Natural Language Processing (2024), 1–20
doi:10.1017/nlp.2024.4

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Evaluating NMT using the non-inferiority principle

María do Campo Bayón and Pilar Sánchez-Gijón

Universitat Autònoma de Barcelona, Barcelona, Spain
Corresponding author: María do Campo Bayón; Email: maria.docampo@autonoma.cat

(Received 6 March 2023; revised 4 December 2023; accepted 12 February 2024)

Special Issue on '**The Role of Context in Neural Machine Translation Systems and its Evaluation**', guest-edited by Rebecca Knowles and Sheila Castilho

Abstract

The aim of this article is to propose a new neural machine translation (NMT) evaluation method based on the non-inferiority principle. In order to do that, we evaluate raw machine translation (MT) in terms of naturalness, which for this research is defined as not just the lack of fluency errors but also meeting the linguistic expectations of Galician end users when reading original texts in Galician. Our main objective is, in the first place, to validate the new methodology presented in our previous study by evaluating an NMT engine from Spanish into Galician for the social media domain that was retrained with a new Twitter corpus. This new methodology and NMT engine were applied after analyzing the conclusions of a pilot survey conducted among Twitter users to evaluate their perception of tweets translated from Spanish into Galician with our NMT engine created with a corpus of tweets. As in our preliminary study, our aim is to propose a robust quality approximation method based on the reception parameters of end users' perceptions. This new survey was conducted in December of 2022 with the participation of 228 Galician-speaking Twitter users. Among the main changes proposed are the inclusion of more information about the participant profile, so the non-inferiority principle can be also evaluated according to these parameters; the inclusion of a new typology of tweets, the threads; the provision of context by means of presenting the tweets in their original display as shown in the Twitter app; a change in the number of tweets evaluated and the number of different questionnaires; the change in the distribution of the questionnaires; and the inclusion of an error classification human evaluation conducted by professional linguists to correlate the findings. We will present the steps carried out following the conclusions of the pilot study, describe the new study's design, analyze the new findings, and present the final conclusions regarding the engine and the evaluation method based on the non-inferiority principle. Finally, we will also provide some examples of the use of this new methodology in the translation industry.

Keywords: non-inferiority principle; evaluation; machine translation; NMT evaluation; naturalness evaluation

1. Introduction

Many of the fluency issues that plagued earlier machine translation (MT) systems have been overcome by neural machine translation (NMT) systems (Hassan *et al.,* 2018), hence why the arrival of NMT was accompanied by claims of parity with human translations. One of the first articles claiming to have bridged the gap between human and MT was Wu *et al.* (2016). This paper analyses parity by using automatic quality evaluation metrics based on a comparison of MT output with human translation from the same source (Chatzikoumi, 2020). Differences between two translations, however, can be the result of one of the following two issues: either one of the translations contains errors or the two translations are completely correct despite their wording differences.

According to Leppänen *et al.* (2016), the tweet is a hybrid genre that combines characteristics of literary writing and oral communication and has developed its own linguistic and discursive conventions. Being a hybrid genre, with nuances of written and oral text, evaluation based solely on the absence of fluency errors is insufficient. Since MT is also employed in such contexts to translate tweets, evaluating the automatic translation of this genre solely for the absence of fluency errors is likewise inadequate. Therefore, it becomes imperative to move beyond a mere assessment of fluency errors and delve into evaluating the naturalness of the translated text within the context of the target speaker. The results obtained in this study stem from a comparison of texts written in Galician and texts machine translated into Galician. Unlike studies focusing on the description of translationese (which compare types of translations with the source), our approach involves comparing the machine-translated target texts (henceforth MT outputs) with texts written in the target language (henceforth Human original texts). The use of the non-inferiority principle is particularly appropriate for the evaluation of naturalness when comparing MT outputs *versus* human original texts; accordingly, this article is based on the premise that translations of the same text can differ, and hence, MT quality should be evaluated using a different comparison principle. The particular focus of this article is to propose evaluating MT using the statistical principle of non-inferiority.[a] This principle should make it possible to determine whether or not texts obtained through MT are less natural than texts created in that language. This method goes beyond measuring fluency errors and additionally measures naturalness. Naturalness should be understood as a habitual use of the language, free of grammatical errors, fluid in style, and without expressions that are strongly influenced by other languages, such as Spanish.

Non-inferiority analyses should be employed when the aim of a study is not to detect differences, but rather to establish the similarities of the studied parameters. In a standard t-test analysis, when the groups are not significantly different, the resulting p-value is high. This outcome only allows us to conclude that there is insufficient evidence to claim the groups are different, without being able to provide a definitive affirmative answer to the posed research question.

This MT evaluation methodology was validated by analyzing the translation of an NMT engine from Spanish into Galician which has been specially trained to translate texts for the social media domain using a new Twitter corpus and augmented using back-translations. The development of this NMT engine is part of a larger project to develop an NMT system for a low-resource language combination.[b]

The non-inferiority statistical approach was analyzed via survey responses carried out in December 2022, which gathered responses from 228 Galician-speaking Twitter users. The survey collected information from participants to provide a more accurate interpretation of the results. The analysis was performed by means of a generalized linear mixed model considering a binomial distribution with a logit link. Human evaluation of MT fluency errors, which generally evaluates the absence of errors in the translation equivalence (accuracy) as well as the lack of errors in the target text, was performed by 3 professional linguists (do Campo & Sánchez-Gijón 2023). Nevertheless, this type of evaluation is not enough to evaluate the end-user experience when dealing with raw MT output. Consequently, this research undertakes a study based on the end users' perception of naturalness, understood as not only the lack of fluency errors but also meeting the linguistic expectations of Galician native speakers. The hypothesis underlying this research is that the raw MT output is not inferior to texts originally written in Galician in terms of user expectations. Therefore, we decided to use the statistical approach of the non-inferiority principle.

[a]The authors will use the term "non-inferiority principle" throughout this paper to be consistent with the existing literature. However, in this context, it may be more appropriate to refer to it as a statistical method.

[b]do Campo Bayón, María. 2023. Neural machine translation for low-resource languages. User evaluation according to the non-inferiority principle. Universitat Autònoma de Barcelona.

