Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation

Eva Vanmassenhove[α] **Dimitar Shterionov**[α] **Matthew Gwilliam**[β]
[α] Cognitive Science and AI, Tilburg University, The Netherlands
[β] University of Maryland, College Park
{e.o.j.vanmassenhove, d.shterionov}@tilburguniversity.edu
mgwilliam@umd.edu

Abstract

Recent studies in the field of Machine Translation (MT) and Natural Language Processing (NLP) have shown that existing models amplify biases observed in the training data. The amplification of biases in language technology has mainly been examined with respect to specific phenomena, such as gender bias. In this work, we go beyond the study of gender in MT and investigate how bias amplification might affect language in a broader sense. We hypothesize that the 'algorithmic bias', i.e. an exacerbation of frequently observed patterns in combination with a loss of less frequent ones, not only exacerbates societal biases present in current datasets but could also lead to an artificially impoverished language: 'machine translationese'. We assess the linguistic richness (on a lexical and morphological level) of translations created by different data-driven MT paradigms – phrase-based statistical (PB-SMT) and neural MT (NMT). Our experiments show 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).

1 Introduction

The idea of translation entailing a transformation is widely recognised in the field of Translation Studies (TS) (Ippolito, 2014). Translations are specific communicative acts occurring in a particular context governed by their own laws. Some of the features that characterize translated texts are defined as simplification, explicitation, normalization and leveling out (Baker, 1999). The fingerprints left by the translation process and the language this results into, have been referred to as 'translationese' (Gellerstam, 1986). Empirical evidence of the existence of translationese can be found in studies showing that machine learning techniques can be employed to automatically distinguish between human translated and original text by looking at lexical and grammatical information (Baroni and Bernardini, 2006; Koppel and Ordan, 2011). Translationese differs from original texts due to a combination of factors including intentional (e.g. explicitation and normalization) and unintentional ones (e.g. unconscious effects of the source language input on the target language produced). Unlike other work on (human) translationese (or even related work on 'Post-editese'), we delve into the effects of machine translation (MT) algorithms on language, i.e. 'machine translationese'.

So far, generating accurate and fluent translations has been the main objective of MT systems. As such, maintaining the richness and diversity in the outputs has understandably not been a priority (Vanmassenhove, 2020).[1] However, as MT systems have reached a quality that is (arguably) close to that of human translations (Läubli et al., 2018; Toral et al., 2018) and as such are being used widely on a daily basis, we believe it is time to look into the potential effects of (MT) algorithms on language itself.[2]

The main motivations behind this work are: (i) if algorithmic bias is indeed a by-product of our algorithms, a statistically biased MT system might prefer frequently occurring words (or sub-words) over others. Since MT systems do not necessarily distinguish between different synonymous translations (lexical richness) and morphological variants (grammatical richness), algorithmic bias could lead to the loss of morphological variety (and thus interfere with the ability of our systems to generate

[1] One might argue that for some tasks and domains, diversity can be harmful (specific in-domain translations would prefer consistency over e.g. lexical diversity).

[2] Google Translate alone translates more than 100 billions words per day and is used by at least 500 million people according to estimates (https://www.blog.google/products/translate/ten-years-of-google-translate/).

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2203–2213
April 19 - 23, 2021.

at all times a grammatically correct option); (ii) the sociolinguistic perspective of machine translationese since it has been established that language contact (e.g. via translationese) can entail language changes (Kranich, 2014). If machine translationese (and other types of 'NLPese') is a simplified version of the training data, what does that imply from a sociolinguistic perspective and how could this affect language on a longer term?

The main objective of the presented research is to establish whether there is indeed a quantitatively measurable difference between the linguistic richness of an MT system's training data and its output in terms of morphological and lexical diversity. To do so, we conduct an in-depth analysis that goes beyond frequently used standard lexical diversity metrics such as TTR, Yule's I and MTLD. We assess the lexical and morphological diversity through an adapted version of the Lexical Frequency Profile used to assess language acquisition, a measure of morphological diversity based on Shannon and Simpson Diversity and an novel automatic evaluation of synonym frequency. We focus on the most prominent data-driven MT paradigms: Neural MT (NMT), both LSTM (Bahdanau et al., 2015; Sutskever et al., 2014) and Transformer (Vaswani et al., 2017), and Phrase-Based Statistical MT (PB-SMT). Up to our knowledge this is the first research on lexical and morphological diversity of machine translation output, i.e. machine translationese.

The contributions of this work can be summarised as: (i) a detailed analysis of lexical and morphological diversity of machine translationese and the loss thereof to quantify the effects of algorithmic bias; (ii) the adaptation of a metric used in language acquisition for assessing lexical sophistication in MT[3]; (iii) the use of Shannon entropy and Simpson diversity to measure morphological richness, and (iv) a novel, automatic evaluation of synonym frequency.

