Richness Lost in Machine Translationese: Lexical Richness in Human Translation versus Neural Machine Translation from Arabic into English

Radwa Kotait

English Department, Faculty of Al-Alsun, Ain Shams University, Cairo, Egypt
radwa_kotait@alsun.asu.edu.eg

Abstract: ***Neural Machine Translation (NMT) might have been pronounced as faster and better than human translation. However, NMT inherently overgeneralizes the more frequently appearing patterns detected in their training data at the expense of the less frequently appearing ones in a phenomenon dubbed "machine translationese". This machine translationese has been noticed to reflect some controversial asymmetries. One usually overlooked facet of this machine bias is the loss of "lexical richness". The generated translations have only recently been noticed to be disproportionately deformed and impoverished, negatively impacted with the NMT's tendency to overgeneralize. Lexical richness, notwithstanding its worth, has not received the same attention that lexical accuracy and error-measuring have received, and more important, it has not received any attention at all in under-researched language pairs, such as Arabic–English. This study aims to shed light on lexical richness in the output of Arabic-into-English NMT as opposed to human translation (HT), answering the question: Does HT exhibit more lexical richness than NMT does? The study adopts the most agreed-upon definition of lexical richness as a superordinate term that includes "lexical diversity", "lexical density", and "lexical sophistication"; all three are statistical metrics that gauge the lexical richness of a text. The study analyses the outputs of two NMTs, Google Translate and Microsoft Translator, in terms of lexical richness, using both quantitative and qualitative methods, and then compares the results to those of the HT output. The corpus of the study is comprised of a news subcorpus and a literary subcorpus.***

Keywords: ***lexical density; lexical diversity; lexical richness; lexical sophistication; neural machine translation; machine translationese.***

1 INTRODUCTION

Translation studies have discovered a special dialect within language; a third code, as Frawley [1] calls it; a language of translation that has its own characteristic features which distinguish translated texts from originally written ones. Gellerstam [2] was the first to call it "translationese". Baker [3] has attempted to define these features and has called them "translation universals". Ever since then, researchers have been looking for empirical methods to locate and measure these fingerprints left on the translation product. Some have even developed machine-learning algorithms to perform automatic text-classification tasks that distinguish translated texts from originally written ones [4, 5, 6].

Recent studies on machine translation (MT) output have come to notice that machine translations have their own translationese. It is a set of distinct features that set machine translation output apart from human translation; it is a footprint of sorts of an MT system. There have been several that conduct a set of computational or corpus analyses of MT output in an attempt to classify these frequently observed patterns, or machine translation universals in a phenomenon dubbed "machine translationese" ([4], [7], [6], [8], [9] to mention a few). They have attempted to establish or negate the presence of inclinations such as simplification, explicitation* and interference.

Some of these studies, attempting to identify MT markers or what constitutes "machine translationese", have noticed that translation generated by MT systems, especially the neural ones, exhibit some controversial asymmetries, or "algorithmic bias". It is "the phenomenon by which trained statistical models unbeknownst to their creators grow to reflect controversial societal asymmetries" [10, p.3]. The most famous of these asymmetries is gender bias, where neural machine translation output is noticed to be particularly prone to producing gender stereotypes that are "sexist". Another usually overlooked facet of this machine bias, which has come to the attention of researchers only recently, is the loss of lexical richness.

The quality of machine translations has been frequently evaluated on the basis of accuracy, or by using error-measuring metrics. Only recently have researchers raised serious concerns about the lexically impoverished language generated by MT, a language that ominously lacks diversity, and suffers from severe loss of lexical richness. The generated translations have been noticed to exacerbate the dominant, already frequent patterns at the expense of less frequent ones, jeopardising any semblance of language richness. Very few studies have attempted to measure the lexical richness of MT and compare it to that of human translation, and those few ones

* It is a term in Translation Studies defined as "a translation shift from the implicit to the explicit"

have surprisingly reached controversial results concerning which of the two exhibits stronger variation. Next to none have conducted similar studies on less resourced languages such as Arabic. Hence, comes the significance of the present study. Is neural machine translation indeed less varied and less creative than human translation, or contradictory to the general opinion, neural machine translation has the ability to learn and to improve its output and make it more diverse?

2 LITERATURE REVIEW

A. Neural Machine Translation

Due to the inherent vagueness and adaptability of human language, the task of machine translation is not an easy one. Automatically converting one language into another, especially if they are not from the same family as the case is in the language-pair English/Arabic, is rather challenging. Classical machine translation methods, also known as Rule-based Machine Translation (RBMT), typically employ manually created linguistic rules and representations, developed by linguists, to convert a text from the source language to the target language [11, p. 133]. With the growing difficulty of hand crafting translation rules for all language pairs on the one hand and with the astounding availability of data, especially large-scale parallel corpora, RBMT has given way to Statistical Machine Translation (SMT), a data-driven approach to MT which, through alignment, learns latent structures [12, p. 5].

With the advent of deep learning, Neural Machine Translation (NMT) has emerged as a new paradigm, quickly assuming the place of SMT as the mainstream approach to MT [12, p. 5]. "The training of NMT is end-to-end as opposed to separately tuned components in SMT", directly mapping the input text to the output text without intermediate processing steps [12, p.5]. NMT involves utilizing a massive artificial neural network approach to forecast the probability of a word sequence, frequently in the shape of complete sentences [13]. It typically proceeds in two stages: (1) the modelling, or encoding, stage, where "a conditional language model (using neural networks) is trained by optimizing a probabilistic objective", and then (2) the decoding stage, where predictions are produced by searching for the mode (or the most frequent) of the conditional distribution [14, p. 1]. NMT has become the most popular technology behind many commercial MT systems, the most famous of which are *Google Translate* and *Microsoft Translator*. Both use a large dataset for training their algorithms, along with the end-to-end design of NMT to learn over time and create better and more natural translations.

B. Algorithmic Bias

As explained above, NMT leans upon machine learning algorithms based on big data which have come to control the tiniest details of our lives; "[f]rom the ads we are served, to the products we are offered, and to the results we are presented with after searching on-line, algorithms, rather than humans sitting behind the scenes, are making these decisions" [15, p. 16]. The output of these trained statistical models has been noticed to reflect controversial asymmetries, resulting in what has come to be known as "machine bias" or "algorithmic bias" [10, 15–18]. These biases are sometimes inherent in the data itself that the machine learning algorithm is trained upon, and other times, even when these biases are cleaned from the input data, discriminatory output still exists because of the correlations that a clever algorithm can still detect [17, p. 2125]. The growing concern over these asymmetries has led to systematic efforts to "de-bias" these algorithmic discriminations, be it gender bias [10, 16, 19-22 to mention a few], or racial and ethnic bias [15, 23].

This biased behaviour in MT output, which inherently optimizes more frequently appearing patterns detected in their training data, has extended beyond gender and racial biases to a linguistic one where the MT output lacks diversity on several levels [24, p. 57]. Linguistic richness, or rather the loss thereof, is an area of research that has not received the same attention as gender bias in MT research. Understandably, accuracy in MT output has assumed higher priority than diversity and richness. However, as Roberts et al. [14] argues, despite the general opinion, "[d]iversity in NMT is valuable" (p.1).

C. Lexical Richness

Lexical richness is originally a term used in the fields of language acquisition and language assessment to refer to the number of words in an author's mental lexicon, or the "wealth of words at [one's] command" [25 p. 83]. It has been extended to the assessment of second language (L2), oral and written, and over the years, it has become of great importance in fields such as cognitive science and artificial intelligence, especially where language production and language generation are concerned. As the case is with almost all linguistics and

translation studies terminology, the term "vocabulary richness", or "lexical richness", has suffered "a great deal of terminological drift", lacking a clearly articulated definition and a clearly defined set of parameters [26, p. 38].

"Lexical richness" is sometimes used as a synonym of "lexical diversity", or "lexical variation" [27]. Some other times, it is used as an umbrella term that includes lexical diversity as well as other lexical indices [see 28, 29]. The most agreed-upon and often quoted definition is the one considered by Read [30], where lexical richness is a superordinate term that includes the four dimensions: "lexical diversity" (lexical variation), "lexical density", "lexical sophistication" and "lexical errors". All four are statistical measures that gauge the lexical richness of a text. The present study adopts the first three only as lexical errors does not lend itself to the scope of the study.

1) Lexical Variation: "Lexical Variation" (LV), also known as "lexical diversity", refers to the use of a variety of different words rather than a limited number of words used repetitively, as Read [30] identifies it. It is "the range of different words used in a text" [31, p. 381]. So, the greater that range, the higher the diversity. It has been traditionally used as indicative of writing quality, vocabulary knowledge, and speaker competence whether in the native or second language. There are different indices for measuring lexical variation; the most famous of which is type-token ratio (TTR).

In a sample text, all the words are "tokens", but an orthographically unique word is a "type". Therefore, statistically, the "token count" for a text is the total number of words it contains, while the "type count" is the number of different words in that text. By comparing the number of different words to the total number of words, vocabulary variability is calculated [32, 33]. The first to propose it was Johnson [34] and the first to use it to calculate LV was Ure [35], who used it as a dimension to discriminate between written and spoken language. TTR is a validated measure to assess lexical variation, but on the one condition that the length of the texts compared remain constant [36]. It has been noticed that the longer the text is, the more likely will repeated words be used, and logically, the smaller the TTR will become.

This has led to improved variations of TTR that have attempted to reduce the effects of the length of the texts on the measurement results. Root TTR (RTTR) is computed as t/sqrt(w), where t is the number of unique terms/vocab, and w is the total number of words. It is also known as Guiraud's R and Guiraud's index (Guiraud 1954, 1960). There are also Herdan's C, which is computed as log(t)/log(w) [37] (for detailed formulas, see [38]). Log values are used as a corrective factor that solves the text length problem [31, p. 384].