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

The analysis included the source nature of the tweet (originally written in Galician or machine-translated), the length of the tweet, and how the tweets were perceived by the participants as fixed effects. Other factors such as age, or social network uses were also tested to explore if these variables could explain the perception of naturalness. The model included subject and tweet-specific random effects to take into account the repeated measures' nature of the experiment, since different tweets are evaluated by different subjects. Finally, a specific comparison between the naturalness perception of machine-translated tweets into Galician and tweets directly written in Galician was performed using a one-sided non-inferiority test with a non-inferiority limit of 10 percentage points.

The results obtained show that the overall acceptance of the machine-translated tweets is slightly higher than the tweets originally written in Galician, since the overall non-inferiority analysis shows that the acceptance rate of the machine-translated tweets is not inferior to the acceptance rate of the tweets originally written in Galician. Results also show how important context is to perceive the naturalness of the machine-translated tweets, especially if we take into account that the brevity and density of language in tweets pose a major challenge for automatic translation (Zappavigna, 2012). As far as context is concerned, not only was the linguistic content taken into account but also the paratextual context, which is of great importance when talking about social media.

The concept of context intervenes on two different levels in this study. First, the importance of context in MT by comparing unrelated texts (short, one-sentence tweets) with more complex texts (either paragraphs or threads). Second, the importance of paratextual context in the perception and acceptance of MT of social media texts such as tweets.

In summary, this article demonstrates the robustness of our evaluation method in assessing the naturalness of NMT systems and draws on some conclusions about the relevance of statistical models and their applicability in the translation industry as well as proving that our NMT engine for low-resource languages is effective and fit for purpose in social media, especially when the paratextual context of the tweet is taken into consideration.

2. Related work

The method proposed in this article can be described as a human extrinsic reference-less contextual MT evaluation method. Unlike evaluation methods that rely on a source text or standard golden translation (such as BLEU), this evaluation is based solely on the source-less (i.e., extrinsic) assessment of the target text in its context: the paratextual elements of Twitter messages.

Measuring the quality of MT has occupied and troubled both the translation industry and academia (Sánchez-Gijón, 2014). Common quality metrics compare the translation produced by an MT engine to a human translation of the same source text. This sentence-based method permits obtaining quality scores, such as BLEU (Papineni *et al.,* 2002). In genres in which consistency is a translation requirement, such as technical documentation, a sentence-based comparison appears to be very appropriate (López-Pereira, 2018). Nevertheless, sentence-based metrics can prevent the detection of over-sentential and textual coherence and cohesion.

The breakthrough of NMT resulted in MT output that differed slightly from that of previous MT systems: it makes fewer errors than previous systems (Temnikova *et al.,* 2019) and provides translations that are closer to being human-like in their fluency (Bhardwaj *et al.,* 2020). Studies like the ones mentioned only measure the lack of fluency errors, and therefore, the human-like quality should be confirmed by exploring alternative quality evaluation methods. Läubli *et al.* (2018) suggest that sentence-based evaluating methods and metrics do not accurately reflect the quality of translated texts. Tan *et al.* (2022) propose a method for evaluating discourse cohesion based on four cohesive approaches: reference, conjunction, substitution, and lexical cohesion. This concept of comparing texts rather than sentences is also becoming popular within the translation industry (Gino, 2022).

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

Additionally, human evaluation of NMT allows it to surpass the sentence-based model. Human evaluation can be conducted in various ways, including post-editing, ranking, and rating (Chatzikoumi, 2020). Post-editing is the process of having human translators edit machine-translated text in order to improve its quality (Sánchez-Gijón *et al.,* 2019). Ranking involves comparing the quality of translations produced by different systems or methods using human evaluators (Görög, 2014). The rating method involves human evaluators rating the quality of translations on a scale or based on specific criteria (Castilho *et al.,* 2017). In each of these instances, the comparison is conducted between translations of the same source text, although there are also methods for evaluating NMT without a reference translation (Zheng *et al.,* 2019). Regarding Natural Language Processing (NLP) and low-resource languages, Ranathunga *et al.* (2023) provide a survey of the techniques used to create data and train NMT engines, as well as their results, but no direct human evaluation is included.

This article agrees with the proposal of Martindale *et al.* (2021), who argue that MT should be evaluated not only on the basis of fluency and adequacy but also on the basis of end-user perception. The authors propose to evaluate the believability of MT, that is, "the user's perception of the likelihood that the meaning of a given MT output matches the meaning of the input, without understanding the input." The focus of the research presented in this paper is an approach to MT evaluation by means of naturalness. Naturalness is a text dimension that has recently been used as an approach to MT quality (do Campo & Sánchez-Gijón 2022; Freitag *et al.,* 2022). The latter authors consider a translation to be natural if it is "adequate and fluent." For the purposes of this paper, and given the language pair in question (Casares *et al.,* 2021), the presence of target language stylistic features rather than source language features is considered part of the naturalness (see previous section).

The evaluation of naturalness, based on an extrinsic evaluation of the target text, is carried out on the basis of the statistical method commonly referred to in the literature as the non-inferiority principle. The purpose of a non-inferiority analysis is simply to reframe the analysis to draw affirmative conclusions, such as evaluating whether there is evidence to support the claim that the differences are smaller than a certain magnitude. From a technical perspective, non-inferiority analyses can, in fact, be interpreted as a shifted t-test. Non-inferiority trials are commonly employed in the realm of health sciences, and these trials are designed to demonstrate that a treatment is at least not appreciably less effective than a reference treatment. Such study designs are often used when comparing a new treatment to an established medical standard of care, particularly in situations where the new treatment offers advantages such as cost-effectiveness, safety, or convenience, and would therefore be preferred if not appreciably less effective.

The statistical principle of non-inferiority is widely applied in pharmacology (Molina Nadal, 2020; Althunian *et al.* 2017), particularly in the case of determining the validity of generic drugs. Their results are then compared to those of on-patent drugs for this purpose. Generics are accepted when it can be demonstrated that their use produces non-inferior results to their on-patent equivalent. This paper investigates the possibility of applying the same approach to human extrinsic contextual MT evaluation based on naturalness.

3. Background

This article presents the most significant findings from a thesis on low-resource languages and NMT as a means of promoting and using a minority language in the context of social media.[c] This study is based on a previous pilot completed in 2022, which focused on evaluating a Spanish into Galician NMT engine for social media and proposed a new methodology for evaluating this type of NMT engine (do Campo *et al.,* 2022).

