information

MDPI

Article

Lexical Diversity in Statistical and Neural Machine Translation

Mojca Brglez * and Špela Vintar

Faculty of Arts, University of Ljubljana, Aškerčeva Cesta 2, 1000 Ljubljana, Slovenia; spela.vintar@ff.uni-lj.si
* Correspondence: mojca.brglez@ff.uni-lj.si

Abstract: Neural machine translation systems have revolutionized translation processes in terms of quantity and speed in recent years, and they have even been claimed to achieve human parity. However, the quality of their output has also raised serious doubts and concerns, such as loss in lexical variation, evidence of "machine translationese", and its effect on post-editing, which results in "post-editese". In this study, we analyze the outputs of three English to Slovenian machine translation systems in terms of lexical diversity in three different genres. Using both quantitative and qualitative methods, we analyze one statistical and two neural systems, and we compare them to a human reference translation. Our quantitative analyses based on lexical diversity metrics show diverging results; however, translation systems, particularly neural ones, mostly exhibit larger lexical diversity than their human counterparts. Nevertheless, a qualitative method shows that these quantitative results are not always a reliable tool to assess true lexical diversity and that a lot of lexical "creativity", especially by neural translation systems, is often unreliable, inconsistent, and misguided.

Keywords: machine translation; neural translation systems; lexical diversity; type-token ratio; measure of textual lexical diversity

Citation: Brglez, M.; Vintar, Š. Lexical Diversity in Statistical and Neural Machine Translation. *Information* **2022**, *13*, 93. https://doi.org/10.3390/info13020093

Academic Editor: Marcos Zampieri

Received: 27 December 2021
Accepted: 10 February 2022
Published: 15 February 2022

1. Introduction

In the past couple of years, an abundance of automatic systems for translation have emerged, a lot of them available to the general public. The older phrase-based systems have given way to newer, "cleverer" neural machine translation systems that have been considered state-of-the-art for some years. These general translation systems offer translation on-the-go and can supposedly handle a wide range of texts and genres, purportedly excelling at newer contexts and unseen data (out of vocabulary words). Not only are they considered faster and better, for some well-resourced languages, they have already been claimed to achieve human parity [1]).

Machine translations have been frequently evaluated on the basis of automatic quality and error-measuring metrics, such as BLEU [2] or METEOR [3], as well as various human evaluation methods. Studies have shown these do not always correlate [4,5], which shows that excelling at automatic metrics is not always the best indicator of quality. Furthermore, researchers have also raised serious concerns about machine translation (MT), such as loss of lexical variation in the target text [6–8], warning of a potential lexical impoverishment of the target language [9] and the dangers of language learners developing a "warped exposure" to that language through neural machine translation (NMT) [8].

1.1. Related Work

Akin to human translations that have been studied in terms of their specific "translationese" that manifests in recurring tendencies in the language of translation (dubbed by some as "universals" [10]), machine translations are being investigated for "machine translationese", primarily on the basis of quantitative measures. In studies of both human and machine translations, researchers try to (dis)prove the existence of tendencies such as simplification, explicitation, and inference.

Toral [11] studied lexical density and diversity and found that post-edited machine translations, compared to human translations, were more simplified, normalized, and exhibited more "shining-through" from the original text. Machine translations, in general, exhibited a lower lexical density than human translations, while the neural systems' translations had lower lexical density than those produced by statistical systems. On the subject of lexical variety (measured by type–token ratio), he finds that human translations are, consistently, lexically the richest, followed by statistical and neural translations. Castilho, Resende, and Mitkov [12] obtain different results with regard to the observed genre. While machine translations of news exhibit slightly higher lexical density and lexical richness compared to human translations (HT), machine translations of literary texts have a slightly lower lexical density than HT but a similar lexical richness ratio.

The specific topic of lexical richness is addressed by Vanmassenhove, Shterionov, and Way [6] who study the output of 12 different machine translation systems with original and back-translated data. They observe the effect of loss of lexical richness, the increase in frequency of more frequent words, and the decrease in frequency of less frequent words. They also compare phrase-based to neural systems, where the former exhibit higher lexical variety than neural systems. They hypothesize that the even greater loss in linguistic variation of neural translations is due to their larger susceptibility to (over)generalization, always choosing only the most probable solutions and ignoring other rarer words [6] (p. 224). In another experiment specifically studying algorithmic bias in training machine translation systems, Vanmassenhove, Way and Gwilliam [7] compare training data (human translations) and the output of machine translation systems trained on the same dataset using different architectures (phrase-based statistical system, neural long short-term memory (LSTM) network, and neural transfomers). They find that machine translations by phrase-based systems are the least diverse. However, contrary to the study of [6], the neural systems include the byte-pair encoding functionality (BPE), which enables them to translate rare or unseen words by segmenting them into smaller parts. As a result of this, the neural transformer models now consistently showed higher lexical diversity than neural LSTM and phrase-based statistical models. They also note that on average, the diversity metrics correlate with translation quality metrics [7] (p. 2212). Their findings on lexical diversity are confirmed by de Clercq et al. [13] for the English–French pair on news texts. Their study expanded the investigation to 22 specific linguistic features (including TTR, *hapax legomena*, frequency of word classes, and counts of n-grams), which were compared between various translation systems as well as to an original (non-translated) French language corpus. Other studies, such as the one by Čulo and Nitzke [14], address terminology variation and cognates in human, machine, and post-edited translations. Their study shows that MT translations offer multiple solutions less often, but when they do, they exhibit much stronger variation, which is an effect also carried over to post-edited texts. Compared to HT, MT was also shown to much more frequently propose cognates as translation solutions, meaning choosing an orthographically similar word over an orthographically different but more appropriate translation solution for the target language.