Mean Segmental Type-Token Ratio (MSTTR) is an improved version of TTR where the text is split into several equally proportioned segments, and the mean of all the TTRs indicates the overall lexical diversity in the text [34]. The Moving Average Type-Token Ratio (MATTR) adopts the same approach of segmenting the text using the sliding-window approach where the text is divided into smaller sections of a fixed length, and then when the TTR is calculated, it slides to the next section, overlapping with the previous one until the end of the text is reached and the average TTR is calculated [39].

Yule's K, and its inverse (Yule's I), is another means to calculating LV [25]. It is considered "more resilient to fluctuations related to text length than TTR" [40, p. 2208]. His suggested constant K does not depend on text length. The larger the Yule's K result, the higher the ratio of repeated words and the less the lexical variation is detected in a text, and vice versa [41, p. 4].

Vocd-D is another approach to calculating LV [42]. It starts by taking 100 random samples of 35 tokens for which the TTR for each of these samples is calculated, and the mean TTR is stored. The same procedure is then repeated for samples from 36 to 50 tokens, and so on. An empirical TTR curve is created from the means of each of these samples; "[f]inal values tend to range from 10 to 100, with higher values indicating greater diversity" [31, p. 383]. HD-D index is suggested by McCarthy and Jarvis [43] to reduce the complexity of the vocd-D. "For each lexical type in a text, the probability of encountering any of its tokens in a random sample of 42 words is drawn from the text" and "[t]he probabilities for all lexical types in the text are then added together, and the sum is used as an index of the text's [LV]" [31, p. 383].

The Measure of Textual Lexical Diversity (MTLD) is one of the most famous means to calculating LV [44]. It is evaluated sequentially by calculating "the mean length of sequential word strings in a text that maintain a given TTR value" and "each word of the text is evaluated sequentially for its TTR" [31, p. 384]. It "iterates over words until TTR scores fall below a threshold, then increase factor counter by 1 and start over" [38]. The threshold is a default factor size of 0.72. Its efficacy as one of the most informative calculations of LV has been proven in several studies.

2) *Lexical Density*: While lexical variation, or diversity, accounts for how many different words are used in a text, lexical density (LD) measures the ratio of content words, as opposed to grammatical (functional) words; namely, nouns, verbs, adjectives and adverbs, to its total number of words in the text. The term "lexical

density" was coined by Ure [35] who used it to compare the density of written and spoken texts. The formula suggested by Ure (1971) is total number of content words/tokens (total number of words in the text). Bates et al. [45] use other variations of LD calculations in their study of individual differences in language development. Adjective Density = open class adjectives / content words; Verb Density = open class verb types / content words; Noun Density = common nouns / content word; and Adverb Density = open class adverbs/ content words (pp. 97–8).

3) *Lexical Sophistication:* Defining what really a "sophisticated" word is has never been an easy task. According to Laufer and Nation [29], it is generally related to the relative difficulty of the lexical items in a text, which is often calculated based on reference corpus frequency counts. Word frequency has been for long the main indicator of the degree of sophistication. Less common terms such as "solidification" or "oroctogenarians" are generally perceived as sophisticated, while more commonly used words such as "people" or "place" are considered less sophisticated [46].

However, recent studies have widened the range of lexical sophistication indices, in addition to word frequency. Range, for one, has been added as a measurement of the number of texts in a reference corpus in which a word occurs. Words that are used extensively across various texts and contexts are indicators of less lexical sophistication in comparison to words with low range values that are typically limited to a smaller number of texts and contexts [47, p. 14]. N-grams, or multiword expressions (MWEs), have also been considered as indicators of lexical sophistication. Using them as indicators of sophistication has been of growing interest over the last two decades. Texts with higher frequency of n-grams (such as "as well as" and "as a result of") have been found to display higher lexical sophistication than those with infrequent n-grams.

There have been several attempts at creating tools that automatically calculate the rate of lexical sophistication of a sample text, or, in other words, that quantitatively measure the Lexical Frequency Profile (LFP), such RANGE [48], VocabProfile [49]; Coh-Metrix [50]; and Linguistic Inquiry and Word Count (LIWC) [51]. However, each of these tools uses a limited set of indices and functions on small texts [46, pp. 757-58]. That is why "The Tool for the Automatic Analysis of LExical Sophistication" (TAALES) has been created; it is a tool that analyses more than 400 indices of lexical sophistication of a sample text that focus on the above-mentioned indices and more. They use several reference corpora (British National Corpus (BNC), Corpus of Contemporary American (COCA), SUBTLEXus Corpus of American Subtitles, and Brown Corpus) as well as several lists of words (such as the Academic Word List (AWL) and Academic Formulas List (AFL)). The indices are calculated for all words (AW), content words (CW) and function words (FW).

TAALES also offers different metrics that calculate n-grams. One of them is n-grams that are register-specific derived from the COCA. Being register-specific guarantees that results returned are relevant to the genre understudy. In addition to raw counts of n-grams, they also offer n-gram strength-of-association metrics (such as Mutual Information (MI) score which emphasizes infrequent items, MI2 score and t-score which emphasize frequent items, Delta P score which is directional in nature emphasizing the order of the words). These strength-of-association norms calculate "the relative frequencies of the words that comprise n-grams by measuring the conditional probability of word co-occurrence", thus, demonstrating a stronger relationship between the words in bigrams such as "optimistic about" compared to the words in phrases such as "and the" or "in the" [52, p. 1035).

D. Lexical Richness in NMT Research

Vanmassenhove et al. [24] have been the first to bring to the spotlight the artificially impoverished language in the translations produced by MT paradigms, or what they dub "machine translationese" (p. 2203). The translations generated have been noticed to be disproportionately deformed and impoverished, negatively impacted with overgeneralizations. This impoverishment, with the increased dependency of humans on MT output, might very well kill all elements of lexical creativity, which does not bode well for the future of the translation industry. There is also the danger of "language learners developing a 'warped exposure' to that language through neural machine translation" [53, p. 1].

NMT's inability to generate diverse output has been only recently brought to light in MT research. Toral [54] has conducted quantitative analyses of Human Translations (HTs) and Post-editing (PEs) MTs as well as MT outputs, from the MT systems that were the starting point to produce the PEs. His datasets cover five languages: English↔German, English→French, Spanish→German and Chinese→English. He has studied lexical density and diversity; he has reached the conclusion that in terms of lexical density, MTs in general, and NMTs in particular, exhibit lower rates than HTs, and in terms of lexical diversity, HTs are lexically richer than MTs, statistical or neural. Castilho et al. [55] have also conducted a study on PE, looking for "post-editese" features, this time using a collection of articles from *The New York Times* human-translated into Brazilian Portuguese. The

corpus was then translated using Google Translate and post-edited by four translators. In their search to (dis)prove simplification as a translation universal, they calculated lexical density (content words/ total words ratio), lexical richness (type/token ration), as well as sentence count and mean sentence length. Surprisingly, they have found that in the news domain, "a greater loss in lexical richness and lexical density was present in HT and PE than in MT texts", which contradicts Toral's results [54, p. 26].

Vanmassenhove, et al. [24] have conducted an experiment to quantify the loss of lexical richness in MT versus HT using a number of Lexical Diversity metrics on the output of 12 different machine translation systems for English-French and English-Spanish with original and back-translated data. They conclude that "the process of MT causes a general loss in terms of lexical diversity and richness when compared to human-generated text" (p. 230). Their explanation is that MT paradigms indeed overgeneralize in a form of algorithmic bias, increasing the frequencies of more frequent words and decreasing that of the less frequent ones (p. 230). In another experiment, Vanmassenhove, et al. [40] study the effects of algorithmic bias on linguistic complexity in MT, in what they state is the first study of the lexical and morphological diversity of machine translationese (p. 2204). They assess the linguistic richness, lexically and morphologically, of translations created by different MT paradigms – phrase-based statistical and neural machine translation using 9 lexical and grammatical diversity and lexical sophistication metrics. They conclude that "there is a loss of lexical and morphological richness in the translations produced by all investigated MT paradigms for two language pairs (EN↔FR and EN↔ES)" (p. 2203).

Brglez and Vintar [53] analyse both quantitatively and qualitatively the outputs of three English-to-Slovenian MT systems (one statistical and two neural) in terms of lexical diversity in three different genres: information technology, culinary arts, and literature. They have built their study on the hypothesis that "machine translations exhibit lower lexical diversity than human translations but that neural machine translations have a higher lexical diversity than statistical machine translations" (p. 3). They use TTR and MTLD for the quantitative approach, and then they add the qualitative approach for a closer examination of lexical diversity, where they analyse the translation equivalents for selected keywords and multi-word expressions, and compare the number of translations solutions in HT versus MT. They have found that automatic metrics (quantitative approach) measuring lexical diversity show divergence on a case-by-case basis, yielding occasionally contradicting results. Surprisingly again, both metrics (TTR and MTLD) put HT "at the very bottom of the lexical diversity ladder" in the majority of cases (p. 12). They explain the contradictory results as due to two contradictory tendencies on part of MT; one is the inherent tendency of "overgeneralization" of more frequent words, which results in the loss of lexical diversity, and the other, is the "undergeneralization" of less frequent words, which results in strange, made-up, inconsistent, miscellaneous, unreliable and misguided translations, or what they dub a "mock" lexical diversity, which might very well impede post-editing later on (p. 13).

3 MOTIVATION

As the review of literature addressing lexical richness in NMT shows, lexical richness, notwithstanding its worth, has not received the same attention that lexical accuracy and error-measuring have received in the literature on MT. More important, it has not received any attention at all in under-researched language pairs, such as Arabic–English. Machine translation between language pairs from different families as the case is with English and Arabic is rather problematic because while English is an analytical language, Arabic is an agglutinative one where words are formed by combining morphemes, or units of meaning, where each morpheme typically corresponds to a distinct grammatical or semantic function, such as tense, aspect, number, or person. Studies comparing Arabic→English MT versus HT are few and far between. Therefore, in light of the growing volume of machine-translated Arabic-English texts in recent years, research on the lexical richness, or lack thereof, of Arabic→English MT vs HT is of essence.

