Hindawi
Complexity
Volume 2021, Article ID 2806108, 15 pages
https://doi.org/10.1155/2021/2806108

Research Article

MT Evaluation in the Context of Language Complexity

Dasa Munkova,[1] **Michal Munk**,[1] **Ľubomír Benko**,[1] **and Jiri Stastny**[2,3]

[1]*Department of Computer Science, Constantine the Philosopher University, Nitra, SK-949 01, Slovakia*
[2]*Institute of Automation and Computer Science, Brno University of Technology, Brno, CZ-619 69, Czech Republic*
[3]*Department of Informatics, Mendel University in Brno, Brno, CZ-613 00, Czech Republic*

Correspondence should be addressed to Ľubomír Benko; lbenko@ukf.sk

Received 18 May 2021; Revised 3 November 2021; Accepted 1 December 2021; Published 17 December 2021

Academic Editor: Wen-Long Shang

The paper focuses on investigating the impact of artificial agent (machine translator) on human agent (posteditor) using a proposed methodology, which is based on language complexity measures, POS tags, frequent tagsets, association rules, and their summarization. We examine this impact from the point of view of language complexity in terms of word and sentence structure. By the proposed methodology, we analyzed 24 733 tags of English to Slovak translations of technical texts, corresponding to the output of two MT systems (Google Translate and the European Commission's MT tool). We used both manual (adequacy and fluency) and semiautomatic (HTER metric) MT evaluation measures as the criteria for validity. We show that the proposed methodology is valid based on the evaluation of frequent tagsets and rules of MT outputs produced by Google Translate or of the European Commission's MT tool, and both postedited MT (PEMT) outputs using baseline methods. Our results have also shown that PEMT output produced by Google Translate is characterized by more frequent tagsets such as verbs in the infinitive with modal verbs compared to its MT output, which is characterized by masculine, inanimate nouns in locative of singular. In the MT output, produced by the European Commission's MT tool, the most frequent tagset was verbs in the infinitive compared to its postedited MT output, where verbs in imperative and the second person of plural occurred. These findings are also obtained from the use of the proposed methodology for MT evaluation. The contribution of the proposed methodology is an identification of systematic not random errors. Additionally, the study can also serve as information for optimizing the translation process using postediting.

1. Introduction

Tasks play a crucial role in human behaviour or performance. Liu and Li ([1], p. 553) stated that human behaviour and/or performance depend on the interaction among task characteristics (such as complexity, which has a significant influence on behaviour and predicting human performance), task performer characteristics (such as performer's competencies), and environment characteristics. When the task is translation, especially machine translation (MT) or postediting of MT output (PE), one of the essential task characteristics is a complexity of MT output or postedited MT output (PEMT). Complexity is an intrinsic property (qualitative) of a translation task, which is given by internal textual structure and represents the objective characteristic of a task ([2], p. 2). Liu and Li ([1], p. 559) define task complexity as the aggregation of any intrinsic task characteristic that influences the performance of a task (translation). Complexity ([3], p. 40) can be understood as (a) primarily a psychological experience (purely subjective psychological approach), (b) an interaction between task and person characteristics (tasks are more or less complex relative to the capabilities of the individuals who perform the task), and (c) a function of objective task characteristics (in terms of objective task qualities).

Complex tasks are characterized [3, 4] (a) by unknown or uncertain alternatives of action (there is not only one alternative of translation; multiple outcomes), (b) by inexact or unknown means-ends connections (there are many ways to express the same meaning in translation; multiple potential ways), and (c) by the existence of a number of subtasks which may or may not be easily factored into nearly

independent parts (analysis of source text and analysis of target text in terms of intratextual and extratextual factors; conflicting interdependence among ways to outcomes, and also uncertain or probabilistic links among ways and outcomes). These objective task qualities contribute to complexity which placed high demands on the translator and also on the posteditor. Linguistic complexity contributes not only to the difficulty in solving word problems but also to the difficulty in translation tasks ([5], p. 1). There are many factors hidden by word complexity (e.g., patterns, grammatical constructions, lexicon, or physical circumstances in the given language); therefore, there is no general definition of linguistic complexity. Ma and Wang ([6], p. 3) state qualitative characteristics of language complexity such as uncertainty, incompleteness, sensitivity to initial conditions, dynamicity, nonlinearity, instability, path dependency, openness, and adaptivity, while the core of language complexity is nonlinearity. Nonlinearity consists of imbalance, emergence, and interactivity features ([6], p. 5), which means that expression of language patterns or their combinations deviates from linearity ([7], p. 53–54). Besides the qualitative characteristics of language complexity, there are also quantitative characteristics such as high organizational depth features, since quality and quantity are a pair of interdependent contradictions ([6], p. 13). In terms of language systems, high organizational depth refers to multilevel, which is an essential way to organize complex systems [8, 9]. Levels of language are used to understand the language complexity in oral or written form [10]. Ma and Wang ([6], p. 26) state that the higher the language complexity, the longer the minimum description length of the information, and the greater the resource/cost consumption.

Only the use of computer programming technology, e.g., oral or written corpora, to analyze and study language complexity will help improve the ability of language processing [11]. Complexity of text can be measured at word- or sentence-level using corpus-based methods—readability or lexico-grammatical features: word/sentence length and frequency of part of speech [12].

Complexity of text features always implies independent variables, i.e., textual elements (word-level, sentence-level, and discourse-level variables) that can be examined and analyzed ([13], p. 236). Based on the complexity of textual elements (words, syntax, and discourse structure), we can examine performance, such as translation or PE tasks.

1.1. Evaluation of MT System. The invention of neural machine translation (NMT) brought several fundamental changes to the translation industry, both from the point of view of the translation process or task [14] and of the business model [15]. Current NMT systems provide fluent translations of fairly good quality [16], but often, this fluency is at the expense of accuracy or intelligibility [17]. NMT and its predecessor, statistical MT (SMT), were commonly used not only for personal use but also to reduce the cost of translation in the translation industry for many years. NMT and SMT operate on a statistical basis using a corpus-based approach to MT. NMT took a huge leap forward when sequence-to-sequence models were first introduced [18]. So far, it has already achieved excellent performance on a great volume of translations from English into French [12, 19] as well as from English into German [20]. MT systems for translating dialectal sentences into their standard language form can be of great benefit as Farhan et al. [18] have shown. Translation technologies have become an integral part of the translator's work, so it is very important to know what machines can and, conversely, what they cannot adequately translate. They also serve to alert users to errors that occur in MT [21].

MT system is a complex natural language processing system composed of a large number of heterogeneous modules [22]. MT systems, like language, can be considered a complex adaptive system, which involves multiple agents (both natural and artificial) interacting with one another to achieve a common goal—a translation task ([23], p. 261). Since a natural language is not static but dynamic, i.e., previous behaviour influences the current and future behaviour of natural language, it can be considered as a complex adaptive system [23]. Complex systems have shifted from reductionist analyses of component parts (agents) and simple linear change to the study of interconnected elements ([24], p. 2).

The behaviour of agents in the language system is influenced by several elements at different levels, whether internal or external (locutionary, illocutionary, and perlocutionary acts). If we add a translation task as an act of communication, it brings a new level of complexity—a metalevel. The complexity of the system together with translation errors increases, since at least two languages are considered in the translation process—both source and target languages. Systems may differ from each other not because of differences in their features, but because of differences in how these features depend on and affect one another ([25], p. 2). Siegenfeld and Bar-Yam [25] liken it to steam and ice, both of which are made of the same water molecules but have different properties due to differences in the interactions between molecules, like MT output and its postedited MT output (PEMT). Both translations come from the same original, but due to differences in interactions, they have different properties, i.e., different translation qualities.

In MT systems, at the locutionary level, a language model of the source text is created using a neural network, i.e., the neural network is trained on a large amount of source text data. It creates patterns and identifies grammatical constructions and a lexicon to be able to assign corresponding patterns to them in the target language, and the same is done in the target language, in which another neural network is also trained. The key concept behind MT is to capture the linguistic knowledge of the locutionary level of the languages involved by means of translation pairs linking constructions across languages ([23], p. 269). The illocutionary level lies in the transfer itself, from source to target language. At the perlocutionary level, it is the quality of MT output from the MT system that plays a key role in the given communication.

Progress in MT depends on the results of the evaluation of MT quality. NMT, as a metric to evaluate the development of artificial intelligence, plays a crucial role in the current

natural language processing (NLP) community [26]. Many experts [27, 28] have sought various ways to assess MT quality, whether in the form of manual evaluation, automatic evaluation or both, in the form of a framework (e.g., dynamic quality framework or multidimensional quality metrics).

MT output can be evaluated manually or automatically with intrinsic and extrinsic methods applied [29]. Castilho et al. [30] distinguish manual methods according to six criteria: (1) adequacy and fluency; (2) readability and comprehensibility; (3) acceptability; (4) ranking; (5) usability and performance; and (6) evaluators. Adequacy and fluency are the most commonly used measures in translation assessment [31]. Alongside the standard criteria of fluency and adequacy, some researchers have focused on examining the linguistic features of a text, specifically on the identification of differences among original text and different translation outputs, i.e., human translation (HT), MT output, or PEMT output [32–34]. Methods are typically based on linguistic (e.g., word frequency) and extralinguistic features (formatting) [30]. Vanmassenhove et al. [34] have shown (for translation directions—English into French or Spanish) that MT texts contain lesser lexical variety compared to their source English texts or compared to their human translations in French or Spanish. Loock [12] has shown that the linguistic characteristics of MT texts from English into French differ from the original French texts.

