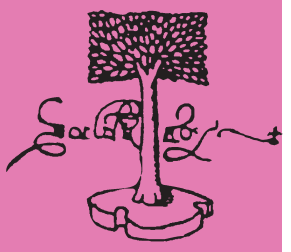
Año LVII. urtea
140 - 2025
Uztaila-abendua
Julio-diciembre



FONTES
LINGVÆ
VASCONVM
STVDIA ET DOCVMENTA

SEPARATA

Exploring lexical diversity in Basque news: original vs. machine-translated texts

Amaia Solaun Martínez, Nora Aranberri Monasterio

Gobierno de Navarra
Nafarroako Gobernua

Sumario / Aurkibidea

Fontes Linguae Vasconum. Studia et Documenta

Año LVII. urtea - N.º 140. zk. - 2025

Uztaila-abendua / Julio-diciembre

ARTIKULUAK / ARTÍCULOS / ARTICLES

Sumario / Aurkibidea

Exploring lexical diversity in Basque news: original vs. machine-translated texts

Euskarazko albisteetako aniztasun lexikoaren azterketa: jatorrizko testuak eta automatikoki itzulitakoak

Análisis de la diversidad léxica en euskera: noticias originales y traducidas automáticamente

Amaia Solaun Martínez
HiTZ Center, University of the Basque Country UPV/EHU
amaia.solaun@ehu.eus
https://orcid.org/0009-0004-6474-7979

Nora Aranberri Monasterio
HiTZ Center, University of the Basque Country UPV/EHU
nora.aranberri@ehu.eus
https://orcid.org/0000-0003-3719-9167

DOI: https://doi.org/10.35462/flv140.2

This work has been partially supported by the Basque Government (Research group funding IT1570-22). We are also thankful to several projects funded by MCIN/AEI/10.13039/501100011033: TRAIN (PID2021-123988OB-C31) and by FEDER, EU; and DeepR3 (TED2021-130295B-C31) and by the European Union NextGeneration EU/PRTR. Amaia Solaun was supported by the Basque Government through the Predoctoral Grant (PRE-2024-1-0147).

Received: 09/01/2025. Provisionally accepted: 16/06/2025. Accepted: 05/11/2025.

ABSTRACT

The rapid advancement of machine translation (MT) has led to increased exposure to MT-generated content. Given the potential influence of this technology on language use, this study compares the lexical diversity of machine translated and originally-produced online news articles in the Basque language. Combining established automatic metrics with detailed manual analysis, we conduct a comprehensive evaluation of multiple dimensions of lexical diversity. Our findings indicate a striking similarity in lexical richness between the two text types, with few aspects in which original texts surpass the lexical diversity of machine-translated texts.

Keywords: machine-translation; lexical diversity; Basque.

LABURPENA

Itzulpen automatikoaren (IA) aurrerapen azkarraren ondorioz, IAren bidez sortutako testuekiko dugun esposizioa handitu da. Teknologia horrek hizkuntzaren erabilera eraldatzeko potentziala duenez, azterketa honetan automatikoki euskaratutako eta jatorriz euskaraz sortutako albiste digitalen aniztasun lexikoa alderatzen ditugu. Finkatutako metrika automatikoak eta eskuzko azterketa xeheak bateratuz, aniztasun lexikoaren hainbat dimentsio zabalki aztertzen ditugu. Gure aurkikuntzen arabera, bi testu-motek antzekotasun nabarmenak dituzte aberastasun-lexikoaren aldetik. Izan ere, alderdi oso espezifiko gutxi batzuetan ikusten dugu jatorrizko testuen aberastasun-lexikoa apur bat handiagoa dela.

Gako hitzak: itzulpen automatikoa; aniztasun lexikoa; euskara.

RESUMEN

El rápido avance en la traducción automática (TA) ha dado lugar a una mayor exposición a contenidos generados por TA. Puesto que esta tecnología podría influir en el uso del lenguaje, en este estudio comparamos la diversidad léxica de las noticias digitales creadas originalmente en euskera frente a aquellas traducidas automáticamente. Combinando métricas automáticas consolidadas y análisis manuales detallados, realizamos un extenso análisis de varios aspectos de la diversidad léxica. Los resultados indican una notable similitud en la riqueza léxica entre ambos tipos de texto, solo en unos pocos aspectos específicos observamos un mínimo aumento de la riqueza léxica en los textos originales.

Palabras clave: traducción automática; diversidad léxica; euskera.

1. INTRODUCTION

Languages naturally evolve and adapt, influenced by a range of factors. This study explores a relatively new potential influence: machine translation (MT). The use of MT systems is rapidly growing in both professional and everyday settings, driven by improvements in quality and accessibility. Consequently, more people are engaging with MT-generated content, even in less widely spoken languages like Basque (Aranberri & Iñurrieta, 2024). Recent studies have highlighted distinctive linguistic patterns in MT output, arising from different underlying factors (Vanmassenhove et al., 2021). In this context, we deem important to examine whether these observations hold true for Basque and to identify the language features that are potentially open to change as a result.

Admittedly, the act of translation itself affects resulting texts. Previous research in translation studies has shown that translated texts possess distinct characteristics that set them apart from original works. Scholars classify these features into two categories: those arising from source-language interference (Toury, 1980, 2012) and those considered universal, independent of the source language (Baker, 1993; Blum-Kulka, 1986; Blum-Kulka & Levenston, 1983). These differences are pronounced enough that machines can accurately recognize whether a text is a translation (Baroni & Bernardini, 2005; Volansky et al., 2015).

Research has found that MT texts may also display distinct differences, not only when compared to original target-language texts but also in relation to human translations. Bizzoni et al. (2020) analysed differences between interpreted speeches, human

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

translations, and MT texts, concluding that while the three are similar, the latter exhibit notable differences from the other two. There is also evidence suggesting that MT output tends to be more closely aligned with the source text, showing higher linguistic interference (Castilho et al., 2019; Green et al., 2013). For instance, sentence length and part-of-speech (POS) sequences in MT texts often resemble those of the source language more closely than in human translations (Toral, 2019).

Although research on MT output is still limited, scholars have begun to outline its specific features. Evidence suggests that MT can reduce lexical and morphosyntactic diversity while amplifying frequency biases present in the training data due to algorithmic bias (Vanmassenhove et al., 2019, 2021). These amplified features persist even in post-edited texts, after professional translators revise the MT output (Toral, 2019). In a focused study, Gamallo and Labaka (2021) examined the use of passive constructions in English-to-Spanish MT and found that, although Spanish offers three passive structures, MT systems overwhelmingly favoured the option most similar to the English passive construction. Despite the apparent general tendency for MT systems to produce less diverse output, it is important to note that as MT systems continue to improve, they may in certain scenarios obtain competitively diverse output. This has been observed by Shaitarova et al. (2023), who report that, in certain corpora, some commercial MT systems are capable of achieving results in commonly used lexical diversity metrics that are comparable to those of human translations.

The features found in human translations reflect the nature of the translation process and should not be seen as a weakness in the text (Chesterman, 2010; Gellerstam, 1986). However, the amplification of such features in MT, probably stemming from the system's inability to fully replicate the complexity of the target language, can be considered a flaw, especially when it leads to incorrect, inadequate, or unnatural language use. This raises concerns about the potential impact of widespread MT usage on language development, particularly in the case of minority languages.

Basque serves as a compelling example. The language exists in a diglossic environment, undergoing a normalization process with a recent standardization (Zabaleta, 2019). These circumstances lead to varied levels of language competence among speakers and a cultural reliance on translations from dominant regional languages, such as Spanish and French.

From a technological perspective, Basque is considered a low-resource language (Hernáez et al., 2012; Sarasola et al., 2022), resulting in MT systems of lower quality compared to high-resource languages. Nevertheless, a recent survey highlights the increasing adoption of MT by the Basque-speaking community, coupled with positive attitudes toward its usefulness, which suggests that MT use may grow even further (Aranberri & Iñurrieta, 2024). Given Basque's vulnerable position and the improvements in MT systems, the language may become increasingly susceptible to the linguistic influence by MT. Therefore, we propose to analyse MT output and its potential effects on Basque to develop strategies to mitigate any negative consequences if necessary.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

As a step in this direction, this study focuses on the lexical diversity of Basque in original texts and MT-generated texts. We aim to analyse whether the breadth of lexical items is indeed reduced in MT output, as claimed by the majority of previous research, as this could result in an impoverished and limited language. Using the EiTB comparable corpus (Etchegoyhen & Gete, 2020), a collection of Basque and Spanish news texts, we translate the Spanish texts into Basque using the Elia translation engine (https://elia.eus/itzultzailea) and compare this output to the original Basque versions. We employ several widely used automatic metrics to measure lexical richness and complement them with manual analysis where relevant.

The remainder of this article is structured as follows: Section 2 describes the characteristics and the preprocessing of the corpus used for the study, Section 3 presents the results of the lexical diversity metrics, and Section 4 offers the final conclusions.

2. SELECTION AND PREPROCESSING OF THE COMPARABLE CORPUS

Our study relies on a specific type of textual data for analysis. While previous research on linguistic diversity in MT has focused on comparing MT with human translations using parallel corpora –source texts aligned with their translations (Baker, 1995)–, this approach is not suitable for our objectives, as it does not allow us to study texts directly produced in the target language. To address this, we use a bilingual comparable corpus, which consists of independently produced texts in two languages that are selected based on shared criteria, such as topic and genre (Bernardini, 2022). This method is crucial as it allows us to (1) analyse original Basque texts within a specific domain and (2) generate Basque MT versions of Spanish texts in the same domain, ensuring a fairer comparison and more accurate results.

We chose to work with the EiTB Corpus (Etchegoyhen & Gete, 2020), a comparable corpus of news between 2009 and 2019 published on the website of Euskal Irrati Telebista (EiTB), the Basque Country's public broadcast service. According to the corpus documentation, the news articles in Spanish and Basque were independently created by journalists yet covered the same events. The articles cover a wide range of topics –politics, sports, culture, international events, economy– providing a rich and diverse vocabulary that is particularly relevant for our study.

The EiTB Corpus was originally designed with MT development in mind, and therefore, the texts were processed to identify and align similar segments in Spanish and Basque based on their content. Segments without clear counterparts were excluded. Specifically, only those that surpassed an alignment threshold of 0.17 in the lexical variant of the STACC metric (Etchegoyhen & Azpeitia, 2016), and a 2.0 threshold in length-based filtering were retained (see Etchegoyhen & Gete, 2020 for further details). As a result, the EiTB Corpus is classified as highly comparable. This is advantageous for our study, as the original Basque texts and their MT counterparts are not only from the same domain but also correspond to the same information at the segment level. Additionally, the large size of the corpus enhances the reliability of our

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

findings. As shown in Table 1, the corpus contains over 600,000 aligned sentences, with more than 10 million tokens in Spanish (ES-O) and over 8.5 million tokens in Basque (EU-O).

Table 1. Quantitative description of the corpus

	Sentences	Tokens	Types
ES-O	637,182	11,690,995	256,547
EU-O	637,182	8,550,695	225,593
EU-MT	637,182	8,080,937	231,534

After selecting the corpus, we obtained the Basque MT version by automatically translating the Spanish sentences (EU-MT). Various publicly available MT systems exist for this language pair, such as Elia, Batua.eus, Itzuli, and more. For this study, the MT output was provided by Elhuyar, the developer of Elia, a neural system based on a seq2seq Transformer architecture (Vaswani et al., 2017). As noted by Shaitarova et al. (2023), publicly available parallel corpora like the EiTB Corpus, are likely to be part of the training data of general MT systems, which could potentially introduce bias. However, Elhuyar confirmed that the EiTB Corpus was not included as part of Elia's training data.

Finally, the corpus was processed to perform analyses that consider linguistic elements beyond surface word forms. To this end, we used the Ixa Kat modular chain (Otegi et al., 2016) and automatically analysed the corpus at different linguistic levels. Concretely, we availed of the tokenization and lemmatization modules, the POS tagger and the morphological analyser[1].