4 RESEARCH HYPOTHESIS & RESEARCH QUESTIONS

This study aims to shed light on lexical richness in the output of Arabic-to-English neural machine translation as opposed to human translation. Is human translation more creative than that of NMT? Does it exhibit more lexical richness than NMT does? The study adopts the most agreed-upon definition of lexical richness as a superordinate term that includes the four dimensions: "lexical diversity" (lexical variation), "lexical density", "lexical sophistication" and "lexical errors". However, lexical error will be postponed for further research. It hypothesises that the output of NMT exhibits lower lexical richness than human translation.

In order to prove or disprove this hypothesis, the study attempts to answer the following research questions:

1- Using the automatic lexical variation quantitative metrics, what is the lexical variation of the output of NMT versus HT?

2- Using automatic lexical density metrics, what are the scores of NMTs versus HT?
3- Using automatic lexical sophistication metrics, what are the scores of NMTs versus HT?
4- On a scale from 0% to 100%, what are the agreement percentages between the translations offered by NMT and HT for the top 10 most frequent lemmas in the source corpus?

5 METHODS AND MATERIALS

A. Corpus of the Study

In order to carry out the experiment of measuring and comparing lexical richness between neural machine translation and human translation, the study needs a corpus that is human translated to be used as a reference, and then to have the source corpus translated again using two different neural machine translation systems: the neural Google Translate (GNMT), and the neural Microsoft Translator (MNMT). The study uses two subcorpora representing two different genres: news and literary.

The news corpus is an excerpt from the Arabic News Translation Text Part 1 (LDC2004T18) produced by the Linguistic Data Consortium (LDC) [56]. It contains Arabic news stories selected from various Arabic newspapers, and the English translation was provided by eight translation agencies who translated each Arabic news story once according to clear guidelines [56]. The translations were revised several times to assure conformance and quality. The excerpt has been made available for the researcher by LDC as a trial copy of the original corpus. The Arabic source amounts to 10,144 words. The literary subcorpus is an excerpt, relatively equal in size to the news subcorpus (11,063 words), from Naguib Mahfouz's novel, '*Awlad Haratina* (1959), and its translation into English, *Children of the Alley*, by Peter Theroux published in 1996.

B. Tools and Methods of the Study

Each of the original Arabic source texts, the news and the literary, has been separately translated into English once using Google Translate API (https://translation.googleapis.com/language/translate/v3), and another time using Microsoft Translator API (https://www.microsoft.com/en-us/translator/business/office/), resulting in two subcorpora, each consisting of three translations into English, one humanly rendered and two machine translated, with a total of six English subcorpora.

The study adopts quantitative methods to establish whether there is a measurable significant difference between the lexical richness of NMT output and that of human translation. The quantitative analyses are based on lexical richness metrics discussed above. For measuring lexical variation (diversity), the study uses Shen's [38] Lexical Richness 0.5.0, a small Python module to compute textual lexical diversity measures. For calculating lexical density, the study uses LinguaF 0.1.0, a python package for calculating famous measures of quantitative language analysis, developed by Perevalov & Lopez [57] to calculate average sentence length and average words per sentence. It also uses the Averaged Perceptron Tagger, the default python tagger of the Natural Language Toolkit (NLTK) version 3.1, for the Part-of-Speech Tagging of the 6 subcorpora [58]. The results from the PoS tagging are used to calculate the adjective density, verb density and noun density for each of the translations of the two subcorpora. As for lexical sophistication, the study uses "The Tool for the Automatic Analysis of LExical Sophistication" (TAALES 2.2) [52]. To the best of the researcher's knowledge, this is the first study in machine translation that uses TAALES in its investigations.

For the qualitative analysis, each pair of the 6 subcorpora has been rendered into a translation memory exchange format (.tmx). The segments of the news subcorpus (source text and human translation) have already been aligned by LDC. The literary subcorpus as well as the machine-translated texts of the news subcorpus have been automatically aligned using LF Aligner (https://github.com/xy-cypher/LF-aligner) and manually revised for misalignments. The .tmx files are then uploaded to SketchEngine (https://www.sketchengine.eu//) corpus management platform. SketchEngine's word list generator has been used to identify the most frequent Arabic lemmas in each of the two Arabic source texts separately, and then the parallel concordancer has been utilized to identify the translation equivalents rendered for these lemmas in each of the six subcorpora, in preparation for measuring the inter-agreement rate between the translation equivalents rendered by human translators versus NMTs.

The aim of the qualitative analysis is to compare the number of translation solutions provided by Neural Machine Translation (NMT) with the number of solutions proposed by human translators (HT). The inter-agreement percentages between the translations offered by NMT and HT for these lemmas will be computed on a scale ranging from 0% (indicating no matching translation solutions between NMT and HT) to 100% (indicating

complete agreement between NMT and HT translation solutions). This qualitative analysis has been proposed by Brglez and Vintar [53] for the sake of "a more reliable analysis of lexical diversity and to check the interpretability of quantitative methods" (p.8). This closer look will also help identify what Brglez & Vintar [53] call "mock" lexical diversity, which are in fact inconsistent and misguided translations (p. 13). All codes and corpora used in the study are available at https://github.com/RadwaKotait/LexicalRichness/

6 Discussion & Results

This section presents the results of the quantitative and qualitative analyses of the lexical richness of the two subcorpora understudy. First, the results of the metrics of lexical variation (aka diversity) are discussed, followed by those of lexical density and lexical sophistication. Second, the results of the qualitative approach measuring the inter-agreement percentage of the translation equivalents of 10 of the most frequent lemmas are analysed, shedding light on the different equivalents encountered, and the mean of agreement between those of the human translation and the outputs of the NMTs.

A. Quantitative Analysis

1) Lexical Variation: Lexical variation has been evaluated using ten different widely used metrics calculated by Lexical Richness 0.5.0. The results for lexical variation (see Table I and Table II) show that human translation in both genres, news and literary, has the highest lexical variety compared to both NMTs using different lexical variation measures. Microsoft Translator has higher lexical diversity than Google Translate in both domains.

TABLE I

Lexical Variation Scores – News Subcorpus

Human Translation		Google NMT		Microsoft NMT	
word count	11684.000000	word count	13147.000000	word count	12946.000000
unique word count	2513.000000	unique word count	2531.000000	unique word count	2532.000000
TTR	0.215080	TTR	0.192515	TTR	0.195582
RTTR	23.248595	RTTR	22.073887	RTTR	22.253376
CTTR	16.439239	CTTR	15.608595	CTTR	15.735513
MSTTR	0.876146	MSTTR	0.847848	MSTTR	0.850445
MATTR	0.877108	MATTR	0.848727	MATTR	0.849138
MTLD	75.745633	MTLD	58.503789	MTLD	59.386571
HD-D	0.864822	HD-D	0.822248	HD-D	0.823320
voc-D	117.113985	voc-D	78.933645	voc-D	81.467329
Yule's K	106.451810	Yule's K	163.114507	Yule's K	161.927539
Yule's I	4.318339	Yule's I	2.263634	Yule's I	2.353272

TABLE II

LEXICAL VARIATION SCORES – LITERARY SUBCORPUS

Human Translation		Google NMT		Microsoft NMT	
word count	16149.000000	word count	17166.000000	word count	16531.000000
unique word count	2879.000000	unique word count	2719.000000	unique word count	2725.000000
TTR	0.178277	TTR	0.158395	TTR	0.164842
RTTR	22.655249	RTTR	20.752702	RTTR	21.194196
CTTR	16.019680	CTTR	14.674376	CTTR	14.986560
MSTTR	0.892775	MSTTR	0.875102	MSTTR	0.866929
MATTR	0.891393	MATTR	0.875800	MATTR	0.868056
MTLD	94.816050	MTLD	76.478412	MTLD	73.072420
HD-D	0.872906	HD-D	0.848398	HD-D	0.845656
voc-D	126.808693	voc-D	99.639427	voc-D	98.537522
Yule's K	83.372458	Yule's K	105.456666	Yule's K	110.218917
Yule's I	3.789020	Yule's I	2.368055	Yule's I	2.454101

The analysis of both subcorpora supports the hypothesis that lexical variety is undermined in both machine-translated outputs. These results corroborate the findings reached by Toral [54], Vanmassenhove et al. [24] and Vanmassenhove et al. [40] that the process of MT does cause a general loss in terms of lexical variations when compared to human-generated translation. The qualitative analysis in the coming section will help shed better light on the details of this diversity in all six translations.

2) Lexical Density: Lexical density is measured using both Ure's [35] and Bates et al.'s [45] calculations of LD. Total LD is calculated by dividing the total count of content words (nouns, verbs, adjectives and adverbs) by the total count of tokens in the corpus. Variations of LD are calculated by dividing the count of each PoS by the total count of content words (see Table III and IV).

TABLE III

PART-OF-SPEECH COUNT IN BOTH SUBCORPORA

Translation	PoS Count	News Subcorpus	Literary Subcorpus
HT	**Noun**	4203	3748
	Verb	930	755
	Adjective	1587	2904
	Adverb	314	951
Total	**Content Words**	**7034**	**8358**
GNMT	**Noun**	4504	3946
	Verb	959	703
	Adjective	1583	2991
	Adverb	260	755
Total	**Content Words**	**7306**	**8395**
MNMT	**Noun**	4517	3706
	Verb	944	653
	Adjective	1545	2927
	Adverb	257	773
Total	**Content Words**	**7263**	**8059**

TABLE IV

LEXICAL DENSITY SCORES FOR NEWS & LITERARY SUBCORPORA

News Subcorpus				Literary Subcorpus			
	HT	**GNMT**	**MNMT**		**HT**	**GNMT**	**MNMT**
Avg. Sent Length	0.8346	0.8411	**0.8422**	Avg. Sent Length	**0.8248**	0.8135	0.8245
Avg. words per Sent.	0.8023	0.8139	**0.8219**	Avg. words per Sent.	0.7766	0.7696	**0.7767**
Total Lexical Density	**60.20%**	55.60%	56.10%	Total Lexical Density	**52%**	49%	48.80%
Noun Density	59.80%	61.60%	**62.20%**	Noun Density	44.80%	**47%**	46%
Verb Density	**22.60%**	21.70%	21.30%	Verb Density	34.70%	35.60%	**36.30%**
Adjective Density	**13.20%**	13.10%	13%	Adjective Density	**9%**	8.40%	8%
Adverb Density	**4.50%**	3.60%	3.50%	Adverb Density	**11.40%**	9%	9.60%

Note. Highest scores are highlighted in bold

The total lexical density scores for both human translations (60.2% and 52%) are higher than those for the NMTs in the news and literary subcorpora, respectively (GNMT 55.6% and 49%; MNMT 56.1% and 48.8%). Google Translate scores higher than Microsoft Translator in the literary translation, whereas Microsoft Translator displays higher lexical density than Google Translate in the news translation. It is worth noting that both NMTs score higher in the noun density than HT, while HT displays exclusively higher adjective and adverb density scores in both subcorpora. As adjectives and adverbs are content words that provide descriptive or modifying information, this indicates that humans are still more capable than machines to use vivid or specific descriptions, attributes or qualities.

