evaluators master the linguistic knowledge necessary to deeply analyze a translation and integrating such a piece of knowledge into an automatic evaluation system undoubtedly would not only be difficult but also costly and suitable for just a limited number of language pairs. Extensive research has been conducted over the decades to measure and improve the effectiveness of human assessment as an evaluation method applicable to the output produced by Machine Translation systems. In doing so, special attention has been devoted not only to estimating human beings' ease, speed, and consistency in performing the evaluation tasks but also to determining the most favourable conditions under which a proper human assessment can be performed. On the other hand, automatic evaluation metrics tend to equally weigh all the words, or n-grams contained in the translation, regardless of the degree of informativeness of their content. Hence, the substitution of articles or interjections risks being equalized by the system to one of highly informative parts of speech, such as nouns and verbs. This kind of evaluation is indeed responsible for penalizing those candidate translations that, although correct, present a low degree of similarity with the golden translations and being unable to properly detect long-distance linguistic relationships to

provide a corpus-level quality assessment ([5]). In addition, by comparing MT systems' output against a corpus of human reference translations, automatic Machine Translation evaluation metrics necessarily include in the evaluation process a subjective human element and limit the scope and accuracy of the assessment itself. Indeed, for a given source sentence, there are a great number of reliable translation variants, undoubtedly more than the ones that can be contained in a limited, although exhaustive corpus. However, not only automated translation evaluation certainly constitutes a fast and low-cost evaluation method but is also able to provide a degree of consistency that cannot be reached by human beings alone. In addition, the recent increasing development of Neural Machine Translation has posed new technical challenges to machine translation evaluation metrics, be them human or automated, as a higher degree of precision is demanded. In fact, for the evaluation techniques to comply with the newly implemented Neural Machine Translation systems and constitute a useful tool for their development, they need to present a deeper sensitivity to linguistic nuances, be able to perform extensive analysis at the document level and be specifically designed to focus on specific linguistic features ([5]). As a consequence, a universal and commonly accepted Machine Translation evaluation method has not been implemented yet, and MTE remains an open question ([4]). For the purposes of our research, we will apply both a human evaluation method, namely the error analysis and an automatic evaluation method, namely the BLEU metric. By combining the two methods, namely the comparative error analysis, which relies on a human assessment of the translation performance, and the BLEU metric, which provides an objective and numeric machine translation evaluation, we wish to achieve a comprehensive analysis of the strengths and weaknesses of the two machine translation tools under examination, as well as an actual comparison of their translation performances.

2.1 Error analysis

The error analysis constitutes a human evaluation method consisting of assessing machine translation quality by annotating each error and marking it with an error-tag. The annotator or the group of annotators is usually provided with additional reference material, such as the source text or a set of reference human translations, or both, and asked to classify the errors encountered throughout the translation according to a number of previously agreed error categories. In order to be comprehensive, error analysis needs careful planning. Firstly, it is necessary to consider the knowledge and background of an annotator or the group of annotators, in case the evaluation is performed by a group of evaluators, to increase inter-annotator agreement by organizing a specific training devoted to clarifying the scope and purposes of the analysis ([27]). Secondly, the choice of error classes requires special attention. In fact, a great amount of error categories undoubtedly contributes to providing an exhaustive error analysis. Nonetheless, in this case, since the boundaries between the different categories are relatively blurry, it may be difficult to properly assign each error to its corresponding category. Moreover, error classes are to be carefully chosen according to the type of analysis that has to be conducted, on which depends the importance given to each error, and the relevant linguistic features of the text under examination. Finally, not only the number and type of error classes have to be considered, but also an exact

and unambiguous definition of each error class needs to be provided. In this way, correct classification is easier to perform by the evaluator or the group of evaluators ([15]).

2.2 The BLEU metric

The Bilingual Evaluation Understudy (BLEU) ([24]) metric undoubtedly represents one of the most popular precision and recall metrics as well as the benchmark to judge other automatic Machine Translation evaluation systems ([5]). Precision and recall metrics evaluate the quality of a Machine Translation output according to the measure of textual similarities between an MT system's output and a set of human reference translations. In particular, precision refers to the ratio of n-grams in the translation under examination that occur in any of the reference translations to the total number of n-grams contained in the translation under examination. The recall is defined by calculating the ratio of n-grams in the translation under examination that occur in any of the reference translations to the total number of n-grams of the reference translations ([5]). BLEU metric's algorithm is composed of two main constituent parts, a numerical metric designed to measure the closeness between the candidate machine translation and the reference translations, and a corpus containing a number of reliable human reference translations. In evaluating the candidate translation, the BLEU metric considers three core factors, namely word choice, word order, and length. A score is indeed assigned to each n-gram of the candidate translation on the basis of its similarity with the reference translations. Afterward, all the scores are averaged over the entire corpus. In this way, a comprehensive corpus-based evaluation indicating the candidate translation's overall quality is provided. The score given to each n-gram is calculated according to a precision measure: the words contained in the n-gram under examination that occur also in any reference translation are counted and their total number is then divided by the number of words contained in the n-gram ([5]). However, since MT systems tend to generate a greater number of words when compared to any reference translation, in order to achieve a reliable evaluation, a modified unigram precision measure is adopted. The modified precision measure takes into account not only whether a candidate translation n-gram appears in any reference translation, but also its maximum number of times of occurrence in any of the reference translations, which is afterward divided by the number of times the same n-gram appears in the candidate translation itself. The size of the n-grams used as units of comparison has a strong effect on the overall evaluation. Shorter n-grams, such as unigrams, i.e. consisting of a single word, are proven to be useful to assess the adequacy, whereas longer n-grams are better in determining the fluency of the candidate translation ([25]). As mentioned above, for the candidate translation to be given a high score, it should correlate well with the reference translations also in terms of length. Moreover, the candidate translation is expected to contain just one of the synonymic words that appear in the different reference translations as variants of the same potential source word. In this case, a candidate sentence could indeed be given a high score as it contains many words that occur also in the reference translations, without being a reliable translation, nor having, in many cases, full meaning. In order to include the length factor in the evaluation process and try to overcome the recall problem, a brevity penalty is introduced. The reference translation that better reflects the candidate translation's length is selected and the same procedure is performed at the sentence

level. Afterward, all the length best matches are summed and compared to the total length of the machine translation output under examination. BLEU metric's scores range from a minimum of 0 and a maximum of 1, which would be assigned to a candidate translation proven to be identical to one of the reference translations contained in the corpus ([3]).