[c]Research supported by the DespiteMT project, grant number PID2019-108650RB-I00 [MINECO/ FEDER, UE; Principal researcher: Dr. Pilar Sánchez-Gijón, Grup Tradumàtica, UAB.

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

For the pilot study, we created an NMT engine based on Joey (Kreutzer *et al.*, 2019) through the online platform MutNMT[d] (Kenny, 2022). The corpus used to train the engine was a mix of two corpora. We used the Paracrawl corpus Spanish—Galician (1,879,651 sentences and 44,626,394 words) as a generic base corpus. Then, we created a Galician monolingual corpus of tweets written in Galician and extracted from Galician institutional accounts, mainly linguistic institutions and Galician universities. After cleaning the corpus, we back-translated the monolingual corpus into Spanish using the generic Google engine to give us a bilingual corpus (4,973 unique sentences and 740,395 words). Once trained, the engine achieved a total BLEU score of 70.63 against the test corpus extracted only from the Twitter corpus with a size of 5000 sentences.[e]

Regarding its design, we created a study based on the non-inferiority principle. The pilot study had two main objectives: validate the method and evaluate our NMT engine. To do this, we planned an empirical study with an incomplete factorial design, with each participant answering a total of 20 items (tweets) from a sample of 80 possible items in four separate questionnaires with shared items. These tweets were presented one at a time to prevent participants from anticipating the existence of a systematic pattern involving machine-translated tweets into Galician or tweets written in Galician directly. The 261 participants answered the following yes or no question: *do you consider that this tweet was directly written in Galician, or do you consider that this tweet was machine translated into Galician?* The survey was promoted directly on Twitter in order to gather participants that were Galician-speaking users of Twitter, but no demographic information was collected.

The sample of tweets was selected following two criteria. First, they were classified according to their origin: original text if the text was directly written in Galician, and MT if the text was machine translated from Spanish into Galician using our NMT engine. Then, they were classified into five different categories according to the linguistic characteristics of the tweets composed as just one sentence (with no further context) or a paragraph (each sentence being the linguistic context of the others): short sentence, long sentence, paragraph composed of short sentences, paragraph composed of long sentences, and mixed paragraph if the paragraph contained both short and long sentences.

In this study, the non-inferiority principle attempted to determine whether tweets generated by NMT are perceived as original (Molina Nadal, 2020) or MT, which was calculated by the acceptance rate. The statistical analysis was conducted with R v4.1. software. To describe the quantitative variables, we used average rate, standard deviation, maximum rate, minimum rate, and number of cases. To describe qualitative variables, we used absolute and relative frequencies.

The preliminary comparison between the acceptance rate and the characteristics of the tweets was carried out by using the x^2 test or the Fisher exact test in compliance with the Cochran criteria. The level of significance for all statistical tests was set at 5% ($p < 0.05$), and to analyze the acceptance rate, we employed a generalized linear mixed model with a binomial distribution, randomly crossing tweets (Item) and participants (ID) and using the tweet characteristics as fixed factors.

According to the results of this pilot study, we were able to draw several conclusions. On the one hand, we validated our analysis method. We have proved that information about end-user perceptions of MT texts processed under the non-inferiority principle allows conclusions to be drawn about the naturalness of the translation compared to originally written texts. On the other hand, we extracted clear information with a significant value for the weaknesses and strengths of the performance of the NMT engine in a low-resource language context. The estimations based on the model indicated the path to improving our engine. Our engine's performance on short sentences, both single and in paragraphs, should be improved by augmenting data similar to the genre it is supposed to translate. Surprisingly, we discovered that our engine was not inferior to tweets composed of long sentences and directly written in Galician. This means that better

[d] Available at: https://www.multitrainmt.eu/index.php/es/formacion-en-ta-neuronal/mutnmt
[e] The engine is available at: https://ntradumatica.uab.cat/

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

perception is achieved through syntactic complexity, which seems to provide a better linguistic context.

Nonetheless, in our pilot, we ran into some claims that were difficult to prove and that we should address in subsequent surveys after the engine has been retrained. To begin, we required more information about the participants to better understand the results, so we decided to include some control questions to detect basic linguistic logic and attention, allowing those participants to be excluded from the analysis. Second, participants' lack of context may appear to be the reason for rejecting short tweets, so we decided to include threads that would help with context and cohesion. Finally, we decided to change the main question because its formulation might have induced participants to search for machine-translated tweets rather than naturalness. Hence, we designed the final study with these conclusions in mind, which are presented in the following section.

4. Retrained NMT engine for social media

The NMT engine was retrained before conducting the final study, and we changed two settings on our first NMT engine: the specific corpus and the NMT technology. We kept the Paracrawl corpus as a base generic corpus but decided to expand our Twitter corpus. To build a larger Twitter corpus, we chose more institutional Galician accounts, such as those associated with Galicia's official television and radio. These accounts were specifically picked, as in the previous study, because they would be more reliable in terms of good use of grammar and spelling, as well as common and natural expressions. After cleaning and tokenizing the monolingual Galician corpus, we translated it into Spanish with a Google generic NMT engine using back-translation. We used this technique because of the positive results obtained in the first evaluation. This bilingual corpus contains 193,072 unique sentences and 5,448,375 words, and we trained our engine with the generic Paracrawl corpus Spanish—Galician (1,879,651 sentences and 44,626,394 words) and this specific corpus.

Regarding the NMT technology, we used the default parameters of the transformer-big configuration for Marian (Junczys-Dowmunt *et al.*, 2018), and we achieved a total BLEU score of 85 against the test corpus extracted only from the Twitter corpus, which was higher than the BLEU score achieved in the pilot study. As we designed a realistic setting, the sentences included in the test corpus were similar to the Twitter training corpus, and therefore, the BLEU score was high.

5. Implementation and analysis

As mentioned before, we replicated the non-inferiority model presented in our first study with some adjustments. We modified the question that the participants had to answer and the data collection by introducing demographic questions and we carried out the study in December 2022.

5.1 Study design

This time, we kept the same empirical approach with an incomplete factorial design, but we modified the number of items answered by each participant. We chose 6 different blocks of 5 complete tweets of different length to compose 9 different questionnaires that included 10 tweets to evaluate. Annex I provides one example of each category. The questionnaires were as follows (Table 1).