2 Related Work

Several studies have exposed the societal biases present in datasets (racial bias (Merullo et al., 2019), political bias (Fan et al., 2019), gender bias (Vanmassenhove and Hardmeier, 2018). Existing NLP technology are likely to pick up biases present in the training data and various explorations of e.g. gender bias in NLP systems have indeed revealed the existence of harmful biases in the output they generate (Bolukbasi et al., 2016; Caliskan-Islam et al., 2016; Garg et al., 2018; Vanmassenhove et al., 2018; Stanovsky et al., 2019; Sun et al., 2019; Habash et al., 2019). Research related to bias has often focused on gender or race. Especially in a field such as MT, the implicit gender in a language such as English and its consecutive translations into morphologically richer languages with gender agreement, makes it relatively easy to expose and study biases related to gender in a contrastive linguistic setting. In the context of this paper, we would like to note that (statistical) bias is not limited to gender or race but can be defined as any systematic inaccuracy in one direction leading to an under (or over) estimation of observations.

A handful of recent work has mentioned the possibility of algorithmic bias on top of the already existing societal biases in the training data (Bolukbasi et al., 2016; Caliskan-Islam et al., 2016; Garg et al., 2018). For instance, Zhao et al. (2017) observe a phenomenon they refer to as 'bias amplification'. They note that in their training data an activity such as 'cooking' is associated 33% times more with women compared to men. After training a model on that dataset, the existing disparity is amplified to 68% times more associations with women.

In the field of MT, Vanmassenhove et al. (2019) address the effects of statistical bias on language generation in an MT setting. They assess lexical diversity using standard metrics –TTR, MTLD and Yule's K– and conclude that the translations produced by various MT systems (PB-SMT and NMT) are consistently less diverse than the original training data. Their approach was conducted on NMT systems that were trained without byte-pair-encoding (BPE) (Sennrich et al., 2016) which limits the creativity of the translation systems.

Toral (2019) measures the lexical diversity of 18 state-of-the-art systems on 6 language pairs, reaching similar conclusions. They do so focusing specifically on post-editese. The experiments indicate that post-editese is simpler and more normalised than human translationese. The post-editese also shows a higher degree of interference from the source compared to the human translations. Daems et al. (2017), like Toral (2019), centers around the automatic detection of post-editese and does not look into properties of unedited machine translationese. In Aranberri (Aranberri, 2020) different freely available MT systems (neural and rule-

[3]In fact, our implementation of the LFP metric can be employed for any NLP tasks.

based) are compared in terms of automatic metrics (BLEU, TER) and translationese features (TTR, length ratio input/output, perplexity, etc.) to investigate how such features correlate with translation quality. Bizzoni et al. (2020) presents a comparison using similar translationese features of three MT architectures and the human translations of spoken and written language.

In the field of PB-SMT, Klebanov and Flor (2013) show that PB-SMT suffers considerably more than human translations (HT) from lexical loss, resulting in loss of lexical tightness and text cohesion. Aharoni et al. (2014) proof that automatic and human translated sentences can be automatically identified corroborating that human translations systematically differ from the translations produced by PB-SMT systems.

Aside from Vanmassenhove et al. (2019), the above discussed related work uses metrics of lexical diversity to compare human translations to (post-edited) machine translations. In this work, we compare how and whether the output of an MT system differs (in terms of lexical and morphological diversity) from the data it was originally trained on. This way, we aim to investigate the effect of the algorithm (and algorithmic bias) on language itself.

3 Machine Translation Systems

MT paradigms have changed quickly over the last decades. Since this is the first attempt to quantify both the lexical and grammatical diversity of machine translationese, we experimented with the current state-of-the-art data-driven paradigms, LSTM and Transformer, as well as with PB-SMT. We used data from the Europarl corpus (Koehn, 2005) for two language pairs, English–French and English–Spanish in both direction (EN→FR, FR→EN, EN→ES and ES→EN). We are interested in both directions in order to verify whether there is a difference in terms of (the potential loss of) diversity when comparing translations from a morphologically poorer language (English) into morphologically richer ones (French and Spanish) and vice versa. Our data is summarised in Table 2.[4]

[4]We ought to address the fact that the Europarl data consists of both human-uttered and translated text which have different properties in terms of diversity. In this work we analyse the impoverishment of data when it passes through the "filter" of the MT system, i.e. the effect of algorithm. As the origin of the data, human-uttered or translated, has no impact on the inherent workings of the MT system we do not take this into account in our analysis.

Lang. pair	Train	Test	Dev
EN-FR/FR-EN	1,467,489	499,487	7,723
EN-ES/ES-EN	1,472,203	459,633	5,734

Table 1: Number of parallel sentences for the training, testing and development sets.

The specifics of the MT systems we trained are:

PB-SMT For the PB-SMT systems we used Moses (Koehn et al., 2007) with default settings and a 5-gram language model with pruning of bigrams. We also tuned each system using MERT (Och and Ney, 2003) until convergence or for a maximum of 25 iterations. During translation we mask unknown words with the UNK token to avoid bleeding through (source) words which would artificially increase the linguistic diversity.