3. A comparative error analysis between DeepL and Yandex

In order to compare the effectiveness and reliability of DeepL and Yandex, we perform a manual comparative error analysis of the translations provided by the two Machine Translation tools. We detect the errors committed by the two translation tools and provide a brief linguistic analysis. Moreover, each error is marked as belonging to one or more of the error categories that are listed above. A universal approach to the problem of properly setting error analysis has not been implemented yet and different scholars proposed their own error typology. A significant step towards error analysis standardization has been carried out with the implementation of the Multidimensional Quality Metrics (MQM) by the German Research Centre for Artificial Intelligence (DFKI) from 2012 to 2014 ([15]). When setting the comparative error analysis discussed in the present paper, we decided to refer to MQM with regard to its tree structure, as well as some of its specific error categories. Indeed, we have selected a number of error categories of the MQM model and developed other specific ones according to the linguistic features of the texts and the language pair under examination. In particular, we added Theme-rheme pattern[1] category, as well as Untranslated elements[2] and Transliteration[3] subcategories. Moreover, some

1 Theme-rheme pattern category refers to the incorrect rendering of the theme-rheme pattern of the source document and its corresponding original communicative function. See an example with its corresponding analysis in Section 3.2.6.

2 Unlike Omission subcategory, which refers to those fragments of the source text that are not rendered in the target text, Untranslated elements subcategory refers to those fragments of the original text that are not affected by a translation process and maintained in the source language in the target text. See the following fragment from the popular-science article Ведение детей с заболеванием, вызванным новой коронавирусной инфекцией (SARS-CoV-2) ([1]) and the corresponding translation by Yandex:

Ведение детей с заболеванием, вызванным новой коронавирусной инфекцией (SARS-CoV-2)

С целью обеспечения детского населения эффективной медицинской помощью в условиях пандемии новой коронавирусной инфекции Минздравом России совместно с профессиональными ассоциациями и экспертами в области педиатрии, инфекционных болезней и реанимации.

Con l'obiettivo di garantire alla popolazione pediatrica efficace aiuto medico in condizioni di pandemia di nuova коронавирусной infezione dal ministero della salute Russia in collaborazione con le associazioni professionali e di esperti nel campo della pediatria, malattie infettive e rianimazione.

3 Transliteration subcategory refers to the violation of the rules of scientific transliteration of Cyrillic alphabet. See the following example from the popular-science text Смертность от COVID 19 -

primary error categories branch into one or more subcategories, which constitute more specific types of issues that fall within a certain error category.

- Grammar
    - Syntax
    - Use of the articles
- Lexis
- Acronym
- Terminology
    - Domain-specific terms
- Culture-specific references
- Theme-rheme pattern
- Accuracy
    - Omission
    - Untranslated elements
- Consistency
- Orthography
    - Spelling
    - Capitalization
    - Transliteration
- Format

3.1 Dataset

For the purposes of our research, we have selected and translated from Russian into Italian three highly specialized and three popular-science medical texts. The titles of the selected texts, along with the number of translated words for each text, are listed below:

1. Биологическая терапия в эру COVID-19 (97 words) ([21])[4]

Взгляд демографа на статистику причин смерти в России и мире (Timonin, 2020) and the corresponding DeepL's translation.

Какие два типа статистических данных должны разрабатываться в условиях эпидемии — на эти и многие другие вопросы отвечают демографы Сергей Тимонин и Анатолий Вишневский.

Quali due tipi di statistiche dovrebbero essere sviluppate nel contesto di un'epidemia - a queste e a molte altre domande rispondono i demografi Sergei Timonin e Anatoly Vishnevsky.

4 The biological therapy in the COVID-19 era.

2. Ведение детей с заболеванием, вызванным новой коронавирусной инфекцией (SARS-CoV-2) (133 words) ([1])[5]
3. Коронавирус SARS-Cov 2: сложности патогенеза, поиски вакцин и будущие пандемии (197 words) ([38])[6]

Moreover, the following three Russian popular-science texts have been selected:

1. Коронавирус завозили в Россию не менее 67 раз (149 words) ([16])[7]
2. Неизвестная летальность - Почему мы не знаем истинных масштабов COVID-19 (173 words) ([28])[8]
3. Смертность от COVID 19 - Взгляд демографа на статистику причин смерти в России и мире (150 words) ([32])[9]

3.2 Qualitative analysis of the results

Once we marked all the errors, two tables have been compiled to provide a visual overview of the quality of the translations provided by the two translation tools. For the mentioned table to be easily readable and understandable to the intended readers, it has been divided according to the translation tool and the text type that is displayed. Therefore, a total of four tables has been inserted. Each table shows how the erroneous fragments are displayed in the source document, in the provided translations, and a specifically made human translation, used as reference translation. Each erroneous fragment has been further associated with one or more of the above-mentioned error categories and then summed according to the error category within which it falls. A number of example relevant translation errors along with their corresponding error categories have been listed and briefly commented in the following lines, with the aim of providing a deeper understanding of the methods of our comparative error analysis as well as DeepL's and Yandex's strengths and weaknesses. For a better understanding of the contextual information, the fragments of the original text containing the translation errors under examination have been inserted at the beginning of each subsection, and the portion of text analysed highlighted in bold and reported in the corresponding table.

5 Clinical management of children with a disease caused by the new coronavirus infection (SARS-CoV-2).

6 Coronavirus SARS-Cov 2: complexities of the pathogenesis, search for the vaccines, and future pandemics.

7 Coronavirus was imported into Russia at least 67 times.

8 Unknown lethality - Why we do not know the true extent of COVID-19.

9 COVID-19 mortality rate - A demographer's perspective on statistics of causes of death in Russia and worldwide.

3.2.1 Grammar - Syntax

> При сравнении этих данных со статистикой по другим странам обращает на себя внимание и вызывает неизбежные вопросы исключительно низкий уровень летальности от коронавируса в России ([32]).

Original document	Yandex's translation	Human translation
обращает на себя внимание и вызывает неизбежные вопросы исключительно низкий уровень летальности от коронавируса в России.	richiama l'attenzione e solleva questioni inevitabili eccezionalmente basso tasso di mortalità da coronavirus in Russia.	attira l'attenzione e solleva inevitabili domande il tasso di mortalità estremamente basso da coronavirus in Russia.

The sentence under examination provides an interesting starting point for discussion as Yandex's version does not result in compliance with Italian grammatical rules, especially with regard to its erroneous word order. Indeed, the main verb of the sentence and its subject are not only displayed in an order that does not respect Italian grammatical rules, according to which the verb, except for in some special cases, usually follows the corresponding subject, but they are also completely disconnected from one another (Treccani, 2020). This undoubtedly contributes to distorting the meaning of the whole translation variant and preventing it from being acceptable.

3.2.2 Grammar - Use of the articles

> Актуальность. Вакцина против коронавируса SARS-Cov-2 рассматривается как наиболее перспективное средство для укрощения вызванной им нынешней пандемии и воспрепятствования возникновению новой ([38]).