1.2. Motivations

Studies for a range of languages have claimed that machine translations exhibit some recurring trends, hinting at translation universals, such as loss of lexical density and lexical variation—with other studies showing quite the contrary. It seems then that these "universals of (machine) translation" are not that universal but largely depend on the text genre, language combination, and translation direction, as well as other factors, such as the specific architectures of machine translation systems involved. While many studies report on other language combinations involving neural machine translation (as well as their comparisons to older statistical translation systems), less resourced languages such as Slovenian are still a relatively poorly investigated territory. For Slovenian, a comparison between neural (NMT) and statistical machine translation (SMT) in terms of automatic quality estimation was made by Arčan [15] and Donaj and Sepesy Maučec [16], and by

Vintar [17], who complemented the study with an analysis of terminology translation. Focusing on karst terminology, she finds improvements in translation accuracy of NMT over SMT for the English–Slovenian direction but no significant difference in the reverse direction. While the most common mistakes in SMT are non-translated words and wrong translations due to unsuccessful disambiguation, NMT more frequently produces "strange", made-up translations and shows a great amount of inconsistency.

In line with previous studies, one of the first questions we pose is whether machine translations differ from human translations on a general, quantitatively measurable level, and in what way. Moreover, questions arise as to what neural systems in particular bring to the table compared to their older statistical counterparts. Are they really more similar to human translations, and do they exhibit more "creativity" in terms of lexical variation? Do they really better adjust their solutions to the context than phrase-based statistical models? In this study, we focus on translations from English to Slovenian and choose to look specifically at lexical diversity in human vs. various machine translations. A lower lexical diversity in MT would indicate a less varied and "creative" output with a smaller set of translation equivalents than that proposed by a human translator. Based on the majority of previous studies [4,6,11], we hypothesize that machine translations exhibit lower lexical diversity than human translations but that neural machine translations have a higher lexical diversity than statistical machine translations. Given that previous studies achieved inconsistent results with regard to genre and the type of MT engine used, we explore these parameters in our study by including three different genres (a literary novel, a technical manual, and a cookbook) as well as two neural and one statistical translation system.

2. Materials and Methods

2.1. Data and Tools

First, we select four texts in English with existing human translations in Slovenian. To be able to carry out a fairly reliable quantitative study of lexical diversity, we choose three different genres and lengthier texts comprised of at least 50,000 words. Our corpus is comprised of an information technology (IT) subcorpus consisting of a printer instruction manual [18,19] and a printer user guide [20,21], a culinary subcorpus (CUL) consisting of a book of recipes [22,23], and a literary subcorpus (LIT) consisting of a popular fiction novel [24,25]. The size of the subcorpora is detailed in Table 1. The first two texts are freely available on the internet (accessed in March 2020), while the last two were obtained with special permission from the publishers and were compiled as a .tmx aligned corpus in the context of a student project aimed at training translation systems in 2019.

Table 1. Text categories and sizes.

	English Original		Slovenian Translation	
Domain	**Title**	**Words**	**Title**	**Words**
Information technology (IT)	*LASERJET PRO 300 COLOR MFP / LASERJET PRO 400 COLOR MFP User Guide* and *Installation Guide*	50,146	*LASERJET PRO 300 COLOR MFP / LASERJET PRO 400 COLOR MFP Uporabniški priročnik* and *Priročnik za namestitev*	45,206
Culinary Arts (CUL)	*The Cooking Book*	127,853	*Dobra kuha*	107,373
Literature (LIT)	*Practice Makes Perfect*	76,951	*Osem let skomin*	72,604

Then, each of the original English texts was separately translated by three different machine translation systems: the statistical Google Translate (GSMT), the neural Google Translate (GNMT), and the neural Microsoft Translator (MNMT). The translations were

obtained through Google API (https://translation.googleapis.com/language/translate/v3, accessed on 15 March 2020) and the built-in Microsoft Office Word Translator (https://www.microsoft.com/en-us/translator/business/office/, accessed on 15 March 2020).

For further analyses, each of the original source texts was separately aligned with each of its translations and exported into the translation memory exchange format .tmx, after which we uploaded them to the corpus management platform SketchEngine (www.sketchengine.eu, accessed on 15 June 2020) [26]. The platform was also used to gather information for all further analyses, including lemmatization, part-of-speech-tagging, frequency counting, and concordance search.

2.2. Analyzing Lexical Diversity

We compute lexical diversity for each of the subcorpora to corroborate the findings by previous studies, namely that lexical diversity is lower in machine translations versus human translations and that translations by neural systems have lower lexical diversity than phrase-based systems [6,11]. Lexical diversity or variety can be computed by various methods, such as voc-D, HD-D, MTLD, TTR, Maas, and others. For a global diversity of the vocabulary, we use the automatic metrics TTR (type–token ratio) and MTLD (measure of textual lexical diversity), which was first proposed by McCarthy [27] and later deemed one of the best measures of lexical diversity [28,29]. TTR is measured as a simple overall ratio between tokens and types, where types are orthographically unique words and tokens represent the total number of words. However, this ratio is very sensitive to text length because very common and function words are bound to repeat, which is why shorter texts tend to have a higher TTR and longer texts tend to have a lower TTR (we use the non-standardized version of TTR. The standardized (sTTR) has been proposed to tackle the sensitivity to length by calculating the TTR for every n-running words and averaging the ratios for a final result). The second metric, MTLD, tries to account for and avoid the influence of text length. The calculation of MTLD is a sequential analysis of text chunks in both directions, the result of which tells us the average length of text that maintains a predefined TTR threshold. TTR was computed on the basis of SketchEngine data, while MTLD was computed using the lexical-diversity Python module [30]. We compute lexical diversity only for Slovenian translations as the lexical diversity is not directly comparable between English and Slovenian in view of the extremely inflectional nature of Slovenian.