NMT For the RNN and Transformer systems we used OpenNMT-py.[5] The systems were trained for maximum of 150K steps, saving an intermediate model every 5000 steps or until reaching convergence according to an early stopping criteria of no improvements of the perplexity (scored on the development set) for 5 intermediate models. The options we used for the neural systems are:

- RNN: size: 512, RNN type: bidirectional LSTM, number of layers of the encoder and of the decoder: 4, attention type: MLP, dropout: 0.2, batch size: 128, learning optimizer: Adam (Kingma and Ba, 2014) and learning rate: 0.0001.
- Transformer: number of layers: 6, size: 512, transformer_ff: 2048, number of heads: 8, dropout: 0.1, batch size: 4096, batch type: tokens, learning optimizer Adam with $\text{beta}_2 = 0.998$, learning rate: 2.

All NMT systems have the learning rate decay enabled and their training is distributed over 4 nVidia 1080Ti GPUs. The selected settings for the RNN systems are optimal according to Britz et al. (2017); for the Transformer we use the settings suggested by the OpenNMT community[6] as the optimal ones that lead to quality on par with the original Transformer work (Vaswani et al., 2017).

For training, testing and validation of the systems we used the same data. To build the vocabularies for the NMT systems we used sub-word units, allowing NMT to be more creative; using sub-word units also mitigates to a certain extent the out of vo-

[5]https://opennmt.net/OpenNMT-py/

[6]http://opennmt.net/OpenNMT-py/FAQ.html

cabulary problem. To compute the sub-word units we used BPE with 50,000 merging operations for all our data sets. In Table 2 we present the vocabulary sizes of the data used to train our PB-SMT and NMT systems.

Lang. pair	no BPE		with BPE	
	EN	FR/ES	EN	FR/ES
EN-FR/FR-EN	113,132	131,104	47,628	48,459
EN-ES/ES-EN	113,692	168,195	47,639	49,283

Table 2: Vocabulary sizes. For completeness we also present the vocabulary size without BPE, i.e. the number of unique words in the corpora.

The quality of our MT systems is evaluated on the test set using standard evaluation metrics – BLEU (Papineni et al., 2002) (as implemented in SacreBLEU (Post, 2018)) and TER (Snover et al., 2006) (as implemented in Multeval (Clark et al., 2011)). Our evaluation scores are presented in Table 3.

English as source

System	EN→FR		EN→ES	
	BLEU↑	TER↓	BLEU↑	TER↓
PB-SMT	35.7	50.7	38.6	45.9
LSTM	34.2	50.9	38.2	45.3
TRANS	**37.2**	**48.7**	**40.9**	**43.4**

English as target

System	FR→EN		ES→EN	
PB-SMT	36.2	47.1	39.3	44.0
LSTM	34.6	48.2	38.1	44.7
TRANS	**37.0**	**46.4**	**41.3**	**41.4**

Table 3: Quality evaluation scores for our MT systems. TRANS denotes Transformer systems.

We computed pairwise statistical significance using bootstrap resampling (Koehn, 2004) and a 95% confidence interval. The results shown in Table 3 are all statistically significant based on 1000 iterations and samples of 100 sentences. All metrics show the same performance trends for all language pairs: Transformer (TRANS) outperforms all other systems, followed by PB-SMT, and LSTM.

For all PB-SMT we replaced marked unknown words with only one token "UNK". While this does not effect the computation of BLEU and TER, it allows us not to artificially boost the lexical and grammatical scores for these MT engines (see Section 4) and assess their realistic dimensions.

4 Experiments and Results

Assessing linguistic complexity is a multifaceted task spanning over various domains (lexis, morphology, syntax, etc.). The lexical and grammatical diversity are two of its major components (Bachman, 2004; Bulté et al., 2008; Bulté, 2013). As such, we conduct an analysis using (i) lexical diversity and sophistication metrics (Section 4.2) and (ii) grammatical diversity metrics (Section 4.3). For lexical diversity, we use the following metrics: an adapted version of the Lexical Frequency Profile (LFP), three standard metrics commonly used to assess diversity –TTR, Yule's I and MTLD–, and three new metrics based on synonym frequency in translations. Up to our knowledge, this research is the first to employ LFP to analyze synthetic data. For grammatical diversity, we focus specifically on morphological inflectional diversity. We adopt the Shannon entropy and Simpson's diversity index to compute the entropy of the inflectional paradigms of lemmas, measuring the abundance and the evenness of wordforms per lemma.

Next, we will discuss the evaluation data and the aforementioned metrics designed in order to compare the diversity of the training data with the machine translationese. Our evaluation scripts are available at `https://github.com/dimitarsh1/BiasMT`; due to its large size, the data is not hosted in the github repository but is available upon request.

4.1 Evaluation data

To observe the effects of the MT algorithm on linguistic diversity, we used the MT engines (Section 3) to translate the source side of the training set, i.e. completely observed data. Data that has fully been observed during training is most suitable for our objectives as we are interested in the effects of the algorithm on language itself. It is also the most favourable translation (and evaluation) scenario for the MT systems since all data has been observed.

