Natural Language Processing (2025), 1–20
doi:10.1017/nlp.2025.10003

ARTICLE

Propagating machine translation traits to predict potential impact on the target language

Nora Aranberri[1] and Jose A. Pascual[2]

[1]HiTZ Center, University of the Basque Country UPV/EHU, Donostia-San Sebastián, Spain and [2]Intelligent Systems Group, University of the Basque Country UPV/EHU, Donostia-San Sebastián, Spain
Corresponding author: Nora Aranberri; Email: nora.aranberri@ehu.eus

(Received 8 December 2023; revised 7 January 2025; accepted 20 May 2025)

Abstract

Research suggests that the texts produced using machine translation (MT) do not fully represent the linguistic traits of the natural language. Yet, the ever-increasing quality and access to MT is resulting in its steady adoption by both language professionals and general users. According to contact linguistic theories, such adoption might result in MT-specific language traits permeating the target languages. This work takes a first step into considering the changes that a language might endure over time by observing the variation of linguistic trends along a series of MT generations. We train ten sequential engines using each to produce the target side of the training corpus of the following and calculate a number of metrics to observe linguistic diversity at a lexical, morphological, and syntactic level for a large, fixed test set. Quantitative results show an initial loss of lexical diversity, which, albeit gradually, only continues at a much slower pace in the following MT generations. In turn, structural variations and, in particular, morphological variations across generations are less marked, which might indicate a more stable behaviour regarding grammatical consistency. Overall, the resulting MT language seems increasingly homogeneous, marked by the reduced presence or disappearance of low-frequency words, and compact, with a decreasing proportion of function words relative to content words.

Keywords: machine translation; linguistic features; feature propagation

1. Introduction

Translation Studies scholars have proposed to consider translation from the perspective of contact linguistics. Among other effects, a text produced through the translation process displays features of the original language known as interference (Toury 1995) and shining through (Teich 2012). This means that a translated text might contain lexical and structural options that follow the source text closely and express the meaning, making a correct use of the target language, but not necessarily in the same manner a native speaker would if directly writing the same content in that particular language. As happens with naturally occurring textual characteristics, according to Kranich (2014), the features present in the translated texts could permeate the language if they occur with sufficient frequency; that is, texts originally written in the target language could start displaying such features (see Kotze 2020).

If translation is a factor in language change, we could argue that so is machine translation (MT) as, after all, it is an act of translation. This technologically mediated modality could contribute its own particular features. To start, research seems to show that the MT systems exacerbate certain translation process-related effects such as homogenisation, simplification, and interference (Toral 2019; Vanmassenhove, Shterionov, and Way 2019, 2021). MT systems appear to make a more limited use of the linguistic diversity available in the target languages, opt for simpler units, including

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

a tendency to write out elements that can be omitted, and use words, expressions, and syntax of the source text unusually often. The slight degree of uncharacteristic language is amplified by the suboptimal use of MT. As discussed by Sánchez-Gijón and Piqué Huerta (2020), it has been proven that the success of the MT output is dependent on the alignment between the training data and the source text. However, the authors underscore, once deployed, engines—publicly available ones in particular—tend to be used irrespective of source text genre and domain; people often overlook best practices for MT and avail of the technology to address ongoing translation needs regardless of their nature. This can lead to translations that incorporate language that is not typical of their context of use. Adding to this, let us note that MT is more and more widely used (Pitman 2021), not only by language professionals but also by general users with varying competence and awareness of the languages involved. While the former might mitigate the negative effects brought to language by the systems, the latter might not be equipped with the necessary skills to do so.

In this work, we aim to contribute to the emerging area of studies that seek to uncover the impact MT could have on language development. Rather than focusing on the quality of the systems, we try to account for the shaping of a language due to the use of MT. It is important to note that language change takes place within complex systems where many factors are implicated. In such a scenario, MT is an additional element that will be more or less central to such change depending on the status of the languages, societies, and events involved. We focus on the contribution of MT, and as such, the sole factor considered in this experiment is the change introduced in the target language as a result of the language processing mechanisms of an MT system, which would or would not permeate people's use of language in the real world.

In an attempt to study MT impact, we propagate the features of the MT output by training several generations of systems and monitoring the tendencies of metrics that capture linguistic richness at a lexical, morphological, and structural level. This initial exploration seems to show that there is a decrease in lexical and morphological diversity in the language produced by the MT baseline with respect to the original traits of the language and that, even when the decline might persist, in particular for lexical elements, for the successive MT generations, it occurs at a gradual, slower rate. The structural metrics used, however, do not seem to point to considerable divergences for this aspect of the language over generations.

The remainder of this paper is structured as follows: Section 2 reviews the studies that have looked into the linguistic characteristics of MT output. Section 3 puts forward the language change model and the role of the MT engine in that framework and sets out our MT training approach. Section 4 describes the experimental setup, considering the data, the system characteristics, and the training cycles. Section 5 discusses the results, focusing on overall quality, lexical richness, morphological richness, and structural similarity. Finally, Section 6 draws the main conclusions of this work and outlines the main avenues for future work.

2. Related work

Even when most research in the field of MT focuses on new techniques and approaches to improve the quality of the output, throughout the years, a number of studies have paid attention to the linguistic features of the translation candidates generated by those engines. Albeit scarce, the research focusing on the linguistic features of the MT output suggests that MT systems produce output that does not fully reflect the characteristics of the target languages.

In their experiments with English, French, and Spanish, Vanmassenhove *et al.* (2019) found that the language generated by MT systems was less diverse than when produced naturally based on several lexical metrics (TTR (Templin 1957), Yule's K (Yule 1944b), and MTLD (McCarthy and Jarvis 2010)). What is more, based on an analysis of word frequencies and bias, they claimed that recurrent language elements tend to appear even more frequently, while those that are originally less common appear even more rarely.

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

MT seems to distort the target language, but the features of the translation proposal also appear to vary depending on the MT paradigm used to build the systems. When comparing the output of neural machine translation (NMT) and statistical machine translation (SMT), Bentivogli *et al.* (2016), working on the English-to-German combination, reported that NMT outperformed SMT systems, namely, for lexical diversity, percentage of morphological and lexical errors, and word ordering (based on the information extracted with the HTER metric using forms, lemmas, and shifts). The exception was the effect of sentence length, which deteriorated the output quality of both architectures but especially that of NMT systems for sentences over 35 words.

Toral and Sánchez-Cartagena (2017), working on English and Czech, Finnish, German, Romanian, and Russian, built on the previous findings, adding that, indeed, NMT and SMT system outputs are considerably different based on character sequence overlaps. According to the authors, based on the information extracted from the Hjerson metric (Popović 2011), NMT output is more fluent and makes fewer inflection and reordering errors. Using MGIZA++ word alignments and Kendall's tau distance (Kendall 1938) as metrics, they also claimed that NMT seems to display a greater capacity to perform reorderings, which results in proposals that follow the reference translation closely.

It is not necessary to compare two distinct paradigms such as SMT and NMT to encounter differences. For example, Vanmassenhove *et al.* (2019) included two NMT systems built using different neural architectures in their experiments, a recursive neural network (RNN) (Cho *et al.* 2014; Sutskever, Vinyals, and Le 2014) and a transformer (Vaswani *et al.* 2017), and the outputs differed, associating the latter with translation of higher quality. It is important to note, however, that the linguistic tendencies shown by both architectures were similar. More recently, Marchisio *et al.* (2022), working on English–German, revealed style differences between an unsupervised system (a MASS transformer) and a supervised system (a transformer-big) when measuring TER (Snover *et al.* 2006) over part-of-speech (POS) tag sequences and perplexity (Jelinek *et al.* 1977). From their experiments, the authors conclude that unsupervised MT might have more structural diversity as system quality improves and that its output might sound more natural.

Nevertheless, differences between naturally occurring language and MT language should not come as a surprise. Translation scholars agree on the existence of a set of genuine, common features that make translations distinct from texts originally written in that same language, (un)popularly called *translationese* (Jiménez-Crespo 2023), even if they do not concur with the specific features it involves (Laviosa 2002). In this context, some scholars talk about translation universals, namely, simplification, normalisation, explicitation, and interference (Baker 1993; Toury 1995). Empirical research seems to indicate that translated texts show signs of reduced richness and seem "abnormally normal" (Tirkkonen-Condit 2002, 217).