Intrinsic methods involve comparisons of translation quality between MT output and reference (high-quality HT) or a fixed set of references. Manual intrinsic measures determine MT quality through human subjective judgments such as fluency and adequacy. The biggest issue that manual intrinsic methods face is their subjectivity and nonreproducibility, apart from their price and timeliness. Automatic intrinsic measures, such as ranking, compute sentence similarity among MT outputs and a fixed set of references to produce rankings among MT systems [35]. Unlike intrinsic measures which are focused on accuracy and text coherence, extrinsic methods focus on the effectiveness or usability of MT output in terms of the specific task such as PE [36–38]. PE as a specific task directly assesses the MT output in terms of the time and effort needed to correct the MT. It provides information about difficulty, but it does not provide sufficient information about task characteristics such as linguistic complexity. PE is a result of the linguistic complexity of MT output, which is related to posteditor-task interaction. A closer measure than time, which is related to linguistic complexity, is edit distance (error rate). It represents the number of changes within the sentence including insertion, deletion, substitution, or shifts, which required some correction of the MT output.

Availability of reference translations allows us to use not only manual evaluation methods but also measures of automatic evaluation to evaluate the translation quality [28]. Automatic MT evaluation measures provide quick feedback on translation quality, but this feedback is only a score. According to the criterion of lexical concordance, we divided them into automatic metrics of accuracy and metrics of error rate [39]. Metrics of accuracy are based on the closeness of the MT output/hypothesis (h) with the reference (r) in terms of n-grams. They calculate their lexical overlap in (A) the number of common words ($h \cap r$), (B) the length (number of words) of MT output, and (C) the length (number of words) of the reference. The higher the values of these metrics, the higher the translation quality [40]. Metrics of error rate are based on edit distance. They calculate the Levenshtein distance between an MT output/hypothesis (h) and a reference/human translation (r). The higher the values of these metrics, the lower the translation quality [39].

Automatic measures are a good objective indicator of how to improve system performance and are cheap and achieve more consistent results compared to manual. However, their main drawback is that they are not able to sufficiently assess the syntactic and semantic equivalence of translation (linguistic complexity). We are not able to perform a deeper linguistic analysis. Besides the overall scores, it is helpful to have additional information, i.e., the strengths and weaknesses of the system or types of MT errors [28]. Another problem with automatic measures is that its metrics operate mainly at the sentence/segment level and not at the document level, and they do not take context into account when assessing translation quality [30].

1.2. Error Analysis in the Context of NLP and MT Evaluation. According to Popović [28], error analysis and classification provide the basis for determining what type of errors are produced by the system and whether and how they can be eliminated. It can be carried out not only by classification and annotation of erroneous words but also by analyzing words or parts of speech (POS). In the translation industry, the evaluation usually relies on error analysis [30, 41]. Error analysis offers a number of answers to improve the system, better understanding of human or artificial agent behaviour or performance such as translation or PE tasks. However, it is time-consuming and requires extensive knowledge of annotator(s). Feng et al. [26] showed that the performance of an NMT system benefits from POS tag information of target language (Chinese-English and German-English translation datasets). POS tag is more informative and concise than combinatory categorial grammar (CCG) supertag [42]. Loock [12] showed how a linguistic analysis of a corpus of MT texts can also be used in translator education. Hládek et al. [43] aimed at the present alternative view of the task of morphological tagging and focused on Slovak. They proposed a rule-based system using expert knowledge. The system generates an outcome based on the rule that a certain tag was chosen from the match set. They summarized the whole decision process into three phases (matching, maximization, and minimization). The rules were created using the learning process and then were pruned for more specific rules offering better accuracy. They compared their proposed algorithm with the morphological tagger HunPos [44]. Laki et al. [45] presented a novel universal morphological feature schema as a set of features expressed by inflectional morphology across languages. They examined the variability of inflectional morphology by comparing multiple translations of the same source (the Bible). The results

showed that the schema offers potential benefits for NLP and MT by facilitating direct meaning-to-meaning translation between the language pairs, regardless of form-related differences.

It motivated us to apply POS tagging to determine the error rate and linguistic complexity. Just as word and sentence, POS tagging is implemented in text analysis, and it can also be used to compare two texts (MT output and PEMT output) or to determine the linguistic complexity through the quality of MT output, PEMT output, or HT.

1.3. Research Objectives. The study of multiagent behaviour motivated us in our research, in which we focus on the influence of the agent-machine translator on the behaviour of the human agent-posteditor within one complex adaptive system. In other words, we identify the behaviour of the agent-machine translator using POS tagging and association rules found and then identify its influence on the behaviour of the agent-human posteditor, whose task is to achieve the perlocutionary level of natural language, i.e., to postedit MT output to be both fluent and adequate. We focus on examining the behaviour and/or performance of an artificial agent and a human agent, which depends on the interaction between tasks characteristics and task performer characteristics from the point of view of language complexity. We investigate the translation task through language complexity, which is defined by frequent tagsets and rules.

The aim of the study is to present a new approach to the evaluation of MT quality and subsequently to validate the proposed MT evaluation methodology. The proposed methodology is based on the evaluation of frequent MT and PEMT tagsets and also on frequent POS tagsets and rules summarization. The aim consists of two consecutive objectives.

The first objective comprises three tasks. The first is to analyze MT outputs from two MT systems: Google Translate and the European Commission's MT tool as well as their postedited MT outputs based on POS tags in terms of task characteristics and language complexity at word- and sentence-level. The second is to examine the relationships between individual tags and tagsets within the four examined translations (as described in Section 3.1). The last task focuses on the comparison of translation quality based on the summarization of the incidence of frequent tagsets and rules (as described in Section 3.2).

We examine the extent to which the MT quality in terms of language complexity is identical to its PEMT version based on the frequency of tagsets and rules.

For this study, we have set as null hypotheses:

H01: the incidence of frequent tagsets does not depend on the method of translation (machine translation vs. postediting)

H02: the incidence of extracted rules does not depend on the method of translation (machine translation vs. postediting)

The second research objective is to validate the proposed methodology of MT evaluation using baseline methods. We used both manual and semiautomatic MT evaluation measures as criteria for validity (as described in Section 3.3).

1.4. Implications and Limitations. The study offers new insight into the evaluation of MT quality. The results and findings of the research offer one key theoretical contribution and two practical contributions to the field of complex adaptive systems, including MT evaluation.

The theoretical contribution consists of the design and verification of a novel methodology for evaluating MT quality in the context of inflectional languages. The proposed and verified methodology is unique, combining the advantages of using both intrinsic and extrinsic methods, focusing on translation into the inflectional language, which is characterized by a rich morphology and free word order. It analyzes and subsequently compares the translation quality, based on text complexity, i.e., based on the frequent tags and rules and their quantitative evaluation—summarizing the frequent tags and rules incidence. The proposed methodology allows us to identify the complexity of MT outputs, especially errors that are systematic and not random. The principle of the proposed methodology is applicable to any language pair as well as translation directions, but it is necessary to take into account the character of the target language when determining tags and/or part-of-speech tagging. For instance, declension is typical for inflectional languages such as Slovak but not for analytical languages like English, i.e., Slovak uses suffixes for grammatical cases, in contrast to English, in which cases are expressed by prepositions.

Nominative: *auto* (SK)—*a car* (EN)

Genitive: *auta* (SK)—*from a car* (EN)

Dative: *autu* (SK)—*to a car* (EN)

Accusative: *auto* (SK)—*a car* (EN)

Locative: *aute* (SK)—*about a car* (EN)

Instrumental: *autom* (SK)—*with a car* (EN)

We were inspired by the research of Conforti et al. [46] focusing on machine translation to morphologically rich language using POS tagging. We adopted a similar approach to assessing the quality of MT output but using text complexity from the perspective of word and sentence structure. Callison-Burch et al. [47] or Popović [28, 48, 49] have shown that metrics based on POS analysis correlate very well with human evaluation. Popović [50] provided a useful approach for quality estimation based on morphemes and POS tags. The proposed methodology can also be used to evaluate students' translation performance within their translation education or in language learning.

From a practical point of view, the findings offer a closer understanding of the text complexity of MT outputs, i.e., they allow us to reveal the linguistic features of MT texts from an analytical into a synthetic language. The second practical contribution, which follows on from the first, consists in the identification of "machine translationese" [12], what kind of translation task the machine can and cannot do correctly for the given direction of translation and the genre of the text.

The research also has certain limitations in the aspect that (a) the examined texts are not extensive and come from one genre (technical documentation), as well as the posteditor himself/herself, who has subjective sensitivity to errors within the text. However, in the evaluation, specifically, when assessing the adequacy and fluency of MT and subsequently when postediting the MT outputs, a large amount of manual work is required. In our case, it was done by students and translators during one day. Human evaluation is time- and labour-consuming, but it is considered highly reliable. For this reason, it is sometimes better to have a smaller dataset, but with more reliable data. We are working on expanding the dataset, but we are faced with the problem of evaluators' consistencies, as not all participants wanted to continue in the research (to repeat the same procedure with different genres and translation direction or different source language). (b) We focused only on the influence of an artificial agent's behaviour (MT system) on a human agent's behaviour (PE) using a word and sentence complexity. We did not consider the influence of a human agent's behaviour (preediting) on an artificial agent's behaviour (MT system) and subsequently its influence on the human agent' behaviour (PE). For this reason, we want to focus our future work on text volume as well as the diversity of genres, consistency of posteditors, and also on preediting.