3. ANALYSIS OF THE LEXICAL DIVERSITY

This section presents the methodology and the results of the different analyses employed to assess and compare the lexical diversity of MT and original texts. Our goal is to examine the extent of lexical loss in descriptive texts, specifically online news, and help identify features that could be particularly susceptible to change. By analysing different aspects of the lexicon, we aim to highlight notable differences in language use that may arise from the use of automated translations.

We conduct five main analyses, each addressing a different dimension of lexical diversity. In Section 3.1, we begin by measuring overall lexical richness through several generic metrics. Section 3.2 shifts the focus to POS categories, where we examine

1 Automatic parsing has its limitations, as no parser is 100% accurate. However, Eustagger, the underlying parser of the Ixa Kat, has demonstrated a POS tagging accuracy of 95.17% and 91.89% accuracy when considering all morphological information (Otegi et al., 2016). While some errors may be present, the overall results can still be considered reliable.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

POS-based Type-Token Ratio and general frequency distributions. In Section 3.3, we explore the distribution of lexical items based on their frequency, with particular attention to *hapax legomena*. Finally, Section 3.4 examines the use of connectors, while Section 3.5 evaluates the diversity in synonym usage.

3.1. Generic metrics

Automated general lexical richness metrics can be used to gain an initial overview of the lexical diversity in texts. These metrics allow us to quantify the raw lexical variation and provide insights into the extent of potential lexical loss in the MT output. Following the approach proposed by Vanmassenhove et al. (2019), we employ three general metrics: Type-Token Ratio (TTR), Measure of Textual Lexical Diversity (MTLD), and Yule's I. We also include lexical density as a measure of language complexity. This initial examination lays the groundwork for a more detailed exploration of the specific elements that may be affected by MT in the following sections.

TTR measures the ratio of types (unique words) to tokens (total occurrences of types) in a text (Lu, 2014). In other words, it indicates how frequently a new type appears. It provides a general view of lexical variation, with higher TTR scores indicating less repetition and greater diversity in vocabulary. Yet, a known limitation of TTR is that it tends to decrease as text length increases, since longer texts are more likely to repeat words (Lu, 2014). This could potentially affect our corpus, as some segment sections retain their original order, and therefore discuss the same topic, while other segments are presented in a random sequence and therefore change topics –and vocabulary– at a faster rate. That said, for our purposes, the absolute values of TTR are less meaningful than the relative differences between MT and original texts. Given the highly comparable nature of our data, we believe the impact of TTR's limitation will be minimal. Nevertheless, to address the issue of text length sensitivity, we also include MTLD, which is more robust in this regard and offers complementary insights.

MTLD calculates lexical diversity by dividing the text into «factors» or sequences where the TTR score stabilizes at a given threshold (in our case, 0.72). The mean length of these factors is used to compute the MTLD score, with partial factors included where necessary (Lu, 2014). As with TTR, higher MTLD values indicate greater lexical diversity.

The third metric we use is Yule's I, the inverse of Yule's K (Yule, 1944). Yule's K assesses how constant a text is in terms of vocabulary repetition, and Yule's I, inversely, indicates the richness of the lexicon (Vanmassenhove et al., 2021). Although both Yule's K and I are still affected by text length, they are more resistant to its effects than TTR (Oakes & Ji, 2012). As with the previous metrics, higher Yule's I scores suggest a richer, more varied vocabulary.

Alongside these three metrics, we also compute lexical density, which measures the proportion of content words in a text with regards to the total word count (Lu, 2014). A higher lexical density indicates more condensed information, which increases the

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

cognitive load for readers (Liu & Dou, 2012). Though primarily linked to textual complexity, we include lexical density here as an indicator of content word saturation[2].

Table 2 presents the results for the four metrics[3]. When we look at the overall metrics, we find that MT texts consistently score higher than original Basque texts, both in the tokenized and lemmatized cases. In contrast to previous research results, our data do not provide evidence of systematic lexical diversity loss in MT texts. In fact, the MT system shows a slightly higher diversity than human production across all three metrics.

Table 2. Results on General Diversity Metrics

	Tokenized text		Lemmatized text	
	EU-O	EU-MT	EU-O	EU-MT
TTR ↑	0.04	0.04	0.26	**0.28**
MTLD ↑	938.63	**1016.9**	245.0	**261. 91**
Yule's I ↑	0.56	**0.67**	0.8	**0.10**
Lexical Density ↑	0.39	**0.40**	0.39	**0.40**

While most studies on MT output have reported a tendency for MT systems to limit lexical richness, some prior research has demonstrated that MT systems can match or even surpass human translations in terms of lexical diversity. For example, Shaitarova et al. (2023) note that some commercial MT systems achieve competitive results and may surpass lexical variation in human translations on certain corpora. However, they also caution that as MT technology continues to improve, these traditional lexical diversity metrics may become «less informative» (Vanmassenhove, 2021, p. 22) due to their relatively broad approach to measuring diversity. With these words of caution in mind, the following sections present additional, more targeted experiments to either corroborate or refute these initial findings.

According to the results (see Table 3), in terms of lexical density, both the MT and original texts have identical scores, indicating that the information is similarly condensed.

3.2. POS-based occurrence patterns

The previous global metrics consider words at their surface or form level, and as a result, they might hide different occurrence patterns for specific word classes. This section considers two complementary analyses that dig dipper into the behaviour of

2 Vanmassenhove et al. (2019) calculate lexical diversity metrics on tokenized text in morphologically simpler languages (English, French, and Spanish). However, Basque is an ergative and highly inflected language, where suffixation carries much of the information expressed by prepositions in other languages. To account for this, we calculate metrics on both tokenized and lemmatized texts, preventing inflation caused by combining lemmas with their suffixes. All words are lowercased, and punctuation is excluded for consistency.

3 We provide the results of the metrics in their natural scale. The scales for these metrics are not directly comparable with one another, due to each metric calculating lexical diversity differently and emphasizing different characteristics. Nevertheless, they are meaningful when comparing one text to the other.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

POS categories to expand on the initial findings by exploring ratios of individual word categories and their distributions.

3.2.1. POS-based TTR

To compare POS-level ratios of original and MT text, we use the same TTR formula as in the global metrics but consider the different lexical categories separately. For instance, we calculate TTR for nouns by dividing the number of noun types by the number of total nouns in the text. For each category, this allows us to check how many known items we encounter in the text before a new unseen item appears. First, we calculate TTR for content words and for function words, and then individually for the different specific lexical categories. Concretely, the content word category includes nouns, verbs, adjectives, and adverbs, whereas the function word category includes auxiliary verbs, linking words, determiners, and pronouns.

Table 3. Results on POS-based TTR metrics, with scores multiplied by 100

	Tokenized text		Lemmatized text	
	EU-O	EU-MT	EU-O	EU-MT
Generic TTR ↑	41.18	**45.75**	26.39	**28.67**
Content TTR ↑	45.35	**50.55**	27.63	**30.03**
Function TTR ↑	26.74	**28.31**	24.5	**26.45**
Noun TTR ↑	61.03	**69.57**	38.52	**42.92**
Verb TTR ↑	10.42	**10.5**	**1.42**	1.38
Adjective TTR ↑	120.53	**121.35**	**89.4**	86.0
Adverb TTR ↑	**5.87**	4.56	**4.62**	3.64
Auxiliary TTR ↑	7.11	**7.61**	0.03	0.03
Linking word TTR ↑	0.11	0.11	**0.11**	0.10
Determiner TTR ↑	52.12	**54.44**	49.48	**52.53**
Pronoun TTR ↑	9.45	**9.85**	1.15	**1.18**

With regards to POS-based TTR metrics, the MT text displays higher diversity for the two broadest categories –content TTR and function TTR– (see Table 3). Still, for more specific categories, it is evident that certain trends are distinct. MT exhibits higher diversity in nouns with a noticeable gap compared to the original texts[4]. Though with a smaller margin of difference, MT text also presents higher diversity for determiners. For adverbs, on the other hand, the original text is slightly more diverse, especially if we consider the lemmatized text. Similarly, for adjectives both texts exhibit almost similar levels of diversity on a token level, yet the original has a higher diversity when the text is lemmatized. This means that MT text might be benefiting from a more diverse use of morphology to

4 For noun TTR it should be considered that proper nouns are also included in the calculation of the metric. Likewise, the high values for determiner TTR are explained by the fact that many figures are categorized as numeral determiners.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

gain richness, whereas original text might still be richer in terms of the lexicon used, for these categories at least. The remaining categories score very similarly. The most important takeaway from this analysis is that while overall ratios may appear similar, a closer examination reveals differing trends across specific word categories, albeit minimal.

3.2.2. POS frequency distributions

The metrics we just employed allow us to compare the behaviour of each lexical category in isolation, attending to how frequently new types appear, but they do not provide any information over complementarity such as the prevalence of each of the categories with respect to the rest or their internal composition, that is, more detailed information about the elements that make up each category. Therefore, to complement the first analysis, we shift our focus towards these two aspects. By exploring the distribution of the lexical categories and their internal composition, we aim to identify any potential over- or under-representations of specific lexical categories and subcategories in the original and MT texts, which could reveal specific stylistic and structural choices that could cause not only MT texts to differ from original compositions but also reveal different mechanisms to introduce diversity. For these analyses, we rely on the morphological tags provided by Ixa Kat, which distinguish a total of 13 main word categories.

We start with global POS distributions. Figure 1 presents the distribution of the word categories in both the MT and original texts, with counts displayed as percentages for ease of comparison. The differences between the two texts appear minor, as it can be observed in the bar chart. The overall patterns are strikingly similar, suggesting little divergence in POS distribution.

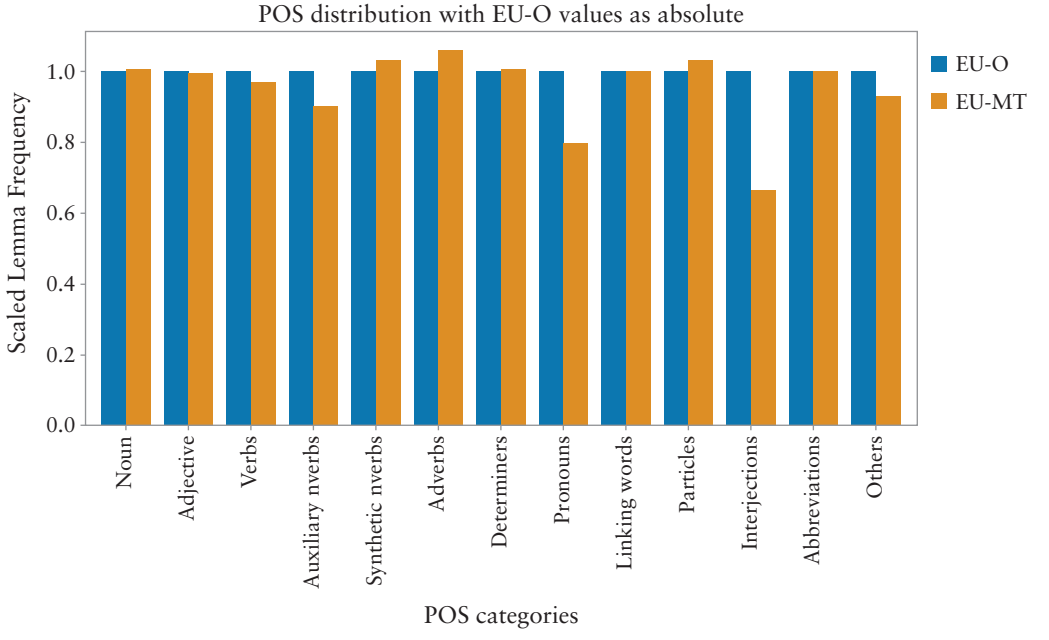


Figure 1. Comparison of POS frequency distribution between EU-O and EU-MT.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

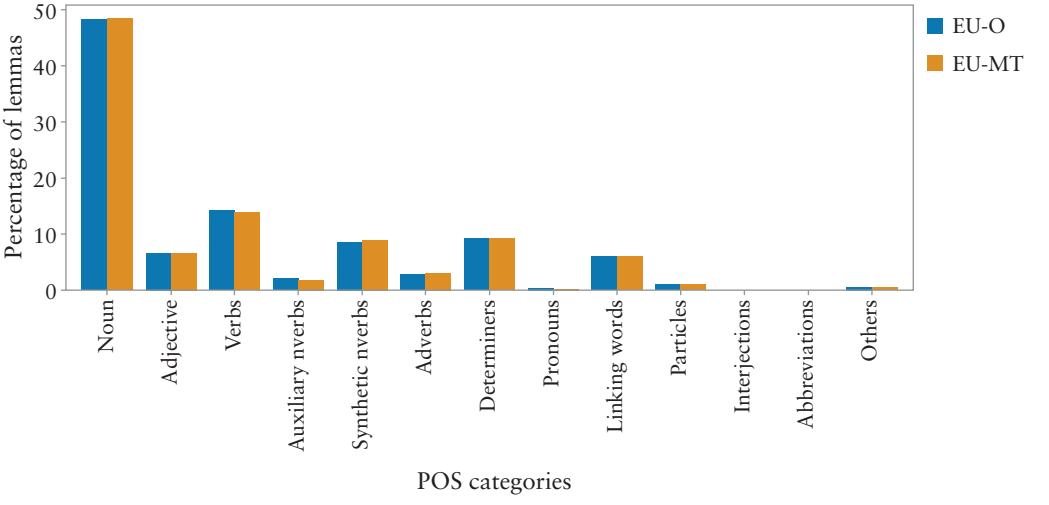


Figure 2. Comparison of the POS frequency distribution using the values of EU-O as reference.