The survey form presented the tweets one at a time to prevent participants from anticipating the existence of some systematic pattern. The questionnaire included textual and paratextual information: the original display of the tweet as seen in the Twitter app was recreated to facilitate the right contextualization of the text. We also kept a yes/no question, but we changed the formulation and focused on the evaluation of the naturalness without asking participants to judge if they were directly written in Galician or MT into Galician. The new question was: *can this tweet or thread be considered a natural tweet or thread in Galician?* Because of this formulation, an

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

Table 1. Distribution of tweets among the nine questionnaires

Questionnaire 1	Questionnaire 2	Questionnaire 3	Questionnaire 4	Questionnaire 5	Questionnaire 6	Questionnaire 7	Questionnaire 8	Questionnaire 9
MSS2	MSS2	MSS2	MTh3	MTh3	MTh3	OSP3	OSP3	OSP3
MSP2	MSP2	MSP2	OSP1	OSP1	OSP1	OSS1	OSS1	OSS1
MSP1	MSP1	MSP1	MSS3	MSS3	MSS3	OSS3	OSS3	OSS3
MLS2	MLS2	MLS2	MTh2	MTh2	MTh2	OTh3	OTh3	OTh3
OLP3	OLP3	OLP3	MTh1	MTh1	MTh1	OLP2	OLP2	OLP2
MLS3	MLP3	MLP1	MLS3	MLP3	MLP1	MLS3	MLP3	MLP1
MSP3	OLS3	OTh1	MSP3	OLS3	OTh1	MSP3	OLS3	OTh1
MLS1	OLP1	MLP2	MLS1	OLP1	MLP2	MLS1	OLP1	MLP2
OTh2	MSS1	OSS2	OTh2	MSS1	OSS2	OTh2	MSS1	OSS2
OLS2	OLS1	OSP2	OLS2	OLS1	OSP2	OLS2	OLS1	OSP2

Legend	
MSS	Machine-translated Short Sentence
MSP	Machine-translated Short Paragraph
MLS	Machine-translated Long Sentence
MTh	Machine-translated Thread
MLP	Machine-translated Long Paragraph
OSS	Originally Written Short Sentence
OSP	Originally Written Short Paragraph
OLS	Originally Written Long Sentence
OLP	Originally Written Long Paragraph
OTh	Originally Written Thread

explanation of the concept of linguistic naturalness was given at the beginning of the questionnaire. The explanation written in Galician said that we understood naturalness as a usual use of the language, with common expressions and a fluid style, free of grammar mistakes and influences from other languages such as Spanish. As the participants spoke both Spanish and Galician (both are official languages in the region, but Spanish is the dominant language), the absence of influence from the major language was willingly emphasized as part of the definition of naturalness. Since the survey targeted all kinds of Galician speakers, we asked Galician associations of all types to promote it among their affiliates and obtained answers from 218 participants.

The questionnaire included five demographic questions and 10 tweet evaluation questions. In the demographic questions, we asked about age, level of Galician linguistic competence, regular use of Galician, level of Twitter use in Galician, and level of Galician education received. To build the tweets sections, we took into account different types of tweets as shown in the previous table, which were classified into two broad categories. First, they were classified according to their origin: original text if the text was written in Galician, and MT if the text was machine translated from Spanish into Galician using our NMT. Second, they were classified according to their length: short sentence, long sentence, paragraph composed of short sentences, paragraph composed of long sentences, and threads.

The end-user perception results do not allow us to conclude whether MT was accurate, so all machine-translated tweets were also subjected to human evaluation. We performed an error evaluation using the MQM-DQF framework (Lommel *et al.,* 2018), which was carried out by three professional linguists with over ten years of experience using the online platform ContentQuo.[f] The goal of this study was to find a link between errors and a negative attitude toward the machine-translated tweet. We are not going to give in-depth details about this evaluation because it is not the objective of this paper, nevertheless, this was carried out this evaluation because we wanted to ensure that the translation perceived as natural does not have critical errors. Although some errors were found, the raw NMT tweets obtained a DQF rate of over 90% (see do Campo *et al.,* 2023).

5.2 The non-inferiority model

Our statistical approximation, the non-inferiority principle (Tunes da Silva *et al.,* 2009), followed the same parameters as the pilot study: a generalized linear mixed model with a binomial distribution, tweets (ITEM), and participants (ID) as crossed random effects, and using the tweet characteristics as fixed factors. Here, p is the proportion of success (positives), α is the independent term of the model, β is the vector of coefficients associated with the explanatory variables, X is the design matrix of the explanatory variables, a_i is the random effect associated with the items, following $N\left(0, \sigma^2_{\text{ITEM}}\right)$ and b_j is the random effect associated with the subjects, following $N\left(0, \sigma^2_{\text{ID}}\right)$.

$$\begin{aligned} \log\left(\frac{p}{1-p}\right) &= \alpha + \beta X + a_i + b_j \\ a_i &\sim N\left(0, \sigma^2_{\text{ITEM}}\right) \\ b_j &\sim N\left(0, \sigma^2_{\text{ID}}\right) \end{aligned} \tag{1}$$

In the final evaluation, we used a generalized linear mixed model with a binomial distribution and a logit link to compare the naturalness perception of the machine-translated tweets into Galician and the tweets directly written in Galician (Agresti, 2015; Zuur *et al.,* 2009). This model randomly crossed the tweets (Item) and participants (ID), using the tweet characteristics as fixed factors. We also tested the perception of naturalness according to age or social network usage. Because different tweets are evaluated by different subjects, the model included subject and

[f] Available at https://www.contentquo.com/

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

tweet-specific random effects to account for the evaluation of different tweets by different subjects. Hence, p is the success proportion (positive), α is the independent term of the model, β is the coefficient associated with the variable Original, γ_j is the coefficient associated with category j of the variable Type of Tweet, $\beta\gamma_j$ is the coefficient associated with category j of the interaction between Original and Type of Tweet, a_i is the random effect associated with items, with variability σ^2_{item}, b_j is the random effect associated with subjects, with variability σ^2_{id}.