The structure of the paper is as follows. Section 2 describes the research methodology, and the subsequent section focuses on the research results based on the association rules analysis and aims at the validation of the proposed methodology for MT evaluation. The penultimate section offers a discussion of the results. The last section comprises research conclusions.

2. Materials and Method

We examined unstructured textual data, namely technical texts—consisting of 606 sentences (more than 6 000 tags). The source texts (ST) written in English were translated into Slovak by two MT systems/engines—Google Translate (GT) and MT@EC.

For our research, the most important step was tokenization, which was done after sentence alignment, since we analyzed two MT engines (MT systems). We also used the TreeTagger tool for tokenization, developed by Schmid [51–53]. It supports morphological annotation of the Slovak language and automatically annotates Slovak texts with POS tagging and lemma information [54].

2.1. Proposed Approach. The applied methodology includes the following stages (in Figure 1):

(1) Acquisition of unstructured textual data: source texts (technical texts)

(2) Data preparation: it consists of multiple tasks:

  (a) Machine translation: translation of the source texts using both MT engines

  (b) Sentence alignment: the generated MT output is aligned with the source text based on the 1-to-1 principle

  (c) Postediting: the MT output is postedited by professional translators and students in M.A. degree

  (d) Evaluation: each MT sentence of both MT outputs is assessed by participants using the scale of fluency and adequacy (scale range is from 1 to 5)

  (e) POS tagging: the MT output and PEMT output are tokenized separately, which generates the tags and lemmas for annotated aligned words (see Supplementary Table 1 for more details on Slovak POS tags)

(3) Data analysis consists of searching the frequent POS tags (tagsets) of MT output (MT@EC_MT or GT_MT) and PEMT output in the examined text. The results were processed by association rule analysis using STATISTICA Sequence, Association, & Link Analysis, which is an implementation of the algorithm using apriori algorithm together with a tree-structured procedure that requires only one pass through data. The support for a tagset is given by a proportion of records in the transactions data set that have the tagset, i.e., for a tagset (A), the support can be calculated as follows:

$$\text{support}(A) = \frac{\text{frequency of } (A)}{\text{number of transactions in the dataset}} * 100. \quad (1)$$

Lift of rules can be similarly calculated. Based on support and confidence, a lift for a rule can be defined and computed (A–tagset, C–tagset)

$$\text{lift}(\textit{if } A \text{ then } C) = \frac{\text{confidence}(\text{if } A \text{ then } C)}{\text{support}(C)}, \quad (2)$$

where

$$\text{confidence}(\text{if } A \text{ then } C) = \frac{\text{support}(\text{if } A \text{ then } C)}{\text{support}(A)} * 100. \quad (3)$$

We focused on frequent tagsets extracted with the minimum support of 10%.

(4) Data understanding based on the results of association rule analysis.

(5) Comparison of found rules and frequent tagsets in examined translations.

We will validate the proposed methodology of MT evaluation, which is based on the evaluation of frequent MT and PEMT tagsets, by manual and semiautomatic MT evaluation.

2.2. Manual and Semiautomatic MT Measures. Adequacy, manual MT measure, represents the extent to which the translation transfers the meaning of the source text into the target language. Fluency, manual MT measure, represents

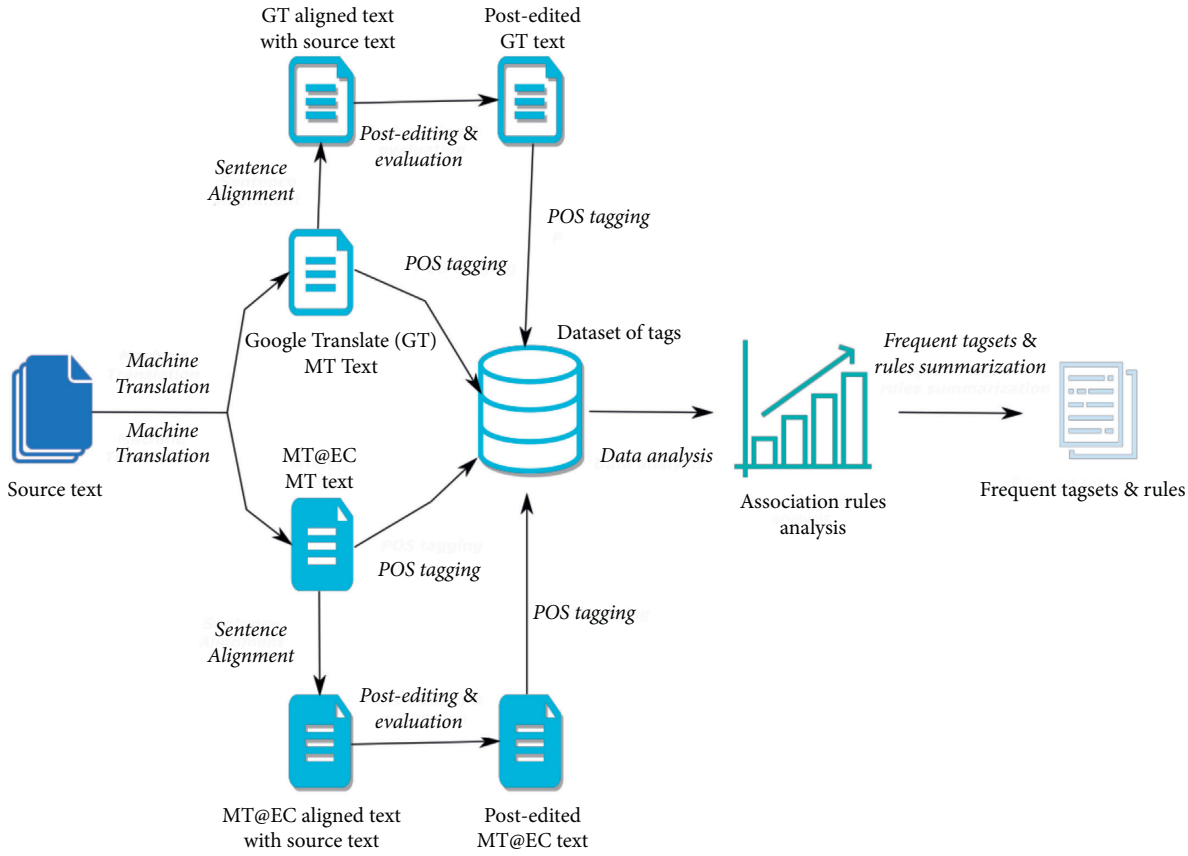

Figure 1: The proposed methodology.

the extent to which the translation follows the norms of the target language. Both measures assess the translation using a five-point Likert scale for each segment, where 1 means "none (adequacy)/incomprehensible (fluency)," 2 means "little meaning/disfluent Slovak," 3 means "much meaning/nonnative Slovak," 4 means "most meaning/good Slovak," and 5 is "all meaning (adequacy)/flawless Slovak (fluency)."

HTER (human-targeted translation error rate) [55] is a more complex semiautomatic MT measure; humans do not score translations directly but rather generate a new reference translation (PEMT) that is closer to the MT output but retains the fluency and meaning of the original reference ([56], p. 259). Specifically, HTER = # of edits (substitutions + insertions + deletions + shifts)/# of reference words.

2.3. Dataset. PE task and evaluation (fluency and adequacy) were conducted in the OSTPERE system [57]. MT outputs were postedited by professional translators and students of translation studies at the master's degree level (extrinsic method). The human translators also assessed each MT sentence using the scale of adequacy and fluency (intrinsic method). Due to the time and laborious complexity of the manual evaluation of translation quality, our dataset is not extensive but specialized on one text type. The data provide the possibility to perform more precise analyzes (e.g., linguistic analysis) for one specific domain. Data were obtained during a one-day workshop in order to keep the consistency of posteditors and evaluators (the average translator translates a maximum of 10 standard pages per day).

The final dataset consists of 24 733 tags: MT outputs (translated by GT and MT@EC) and their corresponding PEMT outputs. Subsequently, we created a program, using C#, to calculate the HTER scores for each MT sentence. Based on the source sentence ID, the corresponding HTER score and scores of adequacy and fluency were merged into a single data matrix. The data matrix was used to create a baseline for analysis.

Each sentence was annotated by the TreeTagger tool. Tokenization produced four files (GT_MT, GT_PEMT, MT@EC_MT, and MT@EC_PEMT) containing annotated tags. The composition of each file (GT_MT, GT_PEMT, MT@EC_MT, and MT@EC_PEMT) is depicted based on two feature types (Table 1) where each file consisted of approximately 6000 tags (including interpunction). It was necessary to adjust the files before merging them into a single data matrix because we wanted to keep a record of each sentence (ID), and the tool works only with text files. For this task, a simple JAVA program was created. This allows us to create a single data matrix incorporating all four files with the corresponding tags and also to compare both translations. A transaction/sequence model [58] was used for text representation. The results were processed by the association rule analysis. Found rules and frequent tagsets were summarized by a Cochran Q test and using multiple comparisons.

Table 1: Dataset composition.