Original document	Yandex's translation	Human translation
Вакцина против коронавируса	Il vaccino contro il coronavirus	Un vaccino contro il coronavirus

The Russian term Вакцина and consequently its Italian equivalent *vaccino* denote a non-specific and still theoretical entity. In other words, since a vaccine against coronavirus had not been developed yet when the article under examination was written, the name describing it refers to a general concept, i.e. the fact that only a vaccine may restrain the existing pandemic to exacerbate and prevent the emergence of a new one. As a result, an indefinite article would more appropriately contribute to conveying the original grade of definiteness meaning ([14]).

3.2.3 Lexis

> **Ведение детей с заболеванием**, вызванным новой коронавирусной инфекцией (SARS-CoV-2) ([1]).

Original document	DeepL's translation	Human translation
Ведение детей	Conduzione di bambini	Gestione dei bambini

The Russian term *ведение* is translated with the Italian noun *conduzione*. Although it can be considered as a possible translation variant for the Russian term under examination (academic.ru), *gestione* more frequently comes in association with the Italian term *bambini* (Tiberii, 2018), i.e., the proper translation of the Russian term детей, and perfectly suits the specific context of the source text. This example undoubtedly confirms that the ability of a Machine Translation Tool in properly rendering contextual information plays a major role in the assessment of the quality of the translation tool itself.

3.2.4 Acronyms

> В обзоре отражены последние рекомендации международных ассоциаций/консенсусов и наблюдения врачей различных специальностей по вопросу прерывания/продолжения терапии **ГИБП** с оценкой последствий в случае прерывания биологической терапии ([21]).

Original document	Yandex's translation	Human translation
ГИБП	GIBP	della terapia dei preparati biologici geneticamente modificati

The Russian acronym ГИБП, which stands for генно-инженерные биологические препараты ([11]), and can be translated into English as genetically engineered biological products, is translated with its transliteration, namely GIBP. Although it does adhere to the transliteration rules, since no equivalents can be found neither in Italian nor in English scientific literature, it cannot be considered a correct translation equivalent. A good variant could be instead the translation into Italian of all the words that constitute the Russian acronym.

3.2.5 Culture-specific references

> Исследовательская группа из Высшей школы экономики и Сколтеха, совместно со специалистами НИИ гриппа им. А.А. Смородинцева в Санкт

> Петербурге и **ИППИ им. А.А. Харкевича РАН** установили, что коронавирус SARS-CoV 2 независимо проникал на территорию России не менее 67 раз, главным образом в конце февраля и начале марта 2020 года ([16]).

Original document	DeepL's translation	Human translation
ИППИ им. А.А. Харкевича РАН	dell'A.A. Kharkevich IPPI RAS	dell'Istituto delle questioni di trasmissione dell'informazione A.A. Harkevič Rossijskaja Akademija Nauk

The translation tool totally fails in properly rendering the Russian acronyms ИППИ and РАН, which stand for Институт Проблем Передачи Информации and Российская Академия Наук. Indeed, neither Italian nor English equivalents for DeepL's translation version can be found. In this case, the English noun that appears on the institute's website (iitp.ru/en/about) as well as a translation into Italian of each word constituting the acronyms may represent good translation options. Moreover, in the translation versions, not only the acronyms but also the Russian proper noun Харкевича are inappropriately transliterated. Indeed, in this case a transcription has been performed and, although a universal rule does not exist, being the document under consideration a scientific paper, the scientific transliteration would undoubtedly be more appropriate.

3.2.6 Theme-rheme pattern

> Проблема: Многие в России считают, что переболели COVID-19 ещё в декабре 2019 года или в январе 2020. Можно ли узнать, когда действительно в России началась эпидемия коронавируса и откуда его к нам завезли? **Ответ дала биоинформатика** ([16])

Original document	DeepL's translation	Human translation
Ответ дала биоинформатика.	La bioinformatica ha dato la risposta.	La risposta è stata data dalla bioinformatica

In the original document, denoting the Russian term биоинформатика a new piece of information, it is located at the end of the last sentence and consequently brought into sharper focus than the other terms. The same does not apply to the translated text, which, although lexically correct, does not reflect the original theme-rheme pattern and consequently fails in fully conveying the source meaning.

3.3 Quantitative analysis of the results

In the present section, the obtained results have been analysed, with regard to specialized texts and popular-science texts, as well as overall translation quality. Afterward, the total number of translation errors has been counted and shown by means of several graphs in order to provide a visual and direct understanding of the data displayed in the previous section. More specifically, before delving into the actual comparison between DeepL's and Yandex's translation performances, careful attention has been devoted to analysing the results obtained for each translation tool with regard to the two text types under examination, namely specialized and popular-science texts. The graphs display the translation errors occurring in each error category separately, in descending order of number of translation errors. Subsequently, the two translation tools' performances have been compared on the basis of the number of errors occurring in the translations of the three specialized texts, the three popular-science texts, and in all texts, considered as a whole. Two different coloured columns have indeed been used to represent the two translation tools, where translation errors have been shown according to the error category to which they belong.

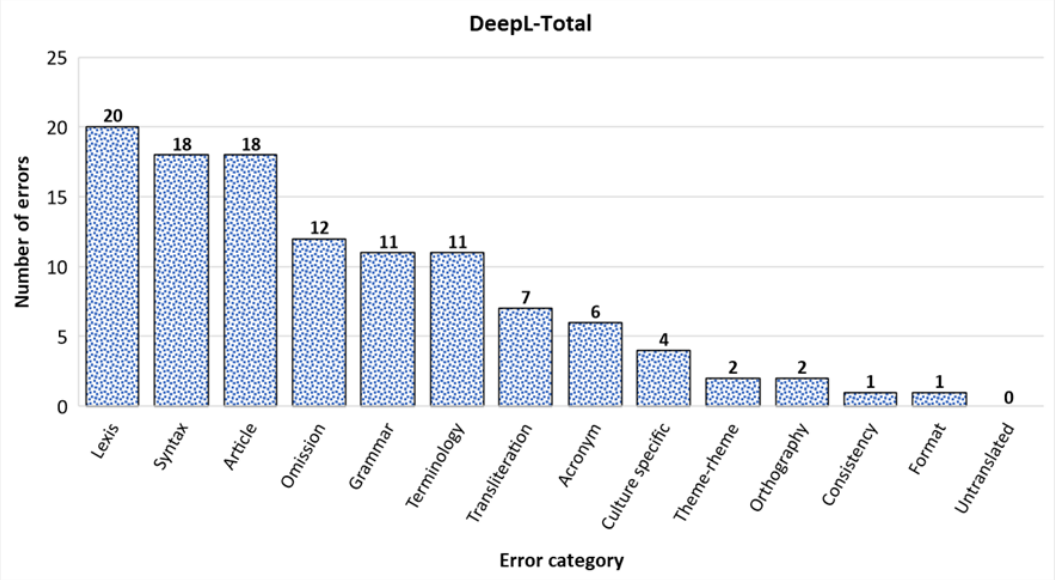

Figure 1: DeepL's translation performance

3.3.1 Analysis of DeepL translation

With a total of 113 errors, DeepL displays an undoubtedly non-uniform error distribution. Indeed, starting from the left side of the graph, three error categories, namely Lexis, Syntax, and