Therefore, because MT is, in essence, an act of translation, the translationese effect is not unexpected and is in fact proven by Vanmassenhove *et al.* (2021). Adding to this, post-editing—the act of revising MT output—has been pointed to amplify the translationese features observed in human translations (Toral 2019) probably because post-editors are primed by the MT output (Green, Heer, and Manning 2013), which seems to already include significant bias. However, it is possible that translationese is not the only effect at play in this context. Research seems to indicate that an algorithmic effect might also be at play; that is, the techniques used to process and learn the translation models might be distorting the language somehow (Zhao *et al.* 2017; Vanmassenhove *et al.* 2019; Prates, Avelar, and Lamb 2020).

All in all, research suggests that, no matter how good its quality, MT generates a target language with linguistic features that differ from those of texts originally written in that language. A long way remains to fully understand the dimension of the distortion, that is, how much the MT language distances itself from the naturally occurring language and how frequently this happens. Similarly, we are yet to pinpoint in what language elements such distortion manifests concretely. For example, we are yet to identify which specific grammatical structures, terms, idioms, and collocations get promoted in terms of recurrence in the MT output and, conversely, determine

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

Table 1. The Replicator model adapted to language evolution (from Steels (2017: 202))

Replicator model	Species	Language
Units (interactors)	Organisms	Language users
Traits (replicators)	Features of organisms	Language components
Sources of variation	Through genetic transmission	Through social learning
Selection	Survival and fecundity	Communicative success, expressive power, cognitive effort

the presence of which is reduced or even discontinued. We might even find that modified structures and meaning allocations surface and appear recurrently or that new ones emerge altogether. Ultimately, it would be relevant to study to what extent the MT language could permeate the natural use of the languages, in other words, how MT could shape the evolution of languages. It is widely accepted that languages (and their speakers) and more prone to assimilate certain variations and reject others. Likewise, the status of a language, among others, its level of normalisation and hegemony, also plays a role in permeability.

3. Language change model

According to Steels (2017), a language consists of three highly complex and significantly language-specific systems, namely, the sound system, the conceptual system, and the vocabulary and grammar system. And it is historical linguists who have made it their goal to account for their evolution by drawing phylogenetic trees that display the temporal and distance relations between languages (Gray and Atkinson 2003) and by examining samples of texts diachronically to identify universal trends of changes, for example, from the perspective of grammaticalisation (Heine and Kuteva 2007). More recently, evolutive features have been considered also from causal and mechanistic perspectives, which consider, among others, the capacity of humans to process information, how they interact and communicate, and how societies are structured (Steels 2011). What we can extract from here is that language change is the result of an extraordinarily intricate process.

To try to understand where MT resides with respect to language change, let us consider how the change happens using the replicator model, a tool evolutionary linguists have borrowed from the field of biology (see Table 1) (Steels 2017). The *units*, in our case language users, carry *traits*, the features of a language (vocabulary, syntactic patterns. . .). The traits *vary* through social learning and are *selected* based on the benefits accrued for the units. Depending on how we view MT—as a tool for communication that language users avail or an entity with its own "voice"—it might be debatable whether MT is a unit or a source of variation. In any case, the consolidation of the MT traits in language would depend on the level of permeation, which would be based on the language users' perception—conscious or unconscious—that they are beneficial.

In this work, we aim to take a preliminary step towards predicting what influence MT might have in the shaping of languages. As a starting point, we study the potential for language change caused by the MT engines in isolation. For our experiment, we take an extreme approach of the replicator model as follows: we start from a scenario where language users interact with each other naturally and construct an MT system with the language that represents that moment. From that instant onwards, interaction with the target language happens through MT alone, and as a result, MT traits completely permeate the target language. At this stage, a new MT system is trained with data that reflects the target language change brought by the first generation of MT traits. We experiment with several generations to observe language trends.

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

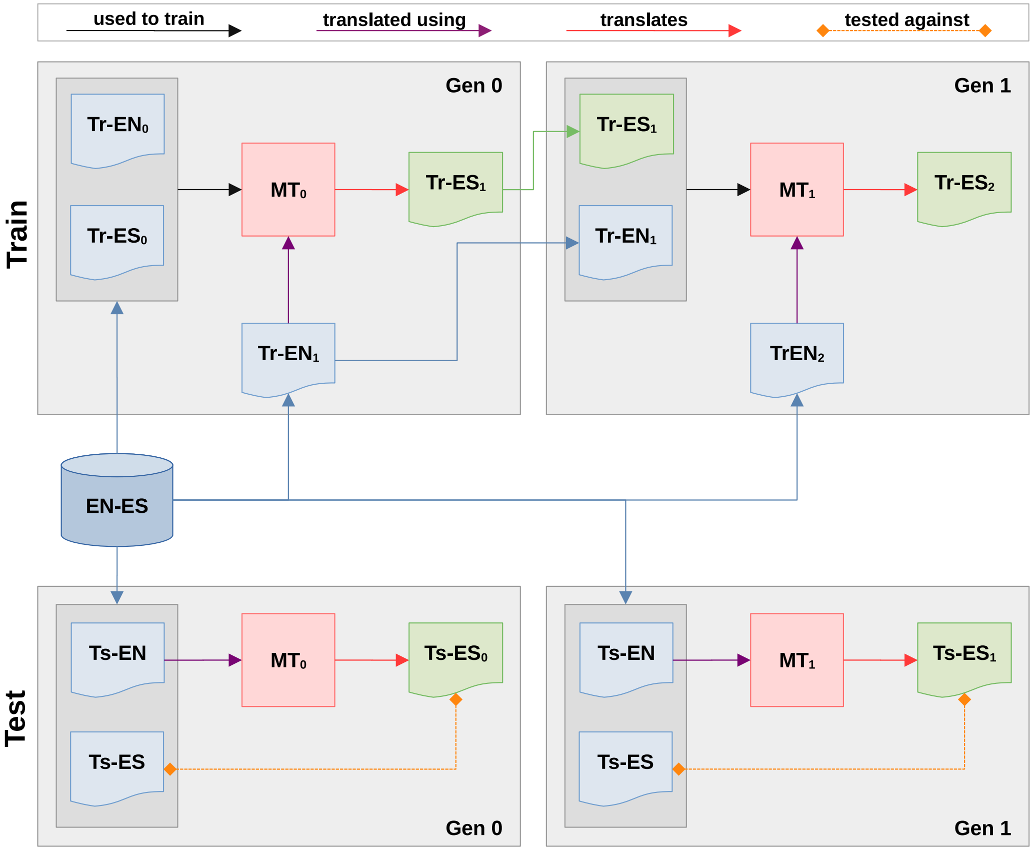

Figure 1. Representation of our MT training (top) and testing (bottom) approach.

Let us consider our particular scenario in more detail (see Figure 1). We focus on the English–Spanish combination where English acts as the source language and Spanish as the target language; in other words, we look into how Spanish could vary influenced by English through automated translation over time. We start by training a system with an extract of a corpus of current web-crawled data (see Section 4). This baseline system (MT_0) introduces the first MT traits within the natural language. We assume that with time, they completely permeate Spanish. Then, we train a second system, MT_1, the first MT generation, which reflects the changed Spanish. That is, we use a new extract of the corpus for which the English source has been translated into Spanish with the previous system. This provides us with the second generation of MT traits. Again, we assume that they completely permeate Spanish. We proceed in this manner to produce several generations of MT traits.

By following this approach, we can monitor the trends in which MT might contribute to the change of Spanish over time. In this constrained approach, the source language remains stable, all traits permeate the target language, and the lapse of time is only partially represented through the different generations of the systems and their different training sets. Yet, all in all, we believe that it allows us to propagate MT influence and capture its gradual contribution. While one-to-one system comparisons cannot be made given the different training sets used, we can study trends through a common test set, which consists of sentences extracted and kept from the same pool of data from where the training sets were obtained.