Feature type	Feature name	GT_MT	GT_PEMT	MT@EC_MT	MT@EC_PEMT
Readability	Average sentence length (words)	8.45	8.81	7.82	8.75
	Average word length (characters)	5.51	5.89	5.70	5.89
	Number of short sentences ($n < 10$)	63.37%	56.44%	67.49%	58.35%
	Number of long sentences ($n \geq 10$)	36.63%	43.56%	32.51%	41.65%
Lexico-grammatical	Frequency of nouns	32.81%	36.82%	31.25%	36.23%
	Frequency of adjectives	8.22%	9.00%	11.81%	9.05%
	Frequency of adverbs	3.16%	2.70%	3.11%	2.73%
	Frequency of verbs	16.57%	16.11%	15.00%	15.87%
	Frequency of pronominals	3.13%	3.03%	3.02%	3.27%
	Frequency of participles	1.63%	1.89%	1.62%	1.97%
	Frequency of morphemes	1.45%	1.32%	1.34%	1.36%
	Frequency of abbreviation	3.01%	2.40%	3.94%	2.44%
	Frequency of numbers	3.87%	3.65%	4.32%	3.53%
	Frequency of undefinable POSs	0.29%	0.23%	1.06%	0.34%
	Frequency of foreign words	6.98%	4.84%	6.02%	5.14%
	Frequency of interjections	0.02%	0.02%	0.02%	0.02%
	Frequency of numerals	0.75%	0.49%	0.70%	0.42%
	Frequency of prepositions & conjunctions	18.10%	17.49%	16.80%	17.63%

3. Results

The section Results is divided into two subsections: the first describes the identified relations between tagsets, and the second represents their quantitative summarization.

3.1. Identification of Relations between Tagsets. The association rule analysis represents a nonsequential approach to the data being analyzed. We will not analyze the sequences but transactions, so we will not include the tag order in the analysis. In our case, a transaction represents a set of tags observed in the MT sentence.

The web graphs (in Figures 2 and 3) depict the discovered association rules for the sentences, namely, the size of a node represents a support of the tag, the thickness of the line represents the support of rule—a pair of tags, and darkness of the line colour represents a lift of the rule.

In the GT_MT output (in Figure 2(a), see also Supplementary Table 2(a) for detailed analysis), the tag (*O*), conjunctions, belongs to the tags with the highest incidence within the text with almost 50% of the support and the tag (%), foreign language citation, with a probability of more than 35%. Other very frequent tags, after conjunctions and foreign language citation (untranslated or domesticated), with less probability of incidence (around 20%) were (*VMdpb+*), i.e., verb in imperative, perfective aspect, second person of plural in the affirmative (stlačte/press, pripojte/connect, vyberte/select, použite/use), and (*Eu4*), i.e., nonvocalized preposition in accusative (na/on, to/k, pre/for), which were tied with substantive in accusative whether in the masculine, inanimate gender (*SSns4*), or in the neuter (SSis4). Furthermore, the verbs in infinitive (*VId+*) were observed (spojiť/connect). The other identified tags (not depicted in Figure 2(a), see Supplementary Table 2(a)) do not meet the minimum support, i.e., the likelihood of occurrence in the identified sentences (transactions) is less than 10% (see Supplementary Table 2). Among the most found pairs (in Figure 2(a), see also Supplementary Table 2(a)), a pair of the tags in the sentence belong (*O*, *VMdpb+*), (*O*, *VKepb+*), and (*%*, *O*) with more than 17% of the *support*, i.e., conjunctions with verbs in imperative or present, in plural, and in affirmative. Subsequently, conjunctions with foreign language citations (použite kábel HDMI alebo ultra HD) use the cord HDMI or ultra HD.

Another large group of pairs, with the probability of around 15%, were (*SSis4*, *O*)—a noun in the inanimate masculine gender or in the neuter, singular, in accusative with conjunctions and also pair (*O*, *VId+*)—conjunctions with verbs in infinitive or infinitives with verbs in present, in the second person of plural (môžete poškodiť modul CAM a televízor/you can damage module CAM and TV, zvoľte a stlačte tlačidlo/select and press the button). Tags not presented in the analysis do not meet the minimum support and confidence of 10%; that is, these tags are identified in sentences with a probability of less than 10% (Figure 2).

The greatest degree of positive correlation (*lift* = 5.11) was identified by the (*SSis6, Eu6*) pair (in Figure 2(a), see also Supplementary Table 2(a)). Lift, in case (SSis6, Eu6), indicates a certain rule, i.e., substantives in the inanimate masculine gender in singular, locative case are tied with nonvocalized prepositions in locative (v prípade/in case, na televízore/on TV), less in case (SSis4, Eu4), where substantives in inanimate masculine or neuter gender in singular, accusative case are tied with nonvocalized prepositions in accusative (na nastavenie/for setting, pre vstup/to enter). Similarly, a greater degree of positive correlation (lift = 3.5) was reached for the pairs (*VId+*, *VKepb+*), i.e., verbs in imperative are tied with verbs in the present, in the second person of plural (môžete pripojiť/ you can connect). Tag pairs (*SSns4, Eu4*), (*SSis4, Eu4*), and (*VKepb+, O*) reached also a positive correlation (*lift* = 2). The remaining pairs, apart from the pair (%, O), achieved the lift degree higher than 1 (see Supplementary Table 2).

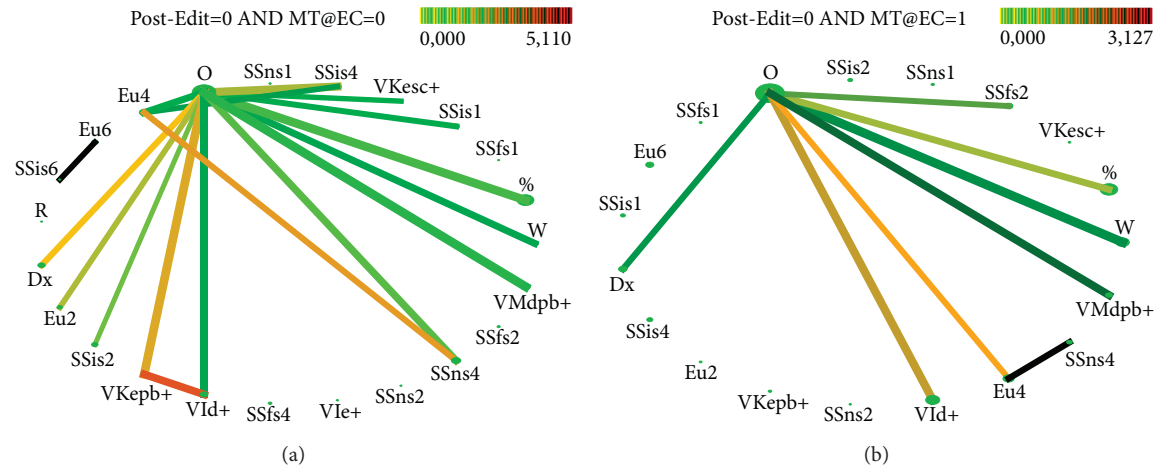

Figure 2: Visualization of tags identified in MT output translated by GT (a) and MT@EC (b).

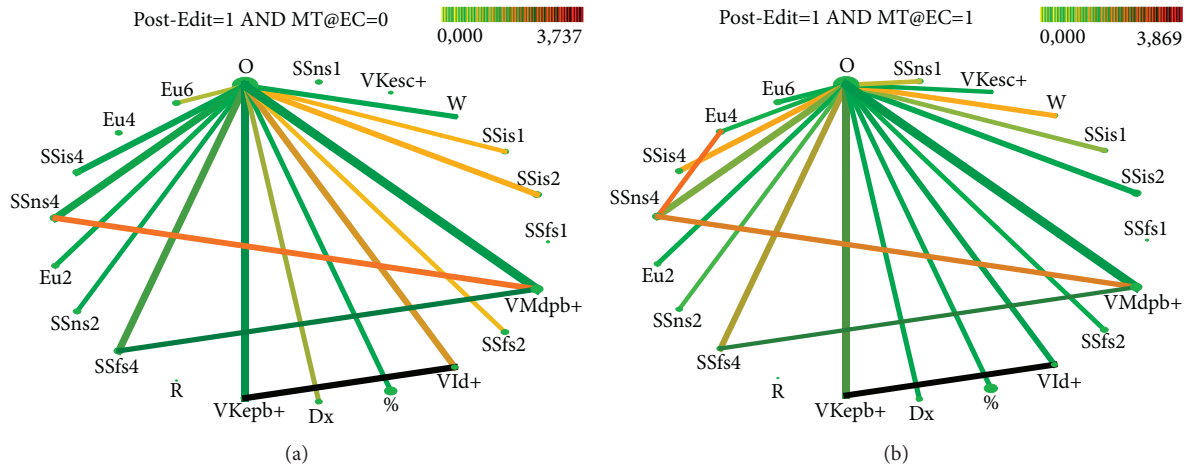

Figure 3: Visualization of tags identified in PEMT output translated by GT (a) and MT@EC (b).

For MT@EC_MT output (in Figure 2(b), see also Supplementary Table 2(b)), tag (*O*), conjunctions, was identified as the tag with the highest incidence in the sentence with the *support* of 45%. Similar to GT_MT output, the tag (*%*), foreign language citation has occurred in a sentence with a probability of more than 30%. The tags, with the probability of more than 10%, are (*VId+*) verbs in the affirmative and infinitive (*pozrieť/to see, poslať/to send*), (*W*) abbreviations (*DVD, HD*), (*Eu4*) prepositions in accusative and locative (*na/on, v/in*), and what is quite interesting, (*SSfs2*) nouns in feminine and in singular genitive cases (*potreby/the need*). The other tags did not again meet the minimum *support* of 10% (see Supplementary Table 2). MT@EC_MT output (in Figure 2(b), see also Supplementary Table 2(b)), with the probability of around 15%, contained mostly the following combinations—(*W, O*), (*O, VId+*), and (*O, VMdpb+*), i.e., abbreviations with conjunctions (DVD a/DVD and), and conjunctions with imperative or with imperative in the second person of plural (zaznamenať/to record, použite/use). It is most unexpected that nouns are not represented here. The rest of tag pairs (*%, O*), (*O, Dx*), (*Eu4, SSns4*), (*O, Eu4*), and (*SSfs2, O*) were identified with a probability of more than 10%. Tags with support of less than 10% are not depicted (in Figure 2(b), see also Supplementary Table 2(b)). For the pair (*SSns4, Eu4*), substantive in neuter gender, in singular, in accusative with the preposition in accusative, the greatest degree of positive correlation ($lift = 3.12$) was found (in Figure 2(b), see also Supplementary Table 2(b)). Slightly less positive correlation ($lift = 1.8$) was reached for (*VMdpb+, O*), imperative in the second person in plural and preposition (vyberte a/select and). Remaining tag pairs, apart from the pair (%, O), achieved the lift degree higher than 1 (see Supplementary Table 2).