Article Usage clearly stand out for their high number of translation errors, followed by Omission, Grammar, Terminology, Transliteration, and Acronym. On the contrary, while Culture-Specific References, Theme-Rheme Pattern, Orthography, Consistency, and Format seem to display a minor, even negligible amount of translation errors, no translation errors are marked as belonging to the Untranslated Elements category. When considering the language pair at stake, consisting of two highly different natural languages, both in terms of syntactical structure and article usage, it does not come as a surprise that the majority of the errors committed by DeepL fall in these two error categories. This may rather mean, generally speaking, that the translation tool under examination still encounters considerable difficulties in properly rendering two of the most challenging aspects of Russian-Italian translation. As for lexis, we will see later on in the present section, when considering DeepL's translation performances with regard to specialized texts and popular-science ones separately, that the highly-specialized terms displayed in the specialized articles are proven to make a major contribution to these results.

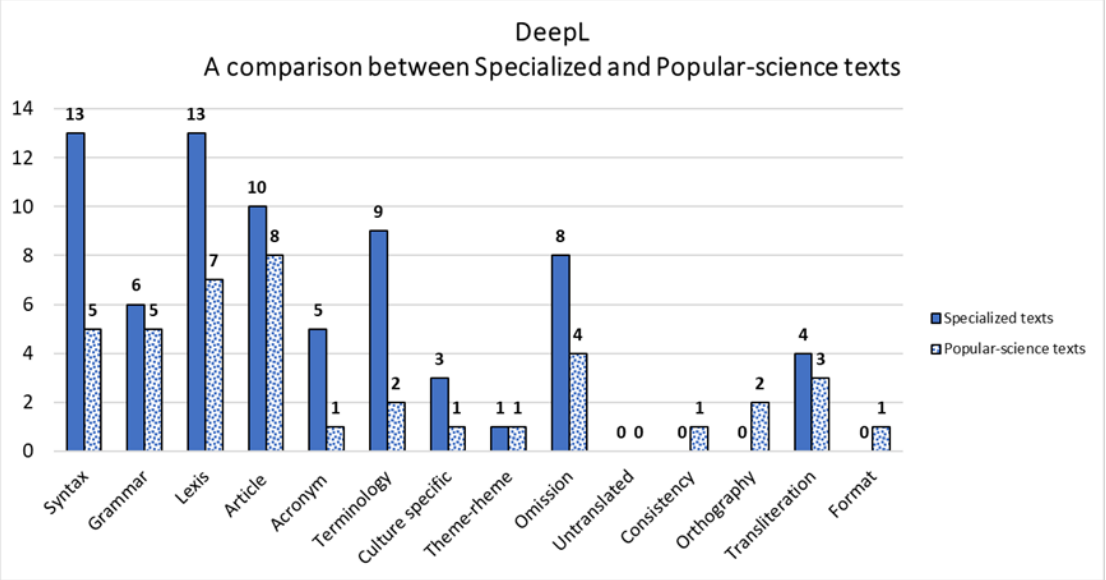

Figure 2: A comparison between DeepL's translations of specialized and popular-science texts

When comparing DeepL's translation performances with regard to specialized and popular-science texts, with a total of 72 and 41 errors, respectively, we can easily notice that it performs significantly better in translating popular-science-texts. It indeed commits a higher number of translation errors with regard to specialized texts in nine out of fourteen selected error categories, except for Theme-Rheme Pattern, Untranslated Elements, Consistency, Orthography, and Format. However, those last categories display a small overall number of translation errors. Moreover, other error categories, namely Grammar, Article Use, Culture-Specific References, and Transliteration feature a narrow gap between the two text types under examination. On the contrary, Syntax, Lexis, Acronym, Terminology, Omission visibly stand out as a significant

disparity can be observed in the renderings of specialized and popular-science texts. When analysing Acronym and Terminology categories, we need to consider that, since the total number of occurrences of these elements throughout specialized and popular-science texts may significantly differ, the results are likely to be biased accordingly. As for Syntax, Lexis, and Omission, we may assume that the translation tool encounters considerable difficulties in rendering the significantly more elaborate syntactical structure and the specific lexis and terminology characterizing Russian specialized medical articles when compared to popular-science ones.