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

4. Experimental setup

4.1 The data

We use CCMatrix (Schwenk *et al.* 2021; Fan *et al.* 2021) as our data source for training and testing. The texts are collected from web data released by the Common Crawl project.[a] We acknowledge that this collection might include both original and translated texts for any of the languages involved. In any case, we can argue that it is a rather representative sample of the current linguistic traits of the languages. The English–Spanish language pair consists of 409.1 million sentence pairs (downloaded from Tiedemann 2012). We set aside a random set of 1 million sentences to be used as the test set. We use a significantly larger sample compared to other MT experiments because we aim to represent the widest breadth of the languages' traits as possible. From the remaining data, we randomly extract batches of 8 million sentence pairs to train the systems (see Section 4.3 for further details as to how these data are used). No further filtering is applied during extraction.

4.2 The system

We use the transformer architecture implementation of the OpenNMT-tf toolkit[b] with the settings suggested by its community as optimal to lead to quality on par with the original transformer work (Vaswani *et al.* 2017) as follows:

- number of layers: 6
- size: 512
- transformer ff: 2048
- number of heads: 8
- dropout: 0.1
- batch size: 4096
- batch type: tokens
- learning optimiser Adam with beta2 = 0.998
- learning rate: 2
- learning rate decay enabled

The cap for system training is established at a maximum of 100,000 steps. We validate the model every 5,000 steps using perplexity as a metric, and we save the 5 best-scoring validation models. We set as early stopping criteria that no improvement higher than 0.001 of perplexity score is reached on the last five validation steps.

The data was tokenised using OpenNMT's tokeniser, and the vocabularies were built based on sub-word units computed using BPE with 60,000 merging operations, based on results from Gowda and May (2020), for the training sets (source and target separately).

4.3 The training cycles

We train a chain of systems as follows: We use a subset of 8 million sentence pairs to train the baseline MT system, MT_0. Next, we use MT_0 to obtain the Spanish translation of a second 8-million subset and use this new parallel data to train the first-generation MT system, MT_1. We proceed in this manner for ten generations.

[a] https://commoncrawl.org

[b] https://opennmt.net/OpenNMT-tf

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

5. Results

To monitor change, we calculate several linguistic characteristics on the original test set, the output of the baseline MT, and the subsequent MT generations and monitor the trends.

5.1 Translation quality

We first look at a number of metrics that are used to assess general MT quality. We calculate the widely used lexical metrics BLEU (Papineni *et al.* 2002), TER (Snover *et al.* 2006), and chrF (Popović 2015) using the sacreBLEU (Post 2018) implementation and, following the advice of Freitag *et al.* (2022), also report on two neural-based metrics, namely, BLEURT (Sellam, Das, and Parikh 2020) (specifically, implementation BLEURT-20-D12, a 12-layer distilled model that is $\sim 3.5X$ smaller and 1 order of magnitude faster than the original BLEURT-20) and COMET (Rei *et al.* 2020) (specifically, the default model WMT$_2$0-COMET-DA), which are reported to correlate better with human judgements.

BLEU considers word-level precision with regard to a reference translation with a penalty for brevity, while chrF works at character n-gram level. In turn, TER, also working at word level, is an edit-distance metric that calculates the minimum number of edits (additions, deletions, substitutions, and shifts) required to transform the MT output into the reference sentence. From a completely different perspective, BLEURT is a pre-trained evaluation model that uses the BERT language model (Devlin *et al.* 2019), synthetic data to represent a diverse range of lexical, syntactic, and semantic dissimilarities and a set of human evaluations. COMET is also a reference-based neural regression model built on top of XLM-R (large) (Conneau *et al.* 2020) and trained on human judgments of translation quality. As can be noted, the chosen metrics cover a considerable set of approaches, which serves to monitor the level of quality from different perspectives for sounder conclusions.

For the outputs of the baseline MT and each of the following generations, we calculate scores using the target side of the test set as a reference, which represents the traits of the original Spanish, and also the source in the case of COMET. Results gathered in Table 2 show that neural metrics and BLEU point to the same tendency: as the MT generations advance, the scores drop slightly and rather steadily.[c] This implies that the output of the systems is more and more different from the reference each time. chrF seems to contradict this. However, the character-based nature of this metric might be having an effect here. As we will see in Section 5.2, the number of types used by each MT generation decreases, in particular, types with low frequencies, which may contain character sequences that are rarer. As a result, the text might be getting more and more homogeneous also in terms of character sequences. Therefore, the quantity of character sequences that are correct (more standard) can be higher in each generation, while at the same time, the text being more and more different from the reference translation. We see another exception with TER, which indicates that, albeit minimally, fewer changes are needed to transform the MT outputs into the reference as the generations advance. Further qualitative research would be necessary at this point in order to understand this behaviour, although the homogenisation effect attributed to MT might be at play again. With regard to the level of quality, it should be noted that absolute numbers seem to point to a rather strong baseline and strong subsequent generations.

5.2 Lexical richness and density

To study linguistic diversity, we first report type and token information and then calculate four widely used metrics, specifically, type/token ratio (TTR) (Templin 1957), Yule's K (Yule 1944a),

[c] Given that the quality of the systems can vary depending on the used seed, we trained three MT$_0$ and MT$_{10}$ systems to check whether the quality remained similar, and therefore, the trends in the experiments are valid.

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

Table 2. Global quality automatic scores as reported by BLEU, TER, chrF, BLEURT, and COMET, where the output of each MT generation (Gen.), MT_0 to MT_{10}, is compared against the Spanish side of the test set. Scores are provided within a 0–100 range, with the best-scoring MT generation for each metric in bold. The ↓ symbol indicates that lower values of the metric correspond to better performance

Gen.	BLEU	TER ↓	chrF	BLEURT	COMET
0	**37.281**	61.653	51.986	**62.045**	**59.67**
1	37.010	61.526	52.156	61.744	58.92
2	36.633	61.318	52.399	61.493	58.02
3	36.333	61.144	52.630	61.281	57.34
4	36.059	60.979	52.871	61.102	56.73
5	35.786	60.815	53.057	60.921	56.14
6	35.548	60.684	53.246	60.778	55.62
7	35.368	60.586	53.334	60.644	55.18
8	35.182	60.474	53.498	60.544	54.81
9	35.010	60.369	53.692	60.442	54.49
10	34.831	**60.262**	**53.862**	60.351	54.06

and the measure of textual lexical diversity (MTLD) (McCarthy and Jarvis 2010), followed by lexical density (LD) (Toral 2019).

The assumption behind these lexical diversity indices, which are based on type (unique words) and token (all instances of the types) distributions and ratios, is that the higher the number of different words in a text, the richer the text is and, as a consequence, better and requiring a higher language competence to produce it. TTR is the simplest of the metrics. It is calculated by dividing the number of types in a text by the number of tokens.

Yule's K focuses on the frequency distribution of words in a text. It measures how constant a text is in terms of the appearance of new words; that is, it aims to capture how often words repeat and to what extent the vocabulary is spread across different frequency levels. It is calculated as follows:

$$K = 10^4 \cdot \left(\frac{\sum_{i=1}^{\infty} i^2 f_i - N}{N^2} \right) \tag{1}$$

where $\sum_{i=1}^{\infty} i^2 f_i$ is the sum of the squares of the frequencies of each word and N is the total number of words in the text.

Lower values of Yule's K indicate a richer vocabulary, meaning the text uses a wider variety of words and has less repetition. Conversely, higher values point to a less rich vocabulary, meaning the text has more repetition of the same words.

In turn, MTLD calculates the average length of word sequences in a text that maintain a given level of lexical diversity. It does this by setting a threshold for the diversity and splitting the text into segments based on when this threshold is crossed (we use the default value of .720 TTR). The text is read word by word, and the TTR is calculated cumulatively. When the TTR falls below the threshold of 0.72, a segment is marked, and the calculation restarts from the next word. Next, the average length of all segments is calculated. This average length is the MTLD score, indicating how many words, on average, can be read before the lexical diversity falls below the threshold. Higher

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

MTLD values indicate greater lexical diversity, as it takes more words to reach the threshold, while lower MTLD values indicate lower lexical diversity, as the threshold is reached more quickly.