4.2 Lexical Diversity

Lexical Frequency Profile To look at the lexical sophistication and diversity in the text produced by the MT systems, we adapted the Lexical Frequency Profile (LFP) method (Laufer, 1994; Laufer and Nation, 1995). LFP is a measure that stems from research in second language (L2) acquisition and student writing methods. It is designed to study the lexical diversity or sophistication in texts produced by L2 learners. It is based on the observation that texts including a higher proportion of less frequent words are more sophisticated than those containing

higher proportions of more frequent words (Kyle, 2019).

The LFP method measures diversity and sophistication by looking at frequency bands. In its original version, the LFP analysis would distinguish between 4 bands: (i) percentage of words in a text belonging to the 1000 most frequent words in that language, (ii) percentage of words in a text belonging to the next 1000 most frequent words, (iii) a list of academic words that did not occur in the first 2000 words and (iv) the remaining words. The lists used to determine the word bands are predefined word lists such as Nation's word lists (Nation, 1984). One shortcoming of the approach is that a mismatch between reference corpus and the target text can lead to misleading outcomes. However, since we are looking into the (side-)effects of the training algorithm, instead of using preset word lists in order to compute the LFP, we use the original training data to generate the word frequency lists. This allows for a better comparison between the original data and the machine translationese while bypassing the potential mismatch issue.

Several studies (Crossley et al., 2013; Laufer, 1994; Laufer and Nation, 1995) have employed the LFP method to assess L2 acquisition in learners. From these studies, it resulted that the less proficient a user of an L2, the more words belonged to the first band and the least words belong to the list of academic words or the remaining words (band 3 and 4 respectively of the original formulation).

The lexical profile mentioned above is a detailed profile, showing 4 types of words used by the learner. Because of interpretability issues, the 'Beyond 2000' is also frequently used to assess the profile of the users. It distinguishes between the first two bands (comprising of the first 2000 words) and the rest. This condensed profile has been found equally reliable and valid as the original LFP having the advantage that it reduces the profile to one single score facilitating a comparison between learners (or in our case MT systems).

Since we are interested in the difference between the original training data and the output of the MT systems, we compute the frequency bands on the original training data instead of based on pre-set word lists used in L2 research. As such, we leave out the third band consisting of a list of academic words. Presenting and computing the LFP this way, will give us immediately the 'Beyond 2000' metric score as well (as we distinguish between three bands only, the last one being anything beyond the first 2000 words).

	FR			ES		
	B1	B2	B3	B1	B2	B3
ORIG	**79.80**	6.59	13.61	**77.80**	6.83	15.36
PB-SMT	81.78	6.48	11.74	79.77	6.86	13.36
LSTM	82.95	6.18	10.88	80.34	6.84	12.81
TRANS	82.01	6.24	11.75	82.35	6.99	10.67
	EN_{FR}			EN_{ES}		
	B1	B2	B3	B1	B2	B3
ORIG	**80.83**	7.10	12.07	**80.81**	7.11	12.08
PB-SMT	82.06	7.04	10.90	82.25	7.01	10.74
LSTM	83.23	6.93	9.81	83.29	6.93	9.78
TRANS	82.25	7.05	10.70	82.35	6.99	10.67

Table 4: Lexical Frequency Profile (French, Spanish, English (EN_{FR} and EN_{ES}) with 3 bands (B1: 0-1000, B2: 1001-2000, B3: 2001-end) for the original data and the output of the MT systems.

The LFP for French, Spanish and English, from the EN→FR, EN→ES, FR→EN (denoted as EN_{FR}) and ES→EN (denoted as EN_{ES}) data is presented in Table 4. It shows that the original data is consistently more diverse than the output of the MT systems as (i) the percentage of text occupied by the 1000 most frequent words (B1) is lower than in the corresponding B1 scores for all MT systems which implies that the 1000 most frequent words take up a smaller percentage of the text in the original training data compared to in the output of the different MT systems; and (ii) the so-called 'Beyond 2000' measure, which in our LFP is equal to the third band (B3), showing us the percentage of text occupied by the words that do not belong to the first two bands, is consistently higher for the original data compared to the MT systems (meaning that the less frequent words occupy a bigger proportion of the original data than they do in its machine translationese variants). Note that it has been established that LFPs are large-grained so small gains in vocabulary are likely to be obscured (Kyle, 2019). The results indicate a consistent and clear difference between the original data and the different types of machine translationese for all language pairs.

Aside from the different LFP scores between the training data and the translations, we also see a difference between the languages themselves. French and Spanish have more variety (higher B1 and lower B3 (Beyond 2000) values) compared to EN_{ES} and EN_{FR}. Since the LFPs are computed on tokens, this reflects the richer morphology in French and Spanish compared to English.

TTR, Yule's I and MTLD For completeness, we also present three more commonly used measures of lexical diversity: type/token ratio (TTR) (Templin, 1975), Yule's K (in practice, we use the reverse Yule's I) (Yule, 1944), and the measure of textual lexical diversity (MTLD) (McCarthy, 2005).