3) Lexical Sophistication: Several indices calculated by TAALES representing a wide range of important aspects related to lexical sophistication have been examined. For lemmatized frequency indices, the indices derived from Thorndike-Lorge corpus of popular magazine articles (TL frequency), and the 1-million written section of the Brown corpus, or the Kučera– Francis written frequencies (KF), have been selected, as they suit the nature of the corpora understudy. For unlemmatized frequency indices and multi-word expressions (n-grams), the Corpus of Contemporary American English (COCA) has been selected, with the News subset of COCA for the news subcorpus and the Fiction subset for the literary subcorpus. The study chooses to focus on content words only as they carry meaning while functional words have little or no semantic content at all. The complete report that comprises all the 400 TAALES indices is available at the GitHub repertoire. An explanation of all the abbreviations and metric titles is available at https://tinyurl.com/44f4knr8.

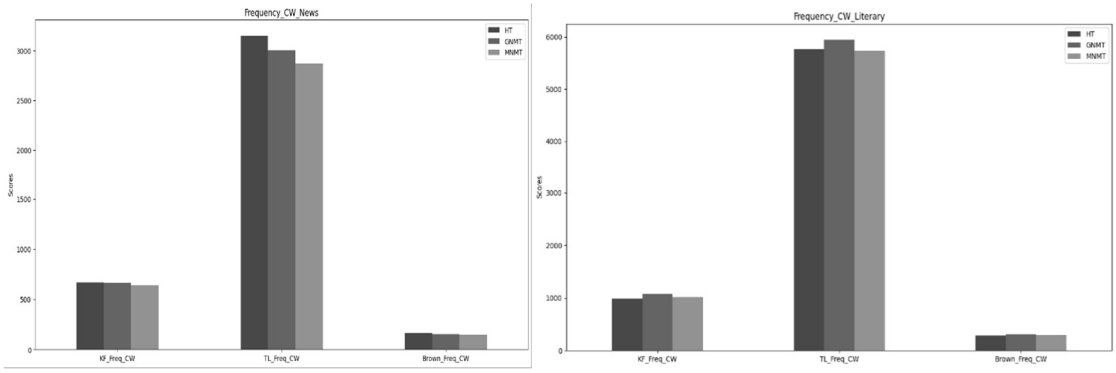

Figure 1: TAALES Word Frequency Scores for Content Words

Focusing on content words (see Figure 1), lexical sophistication scores are the highest for human translation according to KF, TL and Brown corpora (665.79, 3149.89, 161.67 respectively), followed by GNMT (663.76, 3001.91, 151.06 respectively) in the news subcorpus. As for the literary subcorpus, contrary to the previous

results, Google Translate scores the highest lexical sophistication for content words in the three indices (1076.25, 5950.69, 306.21 respectively).

Also, according to COCA subsets, focusing again on the content words, human translation scores the highest in the news subset (554.81) and the fiction subset (1156.1) versus (534.45) for news and (1150.93) for fiction in GNMT, followed at third place by MNMT at (518.87) for news and (1121.62) for fiction (see Figure 2).

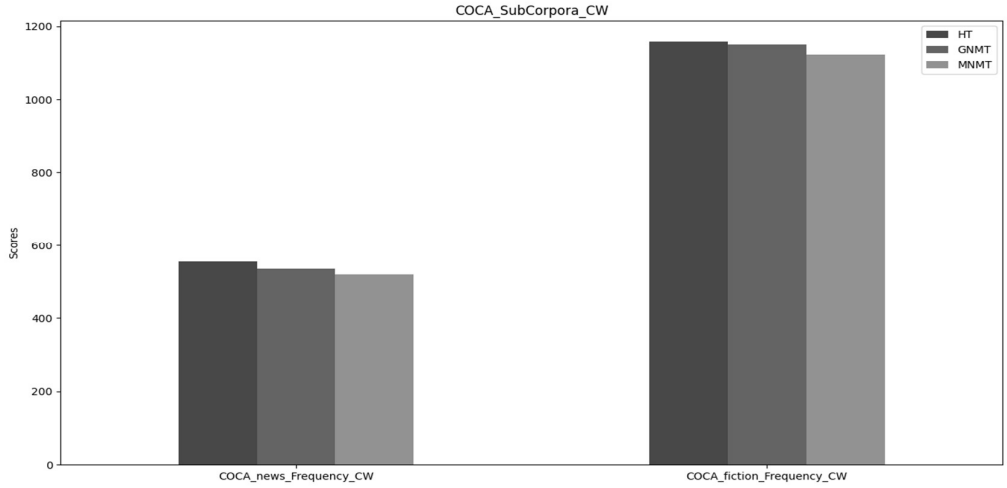

Figure 2: TAALES Unlemmatized COCA Word Frequency Scores for Content Words

The Range indices as mentioned above calculate the average number of text categories in which the words in a text occur; words that occur in all categories are general-purpose ones, while those that occur in only one category are more sophisticated. According to the range indices calculated for content words derived from both Brown and COCA corpora, human translation outperforms in terms of lexical sophistication in both genres understudy at a mean score of 0.846, while MNMT follows at 0.834, with GNMT at 0.822 for the news subcorpus, and at a mean score of 171.991 MNMT for the literary genre according to Brown Corpus, versus 168.823 for GNMT and 165.158 for MNMT. It also achieves the highest sophistication scores according to COCA, at 0.967 for the news genre and 0.419 for the literary genre (see Figure 3).

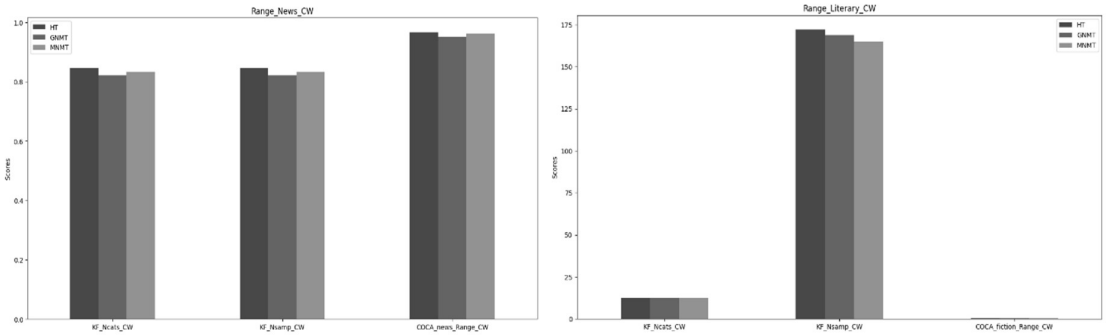

Figure 3: TAALES Range Indices for Content Words

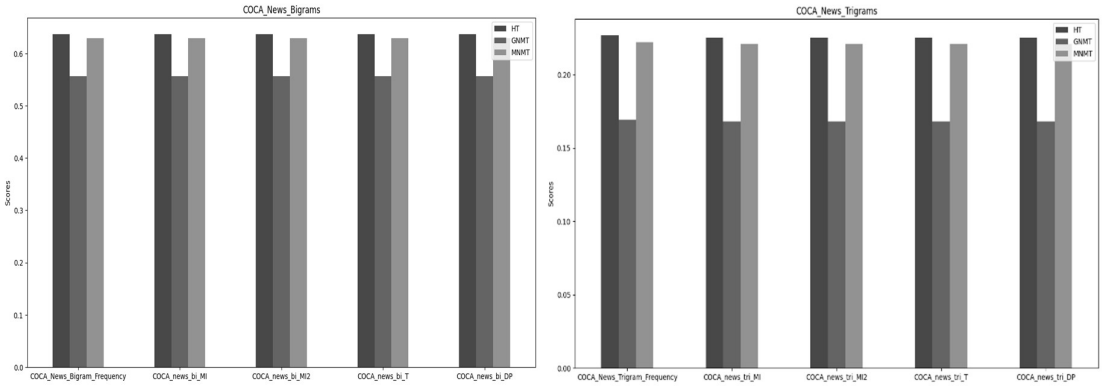

Figure 4: N-gram Scores using COCA News Subcorpus

N-grams calculated using COCA relevant subcorpus as a reference corpus display that in the news translation, human translation outperforms the two MTs in both the bi-grams and the tri-grams scores, whether using the raw count or the strength-of-association metrics (see Figure 4). In the literary translation, the results for the bigrams are different from those of the trigrams. In the bigram score, human translation is outperformed by both Microsoft Translator, and Google Translate, with MNMT scoring 233, and GNMT scoring 225 versus HT scoring 199 at the raw frequency count. However, if the strength-of-association metrics are taken into consideration, a strange discrepancy among the scores comes to the surface (see Table V). According to the tri-gram scores, human translation shows higher lexical sophistication than the two MTs.