According to McCarthy and Jarvis (2010), lexical diversity indices suffer from two key problems: their sensitivity to text length and the assumption of textual homogeneity. The issue with text length lies in the fact that tokens increase linearly as the text expands, whereas types, that is, the number of different words, gradually decrease given the degree of repeated tokens that is necessary to maintain a meaningful and understandable development of a narrative. In fact, some researchers have taken this as an opportunity to account for the representativeness of a topic in a text or corpus (Morse 1995; McEnery 2003). In reference to the assumption of homogeneity, the authors point that the level of diversity required by specific rhetorical purposes and strategies is different, and therefore, an unsteady distribution of types might not necessarily reflect weaker text production skills.

These vulnerabilities lead us to consider their appropriateness for our experimental setup. First, the metrics are not usually tested on texts as large as our test set, and as we mentioned, all metrics are dependent on length to a higher or lower degree. Second, our test set consists of isolated sentences instead of internally coherent full-length texts with a narrative thread that reflects natural type/token distributions.

The previous notwithstanding, we concluded that the two issues discussed above do not seem crucial in our experiments because of the static nature of our test set (we use the same 1 million English sentences to obtain the Spanish translation of the different MT generations) and because our aim is to consider relative changes across generations. However, to reduce the effect the mentioned issues might have, we calculate the scores by applying a batch approach as suggested by Hess *et al.* (1986) and McCarthy and Jarvis (2007), among others. Concretely, we have divided the test set into batches of 1,000 sentences and calculated all metrics for each one of them and report their averages.[d]

The results of the different metrics are displayed in Table 3. Let us focus on the type and token counts first. The former inform about the number of different words used throughout the texts, while the latter account for all instances of such words. These are indicative of the breadth of the vocabulary used, its frequency, and also text length. The original Spanish text, consisting of 1 million sentences, contains 17,345,663 tokens and 299,158 types. The counts for both tokens and types decrease in the Spanish translation produced by the baseline MT (MT generation 0, MT_0). Tokens are reduced by 1.41%, meaning that the length of the text produced by MT is slightly shorter than in the original. The drop in types is more considerable as 17.15% do not appear in the Spanish translation with respect to the original version.

The subsequent MT generations also see a tendency of reduced tokens and types (0.03% and 0.94%, respectively, for MT_1 with reference to MT_0), but the drop is substantially more limited, with even a couple of exceptions where the counts increase. In the case of types, we see a maximum drop of 2,339 units (MT_1) and a minimum of 98 (MT_9). Note that MT_{10} sees a small increase of 68 types. In turn, the number of tokens also tends to shrink, but between a maximum of 31,518 words (MT_2) and a minimum of 4,723 words (MT_1). We observe two cases (MT_4 and MT_9), where the token count increases, even when the types decrease.

Overall, these counts seem to indicate that there is certain diversity or language traits that the MT system cannot learn and/or reproduce, but once the traits of the MT language are determined, the loss is drastically reduced going forward. The Spanish with MT traits is, unsurprisingly, more suitable for the subsequent engines to imitate.

Let us now turn to lexical diversity indices. Results show inconclusive trends: all three metrics, TTR, MTLD, and Yule's K, show a drop from the original to the baseline MT (MT_0)—and an increase in Yule's K, given that it accounts for uniformity, and therefore, it is a lower score that

[d] We recalculated the scores by dividing the test set into extracts of different lengths (100 and 25 sentences), and the results displayed similar trends.

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

Table 3. Lexical diversity and density scores as reported by type and token counts, type-token ratio (TTR), measure of textual lexical diversity (MTLD), Yule's K, and lexical density (LD) for the Spanish side of the test set and the output of each MT generation, MT_0 to MT_{10} (Gen.), and where the generation with the highest diversity for each metric is in bold. The ↓ symbol indicates that lower values of the metric correspond to better performance

Gen.	Types	Tokens	TTR	MTLD	Yule's K ↓	LD
orig.	**299,158**	**17,345,663**	**0.40** (0.01)	**74.45** (3.58)	146.66 (9.94)	**0.55** (0.004)
0	247,845	17,100,251	0.38 (0.01)	69.21 (6.70)	147.87 (9.85)	0.56 (0.005)
1	245,506	17,095,528	0.39 (0.01)	69.67 (7.43)	144.84 (9.87)	0.56 (0.005)
2	244,647	17,064,010	0.39 (0.02)	70.80 (8.01)	142.62 (9.65)	0.56 (0.005)
3	243,731	17,047,171	0.39 (0.01)	70.97 (8.96)	141.04 (9.59)	0.56 (0.005)
4	242,660	17,052,538	0.39 (0.01)	70.82 (9.42)	140.10 (9.42)	0.56 (0.005)
5	242,543	17,028,742	0.39 (0.01)	70.41 (10.28)	138.25 (9.44)	0.56 (0.005)
6	242,083	17,020,396	0.39 (0.01)	69.72 (11.19)	136.85 (9.41)	0.56 (0.005)
7	241,943	17,000,971	0.39 (0.01)	71.68 (10.23)	136.42 (9.25)	0.57 (0.005)
8	241,577	16,992,709	0.39 (0.01)	69.97 (12.25)	135.24 (9.73)	0.57 (0.005)
9	241,479	17,003,902	0.39 (0.01)	69.60 (12.31)	134.98 (9.73)	0.57 (0.005)
10	241,547	16,988,185	0.39 (0.01)	68.61 (13.24)	**132.91** (9.48)	0.57 (0.005)

indicates a higher diversity (see Table 3). From the baseline MT onward, TTR increases slightly (MT_0 to MT_1) but stabilises early. MTLD scores also stay within the 68–71 range, even when the fluctuation is more varied (it increases until MT_3 and then decreases until MT_6; it goes upward for MT_7 and then down again for MT_7–MT_{10}). In contrast to the previous metrics, Yule's K shows a rather steady increase in diversity.

We calculated one final metric for this linguistic level: LD. It is calculated as the ratio of content words (nouns, proper nouns, verbs, adjectives, and adverbs) to total words (tokens) and can be taken as an approximation of the informativeness of a text (Toral 2019). The higher the number of content words, the more information a text is supposed to include. A higher ratio indicates that the text is, indeed, more informative but also that it becomes more and more compact in the sense that fewer function words are used to arrange the content words within the sentence. The results in Table 3 show an apparent slight increase in LD, indicating that, in proportion, the occurrences of function words are actually decreasing as MT generations advance.

What lexical metrics seem to reveal is that there is a tendency to use a reduced vocabulary as MT generations advance. We further inspected this trend by focusing on POS categories. We tagged the original Spanish text and the MT proposals with Universal POS tags[eE] using spaCy[f] model es_core_news_sm-3.7.0. We then counted the occurrences of each category for the original test set and the output of the MT baseline and generations. This allows us to observe whether the vocabulary reduction happens across all grammatical categories similarly or whether there are certain categories that suffer more.