TTR presents the ratio of the total number of *different* words (types) to the *total* number of words (tokens). Higher TTR indicates a higher degree of lexical diversity. Yule's characteristic constant (Yule's K) (Yule, 1944) measures constancy of text as the repetitiveness of vocabulary. Yule's K and its inverse Yule's I are considered to be more resilient to fluctuations related to text length than TTR (Oakes and Ji, 2013). The third lexical diversity metric is MTLD. MTLD is evaluated sequentially as the mean length of sequential word strings in a text that maintains a given TTR value (McCarthy, 2005).[7] We present the scores for TTR, Yule's I and MTLD for our data and MT engines in Table 5.

	FR			ES		
	TTR	Yule's I	MTLD	TTR	Yule's I	MTLD
ORIG	**3.02**	**9.28**	**119.40**	**4.08**	**12.31**	**96.23**
PB-SMT	1.79	3.00	112.00	2.37	4.02	92.01
LSTM	1.56	2.14	104.89	2.03	2.95	86.57
TRANS	2.07	3.82	115.66	2.89	6.23	95.72
	EN_{FR}			EN_{ES}		
	TTR	Yule's I	MTLD	TTR	Yule's I	MTLD
ORIG	**2.89**	**6.64**	**108.70**	**2.88**	**6.61**	**108.46**
PB-SMT	1.74	2.07	94.65	1.82	2.25	93.18
LSTM	1.50	1.53	86.93	1.44	1.42	87.91
TRANS	2.04	3.10	101.95	2.09	3.26	99.62

Table 5: TTR, Yule's I and MTLD scores. For all metrics, higher scores indicate higher lexical richness. For ease of readability and comparison we multiplied TTR scores by 1,000 and Yule's I scores by 10,000.

The scores in Table 5 show that, overall, and according to all three metrics, the original training data has a higher lexical diversity than the machine translationese.

The data for the morphologically richer languages (FR, ES) as well as its machine translationese variants (PB-SMT, LSTM and TRANS) have higher lexical richness than the morphologically poor(er) language (EN).

Synonym Frequency Analysis The objective of synonym frequency analysis is to understand, for words with multiple possible translations, with what frequency the various translations for a given word appear in the translated text. It is called *synonym* frequency in reference to the fact that when translating from one language to another, it is common for a word in the source language to have a corresponding word in the target language to which the source word is typically translated, and that primary translated word can have many *synonyms* that constitute acceptable alternative translation options. Note that we perform this analysis only in one direction: from English into the morphologically richer languages French and Spanish.

	FR			ES		
	PTF↓	CDU↓	SynTTR↑	PTF↓	CDU↓	SynTTR↑
ORIG	**9.666**	**2.725**	**15.10**	**9.131**	**4.539**	**21.13**
PB-SMT	9.715	2.957	11.87	9.236	4.637	17.4
LSTM	9.748	3.154	10.96	9.32	4.782	15.34
TRANS	9.717	3.077	12.25	9.285	4.687	17.15

Table 6: Synonym frequency metrics for our MT systems: primary translation frequency (PTF), cosine distance from uniform (CDU) and TTR modified to only consider words with multiple translation options (SynTTR). The SynTTR scores were multiplied by 100,00 for easier viewing. Higher SynTTR scores indicate greater diversity, while lower PTF and CDU scores indicate greater diversity.

To examine synonym frequency, we first lemmatize the text using SpaCy.[8] Next, we map all nouns, verbs, and adjectives in the source to their possible translation options retrieved from bilingual dictionaries.[9] We then count the number of appearances of these different translation options for the ORIG as well as the MT data. For example, for the English word "look" with translation options in Spanish {"mirar", "esperar", "buscar", "parecer", "dar", "vistazo", "aspecto", "ojeada", "mirada"}, the number of appearances in the TRANS data are as follows: {("mirar": 4002), ("esperar": 3302), ("buscar": 2814), ("parecer": 1144), ("dar": 977), ("vistazo": 182), ("aspecto": 46), ("ojeada": 0), ("mirada": 0)}. From this mapping of translation option to number of appearances we take a vector consisting only of the numbers of appearances for each translation option, and refer to this as a translated word distribution. That is, for the aforementioned example, the distribution vector is: {4002, 3302, 2014, 1144, 997, 182, 46, 0, 0}. We use these counts and distributions as described below.

Our first synonym frequency metric deals di-

[7] In our experiments we used 0.72 as a TTR threshold.

[8] https://spacy.io/

[9] English-Spanish: https://github.com/mananoreboton/en-es-en-Dic, English-French: https://freedict.org/downloads

rectly with the primary translation frequency (PTF), where the "primary translation" is the translation for a given source word that appears in the target text most often. We argue that selecting secondary translation options for each source word less frequently, and selecting the primary option more frequently, indicates a decrease of lexical diversity. We measure the PTF by taking the average *primary translation prevalence* over all source words for each MT system.