However, apart from the quantitative analysis using automatic metrics, we also conduct a quantitative analysis of selected keywords and multi-word expressions. We select 10 keywords and 15 multi-word expressions using the SketchEngine "Extract Keywords and Terms" tool. We limit the extraction to only include words that appear at least 5 times and differentiate between lemposes, e.g., lemmas coupled with their part-of-speech (POS) tag, and exclude proper names and numerals. In the analysis, we look at each keyword and multi-word expression and their translation equivalents, proposed by various translation systems, and analyze their diversity. We look at concordances for each of the keywords and identify their translation equivalent(s) in each of the translations. The expressions are grouped by their lemmatized form in case of keywords and by their canonical form in case of multi-word expressions, meaning we do not consider their (erroneous) declensions.

In the third step of lexical diversity analysis, we also compare the output of different translation systems and measure the agreement between machine translation systems and human translator. Each translation equivalent proposed by machine translation systems is marked as "corresponding" or "not corresponding", according to whether this same translation equivalent is also proposed by the referential translation. This way, we capture the extent of diversity that is in agreement with the human translation and measure to what extent the machine translation systems diverge from the proposed (appropriate) solutions. Finally, we also inspect the most variable translation case among the selected keywords and multi-word expressions per translation system and investigate the nature of the proposed solutions and the reasoning behind them.

3. Results

In this section, we present the results of our quantitative and qualitative analyses of lexical diversity. We first present results obtained from the automatic metrics TTR and MTLD, which is followed by a quantitative analysis of keyword and multi-word translations. In the last part of the section, we also qualitatively analyze each translation system on the basis of its most variable translation case.

3.1. TTR

The results for the measure of lexical variety using TTR in Table 2 show translations produced by automatic systems have a higher lexical variety than their human counterparts in seven out of nine cases. Google's neural machine translation system has the highest lexical variety in all three settings. However, TTR measures for lowest diversity differ according to each specific genre: for IT texts, the human translation has the lowest diversity, for culinary texts, Microsoft's neural translations have the lowest, and for literary texts, Google's phrase-based translation seems lexically poorest.

Table 2. Lexical diversity measured with TTR of human (HT), Google's neural (GNMT), Google's statistical (GSMT) and Microsoft's neural (MNMT) translations.

Corpus	HT	GNMT	GSMT	MNMT
IT	13.07%	13.54%	13.50%	13.40%
CUL	7.76%	9.58%	8.50%	7.57%
LIT	15.41%	16.59%	14.94%	16.06%

3.2. MTLD

MTLD or "measure of textual lexical diversity" tells us the average length of text that maintains a predefined TTR threshold. The results, shown in Table 3, differ from the lexical diversity analysis with TTR. Here, GSMT actually exhibits the highest lexical diversity in two out of three cases. The greatest difference from the previous metric is that GSMT exhibits the greatest lexical diversity in literary texts, while according to TTR, it has the lowest. However, human translations again appear to have the lowest lexical diversity in all but one case (in culinary texts, MNMT has the lowest MTLD).

Table 3. Lexical diversity measured with MTLD.

Corpus	HT	GNMT	GSMT	MNMT
IT	84.53	86.74	91.66	86.46
CUL	164.22	196.80	187.58	162.93
LIT	148.38	158.66	177.53	175.94

Both the analysis of TTR and MTLD seem to contradict the previous studies saying that lexical variety is lower in machine translations than human translations. On the other hand, comparing phrase-based to neural translation systems is not as clear cut, as these two metrics offer opposite findings. While TTR shows the highest diversity for one of the neural translation systems in all three cases, MTLD puts the statistical translation system in first place in two out of three cases. A closer examination of lexical diversity is addressed in the next subsection, where we analyze the translation equivalents for selected keywords and multi-word expressions.

3.3. Diversity of Translations of Keywords and Multi-Word Expressions

To further assess the diversity of translations, we select 10 keywords and 15 multi-word expressions for each of the original texts. As described in the Methods section, we use the built-in SketchEngine tool and limit the search to expressions that appear at least five times. The selected expressions are listed in Table 4.

Table 4. Keywords and multi-word expressions.

Corpus	Keywords	Multi-Word expressions
IT	*fax, cartridge, touch, print (NOUN), setup, tray, button, print (VERB), menu, printer*	*print cartridge, control panel, setup menu, setup button, ok button, document feeder, printer driver, software program, fax number, fax button, scanner glass, wireless network, phone line, print quality, recommended action*
CUL	*tbsp, stir, pan, pepper, saucepan, boil, chop, simmer, tsp*	*frying pan, olive oil, baking tray, cling film, lemon juice, medium heat, low heat, food processor, kitchen paper, large saucepan, black pepper, cold water, greaseproof paper, large bowl, wire rack*
LIT	*deposition, gesture, glance, nod, grin, desk, peer, briefcase, courtroom, sigh*	*making partner, general counsel, voice mail, front door, partnership decision, spare suit, litigation group, opening statement, cocktail hour, class action, suit jacket, law school, coffee shop, partnership spot, deposition transcript*

First, we looked at the number of translation proposed solutions in human versus machine translations. The results are shown in Figure 1; keywords are listed in descending order of diversity in human translation (highest number of human translation equivalents first). In the IT corpus, there are no larger deviations from the human variety of translation equivalents; the machine translations generally stay below the HT threshold. In the culinary corpus, GNMT substantially surpasses the number of human translation equivalents in two cases; in one case, the same thing can be observed for GSMT. In literary translations, the largest deviations are seen for keywords 6 and 7, where both neural translation systems propose a much larger number of translation equivalents than the human- and GSMT-translated texts.