To better show these subtle differences, Figure 2 scales the values of the MT text relative by to the original text, treating the latter as the baseline. This reveals that the most pronounced difference is in the use of interjections, which are less frequent in the MT text. There is also a decrease in the use of pronouns, auxiliary verbs, and other verbs. In contrast, categories such as synthetic verbs, adverbs, and particles, increase slightly. For nouns, adjectives, abbreviations, linking words, and determiners, the differences are negligible. The results unveil that those variations in the use of different lexical categories, while strikingly similar, might be nuanced.

We take one step further to investigate what is happening within the categories by studying the categories in slightly more detail. We aim to see whether a deeper look into the internal composition[5] of the categories can shed light on any preferences for the original or MT texts. We pay particular attention to interjections, pronouns, and auxiliary verbs, given that those are the categories showing more divergence, but we will also consider the remaining briefly (for more details on the results see Appendix A).

Interjections

Interjections are used very rarely in both text types, accounting for less than 0.04% of the total vocabulary in each case. However, we notice a tendency for their use to be slightly more reduced in the MT text. The reason for this decrease remains unclear: does it result from a general reduction across all interjections, or is it due to

5 For the following exploration we study the composition of the different categories attending to the subcategorization provided in Ixa Kat. In cases where there is no possible subcategorization, but the number of lemmas belonging to that category is not too large, we conduct a comparison of the frequency of those lemmas –i.e. in the case of auxiliary verbs, synthetic verbs, particles, and interjections–.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

the underrepresentation of specific interjections? To explore this question, we examine the frequency of individual interjections, leveraging the manageable size of the identified list (59 terms). For each interjection, we analyse its relative frequency per million words.

Table 4 presents the results, sorted in descending order based on the differences of their relative frequencies. Negative values indicate a higher prevalence in the MT text, while positive values reflect a greater frequency in the original text. Overall, we find no evidence that less frequent interjections systematically become rarer in the MT text. Generally, the difference of the relative frequencies is minimal, suggesting comparable usage between the two text types. In some cases –e.g., *horra* ('look'), *bejondeiola* ('God bless him'), *tik-tak* ('tic-tac')–, infrequent interjections disappear entirely from the MT text, but in others –e.g., *hela* ('hey'), *aup* ('hey'), *klik* ('clic')–, they are more prevalent.

No clear pattern emerges in the case of Spanish borrowings either. The MT text uses certain borrowings, such as *jesus* 'Christ' and *amigo* ('friend'), more frequently. It also shows a slight preference for *fuera* ('out') over its Basque equivalent, *kanpora*. However, this does not suggest a general preference for Spanish borrowings in MT, as the use of *bale* ('okay') is reduced, and *beno* ('okay') is preferred over *bueno*.

Table 4. Analysis of interjections: comparison of relative frequencies

Words	Freq. EU-O	Freq. EU-MT	Relative Freq. Difference
no	58.47	43.93	14.54
et	10.99	7.55	3.44
hel	3.04	0.87	2.17
tira	4.68	2.72	1.96
kanpora	3.63	1.98	1.65
ospa	2.46	1.24	1.22
bueno	17.43	16.58	0.85
eskerrak	3.04	2.23	0.81
ut	1.05	0.25	0.80
auskalo	1.29	0.49	0.80
arren	0.70	0.00	0.70
zirt	1.05	0.37	0.68
zart	1.05	0.49	0.56
gabon	1.52	0.99	0.53
ah	0.58	0.25	0.33
ala	0.58	0.25	0.33
bale	9.94	9.65	0.29
bo	1.99	1.73	0.26
bon	8.65	8.41	0.24

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

Words	Freq. EU-O	Freq. EU-MT	Relative Freq. Difference
horra	0.23	0.00	0.23
sast	0.47	0.25	0.22
iso	0.58	0.37	0.21
uf	0.70	0.49	0.21
tik-tak	0.12	0.00	0.12
tori	0.12	0.00	0.12
bejondeiola	0.12	0.00	0.12
urra	0.23	0.12	0.11
danba	0.35	0.25	0.10
atx	0.82	0.74	0.08
biba	1.17	1.11	0.06
bon-bon	0.12	0.12	0.00
blink	0.12	0.12	0.00
ixi	0.12	0.12	0.00
animo	0.12	0.12	0.00
aida	2.46	2.47	-0.01
ai	0.35	0.37	-0.02
eup	0.35	0.37	-0.02
ixo	0.35	0.37	-0.02
arre	0.58	0.62	-0.04
hela	0.00	0.12	-0.12
fini	0.35	0.49	-0.14
fa	0.47	0.62	-0.15
aupa	5.61	5.82	-0.21
asa	0.00	0.25	-0.25
aup	0.00	0.25	-0.25
aitaren	0.12	0.37	-0.25
esti	2.81	3.09	-0.28
amigo	1.75	2.10	-0.35
alde	12.51	12.99	-0.48
fuera	1.87	2.47	-0.60
jesus	3.98	4.58	-0.60
klik	0.00	1.11	-1.11
agur	3.04	4.33	-1.29
kaixo	4.21	5.82	-1.61
a	10.88	13.74	-2.86
ea	9.82	12.75	-2.93
hara	1.75	8.54	-6.79
beno	5.73	15.59	-9.86

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

Pronouns

Pronouns are classified into four categories: personal, indefinite, reciprocal, and others (see Table 5). In the table we include the percentage of use of each category, we also include the absolute counts in parenthesis. As illustrated before, the presence of pronouns is reduced by approximately 20% compared to the original text. Likewise, we also notice a reduction in the diversity of the pronouns employed, since the original text employs 33 different pronouns, and the MT text uses 27. Those 6 different pronouns are *nihaur* ('myself'), *zenbaitzuk* ('some'), *norbaitzuk* ('some [referring to people]'), *nehor* ('anyone'), *berau* ('this one') and *zerori* ('yourself'). In the MT text, a higher proportion of pronouns falls into the indefinite category, whereas the original text employs more personal pronouns. Additionally, we notice a small increase in the proportion of reciprocal pronouns in the MT text.

Table 5. Analysis of subcategories of pronouns. Comparison of relative frequencies with absolute frequencies in parentheses

Category	Examples	EU-O	EU-MT
Personal	*ni*, *zeu*	**63.87** (18,858)	59.88 (13,732)
Indefinite	*norbait*, *nor*	26.67 (7,875)	**30.81** (7,066)
Reciprocal	*elkar*	7.73 (2,283)	**8.04** (1,844)
Others		**1.72** (508)	1.27 (292)

A closer examination of personal pronouns reveals that the proportion of common pronouns is higher in the MT text, while the percentage of reinforced personal pronouns is slightly lower than in the original text; however, this last category remains the least frequent in both cases (see Table 6).

Table 6. Analysis of subcategories of personal pronouns. Comparison of relative frequencies with absolute frequencies in parentheses

Category	Examples	EU-O	EU-MT
Common	*ni*	96.84 (18,263)	**97.89** (13,442)
Reinforced	*zeu*	**3.16** (595)	2.11 (290)

When analysing the subtypes of indefinite pronouns (see Table 7), we observe that the overall proportion of the (non-interrogative) indefinite subcategory is larger than that of interrogative indefinite pronouns. This indicates not only a change in the overall distribution of indefinite pronouns but also a nuanced change in the specific subcategories used.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

Table 7. Analysis of subcategories of indefinite pronouns. Comparison of relative frequencies with absolute frequencies in parentheses

Category	Examples	EU-O	EU-MT
Indefinite	*norbait*	77.26 (6,084)	**85.21** (6,021)
Interrogative	*nor*	**22.74** (1,791)	14.79 (1,045)

Auxiliary and synthetic verbs

Auxiliary verbs are used in combination with main verbs to convey grammatical information, while main verbs primarily provide lexical content. In contrast, synthetic verbs integrate both lexical and grammatical information into a single verbal form.

When comparing POS proportions, we observe that the MT text employs fewer auxiliary verbs but a greater number of synthetic verbs. This suggests that the MT system prefers single-word verbal forms that encapsulate both lexical and grammatical functions.

Table 8 presents the five auxiliary verbs identified in the texts, along with their frequencies per million words ordered by the difference of their relative frequencies. The most notable variation is in the use of **edun*, which is more frequent in the original text[6]. Conversely, auxiliary verbs such as **edin*, **ezan*, and *izan* appear more frequently in the MT text than in the original. When it comes to **iro* a closer inspection reveals that the apparent presence of this historical verb is a mistake, and that it is due to the tagger incorrectly identifying the lemma on certain words that match the conjugate form of the verb.

Table 8. Analysis of auxiliary verbs. Comparison of relative frequencies

Words	Freq. EU-O	Freq. EU-MT	Relative Freq. Difference
**edun*	14,433.45	12,078.18	2,355.27
**iro*	5.61	5.44	0.17
izan	5,349.97	5,364.60	-14.63
**edin*	546.15	655.74	-109.59
**ezan*	1,039.80	1,153.58	-113.78

6 The asterisk in the verbs **edun*, **edin*, **iro* and **ezan* indicates that this infinitive form was reconstructed and is only used for theoretical reference. These infinitive forms are not used in texts but rather their corresponding conjugated ones. Forms like *dut*, *dituzu* come from **edun*; *ezazu* and *dezakezu* from **ezan*; and *nadin* and *zintekeen* from **edin*. *Izan* is the auxiliary for intransitive verbs in the indicative mood, whereas **edun* is used for transitive verbs. **Edin* and **ezan* in the other hand are used for subjunctive, potential and imperative moods, but **edin* is used for intransitive forms and **ezan* for transitive ones. In the case of **iro* it should be noted that this verb belongs to the Eastern dialects, it is a historical form very rarely used in modern Basque.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

For synthetic verbs, the data indicates that their overall frequency is similar in both texts (see Table 9). However, there are four synthetic verbs with substantial differences in usage. The verb *eduki* ('to have') is markedly less frequent in the MT text, whereas *izan* ('to be'), *esan* ('to say'), and *ukan* ('to have') are notably more prevalent. The increased use of these latter three verbs accounts for the higher representation of synthetic verbs in the MT text compared to the original text.