As a second metric we used the cosine distance between a uniform translated word distribution, where each translation option would be equally prevalent, and the actual translation distributions (we denote this metric as CDU). While the ideal distribution of translations for a given word is almost certainly non-uniform (and therefore not perfectly diverse), this metric still gives valuable information about the tendencies of different systems to favor certain translation options over others.

The third metric is a modified TTR which we refer to as Synonym TTR (or SynTTR). Unlike with regular type/token ratio, rather than considering all tokens that appear in the text, we consider as types only translation options from the source-target mappings described above and as tokens we consider only appearances of valid types. This metric exposes where translation systems completely drop viable translation options from their vocabulary.

Table 6 shows the results for these 3 metrics. Interestingly, the MT systems can be ranked in the same order according to all these metrics: PB-SMT $>$ TRANS $>$ LSTM, where $>$ denotes the comparison of lexical diversity (higher to lower). However, across the 3 metrics, for both language pairs, the reference translations (ORIG) appear to be the most lexically diverse in terms of synonym frequency, with the lowest PTF and CDU and highest SynTTR. This reinforces the idea that MT algorithms have a negative impact on the diversity of language.

4.3 Grammatical Diversity

Grammatical diversity manifests itself on the sentence (syntactic complexity) and word level (morphological complexity). With our experiments, we focus on the morphological complexity by averaging the inflectional diversity of all lemmas. To do so, we adopted two measures, originating from Information Theory: Shannon entropy (Shannon, 1948) and Simpson's Diversity Index (Simpson, 1949). The former emphasizes on the richness aspect of diversity while the latter on the evenness aspect of diversity. We used the Spacy-udpipe lemmatizer to retrieve all lemmas.[10]

Shannon Entropy Shannon entropy (H) measures the level of uncertainty associated with a random variable (X). It has been applied in use-cases from economy, ecology, biology, complex systems, language and many others (Page, 2007, 2011). In the study of language Shannon entropy has previously been used for estimating the entropy of language models (Behr et al., 2003). We use it to measure the entropy of wordforms given a lemma. In particular, the entropy of inflectional paradigm of a specific lemma could be computed by taking the base frequency of that lemma (frequency of all wordforms associated with that lemma) and the probabilities of all the wordforms within the inflectional paradigm of that particular lemma. Using such a formulation of entropy allows us to measure the morphological variety (or the loss thereof) for the machine translationese produced by each system – higher values of H indicate higher diversity and vice-versa. We use Equation 1 to compute the entropy of the inflectional paradigm of a lemma.

$$H(\ell) = -\sum_{wf \in \ell} p(wf|\ell) \log p(wf|\ell)) \quad (1)$$

$H(\ell)$ denotes the entropy of the lemma ℓ and, for the wordform wf, $p(wf|\ell)$ is computed as the fraction of the counts of the wordform, $count(wf)$, to the count of all wordforms for the lemma ℓ, i.e. $p(wf|\ell) = \frac{count(wf)}{\sum_{wf* \in \ell} count(wf*)}$. We use $\in$ to indicate wordforms of a given lemma.

Simpson's Diversity Index Like Shannon Entropy, Simpson's Diversity Index (D) is a measure used to determine variation in categorical data. Values close to 1 indicate higher homogeneity, thus lower diversity and values close to 0 indicate higher variability, thus higher diversity.

Following the same reasoning as with Shannon entropy, we compute Simpson's diversity index for each lemma and the corresponding wordforms according to the formula in Equation 2.

$$D(\ell) = \frac{1}{\sum_{wf \in \ell} p(wf|\ell)^2} \quad (2)$$

[10] https://github.com/TakeLab/spacy-udpipe

We average the Shannon entropy and Simpson's diversity index for all lemmas to get an indicative score for each translation system or the original text. We denote these with ***H*** and ***D***, accordingly. To the best of our knowledge, our work is the first to use Shannon entropy and Simpson's diversity index for the study of inflectional richness on a text level. The closest to our application of these two diversity metrics for measuring inflectional richness is the work by del Prado Martín et al. (2004). Their work on morphological processing uses Shannon entropy to compute the amount of information carried by a morphological paradigm.

An illustration of the Shannon entropy and Simpson's diversity index of a lemma is given in Table 7. We list the number of occurrences for every wordform (male singular, male plural, female singular and female plural) appearing in our datasets for the French lemma 'président' (EN: president). We then compute H and D by applying Equation 1 and Equation 2 accordingly. While both Shannon H and Simpson's D scores usually range between 0–1, for ease of readability we multiply the scores presented in Table 7 and Table 8 by 100 to present them in the range of $[0 - 100]$.

	lemma: président				H↑	D↓
	président	présidents	présidente	présidentes		
ORIG	93774	2029	1490	8	**18.11**	**92.95**
PB-SMT	99367	2019	496	1	12.81	95.16
LSTM	95272	2039	291	N/A	12.17	95.3
TRANS	92946	1952	617	N/A	13.86	94.74

Table 7: An illustration of the Shannon entropy and the Simpson's diversity index computed for the occurrences of the different wordforms of the French lemma for 'president' (président).