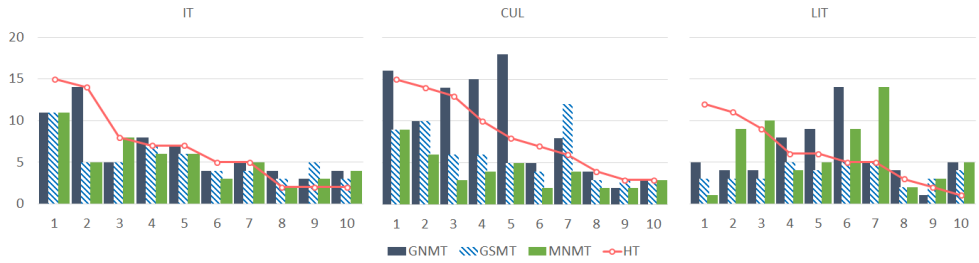

Figure 1. Visual comparison of diversity in translation of keywords: (**a**) Information technology (IT) text, (**b**) Culinary (CUL) text, (**c**) Literary (LIT) text.

We apply the same procedure for multi-word expressions, the results of which are depicted in Figure 2. In the IT corpus, the number of translation solutions is larger than the number of human-proposed solutions in multiple cases. The largest deviations are by GSMT, but both neural systems also surpass the number of reference translation equivalents for some multi-word expressions. In translations of recipes, the diversity of especially neural translations is visibly higher. While neural translators sometimes propose more than four times the number of human solutions, Google's phrase-based system mostly stays below the human referential threshold. In the literary translations, all machine translations generally propose an equal or smaller number of solutions. The only evident exception is for multi-word 15, where the referential translation only proposes one solution compared to three by GSMT, six by GNMT, and eight by MNMT.

The lexical diversity in translations of individual genres seems to follow the results of MTLD: the IT corpus is the least variable, followed by literary texts, while the corpus of recipes shows the most diversity. This diversity is mirrored also in the number of translation equivalents and the extent of deviation, visualized in Figure 3.

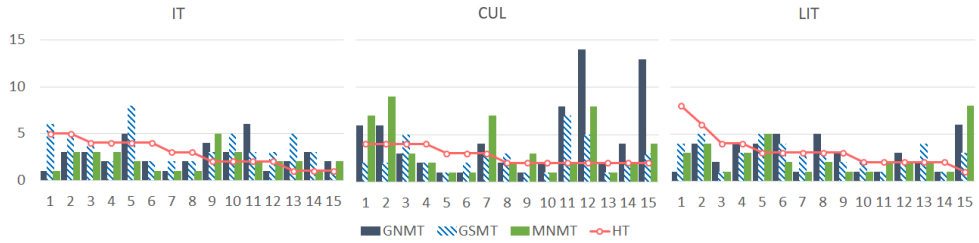

Figure 2. Visual comparison of diversity in translation of multi-word expressions: (**a**) Information technology (IT) text, (**b**) Culinary (CUL) text, (**c**) Literary (LIT) text.

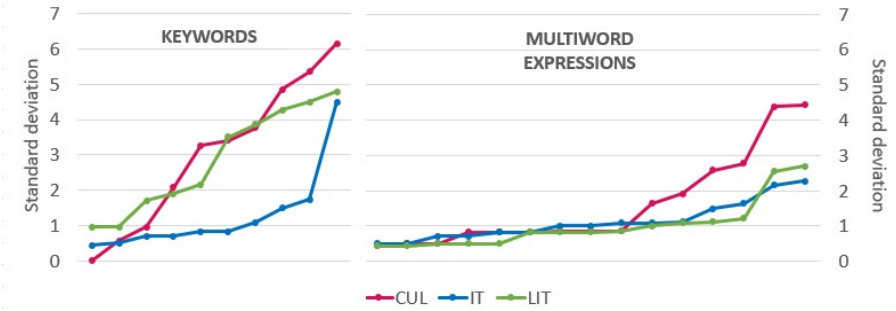

Figure 3. Standard deviations of the number of proposed translation solutions per expression for 10 keyword (**left**) and 15 multi-word expressions (**right**), in ascending order by subcorpus.

3.4. Agreement of Machine and Human Translations

In order to further assess the diversity of machine translations, we compare the proposed solutions for keywords and multi-word expressions to the solutions proposed in the human reference translation. For example, the human translation of the literary text offers six translation equivalents for the verb *nod* in 66 different instances: *prikimati, kimati, pokimati, prikimavati, pomigniti z glavo, pozdraviti*. Microsoft's neural translation system proposes *pokimati, prikimati* in 53 instances that agree with the reference translation but also *premikajoče, prikimanje* in three instances. Hence, the agreement for this example would be 94.5%. We compute the agreement for each of the keywords and each of the multi-word expressions for all three subcorpora and all three machine translations.

Figures 4 and 5 show the agreement percentages for all keyword and multi-word translations on a scale from 0% (no translation solutions match HT solutions) to 100% (all translation solutions match HT solutions), with their mean value at the X marker. The highest overall agreement was achieved for the translations of the printer manual. Agreements for the translations of keywords in culinary and literary texts, on the other hand, are very diffuse and range from 100% to even 0% in some cases.

Not surprisingly, multi-word expressions are, due to their complex, composite nature, even more divergent from the human translation. Here, again, the different machine translations reach the highest agreement in translating IT texts; however, even in this translation corpus, we observe much lower agreement with the human reference translation.

The mean agreement ratios are listed in Table 5. On average, the highest number of translation equivalents corresponding to the human reference translation for one-word expressions was proposed by the neural systems, GNMT and MNMT. The opposite trend can be seen for multi-word expressions, where GSMT reaches the highest agreement in all three corpora. Moreover, all machine translation systems were the most "successful" in translating IT texts, less in cooking recipes, and the least in the literary novel.