Table 9. Analysis of synthetic verbs. Comparison of relative frequencies

Words	Freq. EU-O	Freq. EU-MT	Relative Freq. Difference
eduki	424.29	51.48	372.81
eraman	238.34	202.33	36.01
egin	323.48	308.38	15.10
egon	5,568.79	5,554.43	14.36
jardun	31.93	19.68	12.25
ezagutu	18.13	6.93	11.20
eman	110.75	99.74	11.01
ibili	121.04	110.75	10.29
iritzi	15.91	7.92	7.99
erabili	7.95	2.10	5.85
jario	12.05	8.17	3.88
etorri	703.57	700.29	3.28
erran	3.74	2.60	1.14
utzi	0.82	0.37	0.45
etzan	12.63	12.25	0.38
ikusi	0.35	0	0.35
aditu	1.05	0.99	0.06
atxiki	0.12	0.12	0.00
ekarri	74.15	75.61	-1.46
joan	313.31	315.8	-2.49
entzun	14.50	18.69	-4.19
iraun	10.41	15.47	-5.06
eroan	3.27	9.16	-5.89
**edin*	468.50	477.79	-9.29
ekin	27.13	43.19	-16.06
jakin	340.91	381.39	-40.48
iruditu	129.70	175.85	-46.15
ukan	44,815.07	44,974.49	-159.42
esan	1,467.60	1,944.95	-477.35
izan	30,655.87	33,184.27	-2,528.40

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

These differences in the use of synthetic and auxiliary verbs suggest a potential shift in verb valency, which may indicate broader syntactic differences between the texts. However, as this study focuses on lexical aspects of language, further investigation into these syntactic variations is left for future research.

Remaining categories

We summarise the behaviour of content categories first. For nouns, a closer examination of common nouns, proper nouns referring to people and places, and numbers revealed marginal differences, which, although observable, do not seem substantial (for more detailed information on the results of this subsection see Appendix A). The same holds true for adjectives and adverbs after considering common and interrogative cases. Verbs are categorized into simple verbs, compound verbs, and factive verbs. While the differences of the relative frequencies among these subcategories are minimal, the MT text appears to use factive verbs slightly more frequently, with a corresponding slight decrease in the use of compound verbs and simple verbs. In the case of abbreviations, while both texts use them similarly in proportion, some specific abbreviations become more frequent –e.g., *mug* ('limit') and *pta* ('peseta')– in MT, whereas others become rarer –e.g., *esk* ('right') and *vs* ('versus')–.

Let us consider functional categories next, where we observe peculiar nuances as we analyse more precise categories and lexical items. Determiners display a nuanced situation. They are categorized into four groups: demonstratives, qualifying determiners, numerals, and others. While most subcategories show similar proportions in both texts, numerals and demonstratives exhibit slight shifts, with an increase observed in the MT text. Within demonstratives, a distinction is made between common forms –e.g., *hori* ('that')– and reinforced forms –e.g., *berori* ('that'), *hauxe* ('this')–, with reinforced demonstratives being morphologically marked through prefixes –e.g., *-ber–* or suffixes –e.g., *-xe–* to draw additional attention. Surprisingly, reinforced demonstratives appear more frequently in the MT text, both in proportion and absolute counts, despite being less common and more specific. For qualifying determiners, common forms such as *edozein* ('any') increase slightly in the MT text, while interrogative forms like *zein* ('which') decrease. In the category of numerals, which includes definite *–bi* ('two'), *hiruna* ('three each')–, indefinite *–zenbait* ('some')–, and generalized *–guzti* ('all')– forms, we observe a small increase in the use of indefinite forms and a slight decrease in the use of indefinite forms. Generalized forms, in turn, show similar proportions.

Linking words, or discourse connectors, in turn, are less frequent in the MT text, likely due to the inherent difficulty MT systems face in handling such elements. Connectors explicitly convey relationships between ideas and heavily rely on context, both preceding and following, which MT systems often fail to account for when translating on a sentence-by-sentence basis. Further analysis of connectors is provided in Section 3.5.

Finally, we look at the case of particles. Nine lexical items were identified as particles: *ez* ('no'), *bai* ('yes'), *ahal* ('can'), *bide* ('apparently'), *ba* (this interrogative particle has no equivalent in English), *ei* ('reportedly'), *ote* ('maybe'), *al* (this interrogative particle

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

has no equivalent in English), and *omen* ('allegedly'). While particles are slightly less frequent in the MT text, all particles from the original text are present, indicating their inclusion in the MT system's vocabulary, even if some are used less often. The data suggests a shift in the modulation of certitude, with particles such as *ei*, *ote*, and *omen*, which express uncertainty, appearing less frequently in the MT text. Conversely, the affirmative and negative particles *bai* and *ez* are more frequent in the MT text, pointing to a tendency toward greater explicitation of sentence polarity.

In conclusion, much like the findings from the previous POS-based ratio analysis, this study highlights that while overall POS distributions between MT and original texts are remarkably similar, a closer examination uncovers subtle but meaningful differences in usage trends.

3.3. Frequency bands and *hapax legomena*

The results from the previous section do not point to a notable reduction in lexical diversity within MT texts; rather, they suggest that MT outputs may reach levels of lexical richness comparable to those of original human-produced texts, with only minor deviations. However, while these findings indicate similar richness, they do not reveal whether the specific words contributing to this diversity are consistent across both text types. In other words, it remains unclear whether the MT and original texts use the same vocabulary and if this vocabulary is distributed similarly across both. To address this, the following section examines vocabulary overlap and its distribution across frequency bands to determine if words appear with comparable frequencies in each text.

For this analysis, we focus solely on lemmatized data to assess lexical distribution independently of morphological inflection. Lemmas are organized into six frequency bands (1; 2-10; 11-100; 101-1,000; 1,001-5,000; 5,001-10,000; +10,000) to illustrate how words are spread across the texts. This categorization allows us to compare the size of each band and to observe whether vocabulary remains consistent between the original and MT texts. We report both the total tokens and types within each band. Table 10 presents these lexical frequency distributions, expressed as percentages for clarity, with token counts in parentheses.

Table 10. Frequency band distribution of lemmas. Comparison of relative frequencies with absolute frequencies in parentheses

Frequency		1	2-10	11-100	101-1,000	1,001-5,000	5,001-10,000	10,000+	Total
EU-O	Tokens	1.52 (129,965)	2.99 (255,959)	7.25 (619,843)	19.00 (1,624,276)	**24.42** (2,088,092)	**10.89** (931,479)	**33.93** (2,901,081)	100 (8,550,695)
	Types	57.6 (129,965)	**30.71** (69,290)	**8.78** (19,818)	**2.39** (5,389)	**0.43** (971)	**0.06** (137)	0.03 (77)	100 (225,593)
EU-MT	Tokens	**1.67** (135,259)	**3.19** (257,635)	**7.60** (613,793)	**19.64** (1,586,908)	24,32 (1,965,233)	10.44 (843,283)	33.15 (2,678,826)	100 (8,080,937)
	Types	**58.42** (135,259)	30.32 (70,211)	8.49 (19,646)	2.29 (5,304)	0.40 (923)	0.05 (122)	0.03 (69)	100 (231,534)

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

Overall, even when the MT text is 5.5% shorter (469,758 words shorter), it displays 5,941 more types than the original text. The texts are not fully comparable, as the MT text is a translation of a Spanish text whose specific level of richness might affect results, but the difference is notable.

If we focus on the frequency distribution, the results suggest that both texts organize vocabulary similarly across bands. In both, approximately one-third of the total lemmas fall within the highest frequency category, appearing over 10,000 times. This involves 77 types for original texts and 69 for MT. In fact, overall, the types are slightly higher for the original text than for MT except for those with a frequency range of 1-10, the two lowest frequency bands. Therefore, the contribution of the diversity of the MT text may emerge from *hapax legomena*.

Thus far, we have analysed the distribution of lemma types and tokens. To refine our comparison of vocabulary use, we next examine the overlap of specific lemma types within each frequency band. This analysis enables us to assess the extent to which lemmas within each band are consistent across the texts. Additionally, we identify «mismatched» lemmas –those that appear in both texts but fall into different frequency bands–.

Results indicate that vocabulary alignment increases with higher frequency bands (see Table 11). Even in the lower bands, however, there is substantial overlap, with over half of the *hapax legomena* shared between the two texts. Around one-quarter of shared words appear in different frequency bands, indicating variation in repetition between the two texts.

Table 11. Comparison of lemmatized vocabulary across frequency bands. Comparison of relative frequencies with absolute frequencies in parentheses

		1	1-10	10-100	100-1,000	1,000-5,000	5,000-10,000	10,000+
EU-O	Common Types	28.41 (64,117)	19.98 (45,083)	6.85 (15,459)	2.02 (4,564)	0.35 (798)	0.04 (100)	0.03 (65)
EU-MT	Common Types	27.67 (64,117)	19.46 (45,083)	6.67 (15,459)	1.97 (4,564)	0.34 (798)	0.04 (100)	0.03 (65)
Mismatched Types		13.78 (48,540)						

Although vocabulary overlap is high across all frequency bands, approximately 34% of shared types appear in different bands across text types. The frequency bands are arbitrarily defined and minor differences in usage can result in a word being placed in adjacent bands. These mismatches are not necessarily significant to language use, but larger shifts may suggest a tendency to overuse or underuse specific words.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

As shown in Table 12, nearly all mismatches (96.90%) occur between adjacent bands, while only 2.95% of types shift two bands, and an even smaller percentage shifts three or four bands. No lemmas shift more than four bands. Most frequency shifts occur within the low repetition bands, primarily between the 1 and 2-10 range, and 2-10 and 11-100 range. Notably, words tend to shift toward lower frequency bands, indicating that words more frequently used in the original text appear less often in the MT text.

Table 12. Shifts between frequency bands

	Number of shifted frequency bands					
	1	2	3	4	5	6
Types (%)	96.90	2.95	0.09	0.01	0.00	0.00
Types occurrences	22,727	692	22	2	0	0

In Table 13 we collect the examples that shifted four frequency bands, and in Table 14 we do the same for those that shifted three frequency bands. The two examples in Table 13 correspond to the words *ETBat* and to the declination of *PP* using a hyphen. In the latter case, the issue relies on the incorrect identification of the lemma by our tagger, which also occurs in some examples in Table 14, such as *ura* or *etb-1ean*. In the former, however, the shift is caused by different orthographic choices, where the original text prefers using forms like *EiTBat* over *ETBat*, another example of this is *stee-eilas* in Table 14, for which the MT employs predominantly *Steilas*.

Nevertheless, the most interesting examples, linguistically speaking, are found in Table 14, where we find differences in the use of lemmas that are not proper nouns. For example, we find that the MT text displays terms like *gaztelera* ('Spanish language'), *negoziaketa* ('negociation'), *jazo* ('to happen'), *galdegin* ('to ask'), *ingurugiro* ('entorno'), *helmugaratu* ('reach the finish line'), *eskuduntza* ('competence') and *abagune* ('chance') less frequently. Surprisingly, in a few examples the opposite happens, such as in *teilatu-hegal* ('eaves') or *iruzkingile* ('commentator'). These differences could be due to the occurrence frequency in the training data leading to different lexical preferences between the MT system and the journalists.