For lemmas with a single wordform Shannon entropy and Simpson's diversity index will be $H = 0.0$ and $D = 1.0$, respectively. While this makes sense when measuring the diversity of one morphological paradigm, they actually impact the average scores ***H*** and ***D*** without contributing to the understanding of diversity in a comparative study such as ours. In particular, lemmas with single wordforms may be either an evidence of low diversity, e.g. a translation system will always generate only one form or of high diversity, e.g. rare words that are single wordform for a particular lemma (such as synonyms of more common words) can indicate the ability of a system to generate more diverse language (in terms of synonymy). That is why we computed ***H*** and ***D*** on lemmas with two or more wordforms. For completeness, we also present the number of single wordform lemmas. The Shannon entropy and Simpson's diversity index for French, Spanish and English for all datasets are presented in Table 8. The scores are shown in the range $[0 - 100]$ as noted above.

	FR			ES		
	H↑	***D***↓	Single	***H***↑	***D***↓	Single
ORIG	**75.20**	**56.42**	79k	**78.42**	**54.96**	92k
PB-SMT	69.00	59.64	51k	71.79	58.56	54k
LSTM	69.28	59.48	53k	72.84	58.29	55k
TRANS	73.13	57.70	58k	77.23	56.26	64k
	EN_{FR}			EN_{ES}		
	H↑	***D***↓	Single	***H***↑	***D***↓	Single
ORIG	**59.04**	**63.43**	78k	**59.05**	**63.42**	78k
PB-SMT	55.57	65.80	51k	56.31	65.29	61k
LSTM	53.15	67.02	50k	53.85	66.64	48k
TRANS	55.85	65.43	58k	56.22	65.19	68k

Table 8: Shannon entropy (***H***) for French, Spanish, English (EN_{FR} and EN_{ES}) and Simpson's diversity index (***D***) for original training data and the output of the PB-SMT, LSTM and TRANS systems. Scores are multiplied by 100 for ease of readability.

The ***H*** and ***D*** scores in Table 8 are an evidence of the negative impact of MT on the morphological diversity – the scores for the ORIG indicate a consistent higher diversity. Comparing the MT systems, it results that TRANS retains morphological diversity better than the others. LSTM performs better than PB-SMT for translations into the morphologically richer languages (FR and ES) but PB-SMT seems much better than LSTM for translations into English. While the loss of lexical diversity could, in some cases be a desirable side-effect of MT systems (in terms of simplification or consistency), the uncontrolled loss of morphological richness is problematic as it can prevent systems from picking the grammatically correct option.

5 Conclusions

In this work, we explore the effects of MT algorithms on the richness and complexity of language. We establish that there is indeed a quantitatively measurable difference between the linguistic richness of MT systems' training data and their output – a product of algorithmic bias. These findings are in line with previous results described in Vanmassenhove et al. (2019). Assessing diversity or richness in language is a multifacted task spanning over various domains. As such, we approach this task from multiple angles focusing on lexical diversity and sophistication, morphological variety and

a more translation specific metric focusing on synonymy. To do so, we analyse the results of 9 different metrics including established, newly proposed and adapted ones. The metrics suit we developed is unprecedented in the study of MT quality and we believe it could drive future research on MT evaluation.

Based on a wide range of experiments with 3 different MT architectures, we draw the following main conclusions: (i) all 9 metrics indicate that the original training data has more lexical and morphological diversity compared to translations produced by the MT systems. This is the case for all language pairs and directions; (ii) Comparing the MT systems among themselves, there is a strong indication (for most metrics) that Transformer models outperform the others in terms of lexical and morphological richness. We also ought to note that, on average, the ranking of the systems in terms of diversity metrics correlates with the quality of the translations (in terms of BLEU and TER). This is something that would need to be further explored in future work; (iii) The data for the morphologically richer languages (ES, FR) has higher lexical and (evidently) morphological diversity than the English data both in the original data and in the translations generated by all systems. However, for PB-SMT, LSTM and TRANS the difference in scores is much smaller than the ORIG, indicating that the MT systems have a stronger negative impact (in terms of diversity and richness) on the morphologically richer languages.

Acknowledgement

We would like to thank the reviewers for their insightful comments and feedback.

References

Roee Aharoni, Moshe Koppel, and Yoav Goldberg. 2014. Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL2014), Volume 2: Short Papers*, pages 289–295.

Nora Aranberri. 2020. Can translationese features help users select an MT system for post-editing? *Procesamiento del Lenguaje Natural*, 64:93–100.

Lyle F Bachman. 2004. *Statistical analyses for language assessment*. Cambridge University Press.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA. 15pp.

Mona Baker. 1999. The role of corpora in investigating the linguistic behaviour of professional translators. *International journal of corpus linguistics*, 4(2):281–298.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Frederic H. Behr, Victoria Fossum, Michael Mitzenmacher, and David Xiao. 2003. Estimating and comparing entropies across written natural languages using PPM compression. In *Proceedings of the 2003 Data Compression Conference (DCC 2003)*, page 416, Snowbird, UT, USA.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT2020)*, pages 280–290, Online.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. Massive Exploration of Neural Machine Translation Architectures. In *Proceedings of the Association for Computational Linguistics (ACL2017)*, pages 1442–1451, Vancouver, Canada.