Figures 2 and 3 show the progression of the content words and the function words, respectively. We normalised the results by dividing the counts for each MT output by the counts of the original test. If we look at content words, we see that except for proper nouns (PROPN), all categories tend

[e] https://universaldependencies.org/u/pos

[F] https://spacy.io

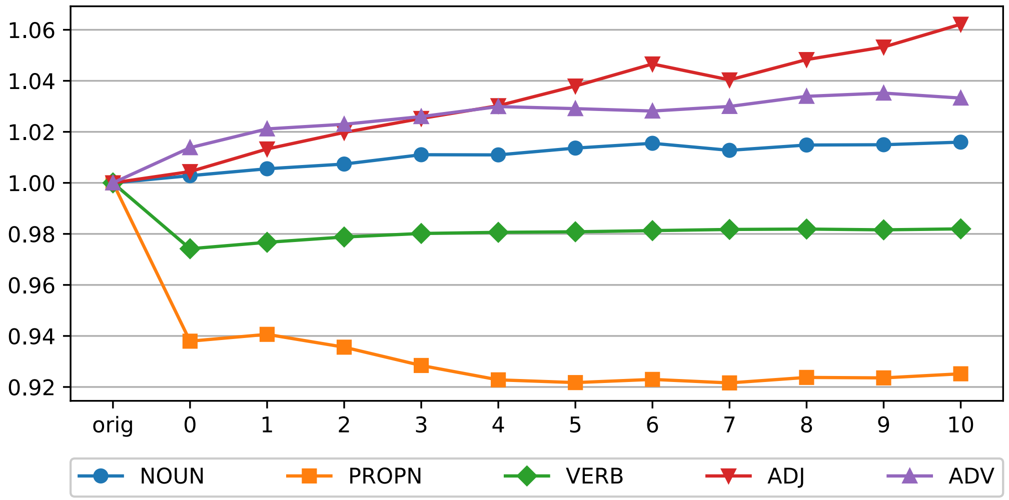

Figure 2. Normalised token counts for content words across MT generations, MT_0 to MT_{10}, where NOUN refers to nouns, PROPN to proper nouns, VERB to lexical verbs, ADJ to adjectives, and ADV to adverbs.

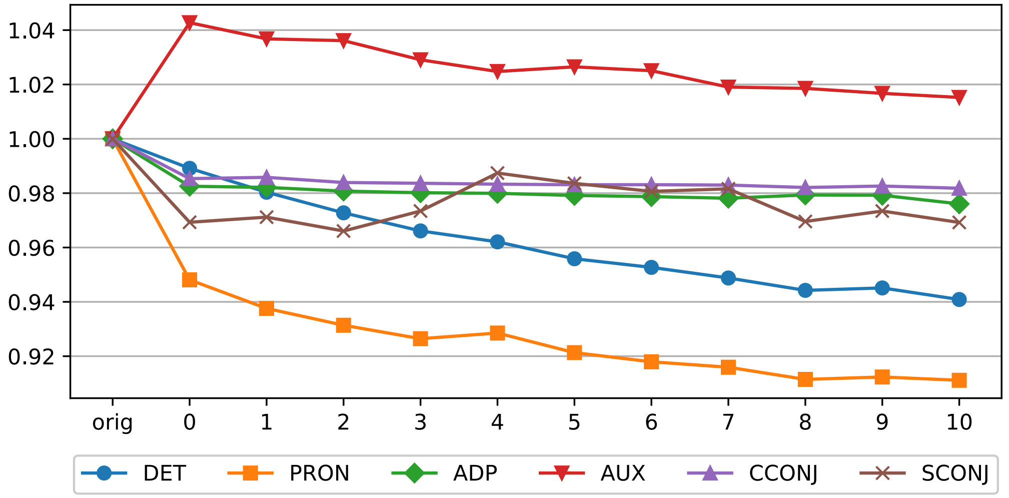

Figure 3. Normalised token counts for function words across MT generations, MT_0 to MT_{10}, where DET refers to determiners, PRON to pronouns, ADP to prepositions, AUX to auxiliary verbs, CCONJ to coordinating conjunctions, and SCONJ to subordinating conjunctions.

to increase, albeit slightly. We see certain nuances when paying attention to each of the categories. Nouns (NOUN) undergo a rather steady increase starting from MT_0. And so seems the case of verbs (VERB) from MT_0 and MT_2, but it seems to level from then onwards. What is interesting about verbs is that this category suffers a drop from the original text to MT_0 before it starts picking up, and ten generations later, it is still behind original counts. Proper nouns experience the same drop, which is even sharper and continues in this direction for a few generations before it starts increasing slowly. Interestingly, adjectives (ADJ) and adverbs (ADV) are more frequent in the machine-translated texts, and adjectives, in particular, increase as generations advance.

If we turn to closed categories, that is, function words, we observe that all categories suffer an initial more pronounced drop from the original text to MT_0 and then decrease at different rates. Prepositions (ADP) and coordinating conjunctions (CCONJ) vary slightly, while determiners

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

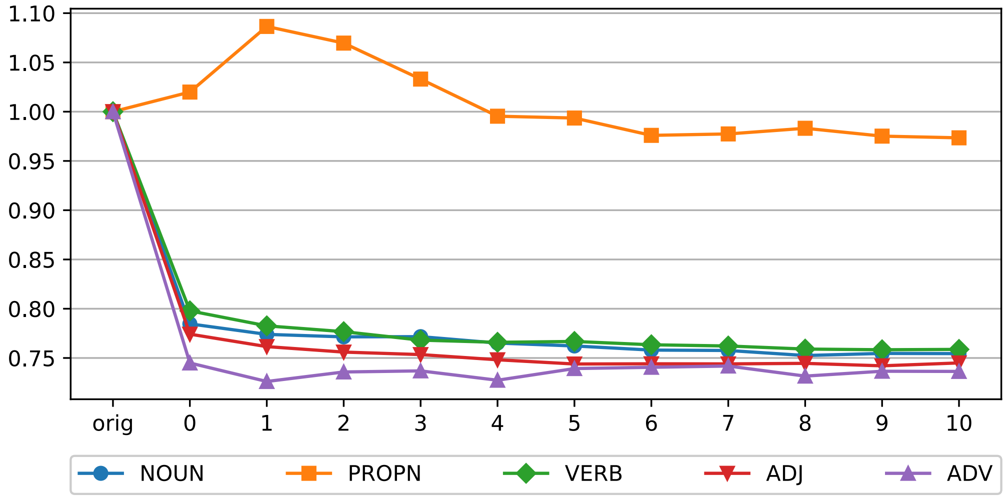

Figure 4. Normalised type counts for content words across MT generations, MT_0 to MT_{10}, where NOUN refers to nouns, PROPN to proper nouns, VERB to lexical verbs, ADJ to adjectives, and ADV to adverbs.

(DET) and pronouns (PRON) do so more sharply. The progression of subordinating conjunctions (SCONJ) is more uneven. Even when it never reaches original occurrences, it fluctuates as MT generations advance. One exception to the downward trend is that of auxiliaries (AUX), that is, the elements that accompany lexical verbs to express grammatical information such as person, number, tense, mood, aspect, voice, or evidentiality. They sustain an increase from the original text to MT_0 and then start decreasing. After ten MT generations, their occurrence rate is still higher than in the original. A possible reason for this increase might be that MT systems might have used verb phrases requiring auxiliaries more frequently compared to the original Spanish.

From the review of POS counts, we can conclude that all open categories except proper nouns increase as the MT generations advance, and function words tend to decrease. This might indicate that the language becomes increasingly compact, which corroborates the LD results in Table 3. This can happen either by favouring syntactic constructions that compress content words and require fewer function words, for example, noun clusters over complements, which include prepositions and determiners for each of the elements involved. It can also mean that language becomes more explicit and lacking referencing by means of pronouns. Of course, the higher density of content words can also be due to incorrect clusterings and omissions of compulsory function words. Yet, each category is affected differently, which calls for a more qualitative study to fully understand the in-category changes.

Token-based counts hide the diversity of the vocabulary in terms of the actual number of different words that are used. Therefore, in order to check whether all POS categories follow the same trends as overall token counts, we investigated word counts further and looked into the counts of types for words grouped by POS tag. Results are displayed in Figures 4 and 5 for content words and function words, respectively. As before, we normalised the results by dividing the counts for each MT output by the counts of the original test set.