$$\begin{aligned} N_{ij}[k] &\sim \text{Bin}(1, p_{ij}[k]) \\ \log\left(\frac{p_{ij}[k]}{1-p_{ij}[k]}\right) &= \mu + \alpha_i^{Id} + \alpha_{j[k]}^{Tw} + \beta O_Y + \gamma_j \text{Type}_j + (\beta\gamma)_j (O_Y \times \text{Type}_j) \\ \alpha_i^{Id} &\sim N(0, \sigma_{(\alpha^{Id})^2}) \quad \text{for } i = 1, \ldots, I \\ \alpha_{j[k]}^{Tw} &\sim N(0, \sigma_{(\alpha^{Tw})^2}) \quad \text{for } j = 1, \ldots, J, \quad k = 1, \ldots, K_j \end{aligned} \tag{2}$$

Finally, a detailed comparison of the naturalness perception of machine-translated tweets into Galician and directly written tweets was made, considering a one-sided non-inferiority test with a non-inferiority limit of 10% points. The selection of this margin is critical and represents one of the main challenges in non-inferiority trials, as there is no universal value. Regulatory agencies provide guidance on how to choose this margin but do not give specific values. Molina Nadal (2020) states that in most of the cases, a 30% margin is decided, while other authors established a 5% margin (Althunian *et al.,* 2017) or 10% (Senn, 2000). So for the purpose of this study, we decided to reduce the non-inferiority margin used in the pilot study to 10%, to be as close as most pharmaceutical trials.

6. Survey results

The setup of the survey has already been described in the previous section and as already mentioned, it was critical for the model's viability to have the same responses in each of the questionnaires to avoid effecting the analysis. As a result, we received a similar number of responses across the nine questionnaires without major differences between them (see Table 2).

Table 2. Distribution of responses among the nine surveys

Survey number	Total of responses
1	29 (13.3%)
2	24 (11.0%)
3	25 (11.5%)
4	23 (10.6%)
5	26 (11.9%)
6	21 (9.63%)
7	20 (9.17%)
8	30 (13.8%)
9	20 (9.17%)

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

6.1 Participants' description

From the 228 Twitter users' responses, some were excluded. Two participants were disqualified because their response time exceeded 70 minutes. Four participants were disqualified because they did not recognize any of the tweets as being original. Four more participants were ruled out because they were under the age of 18. The texts used in the surveys covered topics familiar to adults, for example, bureaucracy, which would be unusual to young people. This meant that any answers given by minors were discarded. Finally, 218 people contributed to the analysis database.

In terms of participant age, we found a similar distribution among age groups except for people over 65 years old, which is likely due to this group's lack of experience or usage of Twitter. We can then say that we have a true picture of the social media age engagement and is in line with our study as we want to evaluate end-user perceptions (see Table 3).

Table 3. Distribution of responses among age groups

Age group	Total responses
From 18 to 25	30 (13.8%)
From 25 to 35	50 (22.9%)
From 35 to 45	53 (24.3%)
From 45 to 55	46 (21.1%)
From 55 to 65	34 (15.6%)
More than 65	5 (2.29%)

In terms of Galician linguistic competence, all participants were native speakers. In terms of written competence, we considered their level of certified education. Only 3% of the sample population had no accredited education. 15.6% of participants were accredited with primary education. Participants with secondary education or a university degree predominated (61.9%), followed by participants with a specific language training such as professors, reviewers, etc. (19.3%). When asked about Galician in their daily lives, the majority of responders claimed that they almost always use Galician (40.8%). Then, there is another significant proportion of people who use primarily Spanish (22.5%), followed by those who use both languages interchangeably (17.9%) or primarily Galician (15.6%). Only a few people never use Galician (3.21%). These percentages are similar to those obtained when participants were asked in which language they write their tweets. Finally, when asked about their education in Galician, only 12.4% of the participants had not received any formal education in Galician. These were control questions used to determine the profile of our respondents, as well as to validate their answers based on their knowledge.

6.2 Description of the tweets and responses collected

Tweets originally written in Galician and tweets translated with our NMT engine were distributed equally but in different proportions across the 9 questionnaires. In Table 4, we can advance the direct responses collected regarding the naturalness of each of the tweets.

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

Table 4. Tweets judged natural

Type of Tweet	NMT in Galician	Originally written in Galician
Short sentence	76.3%	71.4%
Thread	81.9%	83.7%
Long sentence	88.7%	81.9%
Short paragraph	72.4%	82.0%
Long paragraph	80.7%	67.1%

A priori, we might expect that tweets written in Galician (red column of Figure 1) would be perceived as more natural. However, it is noteworthy that they are never perceived as 100% natural and that the NMT tweets figures (blue columns) are close to the originally written texts, sometimes exceeding them.

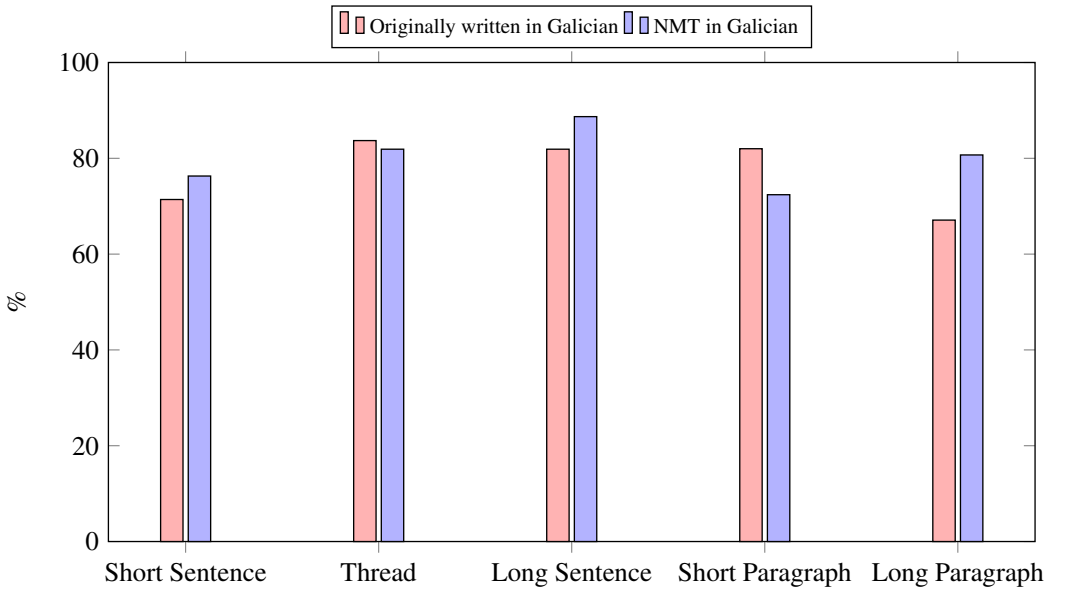

Figure 1. Tweets judged natural.