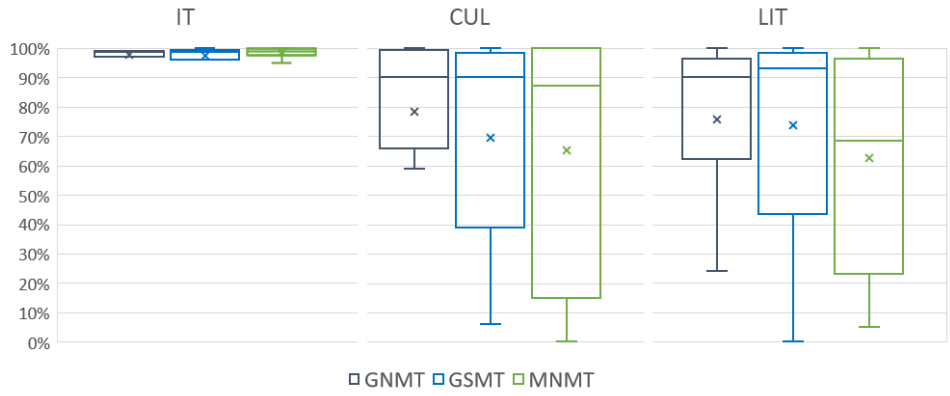

Figure 4. Agreement of translation systems to human translations of keywords: (**a**) IT corpus, (**b**) Culinary corpus, (**c**) Literary corpus.

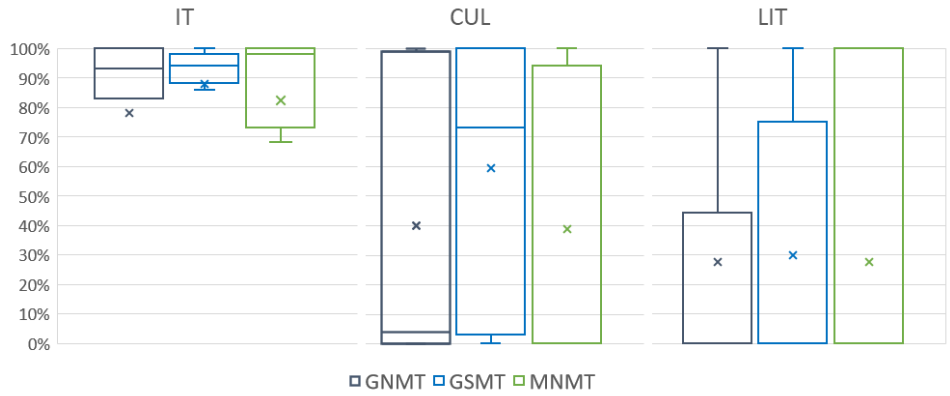

Figure 5. Agreement of translation systems to human translations of multi-word expressions: (**a**) IT corpus, (**b**) Culinary corpus, (**c**) Literary corpus.

Table 5. Mean agreement with human reference translation for proposed translation equivalents for keywords (KW) and multi-word expressions (MWE).

Corpus	Translation Unit	GNMT	GSMT	MNMT
IT	KW	97.9%	97.3%	**98.6%**
CUL	KW	**78.5%**	69.4%	65.1%
LIT	KW	**75.7%**	73.9%	62.4%
IT	MWE	78.1%	**87.7%**	82.2%
CUL	MWE	40.0%	**59.2%**	38.5%
LIT	MWE	27.6%	**29.8%**	27.5%

3.5. A Closer Look at Translation Diversity

To perform a more reliable analysis of lexical diversity and to check the interpretability of quantitative methods, we perform a partial qualitative analysis as an additional step. In this section, we look at the most variable translation case per machine translation system to observe whether any peculiarities emerge. For each system, we choose a single- and multi-word expression where the system diverged the most from the number of HT-proposed solutions, i.e., proposed the largest number of translations.

3.5.1. Google's Phrase-Based Statistical Model

The statistical system showed the most abounding diversity, compared to HT, in the case of translating the verb *chop* in the corpus of cooking recipes (Table 6). The referential

HT offers six different solutions, while the automatic system offers twelve, five of which match in 92.88% of total occurrences.

Table 6. Translation equivalents proposed for the verb *chop* in the CUL corpus.

HT	Occurrences	GSMT	Occurrences
sesekljan	322	sesekljan	310
sesekljati	44	sesekljati	31
ELLIPSIS	16	narezan	17
narezati	6	nasekljan	17
narezan	4	narezati	6
nastrgan	1	rezan	3
		chop	3
		nasekljati	2
		sekljanje	1
		zrezan	1
		ELLIPSIS	1
		sekanje	1

Among the unmatched solutions proposed by GSMT, we can argue that some of them are still appropriate even if not identical to the referential translation. Moreover, except for two cases (the solutions *chop, sekanje*), the proposed translation equivalents are semantically similar. The three instances of *chop* are considered non-translations, and the one instance of *sekanje*'woodchopping' does not suit a culinary context.

Among multi-word expressions, the GSMT deviates the most in translating the phrase *control panel* (Table 7). Here, the HT only offers one possible solution, while the automatic system proposes four more. These are very isolated cases, as otherwise, the translations agree in 97.14% of cases. In these four non-matching cases, the translation system seems not to have identified the compositional nature of the phrase, as it translated each word separately, which is observed in non–agreement of grammatical cases and the distance (other inserted words) between the two units marked by [...].

Table 7. Translation equivalents proposed for the multi-word expression *control panel* in the IT corpus.