In the case of PEMT outputs, both are very similar, because they are translations of the same source texts regardless of the MT system used. The only factor that plays a key role is the posteditor, i.e., the extent of his/her intervention and his/her lexical and stylistic preferences in postediting. The evidence lies in the very similar rules found in the PEMT outputs (in Figures 3(a) and 3(b), see also Supplementary Table 3).

The tag (*O*), conjunctions, also belonged to the tags with the highest occurrence with a probability of 50%. Tags (%), (VMdpb+), and (SSns4) were imperatives in the second person of plural, occurred in the sentences with the support of around 25%. Pairs of tags (*O, VMdpb+*), (*O, VKepb+*), and (*O, SSns4*) were identified with the *support* of around 20% in GT_PEMT output (in Figure 3(a), see also Supplementary Table 3(a)). In the case of MT@EC_PEMT output (in Figure 3(b), see also Supplementary Table 3(b)), (O, VMdpb+) with more than 20% of the support and (O, VKepb+) or (O, SSns4) with a probability of around 15% were found.

Based on the lift, we can claim that PEMT outputs are characterized by more frequent pairs (VId+, VKepb+)—verb in infinitive with modal verb (môžete použiť/can use) or a verb itself (použite/use). The highest interestingness of rule was found for the pair of tags (VKepb+, VId+) with lift = 3.74 in GT_PEMT and in MT@EC_PEMT for the same pair with lift = 3.87 (see Supplementary Table 2).

3.2. Frequent Tagsets and Rules Summarization. Based on the *Q* test results (Tables 2 and 3), the zero hypothesis, which reasons that the incidence of frequent tagsets does not depend on a way of translation (task), is rejected at the 0.001 significance level. The most frequent tagsets (almost 85%) were identified in MT@EC_PEMT, the lowest (almost 53%) in MT@EC_MT (Table 4).

From multiple comparisons (Table 4), two homogenous groups (MT@EC_MT) and (GT_PEMT, GT_MT, MT@EC_PEMT) were identified in terms of the average incidence of found frequent tagsets. Statistically significant differences were proved at the 0.05 significance level in the average incidence of frequent tagsets found between MT@EC_MT output and others.

Based on the *Q* test results (Table 3), the zero hypothesis, which reasons that the incidence of extracted rules does not depend on a way of translation (task), is rejected at the 0.001 significance level. The most extracted rules were found in translation MT@EC_PEMT output (almost 92%), the lowest in MT@EC_MT output (almost 34%) (Table 5).

From multiple comparisons (Table 5), three homogenous groups (MT@EC_MT), (GT_PEMT, GT_MT), and (MT@EC_PEMT) were identified in terms of the average incidence of extracted rules. Statistically significant differences were proved at the 0.05 significance level in the average incidence of found rules between MT@EC_MT output and others as well as between translation MT@EC_PEMT output and others. On the other hand, in both cases (Tables 4 and 5), a statistically significant difference between GT_MT output and GT_PEMT output was not found.

3.3. Validation of the Proposed Methodology. We have validated the proposed MT evaluation methodology based on the evaluation of frequent POS tags (tagsets) of MT outputs (MT@EC_MT or GT_MT) and PEMT outputs using the baseline methods. We used both, manual and semiautomatic MT evaluation measures as criteria for validity. In the case of manual evaluation, the criteria for validity are the scores of fluency (*F*) and adequacy (*A*). In the case of semiautomatic evaluation, we apply the HTER metric. Due to deviations from normality for testing differences between dependent variables, we used (Table 6) the Wilcoxon matched-pairs test.

Statistically significant differences were proved in the case of manual MT evaluation. The null hypotheses are rejected at the 0.001 significance level. We can see (in Figure 4(a)) differences in adequacy of MT output in favour of GT_MT output. Differences can be seen in the quartile range where 50% of the central values, for GT_MT output, were from the range [2, 5], contrary to MT@EC_MT output, where 50% of the central values were from the range [2, 4]. Similarly, in the case of the fluency of MT output (in Figure 4(b)), there are differences in favour of the GT_MT output. The differences can be seen in the median, where the estimation of the central value was 3 for the GT_MT output and 2.5 for MT@EC_MT output. In the case of both MT outputs, human translators used a range of the whole scale [from 1 to 5] to assess individual sentences, which indicates the heterogeneous quality of the examined MT sentences in case of adequacy and fluency.

Statistically significant differences were also shown in the case of the semiautomatic evaluation of MT, where H0 is rejected at the 0.001 significance level. We can see (in Figure 5(a)) differences in the HTER score in favour of GT_MT output. Based on the comparison of MTs with their corresponding PEMTs, a statistically significant lower error rate of MT output translated by GT compared to MT@EC_MT output was achieved.

Similar to manual evaluation, in the semiautomatic evaluation of individual sentences, the implemented HTER metric achieved the values of the whole range [0, 1], which refers to the heterogeneous quality of the examined MT segments for both MT outputs.

After rejecting the zero hypothesis, we are interested in MT segments with the highest differences in error rate (HTER) given to the used MT engine (MT@EC or GT). To identify segments, we use a method drawn from the residual analysis [59, 60]. We used this method to compare the results of semiautomatic MT evaluation of error rate between MT@EC_MT and GT_MT output (segment by segment). The aim of the analysis is to identify the segments (sentences) in which significant differences were found in the score of HTER of MT output (MT@EC and GT) from English into Slovak

$$\begin{aligned}(\text{residual value})_i &= (\text{value of } MT@EC_{MT})_i \\ &\quad - (value\ of\ GT_{MT})_i,\ i = 1, 2, \ldots, I,\end{aligned} \tag{4}$$

where I is a number of examined segments (sentences) in the dataset.

To identify extreme values (in Figure 5(b)), we use a rule $\pm 2\sigma$, i.e., residual values outside the interval we consider as extreme values

$$\begin{aligned}&\text{mean of residuals } (MT@EC_{MT} - GT_{MT}) \\ &\quad \pm 2st.\text{dev.of residuals}(MT@EC_{MT} - GT_{MT}).\end{aligned} \tag{5}$$

Table 2: Cochran Q test for incidence of frequent tagsets in examined translations.

	GT_MT		MT@EC_MT		GT_PEMT		MT@EC_PEMT	
Frequent tagset	Sup	Inc	Sup	Inc	Sup	Inc	Sup	Inc
(R)	11.28	1	10.80	1	10.80	1		0
...	...	...	...	...	...	...	...	...
(*VMdpb+*)	24.71	1	26.25	1	27.24	1	17.08	1
...	...	...	...	...	...	...	...	...
(*VIe+*)	11.77	1		0		0		0
Cochran Q test	$Q = 18.38298$; df = 3; $p < 0.001$							

Table 3: Cochran Q test for incidence of extracted rules in examined translations.

	GT_MT			MT@EC_MT			GT_PEMT			MT@EC_PEMT		
Rule	Sup	Lift	Inc	Sup	Lift	Inc	Sup	Lift	Inc	Sup	Lift	Inc
O = => *SSis1*	11.77	1.21	1			0	10.47	1.08	1	10.80	1.17	1
...	...	...	...	...	...	...	...	...	...	...	...	...
SSfs4 = => *VMdpb+*			0			0	12.29	1.97	1	9.97	2.01	1
...	...	...	...	...	...	...	...	...	...	...	...	...
SSns4 = => *Eu4*	12.11	2.08	1	11.11	3.13	1			0	11.30	2.25	1
Cochran Q test	$Q = 39.90476$; df = 3; $p < 0.001$											

Table 4: Homogeneous groups for incidence of frequent tagsets in examined translations.

Translation	Mean	1	2
MT@EC_MT	0.529		****
GT_PEMT	0.765	****	
GT_MT	0.765	****	
MT@EC_PEMT	0.843	****	

Table 5: Homogeneous groups for incidence of extracted rules in examined translations.

Translation	Mean	1	2	3
MT@EC_MT	0.333		****	
GT_MT	0.708	****		
GT_PEMT	0.750	****		
MT@EC_PEMT	0.916			****

Table 6: Comparison of dependent samples MT@EC output and GT output.