Absolute figures only show a tendency, which has to be proven or disproved. In this study, the principle of non-inferiority is used as the statistical method of analysis for this purpose, and the results are presented in the following sections.

6.3 Bivariate analysis

We compared the tweet naturalness rates with the different variables (questionnaires, age group, language knowledge, language use, language use on Twitter, and education) to search for anomalies that could lead us to failures in the model. We only found significant differences in two variables: the different questionnaires and the group age. Some questionnaires gave a higher rate of acceptance than others, but the rates obtained were not very distant, which could be due to a

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

better translation of the NMT engine or the lack of relevant errors. On the other hand, groups aged 18 to 25 (85.3%) and 35 to 45 (83.0%) were the most likely to accept a machine-translated tweet as the original. People aged 45 to 55, on the other hand, were the most demanding.

6.4 Non-inferiority analysis

The experiment design is reflected in this non-inferiority model. The model assumes that different questionnaires include shared tweets, so some tweets are evaluated in more than one questionnaire. Furthermore, each tweet falls into a unique category of all possible categories (mathematically represented as J). As we wanted to be as thorough as possible in our analysis by taking this into account, with this model, we can state the global non-inferiority of our engine based on a p-value of 0.024 (see Table 5).

Table 5. Global non-inferiority rate

Contrast	Odds Ratio	SE	df	Null	Z-Ratio	p-value
No/Yes	1.266	0.231	Inf	0.884	1.972	**0.024**

Through the model used, we contrasted whether the proportion of MT tweets perceived as natural is non-inferior compared to the proportion of originally written tweets considered as natural. In other words, it consists of demonstrating that the proportion of natural tweets in machine-translated tweets is higher than the proportion of natural tweets in original tweets, minus a non-inferiority margin.[g] The unilateral p-value and the statistic Z confirms the non-inferiority.

The non-inferiority test is designed to show that the observed odds ratio is greater than the odds ratio that would occur under the null hypothesis. The statistic Z presented in the table is the score corresponding to this test of non-inferiority, as is the p-value shown as 0.024. If the p-value is below 0.05, the non-inferiority test refutes the hypothesis that the NMT is worse and supports the hypothesis that the NMT is non-inferior. To complete the interpretation, the odds ratio is shown. This is a measure of the differences between the percentages that result from the model used (see previous section). The odds ratio is also shown under the null hypothesis.

To gain a deeper understanding of our engine's performance strengths and weaknesses, we replicated this analysis while accounting for the variable of tweet type. Taking a closer look at the different length of the tweets, we found a different behavior and two clear groups.

As shown in Table 6, NMT tweets consisting of short paragraphs and threads had the lowest probability of being considered not inferior to originally written tweets (short paragraphs: p-value of 0.952 and z-ratio of −1.669; threads: p-value of 0.589 and z-ratio of −0.226), as can also be seen in Figure 1. Indeed, this figure shows that originally written tweets were perceived as more natural than NMT tweets. One reason for explaining this in tweets formed by short paragraphs could be that those tweets had a more cryptic style, similar to a telegram, jeopardizing the naturalness and fluency, and that paratextual context cannot compensate. In terms of threads, we can expect that even if they were treated as a complete section in our preprocessing, our engine is not context-aware, so cohesion will be disregarded. In other words, providing a longer and more coherent source text that contextualizes each sentence does not necessarily result in a more coherent MT. Moreover, following the MQM framework, more minor errors of fluency, precision, and terminology were found in this category.

[g]We could write this condition as $p_{mt} > (p_{original} - m)$, where p_{mt} is the proportion of natural tweets in the machine-translated set, $p_{original}$ is the proportion of natural tweets from the original tweets, and m is the margin.

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

Table 6. Non-inferiority rate by type of tweet

Contrast	Tweet Type	Odds Ratio	SE	df	Null	Z-Ratio	p-value
No/Yes	Short sentence	1.297	0.505	Inf	0.876	1.007	**0.157**
No/Yes	Thread	0.810	0.341	Inf	0.891	−0.226	**0.589**
No/Yes	Long sentence	2.222	0.935	Inf	0.887	2.181	**0.015**
No/Yes	Short paragraph	0.453	0.183	Inf	0.890	−1.669	**0.952**
No/Yes	Long paragraph	3.079	1.195	Inf	0.866	3.268	**0.001**

In contrast, tweets formed by a long sentence or by a long paragraph were accepted as the most natural (long sentence: p-value of 0.015 and z-ratio 2.181; long paragraph: p-value of 0.001 and z-ratio of 3.268). Surprisingly, tweets formed by a long paragraph were accepted more than the same category of the tweets directly written in Galician, a result which was also found in the previous pilot study. One possible reason for such results could be that participants have more context and a complete text in long NMT sentences and paragraphs, whereas originally written texts could be shortened trying to adjust to the character restriction of Twitter. This would mean that, by translating sentences with a greater degree of syntactic complexity, the engine will achieve a more coherent translation. In other words, each phrase of the sentence is contextualized by the rest of the sentence, so that the result of MT allows for a better perception of naturalness. Tweets consisting of short sentences were closer to being considered non-inferior than tweets consisting of short paragraphs (p-value of 0.157 and z-ratio of 1.007). NMT short tweets were perceived as slightly more natural than originally written texts (see Table 4). The acceptance rate was higher than in the previous study, which could be a direct result of the NMT engine's retraining. Nonetheless, this short-sentence weakness is a well-known drawback of MT systems, which is usually explained by a lack of context. There is no clear indication that the text is natural here, especially if these tweets are presented alone. As a matter of fact, we took a closer look at these tweets to see if we could draw any additional conclusions. These tweets are grammatically correct, but they appear to be closer to typical Spanish syntax than long texts because there are fewer unique expressions to distinguish them. Due to the close relationship between Spanish and Galician, this may be interpreted as less natural in Galician, hence, we can declare the significant non-inferiority only in two categories: long sentences and long paragraphs. For the other three categories, differences are not significant, so non-inferiority cannot be stated.