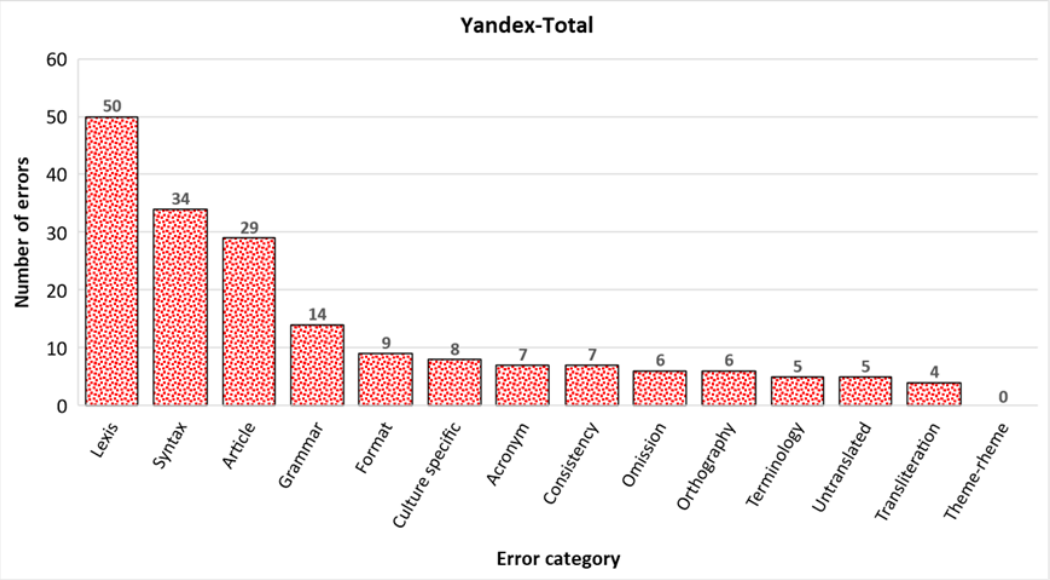

Figure 3: Yandex's translation performance

3.3.2 Analysis of Yandex translation

Yandex seems to distribute its total 184 errors quite uniformly, except for three error categories, namely Lexis, Syntax, and Article Usage, which display a significantly higher number of translation errors when compared to the other error categories. As mentioned with regard to DeepL's translation performance, Syntax's and Articles' great amount of translation errors may be explained by the profound difference in syntactical structure and article usage between Russian and Italian. On the contrary, Lexis constitutes the error category displaying the greatest amount of translation errors and further analysis is needed, in this case, to determine whether this is due to a particular type of text or to the fact that the translation tool under examination encounters serious difficulties in properly rendering Italian lexis, independently of the text type. Apart from that, the other error categories do not seem to feature a great number of translation errors. Moving left to right across the graph, we can indeed easily notice that Grammar, Format, Culture-Specific References, Acronym, and Consistency, although collocated immediately after the Article Usage category, are divided from it by a huge gap, whereas they are clearly closer, by

the number of translation errors, to the error categories occupying the right side of the graph, namely Omission, Orthography, Terminology, Untranslated Elements, and Transliteration, which display a small number of translation errors. Finally, no errors are marked as belonging to the Theme-Rheme Pattern category.

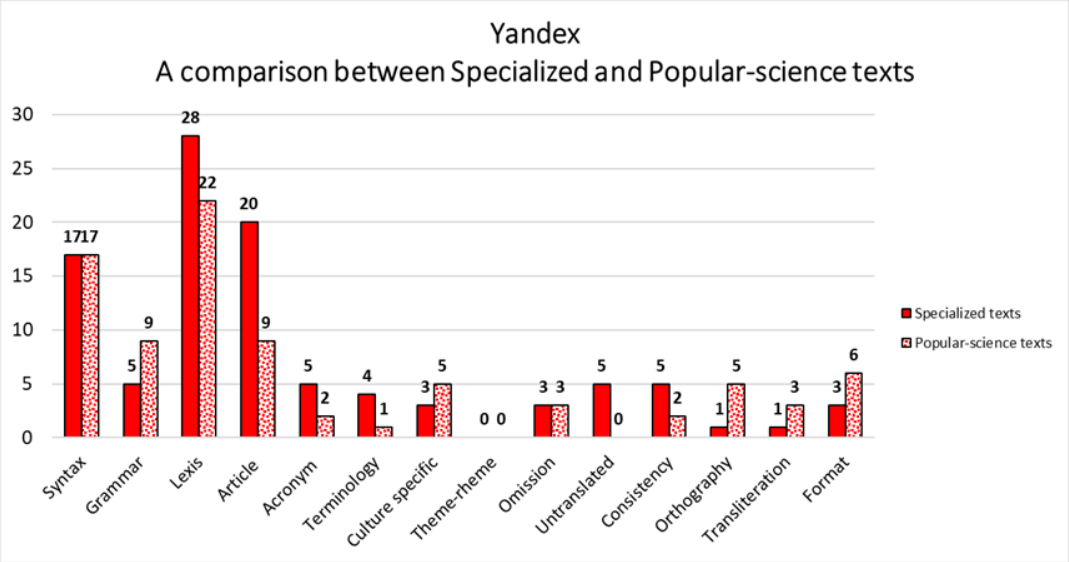

Figure 4: A comparison between Yandex's translations of specialized and popular-science texts

With a total of 100 errors in specialized texts' translations and 84 in popular science's ones, Yandex seems to provide a better translation performance with regard to popular-science texts. Nonetheless, it follows a significantly different pattern when compared to DeepL's translation performances. A small gap is indeed observed in each error category, except for Article Usage, which displays 20 errors in specialized texts and just 9 in popular-science ones. As for the other error categories, a better translation performance with regard to popular-science texts can be observed in Lexis, Acronym, Terminology, Untranslated Elements, and Consistency. On the other hand, Syntax, Omission, and Theme-Rheme display the same amount of translation errors, whereas Yandex seems to better render specialized texts' Grammar, Culture Specific References, Orthography, Transliteration, and Format. Starting from saying that the results obtained in Acronym, Terminology, Culture-Specific References, and Transliteration may be biased by the actual number of times that these elements occur throughout the texts under examination, the ones related to Grammar, Article Usage, Untranslated Elements, Consistency, Orthography, and Format undoubtedly constitute a good starting point for a more detailed analysis of Yandex's behaviour and patterns in translating Russian-Italian specialized and popular-science medical texts.

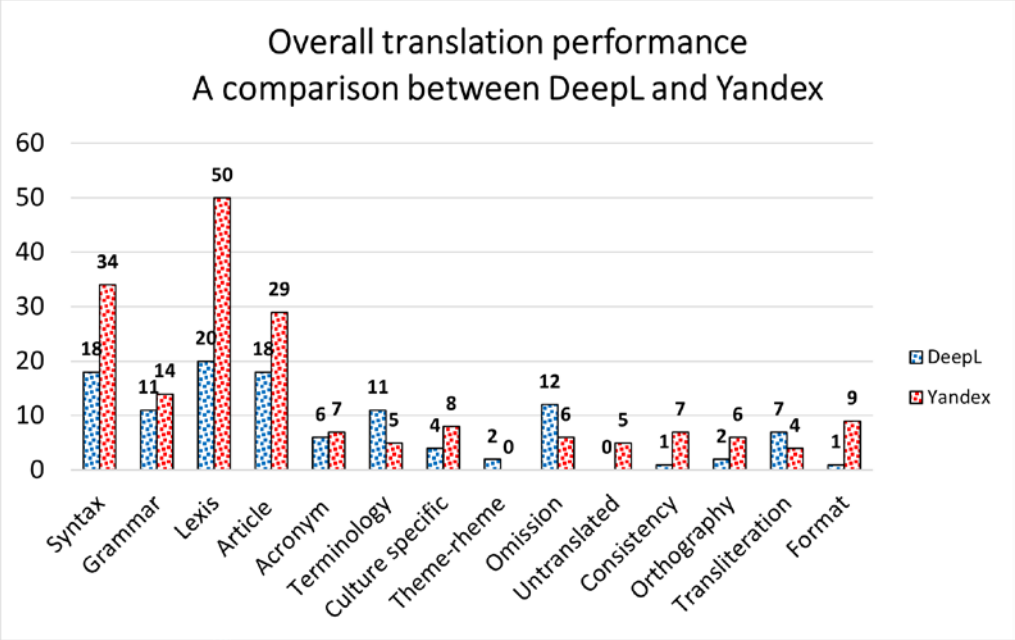

Figure 5: A comparison between DeepL and Yandex regarding the overall translation performance

3.3.3 A Side-by-Side Comparison

When comparing the two translation tools under examination on the basis of their overall translation performances, it can be observed that they both collect a significant amount of translation errors in three error categories, namely Lexis, Syntax, and Article Usage, in order of the number of translation errors. However, these categories, first and foremost Lexis, display a wide gap between DeepL's and Yandex's translation performances. We can easily say that, although Syntax and Article Usage clearly constitute two of the major weakness of both translation tools, in all probabilities because of the profound difference between Russian and Italian syntactical structure and article usage, DeepL seems to be able to better render into Italian the Russian syntactical structure and is proven to remain more compliant with Italian grammatical rules when it comes to properly insert the Italian articles, non-existent in Russian grammar. As for Lexis, DeepL, once again, provides a better translation performance, however, further analysis is needed to detect whether Yandex's high amount of translation errors are related to a specific text type.