HT	Occurrences	GSMT	Occurrences
nadzorna plošča	140	nadzorna plošča	136
		nadzorni	1
		nadzoren [. . .] plošče	1
		plošča [. . .] nadzorni	1
		plošča [. . .] kontrole	1

3.5.2. Google's Neural Model

Translations by GNMT are the most plentiful for the keyword *deposition* in the literary novel (Table 8). The HT offers five translation equivalents, while the neural system offers as many as 14. Three of those (*izjava*,*(odložiti) izjavo* and translation by ellipsis) match the referential solutions, but this accounts only for 13.64% of all occurrences. All the other options proposed by GNMT completely miss the legal context. For instance, its highly preferred solution *odlaganje* signifies either a 'delay, postponement of something' or a 'physical deposit of material'. The translation equivalent *privednik*, on the other hand, is not an existing Slovenian word, and its selection can only be explained through GNMT's handling of subword units.

Table 8. Translation equivalents proposed for the keyword *deposition* in the LIT corpus.

HT	Occurrences	GNMT	Occurrences
zaslišanje	37	odlaganje	23
izjave prič	2	depozit	5
izjave	2	izjava	3
izjave na zapisnik	2	*ELLIPSIS*	2
ELLIPSIS	1	naloga	2
		"privednik"	1
		deponiranje	1
		odpust	1
		odstopil	1
		odložiti	1
		nanos	1
		dejanje	1
		izjava (odložiti izjavo)	1
		odlog	1

In Table 9, we can observe an extraordinary number of proposed translation equivalents for the phrase *food processor*. While the HT only uses two equivalents, the translation by the neural model proposes 14 different options. The preferred solution *kuhalnik hrane* 'food cooker' is not a reasonable solution, but the even more concerning ones are *hranilnik*, *paradižnik*, and *pralni stroj*. The first one, *hranilnik*, can mean either 'storage tank' or 'coin container, piggy bank' in Slovenian, but it was most likely proposed only as a derivative of the root *hrana*, translation for 'food': "food–er". The second, *paradižnik*, is Slovenian for 'tomato', for which we find that it was a translation solution for one of the other words in the sentence that the neural system repeated and with it overrode the translation for 'food processor'. For the third solution, *pralni stroj*, meaning 'washing machine' in Slovenian, we cannot find a sensible explanation, as the original sentence provided more than enough culinary context.

Table 9. Translation equivalents proposed for the multi-word expression *food processor* in the CUL corpus.

HT	Occurrences	GNMT	Occurrences
multipraktik	102	kuhalnik hrane	52
strojček	3	predelovalec hrane	21
		obdelovalec hrane	7
		kuhalnik	5
		hranilnik	5
		predelava hrane	3
		živilski procesor	2
		posodo za kuhanje hrane	2
		predelovalnik hrane	2
		procesor za hrano	2
		živilski predelovalec	1
		paradižnik	1
		pralni stroj	1
		predelava za hrano	1

3.5.3. Microsoft's Neural Model

Microsoft's translation system shows a remarkably varied set of solutions, listed in Table 10. For the verb *grin*, translated with five different equivalents by the human translator, MNMT proposes 14 different solutions, only one of which agrees with HT. Moreover, albeit we might argue that two of those (*režati se, nasmejati se*) could be considered viable solutions, others only demonstrate various errors and missteps of the neural system.

The preferred solution, *zbrusiti se* with 13 occurrences, as well as *zmelje* in one occurrence, can be explained by the neural system mistaking 'grin' for 'grind' (brusiti, mleti). Other solutions such as *zgriniti se*, *zgrniti se* can be explained by their orthographical similarity and the neural system's only graphical adaptation of the word to Slovenian. Other interesting "solutions" are *zasiti se*, *Cerenje*, *zasoviti se*. For the first one, it seems as if the neural model only made a subwords guess with the prefix 'za–' indicating a finite, noncontinuous action, and a regular Slovenian verbal suffix '–(s)iti'. The second, *Cerenje*, could be a reasonable enough translation of the keyword for the Croatian language; however, it is surprising that the word is capitalized. For the third word, we can only speculate that the model "interpreted" 'grin' as 'scowl', as this would mean that it dissected the original word into subwords and combined them, namely by joining the indicator with a perfective aspect (sc– > za–), the root (owl > sov(a)), and the verbal suffix (–il/–iti).

Table 10. Translation equivalents proposed for the keyword *grin* in the LIT corpus.

HT	Occurrences	GNMT	Occurrences
nasmehniti se	20	zbrusiti se	13
zarežati se	8	"zasiti" se	7
zasmejati se	7	"zgriniti" se	3
smehljati se	2	režati (se)	2
sam pri sebi se smehljati	1	nasmehniti (se)	2
		grinned	2
		grinning	2
		nasmejati (se)	1
		"Cerenje"	1
		zasmećen	1
		ELLIPSIS	1
		"zasoviti" se	1
		"zmelje"	1
		zgrniti se	1

Among multi-word expressions, MNMT diverges from the HT the most in translating the phrase *class action* (Table 11). The HT only offers a conventional terminological variant *skupinska tožba*, while the automatic translation proposes eight different solutions. Apart from the plenitude of translation solutions, we observe that the neural system usually produced these translations because it did not recognize the two words as a phrase but translated word-by-word, except in one isolated case. Even though we are in the area of literary discourse, consistency, i.e., a lack of diversity, in translating this terminological expression would be a much more welcome feature.

Table 11. Translation equivalents proposed for the multi-word expression *class action* in the LIT corpus.