	T	Z	p value
adequacy (MT@EC_MT) & adequacy (GT_MT)	7457.0000	6.0682	<0.001
fluency (MT@EC_MT) & fluency (GT_MT)	10870.5000	4.9426	<0.001
HTER (MT@EC_MT) & HTER (GT_MT)	18269.5000	9.3839	<0.001

Figure 5(b) visualizes the residuals for MT outputs (MT@EC_MT and GT_MT). Residual values above the average of the residuals indicate an above-average error rate of MT output produced by MT@EC against MT output produced by GT; residual values below the average of the residuals indicate an above-average error rate of GT_MT output against MT@EC_MT output. It identifies segments where the significant differences in the evaluation of error rate between MT@EC_MT output and GT_MT output exist. In the case of MT@EC_MT output (in Figure 5(b)), we identified 28 segments that showed a significant error rate against GT_MT output. In contrast to GT_MT output (in Figure 5(b)), only 15 segments showed a significant error rate against MT@EC_MT output. The identified segments were subsequently manually analyzed, which resulted in the determination of the main issue of the entire MT process from English into Slovak. The difficulty consists of an incorrect determination of predication (subject, verb, and object) leading to mistranslation or incorrect translation, either grammatically or semantically (different parts of

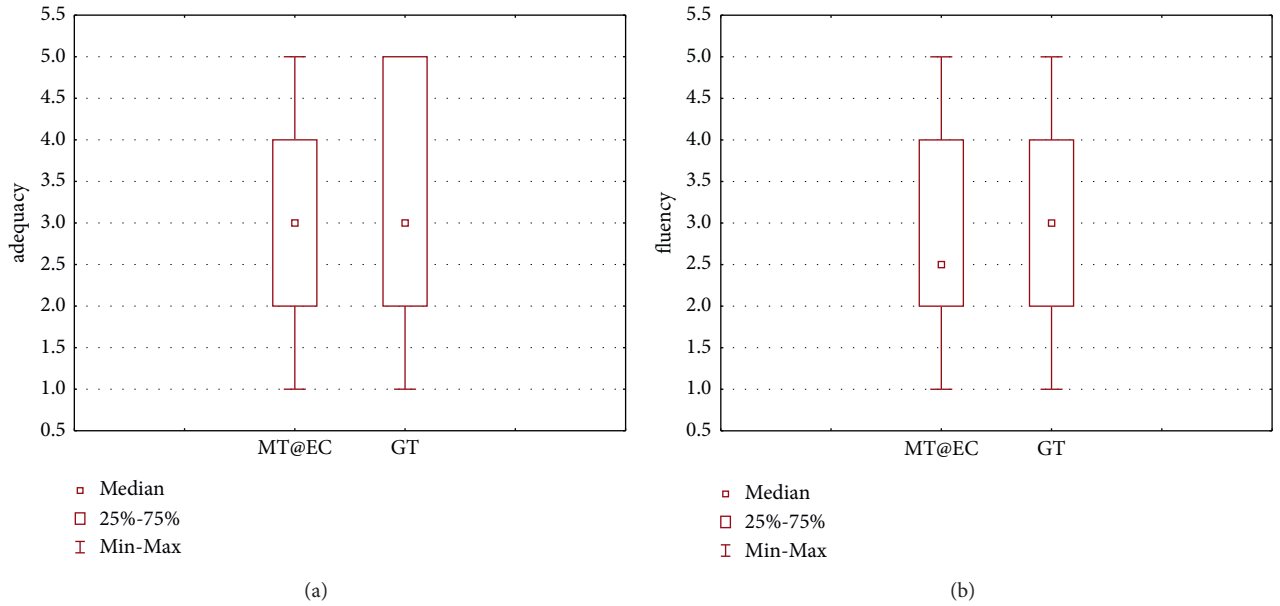

Figure 4: Visualization of descriptive statistics for manual MT evaluation: adequacy (a) and fluency (b).

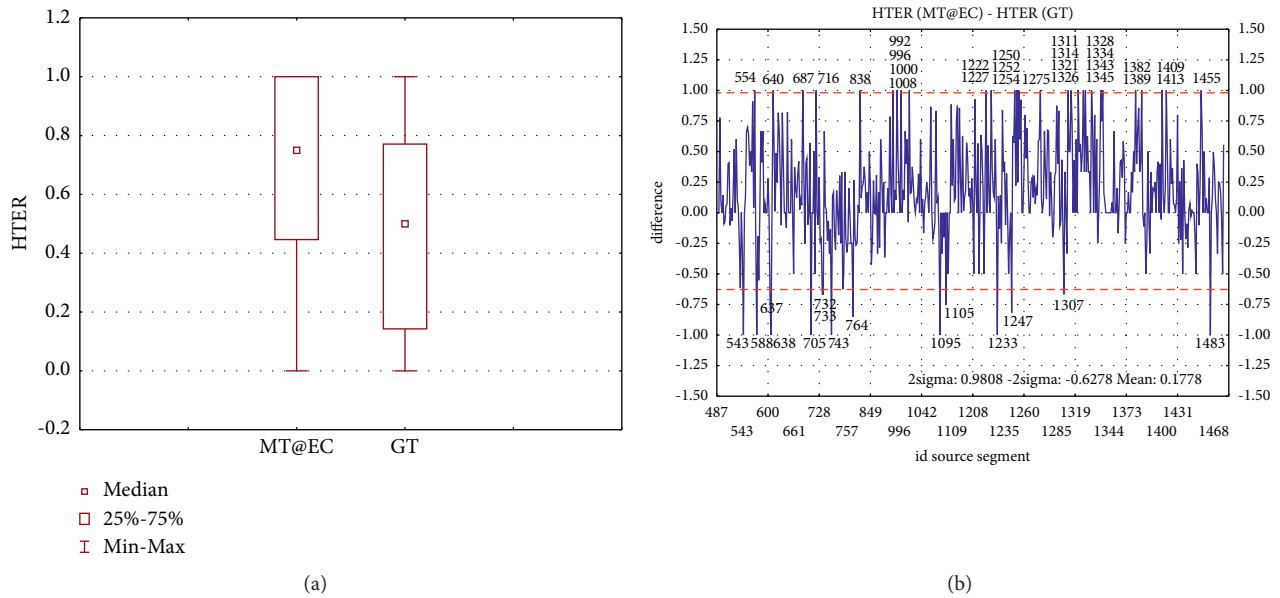

Figure 5: Visualization of descriptive statistics for semiautomatic MT evaluation: HTER metric (a) and residuals for MT output (b).

speech related to incorrect declension). These findings also confirm the results of the proposed MT evaluation methodology (MT@EC and GT) which is based on the evaluation of frequent tagsets.

4. Discussion

We agree with Mesmer et al. [13] that the text complexity must be understood to increase the knowledge about the interaction among the text characteristic, translator, or posteditor, and tasks such as translation or PE. The best approach to understand text complexity is through the analysis of words, sentences, and discourse [10]. For this reason, in the discussion, we will focus on the analysis of words and relationships among them.

In the case of GT, the highest value of the symmetric measure lift (*lift* = 5.11) was reached for a pair substantive in the inanimate masculine gender in the singular locative case with a preposition in locative. When analyzing this pair using the asymmetric measure confidence, we came across a discrepancy in the values (see Supplementary Table 4(a)). For the rule, SSis6 = => Eu6 the confidence value is 1.00, but for the rule, Eu6 = => SSis6 the confidence is only 0.58, which means that more often (100%) a preposition in

locative appears in segments (transactions) that contain substantive in the inanimate masculine gender in the singular locative only, as substantive in the inanimate masculine gender in the singular locative case in segments (transactions) that contain preposition in locative only (58.47%). The confidence measure points us to the fact that in the texts, prepositions are tied only to nouns, but nouns can also be found in a sentence as a subject or object, which do not require the presence of prepositions.

By MT@EC, it is substantive in the neuter gender, in the singular accusative case with the preposition in the accusative. As in the case of GT, a high lift value does not guarantee the same conditional probabilities of both directions of the rule (see Supplementary Table 4(b)). Even in the case of MT@EC, if the sentence contains a noun in the neuter gender, in the singular accusative case, then with 70% of confidence, it also contains the preposition in the accusative. However, the incidence of the noun in the neuter gender and in the singular accusative case is only 49.26% if the sentence contains a preposition in the accusative in the sentence only. A slightly less positive correlation occurred between imperative in the second person in the plural and conjunction (vyberte a/select and), which means if the sentence contains imperative in the second person in the plural, then with 81.55% of confidence, the sentence contains conjunction, but the incidence of imperative in the second person in the plural in the sentence is only 31.46% if there is conjunction only. Again, confidence values point to the fact that conjunctions are tied to either nouns or verbs in technical texts.

We can claim that GT_MT output was of a higher quality with respect to the rules and principles of the examined language. MT output has reached a greater congruence in gender, number, and case of a given language. The MT@EC engine rather translated word-by-word and did not focus on phrases and relations of the language. It literally translated from English (EN) to Slovak (SK) following grammar and the rules of the source language (EN) rather than the target language (SK). It was also confirmed by semiautomated evaluation, namely, by identifying specific segments (#554, #640, #1455), where the above-average error rate of MT@EC_MT output against GT_MT output was identified. These identified segments point to the specific errors that correspond to general (systematic) errors identified just by the analysis of frequent tagsets.

Each MT output either translated by GT or MT@EC was postedited by human translators. The aim of PE was to find out to which extent it was necessary to do PE to obtain a translation of publishing quality and to find out whether PE is more effective than a translation from scratch. We compared the quality of MT output with the quality of PEMT output based on language complexity, i.e., to what extent the tagsets and rules summarization for MT output and PEMT output are similar as well as the relation among tagsets that characterize language complexity.