Moreover, Yandex seems to perform better with regard to Terminology, Omission, and Transliteration, whereas DeepL is observed to commit a smaller number of translation errors in Grammar, Acronym, Culture-Specific References, Untranslated Elements, Consistency, Orthography, and Format, even though the gap in these categories is not comparable to the one observed in Lexis, Syntax, and Article Usage.

Generally speaking, a more comprehensive evaluation of DeepL's and Yandex's translation performances can be made by comparing the two translation tools' behavior concerning specialized and popular-science texts separately.

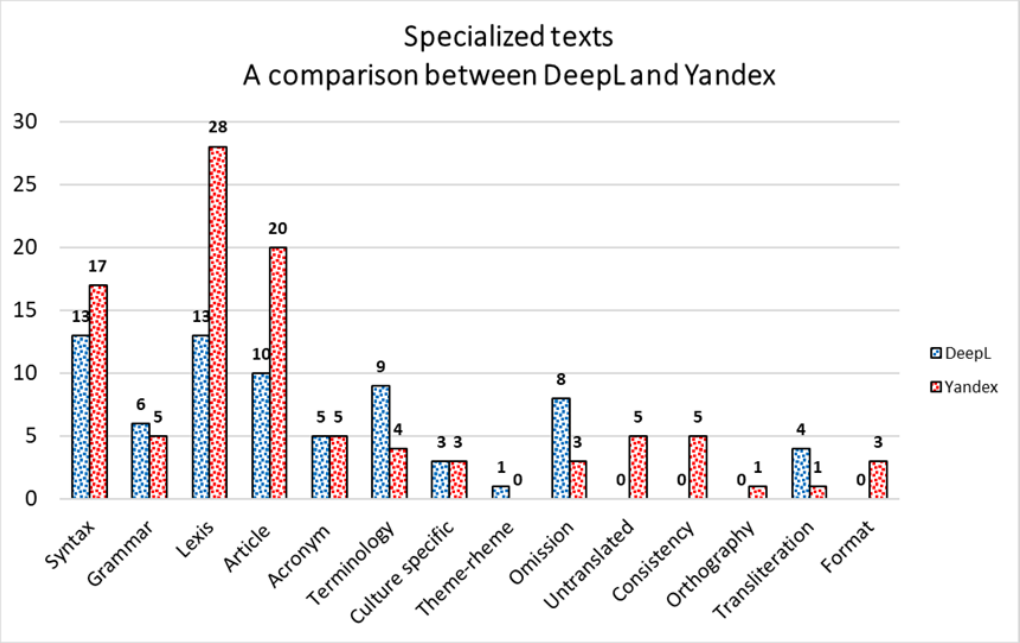

Figure 6: A comparison between DeepL and Yandex regarding specialized texts' translation

When comparing DeepL's and Yandex's translation performances with regard to specialized texts, with a total of 72 and 100 translation errors, respectively, DeepL seems to perform better when compared to Yandex. Indeed, the latter collects a higher number of translation errors in the majority of error categories, except for Grammar, Terminology, Theme-Rheme Pattern, Omission, and Transliteration. On the one hand, Lexis and Article Usage error categories undoubtedly stand out, as not only they display a significantly higher amount of translation errors with regard to both translation tools, when compared to the other error categories but also a great gap can be observed between DeepL's and Yandex's translation performances. On the other hand, Yandex is proven to give a better translation performance when compared to DeepL in Terminology, Omission, and Transliteration, even though the gap is not comparable to the one observed in Lexis and Article Usage.

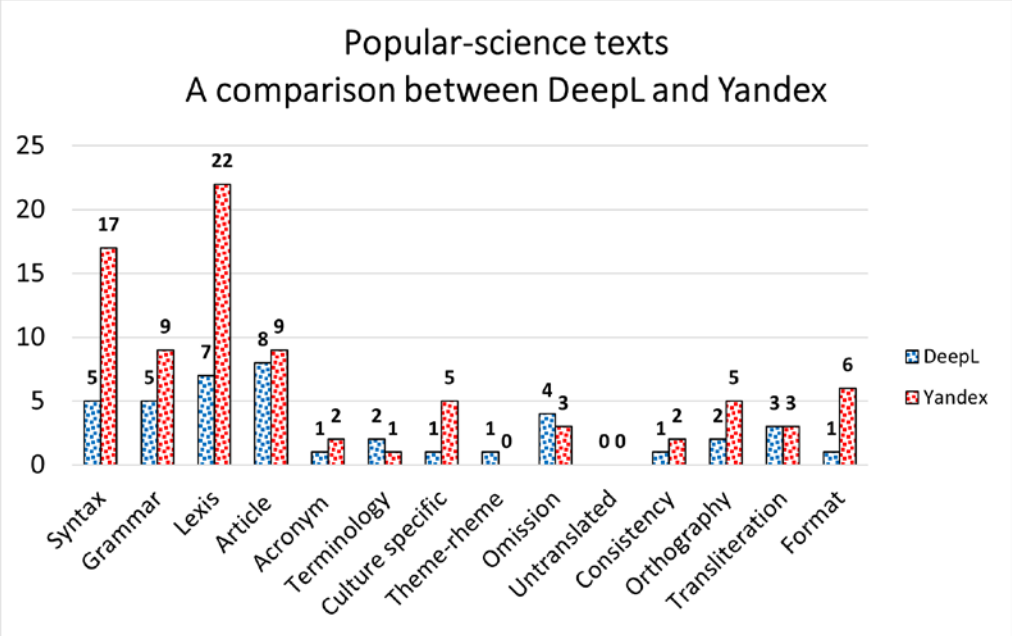

Figure 7: A comparison between DeepL and Yandex regarding popular-science texts' translation