The first thing to notice is that the loss of types, which happens much more markedly from the original test set to the MT_0 baseline, as we already saw in Table 3, occurs across all POS categories. All categories decrease between 10% and 30%. Specifically, in the case of open categories, except for proper nouns, they all decrease between 20% and 28%, and in the case of closed categories between 12% and 29%. The trend for open categories seems to show that noun, verb, and adjective types are reduced very slightly after the initial drop at MT_0 towards a point of stability. Adverbs

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

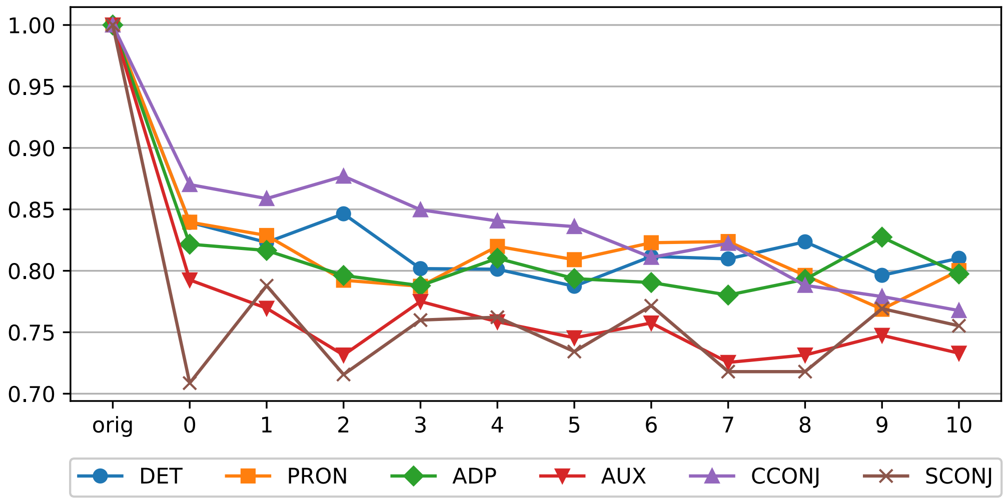

Figure 5. Normalised type counts for function words across MT generations, MT_0 to MT_{10}, where DET refers to determiners, PRON to pronouns, ADP to prepositions, AUX to auxiliary verbs, CCONJ to coordinating conjunctions, and SCONJ to subordinating conjunctions.

drop slightly lower with the baseline MT_0 but then increase slightly until they reach a very similar range as the other categories.

If we combine these trends with those of the token counts (see Figure 2), we can say that in the case of nouns, adjectives, and adverbs, even when the texts display a continuous slightly higher number of occurrences across generations, there is an initial loss of types with the MT_0 baseline, but after that, their number remains rather steady. This means that those types that have got through MT_0 will appear more frequently than in the original test set. The verbs also follow this trend, but we did observe a slight initial drop in token counts too. We see a rather unexpected behaviour with proper nouns, as the frequency of tokens decreases up until MT_4 while types undergo an initial upward trend and only stabilise after MT_4. This could indicate that MT leads to proper nouns being used less frequently while they become more diverse. Further exploration would be necessary to find out how specifically MT handles this word category.

In reference to closed categories, we observe that the type counts for coordinated conjunctions decrease gradually, while the remaining categories stay within a certain range with more marked fluctuations from generation to generation. Remember that in terms of token counts (see Figure 3), we observed that coordinated conjunctions stayed rather stable, which means that certain types appear more and more often and others disappear as generations advance. In the case of pronouns and determiners, which saw a steady drop, we see that there is some stability with regard to types. Therefore, these categories disappear as syntactic elements but not in terms of variety. Auxiliaries undergo an increase in overall occurrences, but their types decrease from the very beginning and there seems to be a tendency towards their reduction, although this is not constant.

We have observed that to a certain degree, all type categories decrease, especially when going through the MT_0 baseline. In order to learn whether the frequency of occurrences for the types changes as MT generations advance, we calculate the number of types across six frequency bands: Band 1 includes the number of words that appear over 100,000 times, Band 2 99,999–10,000 times, Band 3 9,999–1,000 times, Band 4 999–100 times, Band 5 99–10 times, and Band 6 9–1 times. The results, displayed in Table 4, clearly show that the words that have a higher risk to disappear from the language are those in the lowest frequency band. This is in line with the results by Vanmassenhove *et al.* (2019). What is more, it seems that types remain rather stable with frequencies of over 1,000 and some slight loss is observed for those with 100 to 1,000 occurrences. Band 5

Table 4. Counts of types across frequency bands for the Spanish side of the test set and the output of each MT generation, MT_0 to MT_{10} (Gen.), where Band 1 refers to words that appear over 100,000 times, Band 2 99,999–10,000 times, Band 3 9,999–1,000 times, Band 4 999–100 times, Band 5 99–10 times, and Band 6 9–1 times

Gen.	Band 1	Band 2	Band 3	Band 4	Band 5	Band 6
orig.	19	108	1,730	10,186	34,749	247,336
0	19	105	1,708	9,874	35,032	201,017
1	19	105	1,713	9,909	34,907	198,853
2	19	104	1,716	9,903	34,939	197,966
3	19	104	1,721	9,900	34,908	197,079
4	19	106	1,724	9,896	34,871	196,044
5	19	104	1,718	9,916	34,848	195,938
6	19	106	1,718	9,918	34,901	195,421
7	19	106	1,726	9,912	34,916	195,264
8	19	106	1,729	9,918	34,879	194,926
9	19	106	1,729	9,913	34,957	194,755
10	19	107	1,733	9,911	35,025	194,752

with 10–99 occurrences displays a contrasting trend as it shows a higher occurrence for the baseline MT_0 with respect the original language. This then decreases and increases slightly seemingly in a randomly manner, yet never as low as the original. It seems to pick an upward trend after MT_5, even when, within our range of generations, we do not see it surpass the values of MT_0. The obvious decline is sustained by the types with frequencies below 10. Note that in this band, it is the drop between the original test and the baseline MT_0 that suffers considerably (a drop of over 45,000 types), but then, the decline is markedly slower.

5.3 Morphological richness

To account for morphological variety, we consider the inflexional diversity of lemmas. We follow the work of Vanmassenhove *et al.* (2021), where the authors adopt the Shannon entropy (Shannon 1948) and Simpson's diversity index (Simpson 1949) to compute the entropy of the inflexional paradigms of lemmas. This could account for the richness of diversity.[g] The authors claim that, in line with the conceptualisation of Shannon, the entropy of the inflexional paradigm of a specific lemma could be computed by taking the base frequency of that lemma (frequency of all word forms associated with that lemma) and the probabilities of all the word forms within the inflexional paradigm of that particular lemma as shown in Equation 2:

$$p(wf \mid \ell) = \frac{\text{count}\,(wf)}{\sum_{wf^* \in l} \text{count}\,\left(wf^*\right)} \tag{2}$$

where $p(wf \mid \ell)$ is computed as the fraction of the counts of the wordform, count (wf), to the count of all wordforms for the lemma ℓ and $\in$ indicates the wordforms of a given lemma ℓ.

[g] We have replicated the approach used by Vanmassenhove *et al.* (2021) to calculate these metrics, including the Spacy-udpipe lemmatiser to identify the lemmas.

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

Table 5. Morphological variety scores as reported by Shannon Entropy and Simpson's Diversity Index for the Spanish side of the test set and the output of each MT generation, MT_0 to MT_{10} (Gen.), where the generation with the highest diversity for each metric appears in bold. The ↓ symbol indicates that lower values of the metric correspond to better performance

Gen.	Shannon entropy	Simpson's diversity index ↓
orig.	**6.39**	**96.19**
0	5.50	96.74
1	5.69	96.65
2	5.61	96.70
3	5.55	96.74
4	5.57	96.72
5	5.69	96.65
6	5.69	96.65
7	5.61	96.70
8	5.61	96.70
9	5.61	96.70
10	5.61	96.70

Based on this idea, Shannon entropy is calculated as per Equation 3:

$$H(\ell) = -\sum_{wf \in \ell} p(wf \mid \ell) \log p(wf \mid \ell) \tag{3}$$

where $H(\ell)$ denotes the entropy of the lemma ℓ.

This idea can also be followed for Simpson's diversity index (D), which accounts for the evenness of the diversity, and it is calculated as per Equation 4:

$$D(\ell) = \frac{1}{\sum_{wf \in \ell} p(wf \mid \ell)^2} \tag{4}$$

We then calculate the average Shannon entropy and Simpson's diversity index for all lemmas ℓ to obtain a representative score for the Spanish side of the original test set and each MT generation. Both metrics serve as approximations of morphological diversity, with higher values indicating greater diversity and lower values indicating less diversity.