Table 13. Examples of lemmas that shifted 4 frequency bands. Comparison of relative frequencies with absolute frequencies in parentheses

Words	Freq. EU-O	Freq. EU-MT	Relative Freq. Difference
etba	0,12 (1)	362.83 (2,932)	-362.72
pp-r	0,12 (1)	132.16 (1,068)	-132.04

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

Table 14. Examples of lemmas that shifted 3 frequency bands. Comparison of relative frequencies with absolute frequencies in parentheses

Words	Freq. EU-O	Freq. EU-MT	Relative Freq. Difference
oste	607.52 (5,168)	2.47 (20)	605.05
d.b.	534.64 (4,548)	0.25 (2)	534.39
jazo	172.69 (1,469)	0.49 (4)	172.19
abagune	40.67 (346)	0.12 (1)	40.55
ura	29.62 (252)	0.12 (1)	29.50
salneurri	29.62 (252)	0.12 (1)	29.50
negoziaketa	28.57 (243)	0.12 (1)	28.44
konfirmatu	26.57 (226)	0.12 (1)	26.44
gaztelera	25.63 (218)	0.12 (1)	25.50
helmugaratu	23.04 (196)	0.12 (1)	22.92
galdegin	21.87 (186)	0.12 (1)	21.74
ingurugiro	21.28 (181)	0.12 (1)	21.15
etbkn	15.40 (131)	0.12 (1)	15.28
eskuduntza	14.69 (125)	0.12 (1)	14.57
stee-eilas	12.93 (110)	0.12 (1)	12.81
manifestaldi	11.99 (102)	0.12 (1)	11.87
teilatu-hegal	0.12 (1)	12.50 (101)	-12.38
mutiloako	0.12 (1)	12.62 (102)	-12.50
4h	0.12 (1)	15.72 (127)	-15.60
camuna	0.12 (1)	16.71 (135)	-16.59
iruzkingile	0.12 (1)	38.11 (308)	-38.00

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

3.3.1. Hapax legomena

Infrequent or rare words increase a text's lexical diversity score. Our earlier analysis on frequency bands revealed a slightly higher proportion of such rare words in the MT text. However, «noisy» elements –such as meaningless character strings, typographical errors, or invented words, issues commonly associated with MT systems (Macken et al., 2019)– also tend to fall within lower frequency bands due to their lack of repetition, which can artificially elevate diversity scores. Unlike true rare words, these noisy elements degrade overall text quality. This section focuses on analysing the lexicon within the *hapax legomena* category to check whether the higher lexical diversity score of MT output could be hiding some level of excess noise rather than be the result of linguistic richness proper.

We conduct an automated categorization of the lexicon within this frequency band, dividing the lemmatized text into the following groups and calculating the percentage of *hapax legomena* belonging to each group:

1. Proper nouns: lemmas tagged as proper nouns.

2. Other alphabetic lemmas: alphabetic lemmas, excluding proper nouns, including hyphenated compounds.

3. Numbers: lemmas comprised solely of numeric characters.

4. Other non-alphabetic lemmas: lemmas composed of either mixed alphabetic and non-alphabetic characters or non-numeric, non-alphabetic characters.

This categorization allows us to detect discrepancies in noise levels in the corpus and to identify potential rare words within the alphabetic lemma category.

The initial classification of *hapax legomena* (see Table 15) shows a comparable distribution in both texts, though the MT text has a slightly higher proportion of alphabetic lemmas that are not proper nouns. This could imply either a greater level of noise introduced by MT or genuinely higher lexical diversity due to additional rare words. To clarify this, we conduct an in-depth analysis of this category.

Table 15. Automatic categorization of *hapax legomena*. Comparison of *hapax legomena* percentages by category, with absolute frequencies shown in parentheses

	EU-O	EU-MT
Proper nouns	**29.09** (37,472)	28.29 (37,919)
Other alphabetic lemmas	44.82 (58,271)	**46.01** (62,290)
Numbers	0.40 (518)	0.40 (524)
Non-alphanumeric lemmas	**25.68** (33,387)	25.31 (34,265)

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

We begin by examining lexicon overlap, identifying 16,326 types shared between both texts, representing approximately a quarter of the alphabetic *hapax legomena* (excluding proper nouns). A closer inspection shows a higher proportion of hyphenated compounds in the MT text; while about 12% of *hapax legomena* in the original text are hyphenated, this rises to 20.7% in the MT text.

To further examine the alphabetic lemmas in the *hapax legomena*, we manually categorized the lemmas into eight subcategories, namely, spelling errors, proper nouns, tagging errors, rare words, foreign words, adapted foreign words, invented words and noise. Table 16 outlines the tags, definitions, and examples used in this classification.

Table 16. Classification of *hapax legomena*

Tag	Definition	Example
Spelling Error (SE)	It is an actual word, but it is not correctly written or does not conform to the standard spelling.	jokaladian
Proper Noun (PN)	A proper noun that has been incorrectly classified into some other word category.	benneti
Tagging Error (TE)	The word exists in the Euskaltzaindia Dictionary but is incorrectly lemmatized.	herbeherarrekin
Rare Word (RW)	It is a properly spelt word, without lemmatization issues. It has an entry in the Euskaltzaindia Dictionary, or it has been derived from another word following the rules for the standard language.	abertzaletu, birkualifikatu
Foreign Word (FW)	A word that is borrowed without any orthographic adaptation.	ultratrike
Adapted Foreign Word (AFW)	A word that is borrowed but with orthographic adaptation.	txapuzoi
Invented (I)	A word void of meaning in Basque, but that has been used as an actual word within a sentence.	hokagarriak
Noise (N)	A string of letters that has no meaning and does not even look like a possible lemma.	bvl

A random sample of 300 lemmas was annotated by hand following a structured process to ensure accuracy and consistency. Lemmas were first checked in context to determine if they were proper nouns. If not, we verified dictionary entries and standardized derivation rules from the Basque Academy; if a word lacked these criteria and was neither a foreign term nor an adapted foreign word, we classified it as an invented word. Table 17 presents the annotation results with percentages and counts for comparison.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

Table 17. Results of manual categorization of *hapax legomena*. Comparison by category of alphabetic *hapax legomena* percentages, with absolute frequencies shown in parentheses

	EU-O	EU-MT
Tagging Errors (TE)	**27.00** (81)	22.33 (67)
Foreign Word (FW)	5.00 (15)	**6.00** (18)
Foreign Word Adapted (FWA)	2.00 (6)	**2.33** (7)
Proper Nouns (PN)	33.00 (99)	**36.33** (109)
Orthographic Errors (OE)	**13.67** (81)	1.67 (67)
Noise (N)	**1.33** (4)	0 (0)
Rare word (RW)	18.00 (54)	**26.33** (79)
Invented (I)	0 (0)	**5.00** (15)

First, we see that in both cases the noise introduced by the lemmatizer is considerable, around a quarter of lemmas analysed fall in the category Tagging Errors in both texts –e.g. *Obradovicen*– was not properly lemmatized, as this proper name should have been divided into *Obradovic* and possessive particle *-en*. These results highlight some of the limitations of automatic tagging in general, and the tagger used in particular. In the case of foreign words, either with or without orthographic adaptation –e.g.: *powerful*, *mueka* ('gesture')–, we see that the percentages are quite close; something similar happens with proper nouns, where we do not find considerable differences.

When examining orthographic errors, we find a higher proportion in the original text, which aligns with the nature of human-generated content, where such mistakes are more typical. Similarly, the original Basque text exhibits slightly more overall noise, whereas the MT text sample is notably free of such noise, likely due to the MT system eliminating unprocessable elements like hyperlinks.

When it comes to actual rare words, in both cases a large part of these infrequent terms are compound words; however, this proportion is even more pronounced in the case of the MT texts. This trend helps explain why the overall proportion of rare words is larger in the MT text.

Finally, the category of invented words is fully reserved for the MT text, as we capture no examples in the original text. The presence of these new words merits attention. Among the 15 examples analysed, we identified several processes by which MT generates novel words. These include forming lemmas by attaching Basque suffixes to incorrect stems –e.g.: *retwitteado* ('retwitted') translated into *triptatua*–, deriving non-existent words from incorrect target lemmas –e.g.: *embarcar* ('to board') translated into *sakanatzea*–, creating compounds using words from both source and target languages –e.g.: *pistoletazo de salida* ('starting gun') translated into *irteera-pistoladoa*–, or translating Spanish phrases directly but producing words with altered meanings –e.g.: *lograron unas canastas vitales* 'they scored some vital baskets'] translated into *bizi-saskiak*, where *bizi-* can mean 'vital' but not in the sense of 'essential'–.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

Additionally, many hyphenated compounds in the MT text appear to reflect the Spanish noun + preposition *de* + noun structure. For example, *decisiones de consumo* ('comsumption decisions') is rendered as *kontsumo-erabakiak*, and *supervisor de control* ('control supervisor') is rendered as *kontrol-ikuskatzaile*. While this study does not focus on such structural issues, investigating whether MT systems tend to translate this Spanish structure into Basque with hyphenated compounds, or if other structural variations are employed, could provide valuable insights into system preferences.

In summary, the analysis of lexical occurrence frequencies and their distribution across frequency bands reveals strikingly similar patterns in both the original and MT texts. We have established that occurrence frequencies do not differ significantly between the two text types. However, a more nuanced examination of subtle differences uncovers meaningful variations.

The findings suggest that the diversity in MT output might primarily result from a greater use of less frequently repeated lexical items. Manual analysis of *hapax legomena* shows that rare words are used slightly more often in the MT text, although these include a notable proportion of (incorrectly) coined terms. In contrast, the original text displays a higher frequency of spelling errors. Additionally, the MT text is characterized by a greater prevalence of hyphenated words.

3.4. Discourse connectors

The previous sections primarily examined global trends or broad lexical categories in the original and MT texts. Studying the behaviour of each lexical unit would be an extensive and labour-intensive task, but we made an initial attempt by selecting a specific subset of words to assess whether any differences could be discerned at a more detailed level. In this section, we focus on the use of discourse markers –a closed class of words essential for maintaining coherence and supporting comprehension by clarifying the hierarchical structure and logical relationships between ideas. This focus is particularly relevant in the context of MT, as systems frequently struggle with accurately translating these elements, often resulting in mistranslations or omissions (Sim Smith, 2017).

For our analysis, we use the list of Basque connectors in *Multiaztertest*, an open-source tool designed for stylistic analysis and text classification based on reading difficulty (Bengoetxea & González-Dios, 2021). It covers connectors from five categories: 40 causal connectors, 39 logical, 32 adversative, 66 temporal, and 4 conditional connectors (https://github.com/kepaxabier/MultiAzterTest/blob/master/data/eu/Connectives/connectives.txt). Using this list, we extracted the frequencies of the connectors in both the original and MT texts.

In Table 18 we indicate the number of searched items that were found in each of the texts. In all categories, except for conditional connectors, the original text exhibits higher matching rates. This indicates that some connectors that appear in the original

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

text are missing in the MT text, meaning that the variety of connectors in the MT texts is more limited.

Table 18. Count of matches of searched terms

	Causal	Logical	Adversative	Temporal	Conditional	Total
Searched terms	40	39	32	66	4	181
Matched terms EU-O	**22**	**27**	**25**	**54**	4	**132**
Matched terms EU-MT	17	22	19	49	4	111

After comparing the proportion of matched types, we focus on their occurrences. We report these results in Table 19. For better comparison, we present the occurrence of the difference discourse connectors per 1,000 words, with absolute counts provided in parentheses.

According to the results, the overall use of connectors is similar in proportion across both texts. However, the frequency of different types of connectors shows distinct trends. Out of the five possible connector types, three –namely, causals, adversatives, and conditionals– are used less frequently in the MT text. On the other hand, temporal connectors and logical connectors show a different trend, where the MT text shows a higher frequency compared to the original text.

Table 19. Frequency of discourse connectors by category. Comparison of relative frequencies with absolute frequencies in parentheses

	Causal	Logical	Adversative	Temporal	Conditional	Total
EU-O Freq.	**3.72** (31,595)	54.22 (460,267)	**9.14** (77,567)	8.68 (73,660)	**0.61** (5,152)	76.64 (648,241)
EU-MT Freq.	3.34 (26,967)	**54.56** (440,157)	8.56 (69,022)	**10.45** (84,318)	0.39 (3,116)	**77.16** (623,580)

Given these results, the remaining question is how these changes manifest in the use of individual connectors. To this end, we compare the frequencies for each individual connector (see the detailed results in Appendix B).