With a total of 41 translation errors committed by DeepL and 84 by Yandex, both translation tools are proven to perform better in translating popular-science texts than specialized texts. Moreover, the gap between DeepL's and Yandex's translation performances is observed to increase from specialized texts to popular-science ones. On the one hand, DeepL seems to perform better in all error categories, except for Terminology, Theme-Rheme Pattern, and Omission, where Yandex collects a smaller amount of translation errors, even though the gap between the translation performances under examination is quite narrow. On the other hand, Transliteration and Untranslated Elements are characterized by equal translation performances. Worth noting undoubtedly are Lexis and Syntax, which display a considerable gap between Yandex's and DeepL's translation performances. The two error categories, which constitute the major weaknesses of Yandex performance, do not represent an issue in DeepL's one. Moreover, we can easily notice that the two translation tools' performances do not follow the same pattern as with regard to specialized texts. In fact, while Lexis is characterized by a considerable gap for both text types, the same does not hold for Syntax, which displays similar translation performances in translating specialized texts and significantly different amounts of translation errors with regard to popular-science ones, and Article Usage, which presents a wide gap with regard to specialized texts and a narrow one when popular-science texts are considered.

3.3.4 BLEU Metric Evaluation

In order to provide a comprehensive evaluation of DeepL's and Yandex's translation performances, we hereby conduct an automatic evaluation of the provided translations using the

BLEU metric. BLEU, which stands for Bilingual Evaluation Understudy, is a popular and widely used automatic evaluation metric (see Section 2.2), which assesses machine translation quality on the basis of its textual similarity with a number of human reference translations, also called golden translations. More specifically, it is an easily accessible platform where the users may upload the original document, the translations performed by the two machine translation systems under examination, and the corresponding reference translation, which serves as a benchmark to judge the quality of the machine translations. While the original document is optional, the other three texts are mandatory, and their absence may prevent the evaluation metric from properly completing the evaluation process. Once uploaded all the documents, a percentage, defining the quality of each translation, is released by the BLEU metric.

	DeepL's percentage of textual similarity	Yandex's percentage of textual similarity
Биологическая терапия в эру COVID-19 (The biological therapy in the COVID-19 era)	50.31	49.60
Ведение детей с заболеванием, вызванным новой коронавирусной инфекцией (SARS-CoV-2) (Clinical management of children with a disease caused by the new coronavirus infection (SARS-CoV-2))	55.12	43.46
Коронавирус SARS-Cov 2: сложности патогенеза, поиски вакцин и будущие пандемии (Coronavirus SARS-Cov 2: complexities of the pathogenesis, search for the vaccines, and future pandemics)	45.14	51.18
Коронавирус завозили в Россию не менее 67 раз (Coronavirus was imported into Russia at least 67 times)	36.66	52.88
Неизвестная летальность - Почему мы не знаем истинных масштабов COVID-19 (Unknown lethality - Why we do not know the true extent of COVID-19)	71.58	45.34
Смертность от COVID 19 - Взгляд демографа на статистику причин смерти в России и мире (COVID-19 mortality rate - A demographer's perspective on statistics of causes of death in Russia and worldwide)	76.73	52.31

Table 1: BLEU metric's evaluation

As shown in the table, the BLEU evaluation metric provides an automatic evaluation of DeepL's and Yandex's translation performances by means of a percentage, representing the textual similarity of each analysed text with a specifically made human translation, used as reference translation. This undoubtedly constitutes an interesting starting point for our discussion. Indeed, on the one hand, having provided, in the previous sections, a manual error analysis, an automatic evaluation process is hereby displayed. On the other hand, assessing the BLEU metric the provided translations' quality at the article level, DeepL's and Yandex's translation performances with regard to each specific text can be observed. Generally speaking, the BLEU evaluation metric's results seem to reflect the ones shown by our comparative error analysis. In fact, not only popular-science articles display a higher percentage of textual similarity when compared to specialized ones but also DeepL's translations are observed to reflect human reference translations' features more accurately. Moreover, while a significantly wide gap can be noticed between the percentages of textual similarity assigned to DeepL's and Yandex's translations of popular-science texts, a minor difference is observed with regard to highly specialized texts. This perfectly applies to the first two highly specialized and the last two popular-science articles. An exception is undoubtedly constituted by the third specialized and the first popular-science articles. Indeed, by showing a higher percentage of textual similarity with regard to Yandex's translation variants and a significantly wide gap when compared to the other articles belonging to the same text types, they do not follow the pattern established in our previous analysis. In particular, an interesting case is represented by the first popular-science article, Коронавирус завозили в Россию не менее 67 раз (Lvovič, 2020), whose translation performed by DeepL is assigned a rather low percentage of textual similarity when compared not only to the one given to Yadex's translation performance concerning the same article but also the ones achieved by DeepL in the translation of the other highly specialized as well as popular-science texts. Investigating the reasons behind such a low percentage of textual similarity, and the inconsistency, in this specific case, between the manual error analysis and the BLEU metric evaluation of the text under examination undoubtedly represent a necessary starting point for further research in the field of automatic Machine Translation Evaluation itself, as well as regarding the possible combination of human and automatic Machine Translation Evaluation methods.

4. Conclusion

The research hereby conducted undoubtedly reflects the development that neural machine translation systems have experienced over the last decades, reaching ever-increasingly quality standards with regard to Russian-Italian medical translation. However, the results of the comparative error analysis, as well as the BLEU metric evaluation, clearly shed light on the disparities, in terms of translation performance, of the two translation tools under examination, namely DeepL and Yandex. While DeepL provided a generally better overall translation performance, especially evident in its rendering of the context and the syntactical structure of the text at the sentence level, Yandex proved better in transliteration and was able to suggest

undoubtedly more accurate cultural-specific equivalents. One hypothesis is that this may be due to the peculiarities of the architectures used in each tool. On the one hand, the purely neural system, by analysing the source sentences as a whole, and consequently considering the overall context, is able to provide significantly human-like and reliable translations, on the other hand, a hybrid Machine Translation tool seems to benefit from the quality potential of the statistical approach, which is better at translating words and word combinations that rarely occur in the training data. Moreover, both translation tools present a number of weaknesses, especially related to the fields of syntax, lexis, and article usage, that, generally speaking, still prevent machine translation systems from providing natural sounding and accurate Russian-Italian medical translations. By investigating and applying non-automatic and automatic machine translation evaluation methods, the role of the continuous evaluation of machine translation systems' performances has been remarked as crucial to assure their growth and enhancement. Further research in the field undoubtedly is highly needed. Starting from the results obtained in the present research, a possible direction to be taken in the future may regard an in-depth linguistic study of the error categories adopted in these pages as the main criteria to judge machine translation quality, aimed at determining the relevance of each one of them against the background of Russian-Italian medical translation. By doing so, a distinction between major and minor translation errors can be properly made and immediate and effective corrective measures can consequently be proposed in order to enhance machine translation systems performances.