6.5 Variations to the model

We experimented with the way responses and the nature of the tweet were interpreted to gain a better understanding of our analysis model. We repeated our analysis without taking into account the random effects. First, we modified one parameter of our formula: each participant responds to a block of 10 tweets; however, the blocks of tweets are not identical across the nine questionnaires. So, the model still takes into account the participant effect but considers each tweet as unique. Here, p is the proportion of success (positives, in our case the tweets that are perceived as natural), α is the independent term of the model, β is the coefficient associated with the variable Original, γ_j is the coefficient associated with category j of the variable Type of Tweet, $\beta\gamma_j$ is the coefficient associated with category j of the interaction between Original and Type of Tweet, and a_i is the random effect associated with items, with variability σ^2_{item}. Although the p-values in this analysis were lower than in the previous one, the non-inferiority was achieved in the same tweet categories.

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

This is the modified formula:

$$N_{ij} \sim \text{Bin}(1, p_{ij})$$

$$[H] \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + \alpha_i^{Id} + \beta O_Y + \gamma_j \text{Type}_j + (\beta\gamma)_j (O_Y \times \text{Type}_j) \tag{3}$$

$$\alpha_i^{Id} \sim N(0, \sigma_{(\alpha^{Id})^2}) \quad \text{for } i = 1, \ldots, I$$

Second, we went one step further and removed the participant effect from the formula. This way we were able to evaluate our responses using a model that assumes all answers and tweets are unique.

$$N_{ij} \sim \text{Bin}(1, p_{ij})$$

$$[H] \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + \beta O_Y + \gamma_j \text{Type}_j + (\beta\gamma)_j (O_Y \times \text{Type}_j) \tag{4}$$

In this case, the *p*-values were low enough to state non-inferiority in tweets composed of short sentences (see Table 7).

Table 7. Non-inferiority rate per type of tweet without random effects

Contrast	Tweet Type	Odds Ratio	SE	df	Null	Z-Ratio	p-value
No/Yes	Short sentence	1.239	0.337	Inf	0.876	1.273	0.102
No/Yes	Thread	0.815	0.257	Inf	0.890	−0.281	0.611
No/Yes	Long sentence	2.233	0.700	Inf	0.887	2.945	0.002
No/Yes	Short paragraph	0.458	0.132	Inf	0.889	−2.297	0.989
No/Yes	Long paragraph	3.081	0.827	Inf	0.865	4.731	0.000

With these changes, we hoped to demonstrate that our model took into account the nature of our design in order to interpret our responses consistently. Nonetheless, this example provides researchers with additional options for interpreting their database, particularly if they do not have a large sample of answers or tweets. The model can be adapted to the particular needs of their language combination and sample population.

7. Conclusions

This article was based on the assumption that translations of the same text can differ, so MT quality should be evaluated following a different principle than the sentence-based comparison. Hence, we proposed evaluating MT using the statistical principle of non-inferiority to determine whether or not texts obtained through MT were less natural than texts created in that language. Naturalness is understood to mean the usual use of the language, with common phrases, no grammatical errors, a fluid style, and no phrasing strongly influenced by other languages, such as Spanish. We can conclude that the method is effective for gathering end-user perceptions of NMT. Although non-inferiority evaluations are commonly used in pharmacological tests, we have demonstrated that they can also be used in MT tests. When comparing machine-translated tweets to originally written tweets in Galician, our method focuses on identifying perception differences and successfully tested its feasibility. In contexts where the NMT is already correct in terms of accuracy and fluency, the non-inferiority approach is an appropriate method to evaluate the success of the NMT

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

without having to compare the translations among themselves. This method will help determine whether translations are contextually successful not only in correctness but also in naturalness, particularly relevant in creative and communicative situations.

We believe that this method can be used to evaluate naturalness for MT both in high and low-resource languages. In particular in the case of low-resource languages, it would be a valid method when the low-resource language speaker is the final user of the engine. This is especially common in low-resource languages (Ranathuga *et al.,* 2023) where MT is another resource to promote language use and Internet presence through the use of fit-for-purpose engines (Van Edgom and Pluymaekers, 2019). Our method, on the other hand, can be used in the translation industry to help language service providers (LSPs) determine whether an NMT with high fluency and accuracy requires post-editing or can be published unattendedly in multilingual contexts and instant publications (social media, press notes, blog entries, help articles. . .). In creative contexts, in which texts have a higher expressive and phatic function, we believe that it is necessary to combine it with the current methodologies in order to give a holistic view of the quality of the engine.

Our research also discovered that there are differences in naturalness acceptance across age groups. Younger participants showed higher acceptance rates than middle-aged and older respondents. Future research should investigate why younger language users are less reluctant to accept language changes than older language users. However, this detected difference may just be a consequence of the evaluated genre (tweets) and the context, and it might have not been identified if the genre and the context used had been traditional ones (i.e., those of longer and more cohesive texts). Digital genres like typical social media texts (such as Twitter) perceive texts in their linguistic and paratextual contexts. A replication of this test with a genre for which the paratextual elements may be less relevant may not be able to generate such differences between the age cohorts.

Finally, in a low-resource language context, our method of back-translation training allows us to dispel the doubts raised by researchers who have relied only on automatic metrics to measure the effectiveness of back-translation (e.g., Poncelas *et al.,* 2018). Our study not only confirms that careful back-translation (of a homogeneous collection of texts sharing textual characteristics as genre) improves BLEU but can also evaluate and measure MT success in terms of end-user perception in social media contexts. We obtained clear, meaningful information about the NMT engine's strengths and weaknesses. The model-based estimates showed us where we needed to improve our engine or rework the raw MT. Tweets with short sentences, especially when presented in a paragraph, and threads require more work to sound natural. In these cases, the short textual context seems to negatively influence users' perception. At the other end of the spectrum, we would also like to point out that our engine is not inferior to originally written texts in long sentences and paragraphs. Regarding the results of the pilot study and after improving our engine, a second category of tweets (long paragraph) proved to be significantly non-inferior. Our corpus allowed for training our NMT engine to achieve a natural output, especially since no post-editing was performed.

References

Agresti A. (2015). *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons.