The results reveal a gradual decrease in most connectors, which explain the decreases observed in Table 19. Still, we do not observe any dramatic drops in the use of specific connectors. Additionally, this downward trend leads to the disappearance of some of the less frequent discourse connectors. On average, the 17 connectors missing from the MT output (see Table 20) have a frequency of 0.05 occurrences per 1,000 words in the original text.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

Table 20. Discourse connectors that disappear in the MT-text. Comparison of relative frequencies

Discourse connector	Freq. EU-O
Honenbestez	0.0001
Esanak esan	0.0003
Hau guztia dela eta	0.0001
Hori dela medio	0.0008
Horren kariaz	0.0004
Ez eze	0.0001
Osterantzean	0.0005
Berebat	0.0040
Edota	0.2100
Hau egin ondoren	0.0001
Lehenengo eta behin	0.0022
Honen ostean	0.0050
Horratik	0.0002
Hargatik	0.0010
Barren	0.0040
ezezik	0.0009
ostera	0.6200

Despite the general trend, there are several instances where the opposite effect occurs, namely, where the MT system increases the frequency of a discourse connector. In most cases, this increase is very subtle, like in the case of several logical connectors –e.g.: *edo* ('or'), *eta* ('and'), *gainera* ('furthermore')– that help explain the slight increase in the use of this type of connectors. However, there are two notable exceptions where this increase is more pronounced: *lehen* ('before'), and *ondoren* ('after'). For these, their frequency in the MT text nearly doubles that in the original. This substantial rise in their use is the primary reason for the higher occurrence of temporal connectors and overall increase in the occurrence of discourse connectors Table 19). *Ondoren* alone accounts for 23.8% of the total count of the MT text, whereas *lehen* accounts for 22.8%. It is also worth noting that connectors synonymous with *ondoren* do not exhibit a similarly dramatic decrease.

In summary, the analysis highlights several differences in the use of discourse connectors between MT and original text. The MT tends to reduce the frequency of discourse markers, especially those with very low occurrence rates, which in some instances results in their complete omission. However, we also observe that certain discourse connectors, particularly *ondoren*, become overrepresented in MT text.

3.5. Evaluating diversity in synonym use

The previous sections have explored various dimensions of lexical diversity; however, they have not addressed semantic relationships between words. The use of synonyms

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

is a key mechanism for introducing lexical richness into a text. As such, this section focuses on analysing synonym usage by evaluating frequency distributions within synonym clusters.

Vanmassenhove et al. (2021) investigated the ability of MT systems to employ synonyms. Their approach involved extracting all nouns, verbs, and adjectives from the source text and using a bilingual dictionary to identify all possible translations in the target language. For each synonym within a cluster, they calculated its frequency in both the training data and the MT output. To quantify the diversity within these clusters, they introduced three metrics: Primary Translation Frequency (PTF), Cosine Distance of Uniform Synonym Distributions (CDU), and Synonym Type-Token Ratio (SynTTR).

Building on their methodology, we also evaluate synonym frequency using these three metrics, but with modifications to the way synonym clusters are constructed. Instead of starting with Spanish source data, we directly build synonym clusters from the two Basque texts using Basque WordNet (Pociello et al., 2011) (http://hdl.handle.net/10230/22925). Another key difference is that while Vanmassenhove et al. (2021) included all possible translations regardless of differences in word class or sense, we restrict our clusters to monosemic words within the same word class.

These adjustments are necessary because our corpus differs fundamentally from that of Vanmassenhove et al. (2021). Unlike their parallel corpus, which links machine translations to specific source inputs, in our comparable corpus only MT texts are strongly linked to source, whereas the original text does not originate from it. Consequently, we cannot trace synonym variations back to specific source-language inputs. To ensure meaningful comparisons, we prioritize synonyms that are interchangeable and likely to appear in similar contexts, even if this narrower approach excludes polysemic words and cross-category synonyms. Table 21 contrasts examples of synonym clusters derived using the method of Vanmassenhove et al. (2021) with those created under our revised methodology.

Table 21. Comparison of synonym clustering methods

Vanmassenhove et al. (2021)	Word: *look* Synonym cluster: {*mirar*, *esperar*, *buscar*, *parecer*, *dar*, *vistazo*, *aspecto*, *ojeada*, *mirada*}
Our method	Word: *desberdindu* Synonym cluster: {*desberdindu*, *ezberdindu*, *bereizi*}

Using this approach, we created 1,462 noun clusters and 184 verb clusters. Initially, we planned to analyse adjective clusters as well, but due to the limited number of adjective entries in WordNet and our strict clustering criteria, we were able to create only nine clusters. As such, adjectives were excluded from this analysis.

After constructing the clusters and obtaining frequency vectors, we evaluated them using PTF, CDU, and SynTTR. PTF measures the dominance of a primary translation relative to alternative synonyms (Vanmassenhove et al., 2021). A lower PTF score

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

indicates less dominance of the primary translation, reflecting greater lexical diversity. CDU assesses the uniformity of synonym distribution by calculating the distance between the actual distribution and a perfectly uniform distribution (Vanmassenhove et al., 2021). Smaller CDU scores suggest more balanced synonym usage. SynTTR is an adapted version of TTR metric (Vanmassenhove et al., 2021). In this case, a type is each of the unique synonyms, while tokens are their frequencies. The metric is useful to identify words for which the MT system omits the use of alternatives. As in the regular TTR, higher scores indicate a greater diversity.

These metrics were computed individually for each synonym cluster, and aggregate scores were obtained by averaging results across all clusters. Table 22 presents the final scores for noun and verb clusters. The results for these metrics are bound between 0 and 1.

Table 22. Results on synonym diversity metrics

		EU-O	EU-MT
Nouns	PTF ↓	**0.82**	0.84
	CDU ↓	**0.23**	0.24
	SynTTR ↑	**0.12**	0.10
Verbs	PTF ↓	**0.78**	0.80
	CDU ↓	**0.27**	0.28
	SynTTR ↑	**0.49**	0.38

The results indicate that original texts exhibit slightly higher diversity across all three metrics for both noun and verb clusters, though the differences are modest. It is worth noting that our stringent clustering criteria, focusing exclusively on monosemic words, capture only a fraction of synonymy. Broader results might emerge if polysemic words were included.

We observed that the most prevalent synonym choice in MT text is aligned with that of the original text. Specifically, the primary choice matched in 85% of noun clusters and 85.3% of verb clusters. Additionally, the MT system tends to amplify the frequency of the primary choice: in 85.3% of noun clusters and 83.6% of verb clusters where the primary choice aligned with the human preference, the MT system's frequency was 3.76% higher. This supports the notion that MT systems not only favour the primary choices but also exacerbate their dominance.

To delve deeper, we examined synonym clusters with a total frequency of at least 50 occurrences for generalizability. Table 23 provides examples of noun and verb clusters, listing the synonyms in the leftmost column and their frequency distributions in the adjacent columns. The analysis reveals that the MT text exhibits biases not only in its treatment of rare or domain-specific vocabulary but also in its handling of everyday words. For example, we notice in the case of *guraize* and *artazia*, which are two equivalent ways of saying 'scissors' in Basque, the MT system clearly prefers the former over the later.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

In summary, although limited in scope, this analysis of synonym distribution once again demonstrates that considerable differences do not emerge when comparing original text to MT text. However, contrary to other cases, the subtle variations observed tend to favour the original text, which exhibits slightly greater diversity through a more even distribution of synonyms compared to the MT text.

Table 23. Examples of synonym clusters and the corresponding absolute frequencies

		Freq. Distribution EU-O	Freq. Distribution EU-MT
NOUNS	[*guraize*, *artazi*]	[19 25]	[46 1]
	[*margo*, *pintura*]	[81 135]	[2 279]
	[*galtza*, *praka*]	[38 32]	[60 3]
	[*margolan*, *pintura*, *koadro*]	[108 135 94]	[2 279 191]
	[*okela*, *haragi*]	[43 120]	[6 162]
	[*gaztelera*, *gaztelania*, *espainol*, *espainiera*]	[218 550 8 13]	[1 751 3 6]
VERBS	[*muturtu*, *asaldatu*, *sumindu*, *haserretu*]	[2 42 94 79]	[0 15 21 168]
	[*espetxeratu*, *kartzelatu*]	[240 114]	[140 0]
	[*desberdindu*, *bereizi*, *ezberdindu*]	[30 281 69]	[21 574 2]
	[*ondoratu*, *alderatu*, *gerturatu*, *alboratu*, *hurbildu*, *inguratu*, *hurreratu*]	[4 1,168 493 48 1,238 240 20]	[0 838 14 12 1,466 294 0]
	[*prebenitu*, *ekidin*, *saihestu*, *eragotzi*]	[54 543 1,069 307]	[103 33 1,395 608]

4. CONCLUSIONS

This study provides a comprehensive analysis of the linguistic traits of machine-translated (MT) texts compared to original texts in Basque, focusing on lexical diversity in online news. Through an examination of different linguistic dimensions –ranging from generic metrics of lexical richness to POS analysis, frequency distributions, discourse connector usage, and semantic synonym diversity– we have sought to identify patterns in MT output. Contrary to most previous research, our findings reveal a striking degree of similarity between original and MT texts, with only minor differences suggesting a slight advantage for original text or MT text in specific aspects of lexical richness.

One of the most surprising results of this study is the lack of substantial lexical loss in the MT texts. Prior research (Vanmassenhove et al., 2019, 2020) has reported a tendency for MT systems to produce less diverse output. However, as Shaitarova et al. (2023) find, certain MT systems may be able to obtain diversity levels comparable to that of human translators in certain corpora, particularly on a lexical level. Our analysis reports similar results in a low-resource language setting, with MT texts often matching or even exceeding original texts in general lexical diversity metrics such as type-token ratio (TTR), Measure of Textual Lexical Diversity (MTLD), and Yules I. This is true for both tokenized and lemmatized forms.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

However, our exploration of POS-based lexical diversity reveals that while overall lexical richness is comparable, there are subtle differences at the category level. MT texts exhibit higher diversity in the use of nouns and determiners, whereas original texts are slightly more diverse in adjectives and adverbs, especially in lemmatized forms. This suggests that MT systems may rely more on morphological variation to achieve diversity, while original texts employ a wider lexical range within specific categories. These findings highlight the need to consider individual word classes when evaluating lexical diversity, as overall metrics may overlook category-specific trends.

The distribution and composition of lexical categories further underscore these minimal yet meaningful differences. While the overall POS distributions of both text types are nearly identical, subtle variations include a reduced presence of interjections, pronouns, and auxiliary verbs in MT texts, alongside slight increases in synthetic verbs, adverbs, and particles. These discrepancies reflect stylistic and structural differences inherent to MT systems. Such observations are crucial for understanding how MT systems handle linguistic structures and for identifying potential areas of improvement.

In our frequency-based analysis, we observe a slight tendency for words in MT texts to shift toward lower frequency bands, indicating a higher prevalence of rare lexical items in MT output. While this may superficially enhance diversity, manual inspection of *hapax legomena* reveals that a notable proportion of rare words corresponds to incorrectly coined terms, potentially detracting from the quality of the translation. In turn, original texts contain more spelling errors, reflecting the imperfections inherent to human text-generation.

A focused analysis of discourse connectors unveils additional nuances. MT texts tend to reduce the frequency of low-occurrence discourse markers, occasionally omitting them altogether. However, certain connectors, such as *ondoren* ('afterwards'), are overrepresented in MT texts, suggesting a tendency for the system to overuse specific markers.

Our final analysis, examining synonym diversity, supports previous findings by showing small differences between original and MT texts. Original texts exhibit slightly greater diversity, with more balanced distributions across synonym clusters. This suggests that human-generated texts are better at leveraging lexical variety for nuanced expression, whereas MT systems may prioritize consistency or rely heavily on the most frequent choices in the training data.

Our analysis has demonstrated that, for the specific case of Spanish-Basque online news, there is little to no significant difference between original and MT texts in terms of overall lexical diversity. In other words, the use of MT does not seem to cause a loss of lexical representation in the target language, Basque. These results challenge assumptions about MT's detrimental impact on lexical richness, particularly for underrepresented languages. However, when we examine specific word categories and lexical items more closely, subtle nuances begin to emerge, which suggests that while MT can

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

approximate human-produced lexical diversity, there are still areas where it diverges from original texts.