Bram Bulté. 2013. The development of complexity in second language acquisition. *A dynamic systems approach (Unpublished doctoral dissertation)*.

Bram Bulté, Alex Housen, Michel Pierrard, and Siska Van Daele. 2008. Investigating lexical proficiency development over time–the case of dutch-speaking learners of french in brussels. *Journal of French Language Studies*, 18(3):277–298.

Aylin Caliskan-Islam, Joanna J Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *arXiv preprint arXiv:1608.07187*, pages 1–14.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL2011): Human Language Technologies, Volume 2: Short Papers*, pages 176–181, Portland, Oregon, USA.

Scott A Crossley, Tom Cobb, and Danielle S McNamara. 2013. Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, 41(4):965–981.

Joke Daems, Orphée De Clercq, and Lieve Macken. 2017. Translationese and post-editese: How comparable is comparable quality? *Linguistica Antverpiensia New Series-Themes in Translation Studies*, 16:89–103.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP2019)*, pages 6343–6349.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165.

Margherita Ippolito. 2014. *Simplification, Explicitation and Normalization: Corpus-Based Research into English to Italian Translations of Children's Classics*. Cambridge Scholars Publishing.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations: Poster Session*, Banff, Canada.

Beata Beigman Klebanov and Michael Flor. 2013. Associative texture is lost in translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 27–32, Sofia, Bulgaria.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP2004)*, pages 388–395, Barcelona, Spain.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of The Tenth Machine Translation Summit (MT Summit 2005)*, pages 79–86, Phuket, Thailand.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open-Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL2007)*, pages 177–180, Prague, Czech Republic.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL2011), Volume 1: Long Papers*, pages 1318–1326.

Svenja Kranich. 2014. Translations as a locus of language contact. In *Translation: A multidisciplinary approach*, pages 96–115. Springer.

Kristopher Kyle. 2019. Measuring lexical richness. *The Routledge Handbook of Vocabulary Studies*, page 454.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP2018)*, pages 4791–4796, Brussels, Belgium.

Batia Laufer. 1994. The lexical profile of second language writing: Does it change over time? *RELC journal*, 25(2):21–33.

Batia Laufer and Paul Nation. 1995. Vocabulary size and use: Lexical richness in l2 written production. *Applied linguistics*, 16(3):307–322.

Philip M McCarthy. 2005. An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD). In *PhD Thesis, Dissertation Abstracts International, Volume 66:12*. University of Memphis, Memphis, Tennessee, USA.

Jack Merullo, Luke Yeh, Abram Handler, Alvin Grissom II, Brendan O'Connor, and Mohit Iyyer. 2019. Investigating sports commentator bias within a large corpus of American football broadcasts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 6355–6361, Hong Kong, China.

IS Paul Nation. 1984. *Vocabulary lists: Words, affixes, and stems*. English Language Institute, Victoria University of Wellington.

Michael P Oakes and Meng (eds) Ji. 2013. Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research. In *Studies in Corpus Linguistics, Volume 51*, page 361. John Benjamins Publishing Company, Amsterdam, The Netherlands.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics, Volume 29:1*, pages 19–51. MIT Press, Cambridge, Massachusetts, USA.

Scott E. Page. 2007. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press.

Scott E. Page. 2011. *Diversity and Complexity*. Princeton University Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadephia, PA, USA.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT 2018): Research Papers*, pages 186–191, Belgium, Brussels.

Fermın Moscoso del Prado Martın, Aleksandar Kostić, and R Harald Baayen. 2004. Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, 94(1):1–18.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), Volume 1: Long Papers*, pages 1715–1725, Berlin, Germany.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Edward H Simpson. 1949. Measurement of diversity. *Nature*, 163(4148):688–688.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA 2006) 200:6*, pages 223–231, Austin, Texas, USA.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 1630–1640, Florence, Italy.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems (NIPS 2014)*, pages 3104–3112, Montreal, Quebec, Canada.

Mildred C. Templin. 1975. *Certain Language Skills in Children: Their Development and Interrelationships*. Greenwood Press, Westport, Connecticut, USA.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT 2018): Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Eva Vanmassenhove. 2020. On the Integration of Linguistic Features into Statistical and Neural Machine Translation. *arXiv preprint arXiv:2003.14324*.

Eva Vanmassenhove and Christian Hardmeier. 2018. Europarl datasets with demographic speaker information.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, pages 3003–3008, Brussels, Belgium.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII (MT Summit 2019), Volume 1: Research Track*, pages 222–232, Dublin, Ireland.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of The Thirty-first Annual Conference on Neural Information Processing Systems 30 (NIPS 2017)*, pages 5998–6008, Long Beach, CA, USA.

G. Udnv Yule. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), Volume 1: Long Papers*, pages 654–664, Vancouver, Canada.