The MT engine (GT) was relatively accurate (in Figure 3(a), see also Supplementary Table 3(a)) with only 5% of conjunctions inserted into the MT output (text). After a deeper analysis, we discovered that there are many conjunctions such as *and*, *or*, *when*, *after*, *then*, *to*, *before*, *if*. The conjunctions are tied to either the compound sentence or multiple sentence elements (verbs with objects). This is also shown by the high values of confidence measure (see Supplementary Table S5(a)), and despite the fact that the rules VKepb+ = => O and VId+ = => O do not reach the highest values of the lift (1.69; 1.59), they reach the highest confidence (84.38%; 79.46%). Other significant differences between the GT_MT output and GT_PEMT output were that the posteditors had to mainly correct or complete the masculine inanimate nouns in the singular genitive case together with prepositions in the genitive (*pomocou kábla/using cable*) as well as modal verbs (*môžete/you can*). This correction is closely linked with the flexion of the target language. Slovak consists of 4 paradigms of nouns (*S, A, F, U*), 4 genders (*m, f, n*), 2 numbers (*s, p*), and 7 cases (*1-7*). Compared to MT output, the posteditors had to mainly finish the translation by translating words that were not translated yet. It mainly referred to nouns in the neuter gender, in the singular accusative case, i.e., the adequate object was missing.

Based on the lift, we can claim that GT_PEMT output is characterized by more frequent phrase—verb in infinitive with modal verb (*môžete použiť/can use*) or a verb itself (*použite/use*), which is however in contrast with the values of lift in GT_MT output, where the most common phrase was a masculine, inanimate noun in locative of singular. It implies that the posteditors had to mostly correct the object, i.e., inflections and gender of nouns (*SSis6* to *SSns4* or *SSfs4*) corresponding to the agreement in the case of the preposition (*Eu6* to *Eu4*).

In the case of MT@EC, we received different results. In the MT@EC_MT output, the most frequent POSs were verbs in the infinitive (*VId+*), but in the MT@EC_PEMT output, there were imperatives in the second person of plural (*VMdpb+*). We deduce that the posteditors had not only to translate the nontranslated words (especially verbs and nouns in the subject) but also to a large extent modify verbs (the MT engine did not accept the rules of Slovak grammar, it kept the source language with its grammar, and it took only into account the basic form). Compared to MT@EC_MT output, differences have occurred in combination—conjunction and verb, where the verb has changed from infinitive to imperative (*VId* + to *VMdpb+*), as well as a noun which has changed from inanimate masculine to feminine, while the case and number were preserved (*SSis4* to *SSfs4*). The same rule was shown also in the MT@EC_PEMT output, i.e., if the sentence is not simple, then it is a copulative or conditional sentence, without agent expression, expressing only verb and object (*VMdpb+, SSfs4*) or multiple sentence element (*VMdpb+, O*), also one-syllable prepositions, which are tied to nouns in the accusative (*Eu4, SSns4*). The results of our error analysis also confirm the confidence values for MT@EC_MT output and its postedited version, i.e., the highest confidence (81.55%) for machine translation was achieved for the rule VMdpb + ==> O and for its postedited version, and the highest confidence (82.93%) was achieved for the rule VKepb+ = => O (see Supplementary Table 4(b) and Table 5(b)).

5. Conclusions

We focused on investigating the impact of an artificial agent (MT) on a human agent (posteditor) using the proposed methodology, which is based on POS tagging, frequent tagsets, association rules, and their summarization. We examined this impact from the point of view of perlocutionary acts, which includes the evaluation of machine translation. We have shown that the feature of adaptivity of a complex system requires human agents. Through human intervention, in our case by PE, MT systems can also contain the feature of adaptivity.

We proposed a new methodology for automatic MT evaluation using POS tags and association rules (in Figure 1). We compared two different MT engines—Google Translate and MT@EC (European Commission MT engine). We examined technical texts because they are the most frequently machine-translated texts. Based on the results of the analysis and found rules, we are able to characterize not only the text quality but also the text in terms of microstructure (morpho-syntactic relations).

Moreover, we validated the proposed methodology of MT evaluation using both manual and semiautomatic methods of MT evaluation (in Table 6). The results can be considered valid. The contribution of the proposed methodology is an identification of systematic, not random errors. In addition, the proposed methodology takes into account morpho-syntactic relations, which are important in evaluating translations between analytical and inflectional languages. Given that, we examined 4 translations conducted in four different ways—GT_MT, MT@EC_MT, GT_PEMT, and MT@EC_PEMT, and we investigated whether there are differences in the occurrence of frequent tagsets. Based on the Q test result ($Q = 18.38298$; $df = 3$; $p < 0.001$), we discovered that there is a difference in method way of translation (translation process) with respect to tags' occurrence. Using multiple comparisons, we identified two homogenous groups (MT@EC_MT) and (GT_PEMT, GT_MT, MT@EC_PEMT), i.e., there is a statistically significant difference between MT output translated by MT@EC and others. In other words, the GT_MT output was very similar to PEMT whether the translation engine was GT or MT@EC. In the terms of found rules, a statistically significant difference was also proven (based on the result of Q test ($Q = 39.90476$; $df = 3$; $p < 0.001$) and from multiple comparisons) between MT@EC_MT and MT@EC_PEMT, as well as between MT@EC_MT and others and also between MT@EC_PEMT and GT_MT or GT_PEMT output. The GT_MT output was very similar to GT_PEMT output. The posteditors similarly postedited both MT outputs, but to a large extent (statistically significant), the corrections were made in case of MT@EC_MT output.

To sum up our findings, we can state that for technical texts such as manuals, MT systems produce an output with an acceptable level of quality. A statistically significant difference between the GT_MT output and the GT_PEMT output in terms of the meaning or grammar was not proven. Last but not least, to answer the question concerning how to evaluate the MT quality or which methodology to use, we have shown an original and previously unused unique approach using text complexity measures. In our view, it is an objective evaluation of MT output by statistical, NLP, and machine learning methods. It can also be used for the automatic identification of MT errors into the inflectional language (e.g., Slovak).

The proposed methodology can serve as an alternative to the current, which use manual evaluation metrics, and which are not only time but also labour-consuming, but also use standard automatic evaluation metrics such as BLEU. We see the use of the methodology itself not only in evaluating MT quality but also in teaching MT and PE in the study programs of translation studies.

Another interdisciplinary contribution or future work lies in providing information to focus on during the PE process, which can finally improve translators' performance as expected by today's market.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Slovak Research and Development Agency under the contract no. APVV-18-0473. This research was funded by European Commission under the ERASMUS + Programme, KA2, grant number: 2021-1-SK01-KA220-HED-000032095 "Future IT professionals Education in Artificial Intelligence," Ministry of Education of Slovakia, grant no. 004UKF-2-1/2021 "Preparation and development of teaching courses in English with a focus on artificial intelligence in the form of blended-learning," and Ministry of Education of Slovakia, grant number: 2020/8148: 34-A1101 "Support for the development of practical skills of UKF students in Nitra."

Supplementary Materials

Supplementary Table 1: POS tagging for Slovak morphological annotation. Supplementary Table 2: tabulation of tags identified in MT output translated by Google Translate (a) and MT@EC (b). Supplementary Table 3: tabulation of tags identified in PEMT output translated by Google Translate (a) and MT@EC (b). Supplementary Table 4: tabulation of rules confidence in MT output translated by Google Translate (a) and MT@EC (b). Supplementary Table 5: tabulation of rules confidence in PEMT output translated by Google Translate (a) and MT@EC (b). (*Supplementary Materials*)

References

[1] P. Liu and Z. Li, "Task complexity: a review and conceptualization framework," *International Journal of Industrial Ergonomics*, vol. 42, 2012.

[2] R. Pelánek, T. Effenberger, and J. Čechák, "Complexity and difficulty of items in learning systems," *International Journal of Artificial Intelligence in Education*, 2021.

[3] D. J. Campbell, "Task complexity: a review and analysis," *Academy of Management Review*, vol. 13, 1988.

[4] J. G. March and H. A. Simon, *Organizations*, Wiley, Oxford, England, 1958.

[5] G. Daroczy, M. Wolska, W. D. Meurers, and H.-C. Nuerk, "Word problems: a review of linguistic and numerical factors contributing to their difficulty," *Frontiers in Psychology*, vol. 06, 2015.

[6] Q. Ma and X. Wang, "What is language complexity?" *Macrolinguistics*, vol. 7, 2019.

[7] Ö. Dahl, *The Growth and Maintenance of Linguistic Complexity*, John Benjamins Publishing Company, Amsterdam, 2004.

[8] A. Andrason, "language complexity: an insight from complex-system theory," *International Journal of Language and Linguistics*, vol. 2, pp. 74–89, 2014.

[9] X. Guo-zhi, *Systems Science*, Shanghai Technology and Education Press, Shanghai, 2000.

[10] G. Berninger and C. Garvey, "Tag constructions: structure and function in child discourse," *Journal of Child Language*, vol. 9, 1982.

[11] S. Zhou and W. Liu, "English grammar error correction algorithm based on classification model," *Complexity*, vol. 2021, Article ID 6687337, 11 pages, 2021.

[12] R. Loock, "No more rage against the machine: how the corpus-based identification of machine-translationese can lead to student empowerment," *J. Spec. Transl.*vol. 34, pp. 150–170, 2020.

[13] H. A. Mesmer, J. W. Cunningham, and E. H. Hiebert, "Toward a theoretical model of text complexity for the early grades: learning from the past, anticipating the future," *Reading Research Quarterly*, vol. 47, no. 3, pp. 235–258, 2012.

[14] M. L. Forcada, "Making sense of neural machine translation," *Transl. Spaces*, vol. 6, 2017.

[15] O. De Clercq, G. De Sutter, R. Loock, B. Cappelle, and K. Plevoets, "Uncovering machine translationese using corpus analysis techniques to distinguish between original and machine-translated French," *Transl. Q*, vol. 101, pp. 1–21, 2021.