One key question arising from this study is the extent to which these subtle differences are due to algorithmic bias, or whether they are an inherent result of the translation process and the specific source language/text analysed. Algorithmic biases may shape MT outputs by overrepresenting or underrepresenting certain lexical items based on patterns found in the training data, while the translation process might impose additional constraints or simplifications not present in original text creation. Investigating the interaction between these factors could help us understand the origin of these discrepancies and help refine MT systems to better mimic human linguistic behaviour.

Another important area for future research is determining whether these findings extend to other linguistic aspects, such as morphosyntax. While we have focused on lexical elements, morphosyntactic structures are equally critical in capturing the full complexity of language. Analysing variations in syntax, verb conjugation, and word order could provide further insights into MT's possible impact on Basque.

Furthermore, even the subtle differences observed in this work warrant closer examination to determine their significance for users. Are these variations noticeable to readers, and if so, would they permeate the language? Research exploring the perceptual thresholds at which such differences become meaningful for human audiences would be particularly valuable.

Finally, while this study has investigated the occurrence and distribution of lexical items, it does not claim that original and MT texts are identical. Lexical diversity, as measured in this analysis, does not necessarily capture the context of word usage. It does not claim that the translations are correct, either. Therefore, additional studies are needed to assess the adequacy of translations – whether MT texts successfully convey the intended meaning, context, and stylistic nuances of the source material. Such analyses would address a critical question for the application of MT systems: not just whether they preserve linguistic diversity, but whether they produce translations that are contextually appropriate and effective.

5. REFERENCES

Aranberri, N. & Iñurrieta, U. (2024). When minoritized languages encounter MT: perceptions and expectations of the Basque community. *The Journal of Specialised Translation*, *41*, 179-205. https://doi.org/10.26034/cm.jostrans.2024.4718

Baker, M. (1993). Corpus linguistics and translation studies: implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233-250). John Benjamins.

Baker, M. (1995). Corpora in translation studies. *Target. International Journal of Translation Studies*, *7*(2), 223-243. https://doi.org/10.1075/target.7.2.03bak

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

Baroni, M. & Bernardini, S. (2005). 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. https://doi.org/10.1093/llc/fqi039

Bengoetxea, K. & González-Dios, I. (2021). MultiAzterTest: a multilingual analyzer on multiple levels of language for readability assessment. *Computing Research Repository (CoRR)*. https://arxiv.org/abs/2109.04870

Bernardini, S. (2022). How to use corpora for translation. In A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge handbook of Corpus Linguistics* (pp. 485-498). Routledge. https://doi.org/10.4324/9780367076399 (Original work published 2010)

Bizzoni, Y., Juzek, T. S., España-Bonet, C., Dutta Chowdhury, K., van Genabith, J. & Teich, E. (2020). How human is machine translationese? Comparing human and machine translations of text and speech. In M. Federico, A. Waibel, K. Knight, S. Nakamura, H. Ney, J. Niehues, S. Stüker, D. Wu, J. Mariani & F. Yvon (Eds.), *Proceedings of the 17th International Conference on Spoken Language Translation* (pp. 280-290). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.iwslt-1.34

Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In J. House & S. Blum-Kulka (Eds.), *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies* (pp. 17-35). Gunter Narr Verlag.

Blum-Kulka, S. & Levenston, E. A. (1983). Universals of lexical simplification. In C. Færch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 119-139). Longman.

Castilho, S., Resende, N. & Mitkov, R. (2019). What influences the features of postediting? A preliminary study. In *Proceedings of the human-informed translation and interpreting technology workshop (HiT-IT 2019)* (pp. 19-27). Incoma Ltd. https://doi.org/10.26615/issn.2683-0078.2019 003

Chesterman, A. (2010). Why study translation universals. *Kiasm. Acta Translatologica Helsingiensia (ATH)*, *1*, 38-48. http://hdl.handle.net/10138/24319

Etchegoyhen, T., Azpeitia, A. (2016). Set-theoretic alignment for comparable corpora. In E. Katrik & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1 (pp. 2009-2018). Association for Computational Linguistics. https://aclanthology.org/P16-1189/

Etchegoyhen, T. & Gete, H. (2020). Handle with care: a case study in comparable corpora exploitation for neural machine translation. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3799-3807). European Language Resources Association. https://aclanthology.org/2020.lrec-1.469

Gamallo, P. & Labaka, G. (2021). Using dependency-based contextualization for transferring passive constructions from English to Spanish. *Procesamiento del lenguaje natural*, *66*, 53-64. https://doi.org/10.26342/2021-66-4

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In L. Wollin & H. Lindquist (Eds.), *Translation studies in Scandinavia:*

proceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II (pp. 88-95). CWK Gleerup.

Green, S., Heer, J. & 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* (pp. 439-448). Association for Computing Machinery. https://doi.org/10.1145/2470654.2470718

Hernáez, I., Navas, E., Odriozola, I., Sarasola, K., Diaz de Ilarraza, A., Leturia, I., Diaz de Lezana, A., Oihartzabal, B. & Salaberria, J. (2012). *The Basque language in the digital age / Euskara aro digitalean*. Springer.

Liu, Z. & Dou, J. (2023). Lexical density, lexical diversity, and lexical sophistication in simultaneously interpreted texts: a cognitive perspective. *Frontiers in Psychology*, *14*, 1-11. https://doi.org/10.3389/fpsyg.2023.1276705

Lu, X. (2014). *Computational methods for corpus annotation and analysis*. Springer.

Macken, L., Van Brussel, L. & Daems, J. (2019). NMTs wonderland where people turn into rabbits. A study on the comprehensibility of newly invented words in NMT output. *Computational Linguistics in the Netherlands Journal*, *9*, 67-80.

Oakes, M. P. & Ji, M. (Eds.). (2012). *Quantitative methods in corpus-based translation studies: a practical guide to descriptive translation research*. John Benjamins. https://doi.org/10.1075/scl.51

Otegi, A., Ezeiza, N., Goenaga, I. & Labaka, G. (2016). A modular chain of NLP tools for Basque. In P. Sojka, A. Horák & I. Kopeček (Eds.), *Proceedings of the 19th International Conference of Text, Speech, and Dialogue*, *SD 2016, Brno, Czech Republic, Lecture Notes in Computer Science* (pp. 93-100). Springer International. https://doi.org/10.1007/978-3-319-45510-5_11

Pociello, E., Agirre, E. & Aldezabal, I. (2011). Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, *45*(2), 121-142. https://doi.org/10.1007/s10579-010-9131-y

Sarasola, K., Aldabe, I., de Ilarraza, A. D., Grützner-Zahn, R. A. & Giagkou, M. (2022). Project European Language Equality (ELE) Grant agreement no. LC-01641480–101018166 ELE Coordinator Prof. Dr. Andy Way (DCU) Co-coordinator Prof. Dr. Georg Rehm (DFKI) Start date, duration 01-01-2021, 18 months.

Shaitarova, A., Göhring, A. & Volk, M. (2023). Machine vs. human: exploring syntax and lexicon in German translations, with a spotlight on anglicisms. In T. Alumäe & M. Fishel (Eds.), *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)* (pp. 215-227). University of Tartu Library. https://aclanthology.org/2023.nodalida-1.22

Sim Smith, K. (2017). On integrating discourse in machine translation. In B. Webber, A. Popescu-Belis & J. Tiedemann (Eds.), *Proceedings of the third workshop on discourse in machine translation* (pp. 110-121). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4814

Toral, A. (2019). Post-editese: an exacerbated translationese. In M. Forcada, A. Way, B. Haddow & R. Sennrich (Eds.), *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 273-281). European Association for Machine Translation. https://aclanthology.org/W19-6627

Toury, G. (1980). *In search of a theory of translation*. Porter Institute for Poetics and Semiotics, Tel Aviv University.

Toury, G. (2012). *Descriptive translation studies: and beyond*. John Benjamins.

Vanmassenhove, E., Shterionov, D. & Gwilliam, M. (2021). Machine translationese: effects of algorithmic bias on linguistic complexity in machine translation. In P. Merlo, J. Tiedemann & R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2203-2213). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.188

Vanmassenhove, E., Shterionov, D. & Way, A. (2019). Lost in translation: loss and decay of linguistic richness in machine translation. In M. Forcada, A. Way, B. Haddow & R. Sennrichar (Eds.), *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 222-232). Association for Computational Linguistics. https://aclanthology.org/W19-6622

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, V. Vishwanatan & R. Garnett (Eds.), *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, *30*, 5999-6010.

Volansky, V., Ordan, N. & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, *30*(1), 98-118. https://doi.org/10.1093/llc/fqt031

Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge University Press.

Zabaleta, J. (2019). Itzulpengintza eta euskararen batasuna eta normalizazioa: mende erdiko historiaren berrikusketa eta gogoeta batzuk. *Senez*, *50*, 85-109.

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

6. APPENDIX A

Table 24. Results of the analysis of lexical categories and subcategories

Lex. Cat. Level 1	EU-O	EU-MT	Lex. Cat. Level 2	EU-O	EU-MT	Lex. Cat. Level 3	EU-O	EU-MT	Lex. Cat. Level 4	EU-O	EU-MT
Nouns	48.17 (4118879)	48.4 (3910882)	Common	67.23 (2768917)	66.41 (2597210)	Living	5.48 (151640)	5.31 (137871)			
						Non-living	55.14 (1525236)	54.97 (1427764)	Acronym	0.0 (42)	0.0 (37)
									Symbols	0.19 (2843)	0.22 (3135)
									Others	100.0 (1525236)	100.0 (1424629)
						Undertermined	39.44 (1092041)	39.72 (1031575)			
			Person	16.81 (692308)	16.98 (664041)						
			Place	11.79 (485509)	12.09 (472656)						
			Number	2.7 (111230)	3.22 (125949)						
			Others	1.48 (60915)	1.3 (51026)						
Adjectives	6.66 (569348)	6.64 (536704)	Common	99.48 (566380)	99.51 (534067)						
			Interrogative	0.13 (746)	0.09 (502)						
			Others	0.39 (2222)	0.4 (2135)						
Verbs	14.29 (1221516)	13.38 (1119714)	Simple	97.04 (1129943)	96.69 (1039082)						
			Compound	0.77 (8980)	0.55 (5924)						
			Factive	2.19 (25492)	2.76 (29617)						
			Others	4.9 (57101)	4.2 (45091)						
Aux. verbs	2.14 (182771)	1.93 (155619)									
Synthetic verbs	8.59 (734636)	8.87 (716821)									
Adverbs	2.83 (241771)	3.0 (242408)	Common	97.181 (235716)	96.662 (234919)						
			Interrogative adverb	2.477 (6008)	3.192 (7758)						
			Others	0.342 (830)	0.146 (355)						

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

Lex. Cat. Level 1	EU-O	EU-MT	Lex. Cat. Level 2	EU-O	EU-MT	Lex. Cat. Level 3	EU-O	EU-MT	Lex. Cat. Level 4	EU-O	EU-MT
Determiner	9.29 (794335)	9.36 (756049)	Demonstrative	15.869 (126052)	16.771 (126799)	Common	77.36 (97520)	75.23 (95396)			
						Reinforced	22.64 (28532)	24.77 (31403)			
			Calificative	0.81 (6411)	0.81 (6143)	Common	27.94 (1791)	33.19 (2039)			
						Interrogative	72.06 (4620)	66.81 (4104)			
			Numeral	83.22 (661047)	82.31 (622285)	Definite	84.75 (560241)	84.06 (523075)	Distributive	0.16 (915)	0.05 (241)
									Ordinal	26.82 (150257)	75.59 (395415)
									Number	73.02 (409069)	24.36 (127419)
						Indefinite	12.26 (81048)	12.92 (80402)			
						General	2.99 (19758)	3.02 (18808)			
			Other	0.1 (825)	0.11 (822)						
Pronouns	0.35 (29524)	0.28 (22934)	Personal	63.87 (18858)	59.88 (13732)	Common	96.84 (18263)	97.89 (13442)			
						Reinforced	3.16 (595)	2.11 (290)			
			Indefinite	26.67 (7875)	30.81 (7066)	Indefinite	77.26 (6084)	85.21 (6021)			
						Interrogative	22.74 (1791)	14.79 (1045)			
			Reciprocal	7.73 (2283)	8.04 (1844)						
			Reflexive	0.0 (0)	0.0 (0)						
			Other	1.72 (508)	1.27 (292)						
Linker	6.13 (524440)	6.12 (494658)	Conjuctions	80.08 (419964)	82.59 (408515)						
			Connectors	19.43 (101915)	16.88 (83510)						
			Others	0.49 (2561)	0.53 (2633)						
Particles	0.91 (77592)	0.94 (75875)									
Interjection	0.03 (2240)	0.02 (1981)									
Abbreviations	0.06 (5170)	0.06 (4526)									
Others	0.57 (48473)	0.53 (42766)									