The scores of both metrics usually range between 0 and 1, but for ease of readability, we multiply the scores presented in Table 5 by 100 to present them in the range of [0–100]. We can observe that according to these two metrics, similar to what we observed with lexical metrics, there seems to be a drop in diversity for MT_0 compared to the original Spanish (represented as a lower score for Shannon entropy and a higher score for Simpson's diversity index). Then, the MT traits across generations are very stable. This might indicate that while a degree of morphological reduction seems to be occurring when the language goes through the first MT cycle, the grammatical information may not be suffering any further considerable loss as generations progress.

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

Table 6. Structural similarity scores as reported by the perplexity measure and posTER metric, for the Spanish side of the test set and the output of each MT generation, MT_0 to MT_{10} (Gen.), where P:ES represents the perplexity of the the Spanish LM at the word level, P:ES-POS represents perplexity of the Spanish POS LM, and P:EN-POS represents the perplexity of the English POS LM. The ↓ symbol indicates that lower values of the metric correspond to better performance. The generation with the best scores for each metric appears in bold

Gen.	P:ES ↓	P:ES-POS ↓	P: EN-POS ↓	posTER ↓
orig.	27.22	5.52	7.55	-
0	**24.31**	**5.26**	**7.30**	**28.31**
1	24.81	5.28	7.34	28.36
2	25.23	5.30	7.37	28.55
3	25.53	5.32	7.40	28.69
4	25.78	5.34	7.43	28.75
5	26.04	5.35	7.44	28.94
6	26.24	5.37	7.46	29.09
7	26.44	5.38	7.48	29.08
8	26.62	5.39	7.49	29.24
9	26.75	5.40	7.51	29.38
10	26.91	5.41	7.52	29.54

5.4 Structural similarity

In this section, we study to what extent the language of each of the MT generations has the same structural traits as the original Spanish. To do so, we explore the results of two metrics, namely, perplexity and posTER (Marchisio *et al.* 2022).

To start, let us consider perplexity. We first build a language model, both at the word level and at POS level, trained using 8 million sentences from the original Spanish corpus, thus representing unprocessed language, which coincides with the batch used to train the baseline MT system. To train our neural model, we follow the architecture used to build GPT-2 (Radford *et al.* 2019), which is basically a decoder-only transformer. We then score the capacity of the LM to predict the language of the original Spanish test set and the output of the baseline MT and successive MT generations. The lower the perplexity, the better the LM is at predicting the language in the test sets, and therefore, the better it captures the features of the original language.

Results in Table 6 show that, both at word level and at POS level, MT_0 sees an improvement with respect to the original language, and then it degrades slowly and steadily. In ten MT generations, the perplexity does not reach the level of uncertainty it displays with the original language. The fact that the Spanish of the original test set—same language generation with which the LM has been trained—is more difficult to predict for the LM than the subsequent MT generation outputs seems counter-intuitive at first glance.

We hypothesise that the naturally occurring Spanish is more difficult to predict maybe due to its higher diversity. Then, when the language goes through the MT architecture for the first time, it might create a less diverse and more homogeneous language, which might be easier to predict by the LM. The trend of the metric in reference to the MT generations is not surprising if we assume that a level of degradation, which deviates from the original Spanish, is introduced by each system.

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

We also trained an English LM at the POS level to test whether the distortion we observe across MT generations is due to the impact of the source language to some extent; that is, the Spanish output of the MT generations is acquiring traits of the English source through interference. To start, results clearly show that the English LM is worse at predicting Spanish than the Spanish LM, which is a good sign. Next, we also observe that the behaviour of the metric shows the same trend as with the Spanish LMs, that is, the LM is better at predicting machine-translated Spanish than the original Spanish. The perplexity increases gradually for the subsequent MT generations, indicating that it is increasingly difficult for the English LM to predict the POS sequencing of the Spanish outputs. We can derive that the Spanish language is changing across generations and that such change is not identified as resulting in a POS sequencing that resembles English patterns more closely regardless of the level of interference that might (or might not) be happening. Therefore, the distortion introduced by the MT generations cannot be associated to source language interference without further analysis.

Let us focus on the second metric. posTER was proposed by Marchisio *et al.* (2022) as a measure to calculate structural similarity at POS level. The authors calculate the edit distance between translations to account for differences between versions. We apply this measure by using the POS-tagged, original Spanish test set as a reference and the translations of each of the MT systems as hypotheses.

Results in Table 6 show rather gradual yet minimal distortion according to this metric. From MT_0 to MT_{10}, which represent the lowest and highest posTER scores, the difference in the percentage of POS-level changes would be 1.23. Yet, as the authors (Marchisio *et al.* 2022) themselves suggest, the seventeen POS categories used by the Universal Dependency tag set might be too coarse to identify more subtle changes and conceive actual deviations.

6. Conclusions

In this work, we have taken a first step to explore how a language could be shaped by the use of MT. We have done so by applying an extreme approach where language change is solely considered from the MT perspective: we have tried to simulate an environment where the technology dictates which features get permeated and where generational changes are captured. Specifically, we have trained a chain of MT systems, where each training set consists of original English language text and its translation into Spanish obtained from the previous system. This has allowed us to propagate linguistic traits generated by the text processing capabilities of the transformer NMT architecture in a succession of ten MT generations in order to observe diversity trends at a lexical, morphological, and structural level. Needless to say that this approach is limited and not indicative of how exactly the target language would evolve given the complex ecosystems in which languages exist. However, it is a first step towards providing an insight into what happens to lexical and morphological elements and structural patterns of a target language when we put it through a transformer architecture for translation.

Overall quality metrics show that systems degrade generation after generation, albeit slightly, losing around 3 BLEU points and 5 COMET points from the MT_0 baseline to the last MT_{10} generation in a rather steady manner. This means that the translations are getting more and more different from the reference text, that is, the expected language traits.

In reference to lexical richness, type and token information confirms that using MT results in a decrease of diversity. Interestingly, we see that this drop is particularly marked when the language goes through the MT architecture for the first time (MT_0) and that the decline is slower and less noticeable as MT generations advance. We also observe that the number of types with a frequency of occurrence of 10 and over hardly changes, while those which appear more rarely decrease (and even end up disappearing), particularly when going through the MT architecture for the first time. This will bring homogeneity to the text through the increasing use of already frequent words.

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

When inspecting words based on POS categories, we see that the occurrence frequency of nouns, adjectives, and adverbs increases slightly, without an initial drop, as MT generations advance, while verbs remain rather stable after the initial drop. Function words tend to drop, except for auxiliaries, and while conjunctions and prepositions remain stable, determiners and pronouns occur less every time. This points to a language that is more compact every time, with a higher content-to-function word ratio. As per types, for all POS categories except proper nouns and coordinated conjunctions, numbers remain rather stable, with an almost unnoticeable downward trend, after an initial drop. If we combine type and token information, we see that for some POS categories, it is the diversity within them that decreases, whereas for others, it is the category as a whole that is reduced.

In terms of morphology, Shannon entropy and Simpson's diversity index indicate that there is hardly any change in the richness and evenness of the inflectional paradigms of the lemmas after an initial drop. Considering the ranges of the metric scores and our results, we can say that the richness is rather low, but the distribution is even. The fact that there are only small fluctuations within a stable range can be taken as an indication that while lexical items are reduced, the grammatical information—encoded in the inflections—remains in the language as the MT generations advance.

Finally, with regard to structure, a simple posTER metric based on edit-distance counts of POS-tagged test sets indicates a very slow distortion of the language as MT traits propagate through the MT generations. A more complex perplexity metric at a word and POS level based on a neural LM seems to indicate that the language is somewhat distorted after going through the MT architecture for the first time but then starts becoming easier to predict, probably thanks to a homogenisation effect generated by MT. Also, there appears to be no source language interference at the POS level, as the source language LM finds the translations more difficult to predict with each generation.