TABLE V

N-GRAM STRENGTH-OF-ASSOCIATION SCORES USING COCA FICTION SUBCORPUS

Translation	fiction_bi_MI	fiction_bi_MI2	fiction_bi_T	fiction_bi_DP
HT	**1.500572**	9.10666	52.47077	**0.043092**
GNMT	1.469771	**9.18708**	**54.48540**	0.042447
MNMT	1.443556	9.14023	53.96235	0.041759
Translation	**fiction_tri_MI**	**fiction_tri_MI2**	**fiction_tri_T**	**fiction_tri_DP**
HT	**2.578022**	**8.257559**	**18.65283**	**0.006860**
GNMT	2.374315	8.126134	18.32766	0.006656
MNMT	2.342938	8.067431	17.86895	0.006309

Note. Highest scores are highlighted in bold.

To sum up, human translation exhibits higher lexical diversity and higher lexical density than machine translation in both subcorpora, news and literary. Lexical sophistication shows a slight variation; in the news translation, the overall result is that human translation outperforms MTs in terms of lexical sophistication according to word frequency indices with regard to general corpora and register-specific COCA (news subcorpus), as well as according to range indices and n-gram metrics. In the literary subcorpus, HT outperforms MTs according to word frequency indices using COCA fiction subcorpus, according to range indices and trigram indices. GNMT scores highest in word frequency indices that use general corpora as reference and in some bi-gram metrics.

B. Qualitative Analysis

As mentioned in the methodology, a closer look at how humans and machines deal with lexical diversity is essential to examine how humans and machines have really handled lexical diversity and to assess the interpretability of the statistical outcomes achieved by quantitative methods. It can also help detect what Brglez and Vintar [53] dub "mock" lexical diversity, which is when the translation output is a varied set of solutions, yet inconsistent, inaccurate or misguided; hence, lexical diversity is achieved but at the expense of accuracy.

Ten lemmas have been chosen from the list of most frequent lemmas extracted from the each of the two Arabic source texts using *Sketch Engine*'s Wordlist function.

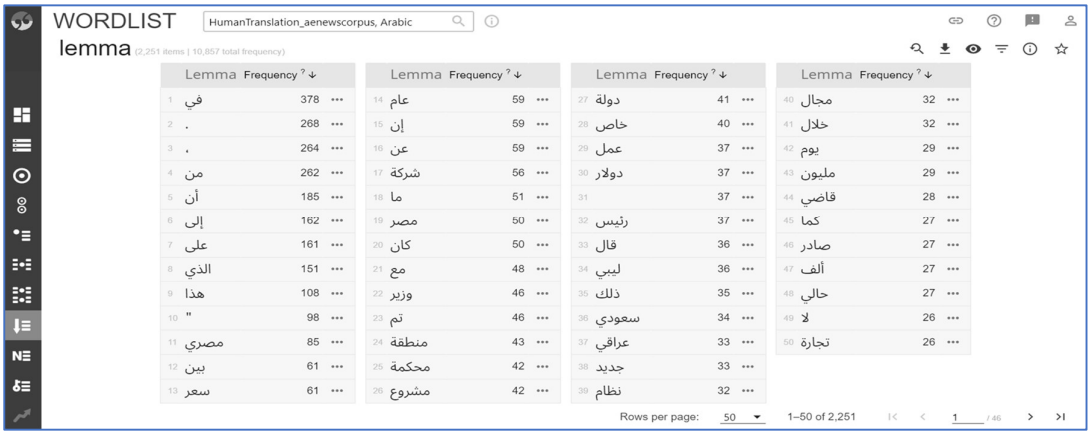

Figure 5: *Sketch Engine* Wordlist (Lemma) Screenshot

As figure 5 shows, not all lemmas returned could be used in the analysis of lexical diversity; for instance, "مصري" will most probably return only "Egyptian" as equivalent. The study chose from the above lemma list 10 lemmas that would probably return more than one equivalent so lexical diversity between human translation and machine output can be gauged. For the news subcorpus, the lemmas of choice are in order of frequency: "سعر", "منطقة", "مجال", "نظام", "عمل", "جهاز", "زيادة", "ارتفاع", "هدف" and "مجموعة". For the literary subcorpus, the lemmas analyzed are: "كبير", "غضب", "وقف", "خلاء", "أرض", "نظرة", "قوة", "حزن", "خير", and "سبيل".

The inter-agreement percentage between human translation and the two NMTs outputs in each of the two subcorpora have been calculated on a range from 0% to 100%, where no matching solutions returned 0% and complete agreement between the machine output and the human translation equivalent returned 100%.

It is clear from figure 6 that inter-agreement percentages are higher in the news subcorpus than in the literary one, which is not surprising due to the technical nature of the news discourse, where the range of equivalents is not as wide as it is in literary translation. For instance, a lemma like "نظام" (32 occurrences) in the news subcorpus, including the word forms "نظام/ النظام/أنظمة/الأنظمة", had only two equivalents in all the translations, human and machine, namely, "system" and "regime". The disagreement between the human translation and the MTs output occurred only when the human translator resorted to ellipsis and the machine translation naturally did not. By contrast, a lemma like "نظرة" (19 occurrences), in the literary subcorpus, with the word forms " نظرة/ نظرات،/النظرة/ النظرات", returned seven equivalents "glimpse", "glance(s)", "look(s)", "gaze", "glare", "stare(s)", and "glint" in the human translation, five equivalents "look(s)", "glance(s)", "stare", "gaze", "glances" in GNMT, and three equivalents "glance(s)", "gaze", "look(s)" in MNMT.

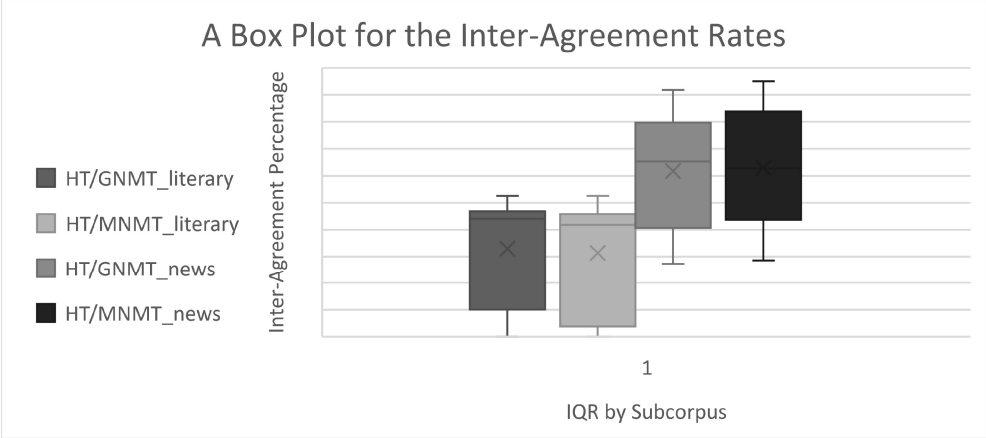

Figure 6: IQR for Inter-Agreement Rates of the Two Subcorpora

1) Qualitative Analysis of the News Subcorpus: A scrutinizing look at the news subcorpus merits that the words with the least inter-agreement rates are word 6 "جهاز" (26 occurrences), word 8 "ارتفاع" (14 occurrences), and word 9 "هدف" (11 occurrences) (see figure 7).

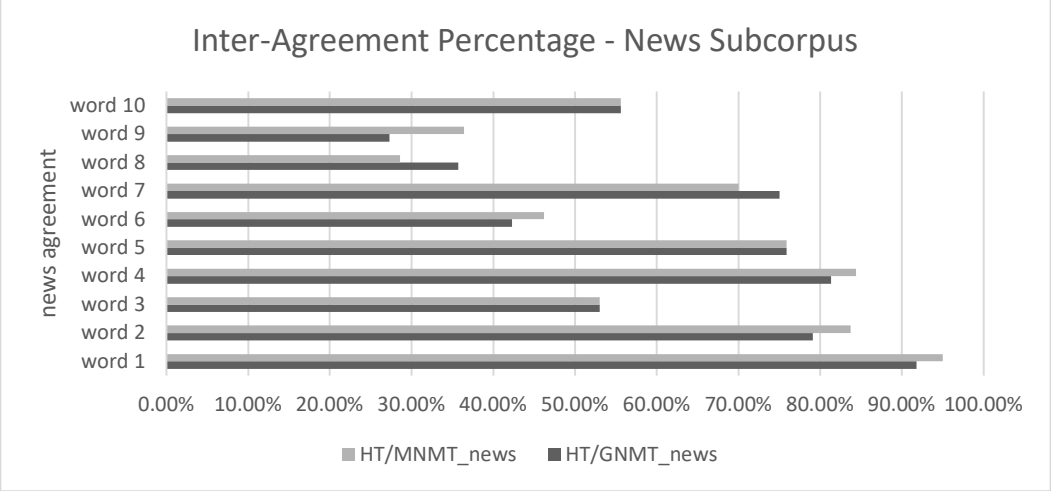

Figure 7: Inter-Agreement Percentages for the News Subcorpus

Concerning the lemma "جهاز", all the three translations agreed on translating "الجهاز المصرفي" as "banking system". The disagreement occurred in translating "جهاز تنظيم الاتصالات", the authority responsible for telecommunications in Egypt. It is formally known as the "National Telecommunications Authority" (NTA). It occurred 12 times in the Arabic source corpus and consistently translated by the human translator as "National Telecommunications Authority", using the abbreviation NTA, or elliptically translated as "the authority". Google Translate translated "جهاز تنظيم الاتصالات" as "Telecom Regulatory Authority" (4 times), "TRA" as an abbreviation of the previous translation (1 time), "National Telecommunication Authority" (1 time), and as "agency" (5 times).