References

[1] Aleksandrovič, Ju.S., Bajbarina, E.N., Baranov, A.A., Višneva, E.A., Zvereva, N.N., Ivanov, D.O., Krjučko, D.S., Konovalov, I.V., Kuličenko, T.V., Lobsin, Ju.V., Mazankova, L.N., Namazova-Baranova, L.S., Petrenko, Ju.V., Prometnoj, D.V., Pšenisnov, K.V., Rtiščev, A.Ju., Sajfullin, M.A., Selimzjanova, L.R., Uskov, A.N., Fedoseenko, M.V., Harkin, A.V., Čumakova, O.V., Efendieva, K.E., Jakovlev, A.V. 2020. "Vedenie detej s zabolevaniem, vyzvannym novoj koronavirusnoj infekciej (SARS-CoV-2)." *Pediatričeskaja farmakologija* 17/2.

[2] Callison-Burch, C., Osborne, M., Koehn, P. 2006. "Re-evaluating the role of BLEU in machine translation research." In *EACL* 6: 249–256.

[3] Castilho, S., Doherty S., Gaspari, F., Moorkens, J. 2018. *Approaches to Human and Machine Translation Quality Assessment. Translation Quality Assessment. From Principles to Practice*. Springer.

[4] Chatzikoumi, E. 2020. "How to evaluate machine translation: A review of automated and human metrics." *Natural Language Engineering* 26: 137–161.

[5] *Collins English Dictionary*, collins.co.uk, HarperCollins.

[6] Guščina, L.N. 2005. "Osobennosti jazika mediciny." *Žurnal GGLLU* 2005 n.1.

[7] Hadiwinoto, C. 2017. "Syntax-Based Statistical Machine Translation" book review. *Computational Linguistics*.

[8] Hutchins J. 2014. *The history of machine translation in a nutshell*. Website
[9] Korotaeva, T.V. 2015. "Immunogennost, vyzvannaja genno-inženernymi biologičeskimi preparatami pri lečenii psoriaza i psoriatičeskovo artrita: vzljad na problemu." *FGBNU Naučno-issledovatelskij institut revmatologii im*. V.A. Nasonovoj. Moscow, Russia.
[10] Lenci, A. 2010. "Modelli distribuzionali del lessico. Modelli computazionali per l'analisi semantica." *Informatica umanistica*. https://www.ledonline.it/informatica-umanistica
[11] Lepschy, L., Lepschy, G. 1993. *La lingua italiana, Storia, varietà dell'uso, grammatica*. Milano: Bompiani.
[12] Lommel, A. 2018. *Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. Translation Quality Assessment. From Principles to Practice*. Springer.
[13] Lvovič, V. 2020. Koronavirus Zavozili v Rossiju ne menee 67 raz. IQ HSE RU https://iq.hse.ru/.
[14] Mićić, S. 2013. "Languages of medicine – present and future." *JAHR* 4.7: 217 –233.
[15] Monti, J., Montella, C. 2015. "About adequacy, equivalence, and translatability in human and Machine Translation." In *Conference: New Horizons in Translation and Interpreting Studies*.
[16] Munkova, D., Hajek, P., Munk, M., Skalka, J. 2020. "Evaluation of Machine Translation Quality through the Metrics of Error Rate and Accuracy." In *Third International Conference on Computing and Network Communications (CoCoNet'19)*. Amsterdam: Elsevier.
[17] Namazova-Baranova, L.S., Myraškin, N.N., Ivanov, R.A. 2020. "Biologičeskaja terapija v eru COVID-19." *Voprosy sovremennoj pediatrii* 19/2.
[18] Osimo, B. 2004. *Traduzione e qualità. La valutazione in ambito accademico e professionale*. Milano: Hoepli.
[19] Papineni, K., Roukos, S., Ward, T., Zhu W.J. 2002. "BLEU: a Method for Automatic Evaluation of Machine Translation." *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*: 311-318.
[20] Papineni, K., Roukos, S., Ward, T., Henderson, J., Reeder, F. 2002. "Corpus-based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French, and Spanish Results" in *Proceedings of Human Language Technology* 2002: 132–137.
[21] Polackova, G. 2001. "Synonymy of medical terminology from the point of view of comparative linguistics." *Bratisl Lek Listy* 102 (3):174-177.
[22] Popovic´, M. 2018. *Error Classification and Analysis for Machine Translation Quality Assessment. Translation Quality Assessment From Principles to Practice*. Springer.
[23] Salmanova, S. 2020. "Neizvestnaja letalnost. Počemu my ne znaem istinnyh masštabov COVID-19." IQ HSE RU https://iq.hse.ru/
[24] Scarpa, F. 2008. *La traduzione specializzata. Un approccio didattico professionale*. Milano: Hoepli.
[25] Tiberii, P. 2018. Dizionario delle collocazioni. Le combinazioni delle parole in italiano. Zanichelli.

[26] Timonin, S. 2020. Smertnost ot COVID-19. Vzgljad demografa na statistiku pričin smerti v Rossii i mire. IQ HSE RU https://iq.hse.ru/. Accessed on October, 15th, 2020.
[27] Van Merrienboer, K.C.B., Bahdanau, D., Bengio, Y. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. SSST@EMNLP.
[28] Wang, R., Finch, A., Utiyama, M., and Sumita E. 2017. Sentence Embedding for Neural Machine Translation Domain Adaptation. Conference paper, In Proceedings of the Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada.
[29] Wołk, K., Marasek, K. 2015. Neural-based machine translation for medical text domain. Based on European Medicines Agency leaflet texts. Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist.
[30] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation, arXiv preprint arXiv:1609.08144.
[31] Xarčenko, E.P. 2020. Koronavirus SARS-Cov-2: složnosti patogeneza, poiski vakcin i buduščie pandemii. Epidemiologija I Vakcinoprofilaktika, 19/3.

Websites

1. Academic.ru https://academic.ru/
2. DeepL GmbH. https://www.linguee.de/
3. One model is better than two. Yandex.Translate launches a hybrid translation system. Yandex company. https://yandex.com/company/blog/one-model-is-better-than-two-yu-yandex-translate-launches-a-hybrid-machine-translation-system/
4. Russian Academy of sciences - Institute for information transmission problems. http://iitp.ru/en/about
5. Traduttore. DeepL Translate. https://www.deepl.com/translator
6. Treccani. Enciclopedia online
7. Yandex company. https://yandex.com/company
8. Zingarelli N. 2020. Lo Zingarelli 2021. Vocabolario della lingua italiana. Zanichelli, online resource

Last access: 8th June 2021.