HT	Occurrences	MNMT	Occurrences
skupinska tožba	8	razredom ukrepanje	1
		razred dejanje	1
		skupinska tožba	1
		dejavnosti	1
		razred akcijske	1
		ukrep () razred	1
		razreda ukrep	1
		akcijski	1

4. Discussion

Our findings from experiments on lexical diversity have shown differing, even contradicting results. The automatic TTR metric presents evidence that (a) machine translations are predominantly lexically more diverse than human translations, and (b) translations by GNMT are the most diverse. MTLD presents evidence of the following: (a) Machine translations are, again, predominantly more lexically diverse than human translations; (b) Contrary to the TTR metric, translations by GSMT are the most diverse in two out of three cases; (c) In the CUL subcorpus, the neural systems take opposite ends (GNMT is the most diverse, and MNMT is the least).

On the one hand, these two metrics show that human translations are, in the majority of cases, lexically the poorest. On the other hand, they cannot yield conclusive results for one machine translation architecture compared to the other. We observe that the results depend on the text type and/or domain, the chosen metric as well as the individual machine translation system.

The general trends found by MTLD are transposed into the keyword and multi-word translation diversity. The highest agreement with the reference translation and the lowest diversity of proposed solutions was observed for the IT corpus, while the translations exhibited the highest level of "creativity" in translating cooking recipes. However, the quantitative analysis did not show a consistent, universal trend. In some cases, the human translation is the least diverse, while in others, the human translation proposes the largest number of different solutions. However, we did observe the phenomenon detected by Čulo and Nitzke [14], that when neural translations exhibit variation, this variation is particularly pronounced. As demonstrated in our more detailed qualitative analysis, GSMT shows the most lexical diversity in translating the word 'fry' with 12 proposed solutions, GNMT proposes 14 solutions for 'deposition', and MNMT offers 14 for 'grin'. Both neural systems seem to exhibit more "creative" solutions; however, while GSMT's translations agree with the human reference translation in 92% of cases, neural systems only achieve an agreement of 13.64% and 5.26% for the observed keyword translations. Moreover, most of the additional solutions proposed by GSMT that do not agree with the human reference translation might be deemed appropriate. On the contrary, GNMT and MNMT may offer a larger number of inventive solutions, but these seem to be mostly based on morphological and subword combinations and adaptations, which prove unintelligible and misguided. This is even more pronounced in the multi-word translations. While it is true that GSMT does not offer an abundance of translations, at least it seems consistent. The three cases where the machine translation system opted for a different solution were most probably due to the fact that the system did not succeed in recognizing the phrase as a whole but tried to translate word by word. However, the neural systems can translate a known phrase, i.e., a recognized collocation such as *food processor*, in several different ways, which has great implications especially for post-editing. While we could argue that lexical richness is an admirable characteristic in literary or general, common texts, translating fixed collocations and terminology nonetheless requires a certain level of consistency, if nothing else for an easier and faster post-editing process.

5. Conclusions

In this study, we addressed the lexical diversity of machine translations by applying qualitative and quantitative methods. First, we found that automatic metrics determining lexical diversity on a global scale show diverging, occasionally contradicting results. However, both metrics put human translations at the very bottom of the lexical diversity ladder in the majority of cases, contrary to previous studies [6,11]. Secondly, looking at particular cases of lexical diversity reveals two types of instability. Translations by automatic systems are not *always* more or less diverse than the human translation; the diversity differs case-by-case. Moreover, the diversity by neural translations, albeit quantitatively substantial, cannot be compared to human lexical diversity and creativity. Those more creative, inventively coined words are oftentimes inappropriate and illogical, acting but

as amusing brainteasers to the logic behind them. Thus, machine translation faces two difficult, seemingly contradictory issues. One is in their inherent overgeneralization with the increase in frequency of more frequent words, resulting in the loss of lexical diversity, and the other, visible for the less frequent words and expressions, in undergeneralization, resulting in inconsistency, miscellaneous translations, and thus a "mock" lexical diversity, in fact acting as a potential impediment in post-editing processes.

The limitations of this study include a focus on only one particular combination of languages and translation direction, a focus on three specific genres, and the lack of large-scale qualitative and human experiments. In further studies, we would opt for a comparison between different directions of translation and include comparable translations from Slovenian to English. To reliably evaluate lexical richness and creativity in translations, we would also employ other lexical resources (synonym bases) and manual evaluation of translations by a larger group of individuals. Additionally, the implications of our findings (the overly diverse and inconsistent MTs) for post-editing can, of course, only be verified in practice, for instance by conducting an experiment measuring the time and effort through the number of edits, post-editing time, eye movements, fixations, etc.

Author Contributions: Conceptualization, data curation, investigation, and writing—original draft preparation by M.B.; supervision, validation, and writing—review and editing by Š.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Slovenian Research Agency by the research core funding P6-0215.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to copyright limitations.

Acknowledgments: The authors acknowledge and thank the students of the 2018/2019 Corpora and Localization course who compiled, aligned, and shared the corpora used for the experiment.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MT	Machine translation
HT	Human translation
NMT	Neural machine translation
SMT	Statistical (phrase-based) machine translation
GSMT	Google's statistical (phrase-based) translation system
GNMT	Google's neural translation system
MNMT	Microsoft's neural translation system
IT	Information technology
CUL	Culinary arts
LIT	Literature
TTR	Type-token ratio
MTLD	Measure of textual lexical diversity
POS	Part of speech
KW	Keyword
MWE	Multi-word expression

References

1. Hassan, H.; Aue, A.; Chen, C.; Chowdhary, V.; Clark, J.; Federmann, C.; Huang, X.; Junczys-Dowmunt, M.; Lewis, W.; Li, M.; et al. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv* **2018**, arXiv:1803.05567.
2. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadephia, PA, USA, 6–12 July 2002; pp. 311–318.

3. Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
4. Castilho, S.; Moorkens, J.; Gaspari, F.; Calixto, I.; Tinsley, J.; Way, A. Is Neural Machine Translation the New State of the Art? *Prague Bull. Math. Linguist.* **2017**, *108*, 109–120. [CrossRef]
5. Shterionov, D.; Superbo, R.; Nagle, P.; Casanellas, L.; O'Dowd, T.; Way, A. Human versus automatic quality evaluation of NMT and PBSMT. *Mach. Transl.* **2018**, *32*, 217–235. [CrossRef]
6. Vanmassenhove, E.; Shterionov, D.; Way, A. Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In Proceedings of the Machine Translation Summit XVII Volume 1: Research Track, Dublin, Ireland, 19–23 August 2019; pp. 222–232.
7. Vanmassenhove, E.; Shterionov, D.S.; Gwilliam, M. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. *arXiv* **2021**, arXiv:2102.00287.
8. Roberts, N.; Liang, D.; Neubig, G.; Lipton, Z.C. Decoding and Diversity in Machine Translation. *arXiv* **2020**, arXiv:cs.CL/2011.13477.
9. Farrell, M. Machine Translation Markers in Post-Edited Machine Translation Output. In Proceedings of the 40th Conference Translating and the Computer, London, UK, 15–16 November 2018; pp. 50–59.
10. Baker, M.; Francis, G.; Tognini-Bonelli, E. Corpus Linguistics and Translation Studies: Implications and Applications. In *Text and Technology: In Honour of John Sinclair*; John Benjamins Publishing Company: Amsterdam, The Netherlands, 1993.
11. Toral, A. Post-editese: An Exacerbated Translationese. In Proceedings of the Machine Translation Summit XVII, Dublin, Ireland, 19–23 August 2019; pp. 273–281.
12. Castilho, S.; Resende, N.; Mitkov, R. What Influences the Features of Post-Editese? A Preliminary Study. In Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019), Varna, Bulgaria, 5–6 September 2019; pp. 19–27. [CrossRef]
13. De Clercq, O.; De Sutter, G.; Loock, R.; Cappelle, B.; Plevoets, K. Uncovering Machine Translationese Using Corpus Analysis Techniques to Distinguish between Original and Machine-Translated French. *Transl. Q.* **2021**, *101*, 21–45.
14. Čulo, O.; Nitzke, J. Patterns of Terminological Variation in Post-editing and of Cognate Use in Machine Translation in Contrast to Human Translation. In Proceedings of the 19th Annual Conference of the European Association for Machine Translation, Riga, Latvia, 30 May–1 June 2016; pp. 106–114.
15. Arcan, M. A Comparison of Statistical and Neural Machine Translation for Slovene, Serbian and Croatian. In Proceedings of the Conference on Language Technologies & Digital Humanities, Ljubljana, Slovenia, 20–21 September 2018.
16. Gregor Donaj, M.S.M. Prehod iz statističnega strojnega prevajanja na prevajanje z nevronskimi omrežji za jezikovni par slovenščina-angleščina. In Proceedings of the Conference on Language Technologies & Digital Humanities, Ljubljana, Slovenia, 20–21 September 2018.
17. Vintar, Š. Terminology Translation Accuracy in Statistical versus Neural MT: An Evaluation for the English-Slovene Language Pair. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; Du, J., Arcan, M., Liu, Q., Isahara, H., Eds.; European Language Resources Association (ELRA): Paris, France, 2018.
18. LASERJET PRO 300 COLOR MFP/LASERJET PRO 400 COLOR MFP User Guide. 2011. Available online: http://h10032.www1.hp.com/ctg/Manual/c02666267.pdf (accessed on 3 March 2020).
19. LASERJET PRO 300 COLOR MFP/LASERJET PRO 400 COLOR MFP Uporabniški Priročnik. 2011. Available online: http://h10032.www1.hp.com/ctg/Manual/c02666724.pdf (accessed on 3 March 2020).
20. LASERJET PRO 300 COLOR MFP/LASERJET PRO 400 COLOR MFP Installation Guide. 2011. Available online: http://h10032.www1.hp.com/ctg/Manual/c02843075.pdf (accessed on 3 March 2020).
21. LASERJET PRO 300 COLOR MFP/LASERJET PRO 400 COLOR MFP Priročnik za Namestitev. 2011. Available online: http://h10032.www1.hp.com/ctg/Manual/c02843079.pdf (accessed on 3 March 2020).
22. Blashford-Snell, V. *The Cooking Book*; Dorling Kindersley: London, UK, 2008.
23. Blashford-Snell, V. *Dobra Kuha*; Mladinska Knjiga: Ljubljana, Slovenia, 2012.
24. James, J. *Practice Makes Perfect*; Berkley Sensation: New York, NY, USA, 2009.
25. James, J. *Osem let skomin*; Hiša knjig, Založba KMŠ: Maribor, Slovenia, 2014.
26. Kilgarriff, A.; Baisa, V.; Bušta, J.; Jakubíček, M.; Kovář, V.; Michelfeit, J.; Rychlý, P.; Suchomel, V. The Sketch Engine: Ten years on. *Lexicography* **2014**, *1*, 7–36. [CrossRef]
27. McCarthy, P.M. An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD). Ph.D. Thesis, The University of Memphis, Memphis, TN, USA, 2005.
28. Mccarthy, P.; Jarvis, S. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behav. Res. Methods* **2010**, *42*, 381–392. [CrossRef] [PubMed]
29. Fergadiotis, G.; Wright, H.; West, T. Measuring Lexical Diversity in Narrative Discourse of People with Aphasia. *Am. J. Speech-Lang. Pathol./Am. Speech-Lang. Assoc.* **2013**, *22*, S397–S408. [CrossRef]
30. Kyle, K. Lexical Diversity. 2020. Available online: https://github.com/kristopherkyle/lexical_diversity (accessed on 1 June 2020).