TABLE VI

EQUIVALENTS FOR THE LEMMA "جهاز" IN THE NEWS SUBCORPUS

HT	Occurrences	GNMT	Occurrences	MNMT	Occurrences
System	1	system	1	system	1
Authority	12	authority	7	authority	10
agency/agencies	5	agency/agencies	6	agency	4
body/bodies	6	body	2	body	2
Appliances	1	appliances	1	appliances	1
Ellipsis	1	services	8	services	8
		devices	1		

Diverse as it is, *Google's* translation showed inconsistency within the same context where any reference to "الجهاز" should be "authority". It also failed to get the right equivalent for the governmental body; however, using an abbreviation of its own creation "TRA" to refer to the authority means it succeeded in recognizing it as a "named entity". Microsoft Translator exhibited the same ability to recognize the governmental body as a "named entity"; however, it also mistranslated it as "Telecommunication Regulatory Authority" and used "TRA" as an abbreviation 4 times. It also exhibited a behavior similar to Google's use of "agency" to refer to the authority in the same context, but not as frequently as GNMT (2 times only). Another reason for the small inter-agreement rate in dealing with "جهاز" is that both MTs prefer using "services" with "security" and "intelligence" in translating phrases, such as "الأجهزة الأمنية" or "أجهزة الاستخبارات" versus the human preference for "bodies" and "agencies". An explanation for the MTs' preference of "services" to "agencies" could be the frequency of usage; consulting the COCA, "security services" returned 865 occurrences versus 251 for "security agencies".

Looking at the lemma "ارتفاع" (14 occurrences), the human translators translated it as "rise" (4 times), "hike(s)" (3 times), "increase" as both a verb and a noun (6 times), and "have gone up" (1 time). *Google Translate* used the translation equivalents "rise" (5 times), "hike(s)" (2 times), and "increase" as both a verb and a noun (7 times), whereas Microsoft Translator used the same equivalents "rise" (6 times), "hike(s)" (2 times), and "increase" (4 times), in addition to "high" (2 times). The three translations, more or less, use the same equivalents, but differently. For instance, "ارتفاع أسعار الخامات" was translated by the human translator as "a hike in raw material prices", GNMT as "increase in the prices of raw iron", and MNMT as "high prices of iron pellets ores". So, in terms of lexical diversity, despite the low inter-agreement rate, they all used the same equivalents, but variantly.

As for the lemma "هدف" (11 occurrences), it included the word forms "بهدف، أهداف، الأهداف، أهدافها". The human translation returned the equivalents "with a view to" (4 times), "goal(s)" (3 times), "aim(s)" (2 times), "with a bid to" (1 time), and "with the objective of" (1 time). Google Translate preferred "with the aim of" (5 times), "goals" (3 times), "in order to" (2 times) and "objectives" (1 time). Microsoft Translator went for "goals" (4 times), "in order to" (4 times), and "with the aim of" (3 times). The human translator showed more lexical diversity in dealing with "بهدف" which accounts for 7 of the 11 occurrences using "with a view to, with the aim of, with a bid to, with the objective of" versus only "in order to" and "with the aim of" for GNMT and MNMT.

The rest of lemmas in the news subcorpus do not exhibit a conspicuous difference in terms of lexical variation. The lemma "منطقة", for instance, was rendered in HT as "region" and "zone", using the latter consistently in reference to "free trade zones". Both GNMT and MNMT rendered it as "region", "zone" and "area", with the latter used interchangeably with "zone" in reference to the "free trade zones". In translating the lemma "زيادة", HT used the equivalents "spread of", "increase", "hike", "raised", and "boost", whereas GNMT and MNMT used only "increase", in one of the very few instances in the news domain of human translation outperforming machine output in terms of lexical diversity.

2) Qualitative Analysis of the Literary Subcorpus: As explained above, the literary subcorpus shows conspicuously less inter-agreement rates between human translation and MT output compared to the news subcorpus, with 9 out of the 10 words below 50% inter-agreement rate. The three words that showed the least inter-agreement rates between the human renditions and the MT outputs are: word 1 "كبير", word 3 "وقف", and word 4 "خلاء".

Concerning "كبير", the reason behind this lack of agreement is that "كبير" occurs 36 times, 28 of which are references to "البيت الكبير", which is central to the events of the literary work. The human translator, Peter Theroux, chose to refer to it as "mansion" consistently all through the novel. *Google* opted for the "big house" all through the excerpt, except for a single instance of "large home". Microsoft used different variations; "big house", "large house", "great house", and "Grand House". While MNMT displayed the highest lexical diversity, this unjustifiable lack of consistency in reference to the same entity, unlike what *Google* did, is what Brglez and Vintar [53] referred to as "mock" lexical diversity. So, MNMT might show higher lexical variation in translating this instance, yet it is inaccurate and misguided.

"وقف", word 3, as a noun is a culture-specific term that plays a pivotal role in the novel as the "mansion" does. "Waqf" in *Merriam-Webster Online Dictionary* is "an Islamic endowment of property to be held in trust and used for a charitable or religious purpose" [59]. Peter Theroux translated "وقف" (27 occurrences) as "estate" (16 times), "property" (8 times), "estate property" (1 time), with a couple of ellipsis. *Google Translate* rendered it as "endowments" (21 times) and "waqf" (6 times), and *Microsoft Translator* as "endowments" (24 time) and "waqf" (3 times). "Waqf" was consistently chosen by both *Google* and *Microsoft* in rendering the phrase "*إدارة الوقف*"; otherwise, it was always "endowments". Lexical diversity was not clear here as all three translations displayed the same tendency of using one of two equivalents.

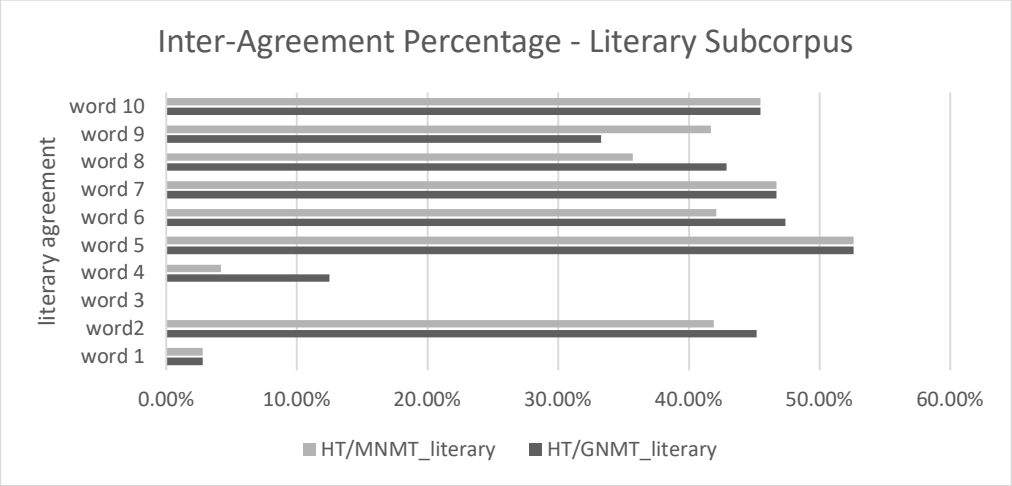

Figure 8: Inter-Agreement Percentages for the Literary Subcorpus

"الخلاء" occurred 24 times in the Arabic source text; it is another key player in the novel symbolizing where Adam ended up after his fall from grace. Peter Theroux translated it using six equivalents: "desert" (15 times), "wasteland" (4 times), "void" (1 time), "dark plain" (1 time), "open land" (1 time), and "desolate land" (1 time). *Google* rendered it as "the void" (6 times), "the open (space)" (5 times), "vacant spot/place/lot/land" (4 times), "the toilet" (3 times), "empty/emptiness" (2 times), and "desert" (2 times), also using six equivalents. *Microsoft* used "the open (space/air)" (8 times), "empty/emptiness" (7 times), "the void" (6 times), "the space" (1 time), and "the vacuum" (1 time). All three translations used diverse equivalents to express what "خلاء" is. Theroux focused on "desert" with a few "wastelands" scattered in between to describe how ugly, barren and devasted was the land where Adam landed. *Google* used more diverse equivalents than MNMT with "void", "vacant" and "open" as the most frequent equivalents.

GNMT offered a misguided translation when it used "toilet" three times, two of which as equivalents to "سيد الخلاء", ending up with the funny phrase "master of the toilet" and one time as an equivalent to " لست اليوم إلا شَقيًا أخبط في الخلاء جارًا ورائي امرأة حبلى" meaning "I hang around the desert hauling this pregnant woman along with me", with an end-product "I am banging a neighbor in the toilet", mistranslating "خلاء" as "toilet" and "جارًا" as "neighbor" instead of "hauling" or "pulling". The missing comma before "جارا" might have helped *Google* decipher the message. But, all in all, GNMT used many equivalents displaying lexical diversity, but failed to heed the context in three instances. *Microsoft Translator* resorted to less diversity than *Google* using equivalents that mostly meant "empty".

The lemma "غضب" (31 occurrences), as well as the lemma "نظرة", are the two instances that display the most lexical variance between HT and NMT (see Table VII).

TABLE VII

EQUIVALENTS FOR THE LEMMA "غضب" IN THE LITERARY SUBCORPUS

HT	Occurr -ences	GNMT	Occurr -ences	MNMT	Occurr -ences
anger/angry/angrily	14	anger/angry/angrily	31	anger/angry/angrily	29
rage/enraged	9			rage	1
Fury	3			wrath	1
Wrath	1				
Mad	2				
Exasperation	1				
Irritably	1				

Comparing the equivalents used by the human translator to those used by the NMT, it is clear that the human translator outperformed the machine in coming up with more colorful equivalents to describe "anger" in the different situations it occurred in. Between the "surge of rage", the "escalating fury", the "times of wrath", and the "waves of fear and exasperation", it is clear that the human translator used more diverse equivalents than

Google's and Microsoft's mere "anger" and "angry". The same tendency was noted in the translation of "نظرة", between catching a "glimpse", exchanging "glances", unbearable "glares", rude "stares" and strange" glints", the human translator clearly used more diverse and colorful equivalents than both MTs did with their mere "looks" and "glances".

7 CONCLUSION

The aim of the experiments, quantitative and qualitative, has been to answer the question: whether lexical richness, manifested in lexical variation, lexical density and lexical sophistication, is adversely affected by the algorithmic bias in neural machine translation. The two subcorpora have been tested to show if these results are consistent in both technical and literary domains. The quantitative metrics for the lexical variation and density have conclusively presented evidence that human translation demonstrates a greater lexical diversity and density than machine translations, in both domains technical and literary. Comparing one neural machine translation to another, Microsoft's MT exhibits greater lexical diversity than Google's, and higher lexical density in the literary domain, with Google having the upper hand in the news domain in terms of density alone.