All in all, there seems to be certain "human" traits of language that the MT engines still fail to capture. As a consequence, the systems do not seem able to reproduce the target language with all its inherent features, which results in a degree of lexical loss and structural distortion. However, once those traits are absent, the MT engines seem to produce and learn the language rather consistently with minor loss. The current results stem from a specific language pair and an initial MT engine of robust quality trained with texts from multiple domains, which leaves other configurations to be tested, including those that consider full, coherent texts rather than isolated sentences. Besides, results point to interesting further research in the change in specific lexical, morphological, and structural elements, as occurrence frequencies and POS categories seem to play a role. However, identifying the specific language traits that the current transformer-based NMT systems cannot capture from naturally occurring languages seems a priority.

Funding statement. This work is partially supported by the DeepR3 (TED2021-130295B-C31), TRAIN (PID2021-123988OB-C31) and INSECTS (PID2023-152390NB-I00) projects funded by MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR/ERDF/FEDER, and the Basque Government through projects IT1570-22, IT1504-22, KK-2023/00012, KK-2023/00090, KK-2024/00030, and KK-2024/00068.

References

Baker M. (1993). Corpus linguistics and translation studies: Implications and applications. In *Text and Technology: In Honour of John Sinclair*, vol. 64, pp. 233–250. John Benjamins Publishing Company.

Bentivogli L., Bisazza A., Cettolo M. and Federico M. (2016). Neural versus phrase-based machine translation quality: A case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 257–267.

Cho K., van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H. and Bengio Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar: Association for Computational Linguistics, pp. 1724–1734.

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L. and Stoyanov V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 8440–8451.

Devlin J., Chang M.-W., Lee K. and Toutanova K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota: Association for Computational Linguistics, vol. 1, pp. 4171–4186. (Long and Short Papers).

Fan A., Bhosale S., Schwenk H., Ma Z., El-Kishky A., Goyal S., Baines M., Celebi O., Wenzek G., Chaudhary V., Goyal N., Birch T., Liptchinsky V., Edunov S., Grave E., Auli M. and Joulin A. (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research* **22**, 1–48.

Freitag M., Rei R., Mathur N., kiu Lo C., Stewart C., Avramidis E., Kocmi T., Foster G., Lavie A. and Martins A.F. (2022). Results of WMT22 metrics shared task: Stop using BLEU –neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation*. Abu Dhabi: Association for Computational Linguistics, pp. 46–68.

Gowda T. and May J. (2020). Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP*, Online. Association for Computational Linguistics, pp. 3955–3964.

Gray R.D. and Atkinson Q.D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**(6965), 435–439.

Green S., Heer J. and Manning C.D. (2013). The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Paris, France: ACM, pp. 439–448.

Heine B. and Kuteva T. (2007). *The Genesis of Grammar: A Reconstruction*, vol. 9. Oxford University Press.

Hess C.W., Sefton K.M. and Landry R.G. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech, Language, and Hearing Research* **29**(1), 129–134.

Jelinek F., Mercer R.L., Bahl L.R. and Baker J.K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America* **62**(S1), S63–S63.

Jiménez-Crespo M.A. (2023). "Translationese"(and "post-editese"?) no more: on importing fuzzy conceptual tools from translation studies in MT research. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, Tampere, Finland*. European Association for Machine Translation, pp. 261–268.

Kendall M.G. (1938). A new measure of rank correlation. *Biometrika* **1**(2), 81–93.

Kotze H. (2020). Translation, contact linguistics and cognition. In Alves, F. and Jakobsen, A. (eds.), *The Routledge Handbook of Translation and Cognition*. London: Routledge, pp. 113–132.

Kranich S. (2014). Translations as a locus of language contact. In *Translation: A Multidisciplinary Approach*. London: Palgrave Macmillan UK, pp. 96–115.

Laviosa S. (2002). *Corpus-based Translation: Studies Theory, Findings, Applications. Series: Approaches to Translation Studies*, vol. 17. Brill.

Marchisio K., Freitag M. and Grangier D. (2022). On systematic style differences between unsupervised and supervised MT and an application for high-resource machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 2214–2225.

McCarthy P.M. and Jarvis S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing* **24**(4), 459–488.

McCarthy P.M. and Jarvis S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* **42**(2), 381–392.

McEnery T. (2003). Corpus linguistics. In Mitkov, R. (ed.), *Handbook of Computational Linguistics*. Oxford: Oxford University Press, pp. 448–463.

Morse J.M. (1995). The significance of saturation. *Qualitative Health Research* **5**(2), 147–149.

Papineni K., Roukos S., Ward T. and Zhu W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318.

Pitman J. (2021). *Google Translate: One Billion Installs, One Billion Stories*. The Keyword. https://blog.google/products/translate/one-billion-installs/

Popović M. (2011). Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics* **96**(1), 59–68.

Popović M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, pp. 392–395.

Post M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation*. Belgium, Brussels: Association for Computational Linguistics, vol. 1, pp. 186–191. Research Papers.

Prates M.O.R., Avelar P.H. and Lamb L.C. (2020). Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications* **32**(10), 6363–6381.

Radford A., Wu J., Child R., Luan D., Amodei D. and Sutskever I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog* **1**(8), 1–24.

https://doi.org/10.1017/nlp.2025.10003 Published online by Cambridge University Press

Rei R., **Stewart C.**, **Farinha A.C. and Lavie A.** (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics, pp. 2685–2702.

Sánchez-Gijón P. and Piqué Huerta R. (2020). Conseqüències de la traducció automàtica neuronal sobre les llengües d'arribada. *Tradumàtica* **18**, 1–10.

Schwenk H., **Wenzek G.**, **Edunov S.**, **Grave E.**, **Joulin A. and Fan A.** (2021). CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics, pp. 6490–6500.

Sellam T., **Das D. and Parikh A.** (2020). BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 7881–7892

Shannon C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal* **27**(3), 379–423.

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

Snover M., **Dorr B.**, **Schwartz R.**, **Micciulla L. and Makhoul J.** (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, pp. 223–231.

Steels L. (2011). Modeling the cultural evolution of language. *Physics of Life Reviews* **8**(4), 339–356.

Steels L. (2017). Do languages evolve or merely change? *Journal of Neurolinguistics* **43**, 199–203.

Sutskever I., **Vinyals O. and Le Q.V.** (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems* **27**, 3104–3112.

Teich E. (2012). *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Berlin, Boston: De Gruyter Mouton.

Templin M.C. (1957). *Certain Language Skills in Children: Their Development and Interrelationships*. Minneapolis: University of Minnesota, The Institute of Child Welfare.

Tiedemann J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 2214–2218.

Tirkkonen-Condit S. (2002). Translationese - a myth or an empirical fact? a study into the linguistic identifiability of translated language. *Target. International Journal of Translation Studies* **14**(2), 207–220.

Toral A. (2019). Post-editese: An exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*. Dublin, Ireland: European Association for Machine Translation, pp. 273–281.

Toral A. and Sánchez-Cartagena V.M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain: Association for Computational Linguistics, vol. 1, pp. 1063–1073. Long Papers.

Toury G. (1995). *Descriptive Translation Studies: And Beyond*. John Benjamins Publishing Company.

Vanmassenhove E., **Shterionov D. and Gwilliam M.** (2021). Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online. Association for Computational Linguistics, pp. 2203–2213

Vanmassenhove E., **Shterionov D. and Way A.** (2019). Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, Dublin, Ireland: European Association for Machine Translation, pp. 222–232.

Vaswani A., **Shazeer N.**, **Parmar N.**, **Uszkoreit J.**, **Jones L.**, **Gomez A.N.**, **Kaiser Ł. and Polosukhin I.** (2017). Attention is all you need. *Advances in Neural Information Processing Systems* **30**, 1–11.

Yule C.U. (1944a). *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.

Yule G.U. (1944b). The statistical study of literary vocabulary. *Mathematical Proceedings of the Cambridge Philosophical Society* **42**, b1–b2.

Zhao J., **Wang T.**, **Yatskar M.**, **Ordonez V. and Chang K.-W.** (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2979–2989.

Cite this article: Aranberri N. and Pascual J. A. Propagating machine translation traits to predict potential impact on the target language. *Natural Language Processing* https://doi.org/10.1017/nlp.2025.10003