[16] H. Hassan, A. Aue, C. Chen et al., *Achieving Human Parity on Automatic Chinese to English News Translation*, http://arxiv.org/abs/1803.05567 ArXiv. accessed, 2018.

[17] L. Macken, L. Van Brussel, and J. Daems, "NMT's wonderland where people turn into rabbits. A study on the comprehensibility of newly invented words in NMT output," *Comput. Linguist. Netherlands J.*vol. 9, pp. 67–80, 2019.

[18] W. Farhan, B. Talafha, A. Abuammar et al., "Unsupervised dialectal neural machine translation," *Information Processing & Management*, vol. 57, 2020.

[19] R. Loock, "Traduction automatique et usage linguistique: une analyse de traductions anglais-français réunies en corpus," *Meta Le J. Traducteurs/Meta Transl. J*, vol. 63, pp. 786–806, 2018.

[20] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, "On using very large target vocabulary for neural machine translation," in *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process*, vol. 1, pp. 1–10, Association for Computational Linguistics, Beijing, China, 2015.

[21] M. Yamada, "The impact of Google neural machine translation on post-editing by student translators," *J. Spec. Transl.*vol. 31, 2019.

[22] P. Kumar, R. Ahmad, B. D. Chaudhary, and R. Sangal, "Machine translation system as virtual appliance: for scalable service deployment on cloud," in *Proceedings of the 2013 IEEE Seventh Int. Symp. Serv. Syst. Eng.*, pp. 304–308, IEEE, San Francisco, CA, USA, Mar 2013.

[23] F. Gobbo, "Machine translation as a complex system: the role of Esperanto," *Interdisciplinary Description of Complex Systems*, vol. 13, no. 2, pp. 264–274, 2015.

[24] A. M. Hayes and L. A. Andrews, "A complex systems approach to the study of change in psychotherapy," *BMC Medicine*, vol. 18, p. 197, 2020.

[25] A. F. Siegenfeld and Y. Bar-Yam, "An introduction to complex systems science and its applications," *Complexity*, vol. 2020, Article ID 6105872, 16 pages, 2020.

[26] X. Feng, Z. Feng, W. Zhao, B. Qin, and T. Liu, "Enhanced neural machine translation by joint decoding with word and POS-tagging sequences," *Mobile Networks and Applications*, vol. 25, 2020.

[27] J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty, Eds., *Translation Quality Assessment*, Springer International Publishing, Cham, 2018.

[28] M. Popović, "Error classification and analysis for machine translation quality assessment," in *Mach. Transl. Technol. Appl.*, J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty, Eds., Springer, Cham, 2018.

[29] D. Shterionov, F. d. Carmo, J. Moorkens et al., "A roadmap to neural automatic post-editing: an empirical approach," *Machine Translation*, vol. 34, no. 2-3, pp. 67–96, 2020.

[30] S. Castilho, S. Doherty, F. Gaspari, and J. Moorkens, "Approaches to human and machine translation quality assessment," in *Transl. Qual. Assessment. Mach. Transl. Technol. Adn Appl.*Springer, Cham, 2018.

[31] K. Hu, S. O'Brien, and D. Kenny, "A reception study of machine translated subtitles for MOOCs," *Perspectives*, vol. 28, no. 4, pp. 521–538, 2020.

[32] G. De Sutter, M.-A. Lefer, and I. Delaere, Eds., *Empirical Translation Studies*, De Gruyter, Berlin, Boston, 2017.

[33] P. Isabelle, C. Cherry, and G. Foster, "A challenge set approach to evaluating machine translation," in *Proc. 2017 Conf. Empir. Methods Nat. Lang. Process.*Association for Computational Linguistics, Stroudsburg, PA, USA, 2017.

[34] E. Vanmassenhove, D. Shterionov, and A. Way, "Lost in translation: loss and decay of linguistic richness in machine translation," in *Proc. Mach. Transl. Summit XVII*, vol. 1, pp. 222–232, Res. Track, European Association for Machine Translation, Dublin, Ireland, 2019.

[35] B. Dorr, M. Snover, and N. Madnani, "Part 5: machine translation evaluation," in *Handb. Nat. Lang. Process. Mach. Transl. DARPA Glob. Auton. Lang. Exploit.*, J. M. Joseph Olive and C. Christianson, Eds., , p. 936, Springer, 2011.

[36] J. Daems, S. Vandepitte, R. J. Hartsuiker, and L. Macken, "Identifying the machine translation error types with the greatest impact on post-editing effort," *Frontiers in Psychology*, vol. 8, p. 1282, 2017.

[37] C. Lo and D. Wu, "MEANT: an inexpensive, high-accuracy, semiautomatic metric for evaluating translation utility via semantic frames," *ACLPPinforma*, vol. 11, pp. 220–229, 2011.

[38] A. Toral, "Post-editese: an exacerbated translationese," in *Proc. Mach. Transl. Summit XVII*, vol. 1, pp. 273–281, Res. Track, European Association for Machine Translation, Dublin, Ireland, 2019.

[39] M. Munk, D. Munkova, and L. Benko, "Towards the use of entropy as a measure for the reliability of automatic MT

evaluation metrics," *Journal of Intelligent and Fuzzy Systems*, vol. 34, no. 5, pp. 3225–3233, 2018.

[40] L. Benkova, D. Munkova, Ľ. Benko, and M. Munk, "Evaluation of English–Slovak neural and statistical machine translation," *Applied Sciences*, vol. 11, 2021.

[41] A. Lommel, "Metrics for translation quality assessment: a case for standardising error typologies," in *Transl. Qual. Assess.*, J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty, Eds., Springer International Publishing, Cham, 2018.

[42] M. Nadejde, S. Reddy, R. Sennrich et al., "Predicting target language CCG supertags improves neural machine translation," in *Proc. Second Conf. Mach. Transl.*Association for Computational Linguistics, Stroudsburg, PA, USA, 2017.

[43] D. Hládek, J. Staš, and J. Juhár, "Rule-based morphological tagger for an inflectional language," in *Cogn. Behav. Syst. Lect. Notes Comput. Sci.*, pp. 208–215, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[44] P. Halácsy, A. Kornai, and C. Oravecz, "HunPos: an open source trigram tagger," in *Proc. 45th Annu. Meet. Assoc. Comput. Linguist. Companion Vol. Proc. Demo Poster Sess.*, pp. 209–212, Association for Computational Linguistics, Prague, Czech Republic, 2007.

[45] L. J. Laki, G. Orosz, and A. Novák, "HuLaPos 2.0 - decoding morphology," in *Advances in Artificial Intelligence and Its Applications*, pp. 294–305, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[46] C. Conforti, M. Huck, and A. Fraser, "Neural morphological tagging of lemma sequences for machine translation costanza Conforti," in *Proc. 13th Conf. Assoc. Mach. Transl. Am. (Volume 1 Res. Track), Association for Machine Translation in the Americas*pp. 39–53, Boston, MA, 2018.

[47] C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder, "Findings of the 2009 workshop on statistical machine translation," *Proc. Fourth Work. Stat. Mach. Transl*, pp. 1–28, 2009.

[48] M. Popović, "Morphemes and POS tags for n-gram based evaluation metrics," *Proc. Sixth Work. Stat. Mach. Transl.*, 2011.

[49] M. Popović, *Machine Translation: Statistical Approach with Additional Linguistic Knowledge*, RWTH Aachen University, Aachen,Germany, 2009.

[50] M. Popović, "rgbF: an open source tool for n-gram based automatic evaluation of machine translation output," *The Prague Bulletin of Mathematical Linguistics*, vol. 98, pp. 99–108, 2012.

[51] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proc. Int. Conf. New Methods Lang. Process*pp. 44–49, Manchester, UK, 1994.

[52] H. Schmid, "Improvements in part-of-speech tagging with an application to German," in *Text, Speech and Language Technology*, S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, Eds., Kluwer Academic Publishers, Dordrecht, pp. 13–25, 1999.

[53] H. Schmid, M. Baroni, E. Zanchetta, and A. Stein, "The enriched TreeTagger system," in *Proc. EVALITA 2007 Work.*, 2007.

[54] V. Benko, "Compatible sketch grammar experiment," in *Proc. Int. Conf. «Corpus Linguist*pp. 21–29, St. Petersburg, 2013.

[55] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. Assoc. Mach. Transl*, pp. 223–231, Am., 2006.

[56] M. Snover, N. Madnani, B. J. Dorr, and R. Schwartz, "Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric," in *Proc. Fourth Work. Stat. Mach. Transl.*, pp. 259–268, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, http://dl.acm.org/citation.cfm?id=1626431.1626480.

[57] D. Munková, J. Kapusta, and M. Drlík, "System for post-editing and automatic error classification of machine translation," in *DIVAI 2016 11th Int. Sci. Conf. Distance Learn. Appl. Informatics, Sturovo, May 2 – 4, 2016*, pp. 571–579, Wolters Kluwer, Sturovo, 2016.

[58] D. Munková, M. Munk, and M. Vozár, "Data pre-processing evaluation for text mining: transaction/sequence model," *Procedia Comput. Sci.*vol. 18, pp. 1198–1207, 2013.

[59] M. Munk and D. Munkova, "Detecting errors in machine translation using residuals and metrics of automatic evaluation," *Journal of Intelligent and Fuzzy Systems*, vol. 34, no. 5, pp. 3211–3223, 2018.

[60] D. Munková and M. Munk, "Automatic evaluation of machine translation through the residual analysis," in *Lecture Notes in Computer Science*, D. S. Huang and K. Han, Eds., Springer, Nitra,Slovakia, pp. 481–490, 2015.