In terms of lexical sophistication, the results do not show a consistent trend. In the news domain, according to word frequency, range and n-gram indices, human translation is predominantly more lexically sophisticated than both NMTs. In the literary domain, the results are not as conclusive. General content word frequency indices show that *Google* outperforms the human translation, whereas according to register-specific content word frequency and range indices the human translator uses more sophisticated content words than both MTs. N-grams indices show a great variation; the general trends show that NMTs are lexically more sophisticated than human translation in the bigrams used in their literary output, but not in using trigrams where humans are definitely lexically more advanced. Hence, human translation is predominantly lexically richer than the machine output in terms of diversity and density in both domains, technical and fiction. In terms of using more complex and advanced language, the quantitative metrics support the human dominance in news translation, but not in the literary one.

An explanation of this deviation from the general trend of the results returned, especially in the analysis of sophistication in the literary subcorpus, might be due to the fact that TAALES will only calculate lexical features for words within its databases; if a word does not exist in its lists, it will be completely ignored. This point has been expressed by the creators of TAALES as one of its limitations. As the excerpt from the novel representing the literary subcorpus is definitely abundant with lexical items that are too cultural-specific in their nature, it is only logical that these items would be overlooked, returning results that might not be accurately representative of the text processed.

The closer, scrutinizing look that the qualitative analysis yields has shown that in the news domain, the inter-agreement rate between human translation and MT output has a mean of 60%. Both human and machine propose almost the same number of equivalents and use them equally variantly. Mismatches between human and machine equivalents are in general not due to mistranslations or inconsistency expect in a very few instances, but rather due to using a different acceptable alternative. In the literary translation, inter-agreement rates are conspicuously low with a mean of 30%. Mismatches in the case of literary translation are in fact due to a clearly impoverished language on part of the machine, or due to misguided translations in most of the lemmas understudy. So, NMT might provide a fairly readable rendering of a literary work; however, until the present moment, human translation exhibits a higher level of "creativity".

The outcome of the study supports that of Toral's [54] study where HTs have been found to be lexically richer than MTs, both statistical and neural. It also supports the findings of both Vanmassenhove et al. [24] and Vanmassenhove, et al [40] that there is indeed a loss of lexical richness in the translations produced by MT paradigms in the language pairs they studied. Unlike the findings of Brglez and Vintar [53] where the automatic metrics they used to measure lexical diversity put HT "at the very bottom of the lexical diversity ladder" in the majority of cases, human translation in almost all automatic metrics in the present study passes with distinction.

The study has addressed lexical richness in its wide sense in the output of both humans and neural machine translation engines from Arabic into English. In general, automatic metrics have been in favour of the human translation whether in technical or fiction domains. The closer look that the qualitative analysis allows shows that in the technical field, machine translation is not far behind human translation in terms of lexical diversity. However, in literary translation, the machine's inherent tendency to overgeneralize using more frequent words result in a clear loss of lexical diversity and colourfulness.

The limitations of this study include a focus on a single translation direction in the language pair Arabic-English, as well as a focus on two genres only, with news as a representative of the technical domain. The corpora under study have been limited in size as well due to the lack of available datasets in the language pair

Arabic/English. For further studies, different genres could be analysed as well as a different translation direction, although the problem with the English-into-Arabic direction is that not all quantitative metrics will be applicable to Arabic.

REFERENCES

[1] W. Frawley, *Translation. Literary, Linguistic & Philosophical Perspectives*, University of Delaware Press, 1984.

[2] M. Gellerstam, "Translationese in Swedish novels translated from English", In L. Wollin and H. Lindquist (Eds.), *Translation Studies in Scandinavia*, pp. 88-95, CWK Gleerup, 1986.

[3] M. Baker, "The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators", *International Journal of Corpus Linguistics*, vol. 4, no. 2, pp. 281–298, 1999.

[4] M. Baroni and S. Bernardini, "A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text", *Literary and Linguistic Computing*, vol. 21, no. 3, pp. 259–274, 2005. https://doi.org/10.1093/llc/fqi039

[5] M. Koppel, and N. Ordan, "Translationese and Its Dialects", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* pp. 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics, 2011. https://aclanthology.org/P11-1132.pdf

[6] V. Volansky, N. Ordan, and S. Wintner. "On the features of translationese", *Digital Scholarship in the Humanities*, vol. 30, no. 1, pp. 98–118, 2015. https://doi.org/10.1093/llc/fqt031

[7] I. Ilisei, D. Inkpen, G.C. Pastor, and R. Mitkov, "Identification of translationese: a machine learning approach". *Proceedings of the 11th international conference on Computational Linguistics and Intelligent Text Processing (CICLing'10),* pp. 503–511, Springer-Verlag, Berlin, Heidelberg, 2010. https://doi.org/10.1007/978-3-642-12116-6_43

[8] N. Aranberri, "Can translationese features help users select an MT system for post-editing?", *Procesamiento del Lenguaje Natural*, vol. 64, no. 11, pp. 93-100, 2020. https://doi.org/10.26342/2020-64-11

[9] A. Frankenberg-Garcia, "Can a corpus-driven lexical analysis of human and machine translation unveil discourse features that set them apart?", *Target. International Journal of Translation Studies*, vol. 34, no. 2, pp. 278 – 308, 2022. https://doi.org/10.1075/target.20065.fra

[10] M.O. Prates, P.H. Avelar, and L. Lamb, "Assessing gender bias in machine translation: a case study with Google Translate", *Neural Computing and Applications*, vol. 32, pp. 6363-6381, 2018.

[11] N. Habash, J. Olive, C. Christianson, J. McCary, "Machine Translation from Text", In J. Olive, C. Christianson, and J. McCary (Eds.) *Handbook of Natural Language Processing and Machine Translation,* pp. 133–397, Springer, 2011. https://doi.org/10.1007/978-1-4419-7713-7_2

[12] Z. Tan, S. Wang, Z. Yang, G. Chen, X.C. Huang, M.S. Sun, and Y. Liu, "Neural machine translation: A review of methods, resources, and tools", *AI Open*, vol. 1, pp. 5–21, 2020. https://doi.org/10.1016/j.aiopen.2020.11.001

[13] DeepAI. (n.d.). "Neural Machine Translation Definition", https://deepai.org/machine-learning-glossary-and-terms/neural-machine-translation, (accessed 1 August 2023).

[14] N. Roberts, D. Liang, G. Neubig, and Z.C. Lipton, "Decoding and Diversity in Machine Translation". *CoRR*, abs/2011.13477, 2020. https://arxiv.org/abs/2011.13477

[15] K. Kirkpatrick, K. "Battling algorithmic bias: how do we ensure algorithms treat us fairly?", *Communications of the ACM*, vol. 59, no. 10, pp. 16–17, 2016. https://doi.org/10.1145/2983270

[16] L. Schiebinger, "Scientific research must take gender into account", *Nature*, vol. 507(7490), no. 9, 2014. https://doi.org/10.1038/507009a

[17] S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16),* pp. 2125–2126, Association for Computing Machinery, New York, NY, USA, 2016. https://doi.org/10.1145/2939672.2945386

[18] M.A. Malek, "Criminal courts' artificial intelligence: The way it reinforces bias and discrimination", *AI Ethics*, vol. 2, pp. 233–245, 2022. https://doi.org/10.1007/s43681-022-00137-9

[19] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. Chang, "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints", *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* pp. 2979–2989, Copenhagen, Denmark, 2017.

[20] M.R. Costa-Jussà, and A. de Jorgem, "Fine-tuning neural machine translation on gender balanced datasets", *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing,* pp. 26–34, Association for Computational Linguistics, Barcelona, Spain (Online), 2020.

[21] D. Saunders, and B. Byrne, "Reducing gender bias in neural machine translation as a domain adaptation problem", *Proceedings of the 58th Annual Meeting of the Association for Computational* Linguistics, pp. 7724–7736, Barcelona, Spain (Online), Association for Computational Linguistics, 2020.

[22] D. Saunders, R. Sallis, and B. Byrne, "Neural machine translation doesn't translate gender coreference right unless you make it", *Proceedings of the Second Workshop on Gender Bias in Natural Language* Processing, pp. 35–43, Barcelona, Spain (Online). Association for Computational Linguistics, 2020.

[23] M. Garcia, "Racist in the machine: The disturbing implications of algorithmic bias". *World Policy Journal*, vol. 33, no. 4, pp. 111–117, 2016.

[24] E. Vanmassenhove, D. Shterionov, and A. Way, "Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation", In *Proceedings of Machine Translation Summit XVII: Research Track*, pp. 222–232, Dublin, Ireland, European Association for Machine Translation, 2019. https://aclanthology.org/W19-6622

[25] G. U. Yule, *The statistical study of literary vocabulary,* Cambridge University Press, 1944.

[26] S. Jarvis, "Defining and measuring lexical diversity". In S. Jarvis and M. Daller (Eds.), *Vocabulary Knowledge: Human ratings and automated measures [Studies in Bilingualism 47],* pp. 13–44, John Benjamins, 2013. https://doi.org/10.1075/sibil.47

[27] H. Daller, R. Van Hout, and J. Treffers-Daller, "Lexical richness in the spontaneous speech of bilinguals". *Applied Linguistics*, vol. 24, no. 2, pp. 197–222, 2003. https://doi.org/10.1093/applin/24.2.197

[28] C.A. Engber, "The relationship of lexical proficiency to the quality of ESL composition", *Journal of Second Language Writing*, vol. 4, no. 2, pp.139–155, 1995.