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

7. APPENDIX B

Table 25. Temporal connectors

List of connectors	Absolute count EU-O	Relative count EU-MT	Relative count EU-O (per 1000 words)	Relative count EU-MT (per 1000 words)	Difference
ondoren	7,649	20,075	0.90	31.51	-30.61
lehen	15,212	19,301	1.80	30.29	-28.49
orain	7,384	6,059	0.87	9.51	-8.64
behin	4,122	4,202	0.49	6.59	-6.1
berriro	2,765	3,822	0.33	6.00	-5.67
gero	6,050	3,941	0.72	6.19	-5.47
bitartean	4,096	3,408	0.48	5.35	-4.87
era berean	2,063	3,091	0.24	4.85	-4.61
gaur egun	1,605	2,771	0.19	4.35	-4.16
azkenik	1,744	2,076	0.21	3.26	-3.05
horrela	4,093	2,134	0.48	3.35	-2.87
orain arte	1,905	1,843	0.23	2.89	-2.66
geroago	1,491	1,629	0.18	2.56	-2.38
orduan	1,999	1,618	0.24	2.54	-2.3
hasieran	1,066	1,380	0.13	2.17	-2.04
amaitzeko	1,169	1,254	0.14	1.97	-1.83
berehala	992	1,008	0.12	1.58	-1.46
lehenengo	3,685	1,200	0.44	1.88	-1.44
laster	671	941	0.08	1.48	-1.4
hasteko	828	695	0.10	1.09	-0.99
lehenik eta behin	153	280	0.02	0.44	-0.42
honela	184	209	0.02	0.33	-0.31
jarraituz	320	206	0.04	0.32	-0.28
ordura arte	142	156	0.02	0.24	-0.22
amaitu baino lehen	70	119	0.01	0.19	-0.18
besterik gabe	44	109	0.01	0.17	-0.16
aurrera egin ahala	188	107	0.02	0.17	-0.15
**. mendean*	113	90	0.01	0.14	-0.13
bigarrenik	26	71	0.0	0.11	-0.11
bukaeran	318	82	0.04	0.13	-0.09
egun batez	33	43	0.0	0.07	-0.07
egun hartan	69	49	0.01	0.08	-0.07
aspaldiko	85	40	0.01	0.06	-0.05
hasi bezain laster	17	35	0.0	0.05	-0.05
bukatzeko	480	62	0.06	0.10	-0.04

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

List of connectors	Absolute count EU-O	Relative count EU-MT	Relative count EU-O (per 1000 words)	Relative count EU-MT (per 1000 words)	Difference
hasi orduko	6	28	0.0	0.04	-0.04
arratsaldero	52	30	0.01	0.05	-0.04
hasiera emateko	86	31	0.01	0.05	-0.04
halako batean	26	20	0.0	0.03	-0.03
honen ondoren	36	20	0.0	0.03	-0.03
antzina	19	16	0.0	0.03	-0.03
lehenengo egunean	39	12	0.0	0.02	-0.02
ezer baino lehen	14	15	0.0	0.02	-0.02
lehen-lehenik	4	5	0.0	0.01	-0.01
bigarrenez	101	10	0.01	0.02	-0.01
azkenez	0	0	0.0	0.0	0.0
hau ikusi eta gero	0	0	0.0	0.0	0.0
hau ikusi ondoren	0	0	0.0	0.0	0.0
honetaz gain	10	2	0.0	0.0	0.0
hau egin ondoren	1	0	0.0	0.0	0.0
aurrera beharrez	0	0	0.0	0.0	0.0
hasteko esan dezadan	0	0	0.0	0.0	0.0
lehenengo eta behin	19	0	0.0	0.0	0.0
eta besterik gabe	3	3	0.0	0.0	0.0
lehenengo batean	0	0	0.0	0.0	0.0
hasteaz batera	0	0	0.0	0.0	0.0
dena hasi zen	2	2	0.0	0.0	0.0
behinola	5	1	0.0	0.0	0.0
behin batean	3	2	0.0	0.0	0.0
bazen behin	2	3	0.0	0.0	0.0
behiala	0	0	0.0	0.0	0.0
behin baten	0	0	0.0	0.0	0.0
gizona mundura baino lehen	0	0	0.0	0.0	0.0
honen ostean	48	0	0.01	0.0	0.01
beranduago	353	12	0.04	0.02	0.02

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

Table 26. Adversative connectors

List of connectors	Absolute count EU-O	Relative count EU-MT	Relative count EU-O (per 1000 words)	Relative count EU-MT (per 1000 words)	Difference
baina	29,845	24,912	3.53	39.1	-35.57
baino	14,062	15,161	1.66	23.79	-22.13
artean	10,344	8,070	1.22	12.67	-11.45
berriz	6,887	6,132	0.81	9.62	-8.81
hala ere	4,047	5,473	0.48	8.59	-8.11
aldiz	3,835	3,030	0.45	4.76	-4.31
horregatik	1,941	2,547	0.23	4.0	-3.77
baizik	1,286	1,638	0.15	2.57	-2.42
ordea	1,648	1,035	0.19	1.62	-1.43
bien bitartean	322	450	0.04	0.71	-0.67
alta	175	273	0.02	0.43	-0.41
aitzitik	191	181	0.02	0.28	-0.26
hala eta guztiz ere	414	58	0.05	0.09	-0.04
haatik	28	10	0.0	0.02	-0.02
*ez*ezen*	0	10	0.0	0.02	-0.02
ez baina	7	6	0.0	0.01	-0.01
bizkitartean	0	0	0.0	0.0	0.0
ezezik	8	0	0.0	0.0	0.0
ez baizik	1	1	0.0	0.0	0.0
alabadere	0	0	0.0	0.0	0.0
ezpada	18	1	0.0	0.0	0.0
badarik ere	0	0	0.0	0.0	0.0
hargatik	9	0	0.0	0.0	0.0
horratik	2	0	0.0	0.0	0.0
barren	4	0	0.0	0.0	0.0
alabaina	340	2	0.04	0.0	0.04
ostera	528	0	0.06	0.0	0.06
dena dela	753	20	0.09	0.03	0.06
dena den	872	12	0.10	0.02	0.08

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

Table 27. Logical connectors

List of connectors	Absolute count EU-O	Relative count EU-MT	Relative count EU-O	Relative count EU-MT	Difference
eta	356,749	346,768	42.18	544.22	-502.04
ere	48,787	36,069	5.77	56.61	-50.84
edo	11,523	15,902	1.36	24.96	-23.6
gainera	10,709	14,519	1.27	22.79	-21.52
*ez*ez*	7,650	7,662	0.90	12.02	-11.12
bestalde	8,075	7,209	0.95	11.31	-10.36
nahiz	2,436	3,681	0.29	5.78	-5.49
halaber	4,987	3,557	0.59	5.58	-4.99
*bai*bai*	480	1,113	0.06	1.75	-1.69
zein	3,405	1,306	0.40	2.05	-1.65
ala	1,305	688	0.15	1.08	-0.93
ez ezik	463	372	0.05	0.58	-0.53
*eta*ere ez*	407	366	0.05	0.57	-0.52
*bai eta*ere*	139	321	0.02	0.50	-0.48
bestela	281	279	0.03	0.44	-0.41
behintzat	398	193	0.05	0.30	-0.25
*eta*ere bai*	259	127	0.03	0.20	-0.17
bederen	15	8	0.0	0.01	-0.01
badere	0	0	0.0	0.0	0.0
bertzalde	0	0	0.0	0.0	0.0
orobat	39	3	0.0	0.0	0.0
ez eze	1	0	0.0	0.0	0.0
ez ezen	0	0	0.0	0.0	0.0
berebat	34	0	0.0	0.0	0.0
Eta ere	0	0	0.0	0.0	0.0
ezta ere	15	1	0.0	0.0	0.0
badarik	0	0	0.0	0.0	0.0
ezpabere	0	0	0.0	0.0	0.0
ezperen	0	0	0.0	0.0	0.0
ezpere	0	0	0.0	0.0	0.0
gainerontzean	0	0	0.0	0.0	0.0
osterantzean	5	0	0.0	0.0	0.0
bertzenaz	0	0	0.0	0.0	0.0
bederik	0	0	0.0	0.0	0.0
behinik behin	13	2	0.0	0.0	0.0
baita ere	244	11	0.03	0.02	0.01
edota	1,848	0	0.22	0.0	0.22

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

Table 28. Conditional connectors

List of connectors	Absolute count EU-O	Absolute count EU-MT	Relative count EU-O (per 1000 words)	Relative count EU-MT (per 1000 words)	Difference
ustez	4,094	2,630	0.48	4.13	-3.65
baldin	946	434	0.11	0.68	-0.57
pentsatuz	24	51	0.0	0.08	-0.08
ez baldin	88	1	0.01	0.0	0.01

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X

Table 29. Causal connectors

List of connectors	Absolute count EU-O	Absolute count EU-MT	Relative count EU-O (per 1000 words)	Relative count EU-MT (per 1000 words)	Difference
hala	7,648	11,618	0.90	18.23	-17.33
beraz	2,638	3,980	0.31	6.25	-5.94
ondorioz	4,942	3,480	0.58	5.46	-4.88
*izan*ere*	8,699	3,278	1.03	5.14	-4.11
bada	2,567	2,696	0.30	4.23	-3.93
zergatik	692	722	0.08	1.13	-1.05
hain zuzen ere	1,948	597	0.23	0.94	-0.71
argi dago	263	268	0.03	0.42	-0.39
funtsean	64	145	0.01	0.23	-0.22
zeren	159	78	0.02	0.12	-0.1
laburbilduz	38	22	0.0	0.03	-0.03
kontuan izan	170	23	0.02	0.04	-0.02
honen arabera	63	17	0.01	0.03	-0.02
ondorio gisa	3	6	0.0	0.01	-0.01
hainbestez	0	0	0.0	0.0	0.0
honebestez	1	0	0.0	0.0	0.0
hortaz bada	0	0	0.0	0.0	0.0
on litzateke	0	0	0.0	0.0	0.0
laburki	1	2	0.0	0.0	0.0
horren kasuaz	0	0	0.0	0.0	0.0
hori horrela delarik	0	0	0.0	0.0	0.0
halatan	0	0	0.0	0.0	0.0
honen ondorioz	54	9	0.01	0.01	0.0
honela ba	0	0	0.0	0.0	0.0
hau guztia dela eta	1	0	0.0	0.0	0.0
hau dela bide	0	0	0.0	0.0	0.0
esanak esan	3	0	0.0	0.0	0.0
horren kariaz	4	0	0.0	0.0	0.0
hori dela medio	7	0	0.0	0.0	0.0
esandakoaren arabera	81	3	0.01	0.0	0.01
horrenbestez	745	16	0.09	0.03	0.06
hortaz	804	7	0.10	0.01	0.09

ISSN: 0046-435X ISSN-e: 2530-5832 ISSN-L: 0046-435X