[29] B. Laufer, and P. Nation, "Vocabulary Size and Use Lexical Richness in L2 Written Production", *Applied Linguistics*, vol. 16, no.1, pp. 307-322, 1995. https://doi.org/10.1177/026553229901600103

[30] J. Read, *Assessing Vocabulary,* Cambridge University Press, 2000. http://dx.doi.org/10.1017/CBO9780511732942

[31] P. M. McCarthy, and S. Jarvis, "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment", *Behavior Research Methods*, vol. 42, pp. 381–392, 2010. https://doi.org/10.3758/BRM.42.2.381

[32] J. W. Chotlos, J. W., "A statistical and comparative analysis of individual written language samples", *Psychological Monographs*, vol. 56, no.2, pp. 77-111, 1944. https://doi.org/10.1037/h0093511

[33] M. Templin, *Certain language skills in children*, University of Minnesota Press, 1957.

[34] W. Johnson, "Studies in language behavior: A program of research", *Psychological Monographs*, vol. 56, no. 2, pp. 1-15, 1944.

[35] J. Ure, "Lexical density and register differentiation", In G. Perren, and J. Trim (Eds.) *Applications of Linguistics: Selected Papers of the Second International Congress of Applied Linguistics,* pp. 443-452, Cambridge University Press, 1971.

[36] J. Treffers-Daller, P. Parslow, and S. Williams, "Back to Basics: How Measures of Lexical Diversity Can Help Discriminate between CEFR Levels", *Applied Linguistics*, vol. 39, no. 3, pp. 302–327, 2018. https://doi.org/10.1093/applin/amw009

[37] G. Herdan, *Quantitative linguistics,* Butterworths, 1964.

[38] L. Shen, "LexicalRichness: A small module to compute textual lexical richness", *MIT license*, 2022. https://doi.org/10.5281/zenodo.6607007

[39] M.A. Covington, "MATTR user manual (CASPR Research Report 2007-05)", Athens, University of Georgia Institute for Artificial Intelligence, 2007.

[40] E. Vanmassenhove, D. Shterionov, and M. Gwilliam, "Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation", *CoRR,* abs/2102.00287, 2021. https://arxiv.org/abs/2102.00287

[41] J. D. Cortés, "Journal titles and mission statements: Lexical structure, diversity, and readability in business, management and accounting research", *Journal of Information Science*, vol. 49, no. 5, 2021. https://doi.org/10.1177/01655515211043707

[42] D. Malvern, B. Richards, N. Chipere, and P. Durán, *Lexical diversity and language development: Quantification and assessment,* Palgrave Macmillan, 2004.

[43] P. M. McCarthy, and S. Jarvis, "Voc-D: A theoretical and empirical evaluation", *Language Testing*, vol. 24, no. 4, pp. 459–488, 2007. https://doi.org/10.1177/0265532207080767

[44] P.M. McCarthy, "An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)", Doctoral dissertation, University of Memphi,. ProQuest Dissertations and Theses Global, 2005.

[45] E. Bates, I. Bretherton, L. Snyder, M. Beeghly, C. Shore, S. McNew, V. Carlson, C. Williamson, A. Garrison, et al., *From first words to grammar: Individual differences and dissociable mechanisms*, Cambridge University Press, 1998.

[46] K. Kyle, and S.A. Crossley, "Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application". *Tesol Quarterly*, vol. 49, no. 4, pp. 757-786, 2015. https://doi.org/10.1002/tesq.194

[47] K. Kyle, and S.A. Crossley, "The relationship between lexical sophistication and independent and source-based writing", *Journal of Second Language Writing*, vol. 34 (December 2016) pp. 12-24, 2016. https://doi.org/10.1016/j.jslw.2016.10.003

[48] I.S.P. Nation, and A. Heatley, A. (1994). Range: A program for the analysis of vocabulary in texts [software], 1994, Available from: http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx, (accessed 12 August 2023).

[49] A. Heatley, I.S.P. Nation, and A. Coxhead, (2002). RANGE and FREQUENCY programs. Available from: http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

[50] A. C. Graesser, D.S. McNamara, M.M. Louwerse, et al. "Coh-Metrix: Analysis of text on cohesion and language", *Behavior Research Methods, Instruments, & Computers*, vol. 36, pp. 193–202, 2004. https://doi.org/10.3758/BF03195564

[51] J.W. Pennebaker, C.K. Chung, M. Ireland, A. Gonzales, and R.J. Booth, "The development and psychometric properties of LIWC2007". *Austin, TX, LIWC. Net,* 2007.

[52] K. Kyle, S.A. Crossley, and C. Berger, "The tool for the analysis of lexical sophistication (TAALES): Version 2.0.", *Behavior Research Methods,* vol. 50, no. 3, pp. 1030-1046, 2018. https://doi.org/10.3758/s13428-017-0924-4

[53] M. Brglez, and Š. Vintar, "Lexical Diversity in Statistical and Neural Machine Translation". *Information*, vol. 13, no. 93, 2022. https://doi.org/10.3390/info13020093

[54] A. Toral, "Post-editese: an exacerbated translationese", *Proceedings of Machine Translation Summit XVII: Research Track*, pp. 273–281, Dublin, Ireland. European Association for Machine Translation, 2019. https://aclanthology.org/W19-6627

[55] S. Castilho, N. Resende, and R. Mitkov, "What influences post-editese features? A preliminary study", *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, pp. 19-27, Varna, Bulgaria, 2019. https://aclanthology.org/W19-8703.pdf

[56] X. Ma, D. Zakhary, and M. Bamba, "Arabic News Translation Text Part 1 LDC2004T17". (Web Download). Philadelphia: Linguistic Data Consortium, 2004. https://doi.org/10.35111/qhv1-1z67

[57] A. Perevalov, and J. Lopez, (2021). LinguaF. [Computer Software]. Available from: https://github.com/WSE-research/LinguaF (accessed 12 August 2023).

[58] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: Analyzing text with the natural language toolkit*, O'Reilly Media, 2009.

[59] Merriam-Webster. (n.d.). "Waqf", In *Merriam-Webster.com Dictionary,* (accessed August 1, 2023), https://www.merriam-webster.com/dictionary/waqf

BIOGRAPHY

Radwa Mohammad Kotait

Associate Professor of Translation & Interpreting Studies

Radwa Kotait (https://orcid.org/0000-0003-4355-6300) is an Associate Professor of Translation and Interpreting Studies at the English Department, Faculty of Al-Alsun, Ain Shams University. With over 22 years of experience as a freelance translator, she brings both academic rigor and practical expertise to her field. Her research, focusing on cognitive and corpus-based approaches to translation, has earned her both an MA and PhD in the field. Her dedication extends beyond her own work; she actively shapes the next generation of translators and interpreters as an editor for *Textual Turning* journal, an annual, academic, peer-reviewed journal, and a mentor to numerous MA and PhD students. Her influence is evident in her high-profile translation projects, including Kwame Appiah's *The Honor Code: How Moral Revolutions Happen* into Arabic published by Hindawy and her recent translation into English, *The Journey of the Holy Family in Egypt: A Journey of Blessings*, published by the National Center for Translation. Her leadership positions, such as heading the Faculty of Al-Alsun's Innovation Hub (iHub), further underscore her commitment to advancing the field. She remains at the forefront of technological innovation, holding a certificate in data analytics and programming, with a keen interest in how Natural Language Processing and Machine Learning can enhance translation practices.

ARABIC ABSTRACT

ضياع الثراء اللغوي في لغة الترجمة الآلية: الثراء المعجمي في الترجمة البشرية مقابل الترجمة الآلية العصبية من العربية إلى الإنجليزية

رضوى قطيط

قسم اللغة الإنجليزية، كلية الألسن، جامعة عين شمس، القاهرة، جمهورية مصر العربية

radwa_kotait@alsun.asu.edu.eg

المستخلص

*قد يذهب البعض إلى أن الترجمة الآلية العصبية **(NMT)** أسرع وأفضل من الترجمة البشرية .ومع ذلك، فإن الترجمة الآلية بطبيعتها تعمم الأنماط الأكثر ظهوراً في بيانات التدريب على حساب الأنماط الأقل ظهوراً في ظاهرة تُسمى "لغة الترجمة الآلية". وقد لوحظ أن هذه اللغة الآلية تعكس بعضًا من أوجه التحيز المثير للجدل. واحدة من أوجه هذا التحيز الآلي، والذي يتم تجاهله في الغالب، هو فقدان "الثراء المعجمي". لم يلحظ إلا مؤخراً أن الترجمات الآلية فقيرة لغويًا ومتأثرة سلباً بميل الترجمة الآلية العصبية إلى التعميم المفرط. وعلى الرغم من قيمة الثراء اللغوي، لم يحظ بنفس الاهتمام الذي حظيت به الدقة المعجمية وقياس الخطأ. والأهم من ذلك، أنه لم يحظ بأي اهتمام على الإطلاق في أزواج اللغات التي لا تحظى بالاهتمام البحثي الكافي، مثل العربية-الإنجليزية. تهدف هذه الدراسة إلى تسليط الضوء على الثراء اللغوي في مخرجات الترجمة الآلية العصبية من العربية إلى الإنجليزية مقارنة بالترجمة البشرية لتجيب على السؤال :هل تظهر الترجمة البشرية ثراءً معجميًا أكبر من الترجمة الآلية العصبية؟ تتبنى الدراسة التعريف الأكثر توافقاً عليه للثراء المعجمي باعتباره مصطلحاً رئيسياً يشمل "التنوع المعجمي" و"الكثافة المعجمية" و"التعقيد المعجمي"؛ وكلها مقاييس إحصائية تقيس الثراء اللغوي للنص. تحلل الدراسة مخرجات نظامين للترجمة الآلية العصبية وهما **Google Translate** و**Microsoft Translator** من حيث الثراء اللغوي، باستخدام كل من الطرق الكمية والنوعية، ثم تقارن النتائج بنتائج مخرجات الترجمة البشرية. تعتمد الدراسة على تحليل ذخيرتين لغويتين، الأولى تتناول نصوص إخبارية والثانية تتناول نصوص أدبية .*

***الكلمات المفتاحية**: الكثافة المعجمية؛ التنوع المعجمي؛ الثراء المعجمي؛ التعقيد المعجمي؛ الترجمة الآلية العصبية؛ لغة الترجمة الآلية.*