Althunian T. A., de Boer A., Groenwold R. and Klungel O. H. (2017). Defining the noninferiority margin and analysing noninferiority: an overview. *British Journal of Clinical Pharmacology* **83**(8), 1636–1642.

Bhardwaj S., Hermelo D. A., Langlais P., Bernier-Colborne G., Goutte C. and Simard M. (2020). *Human or Neural Translation?*. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain, pp. 6553–6564.

Casares Berg H. and Monteagudo Romero H. (2021). La juventud gallega y las pantallas. Una aproximación sociolingüística. In *Quaderns del CAC 47*, vol. XXIV, pp. 39–49.

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

Castilho S., **Moorkens J.**, *et al.* (2017). *A comparative quality evaluation of PBSMT and NMT using professional translators*. In *Proceedings of Machine Translation*, Summit XVI: Research Track.

Chatzikoumi E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, **26(2)**, 137–161.

do Campo M. and Sánchez-Gijón P. (2022). *Evaluating NMT: Superior, Inferior, or Equivalent to Texts Originally Written by Humans*. In *Proceedings of the New Trends in Translation and Technology*, Rhodes, July.

do Campo M. and Sánchez-Gijón P. (2023). *A social media engine for a low-resourced combination*. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, 12-15 June*, Tampere, pp. 269–275.

Freitag M., **Vilar D.**, **Grangier D.**, **Cherry C. and Foster G.** (2022). A natural diet: Towards improving naturalness of machine translation output. In *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland, pp. 3340–3353.

Hassan H., **Aue A.**, **Chen C.**, **Chowdhary V.**, **Clark J.**, **Federmann C.**, **Huang X.**, **Junczys-Dowmunt M.**, **Lewis W.**, **Li M.**, **Liu S.**, **Liu T.**, **Luo R.**, **Menezes A.**, **Qin T.**, **Seide F.**, **Tan X.**, **Tian F.**, **Wu L.**, **Wu S.**, **Xia Y.**, **Zhang D.**, **Zhang Z. and Zhou M.** (2018). Achieving human parity on automatic Chinese to English news translation.

Gino D. (2018). In Human vs. *Machine Translation, Compare Documents, Not Sentences. In Slator, Language Industry Intelligence 2018.*

Görög Attila (2014). Quantifying and benchmarking quality: the TAUS dynamic quality framework. *Tradumàtica 12*, 443–454.

Junczys-Dowmunt M., **Grundkiewicz R.**, **Dwojak T.**, **Hoang H.**, **Heafield K.**, **Neckermann T.**, **Seide F.**, **Germann U.**, **Aji A.**, **Bogoychev N.**, **Martins A. and Birch A.** (2018). *Marian: Fast Neural Machine Translation in C++*. In *Proceedings of ACL 2018, System Demonstrations*, pp. 116–121.

Kenny D. (2022). *Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence. (Translation and Multilingual Natural Language Processing 18)*. Berlin: Language Science Press.

Kreutzer J., **Bastings J. and Riezler S.** (2019). *Joey NMT: A Minimalist NMT Toolkit for Novices*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pp. 109–114.

Läubli S., **Sennrich R. and Volk M.** (2018). *Has machine translation achieved human parity? A case for document-level evaluation*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 4791–4796.

Leppänen S., **Kytölä S.**, **Westinen E. and Peuronen S.** (2020). *Social Media Discourse, (Dis)identifications and Diversities*. Routledge Studies in Sociolinguistics. Routledge.

Lommel A. and Melby A. (2018). *Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the 21st Century)*. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, Boston.

López Pereira A. (2018). *Determining translators' perception, productivity and post-editing effort when using SMT and NMT systems*. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, Alacant, Spain, 28 May 2018, Universitat d'Alacant.

Martindale M., **Duh K. and Carpuat M.** (2021). *Machine Translation Believability*. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, 88-95, Online*, Association for Computational Linguistics.

Molina Nadal A. (2020). Ensayos clínicos de no inferioridad. *FMC - Formación Médica Continuada en Atención Primaria* **27**(7), 345–348.

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

Poncelas A., **Shterionov D.**, **Way A.**, **Maillette de Buy Wenniger G. and Passban P.** (2018). Investigating backtranslation in neural machine translation. In *In arXiv*, 1804.06189.

Ranathunga S., **Lee E.**, **Skenduli M. P.**, **Shekhar R.**, **Alam M. and Kaur R.** (2023). Neural machine translation for low-resource languages: a survey.*In arXiv e-prints.*

Sánchez-Gijón P. (2014). Research in translation and quality, a two way street. *Revista Tradumàtica: tecnologies de la traducció* **12**, 437–442.

Sánchez-Gijón P., **Moorkens J. and Way A.** (2019). Post-editing neural machine translation versus translation memory segments. *InMachine Translation* **33**(1-2), 31–59.

Senn S. (2000). Consensus and controversy in pharmaceutical statistics. *Journal of the Royal Statistical Society. Series D (The Statistician)* **49**(2), 135–176.

Tan X., **Longyin Z. and Guodong Z.** (2022). Discourse cohesion evaluation for document-level neural machine translation. *arXiv preprint arXiv*: 2208.09118.

Temnikova I., **Orasan C.**, **Corpas Pastor G.** and Mitkov. (2019, *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, Varna, Bulgaria, pp. 75-81, September 5-6.

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

Tunes da Silva G., **Logan B. and Klein J.** (2009). Methods for equivalence and noninferiority testing. *Biology of Blood and Marrow Transplantation : Journal of the American Society for Blood and Marrow Transplantation* **15**(Suppl. 1), 120–127.
Van Edgom G. and Pluymaekers M. (2019). Why go the extra mile? How different degrees of post-editing affect perceptions of texts, senders and products among end users. *The Journal of Specialised Translation* **31**, 158–176.
Zappavigna Michele (2012). *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. London: Bloomsbury.
Zheng W., **Wang W. Liu**, **Zhang D.**, **Zeng C.**, **Deng Q.**, **Yang Y.**, **He W.**, **P** and **Xie T.** (2019). *Testing untestable neural machine translation: An industrial case*. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, IEEE.
Zuur A., **Ieno E.**, **Walker N.**, **Saveliev A. and Smith G.** (2009). *Mixed Effects Models and Extensions in Ecology with R*, vol. 574. New York: Springer.

A. Tweet examples

Figure A1. Machine-translated short-sentence tweet.

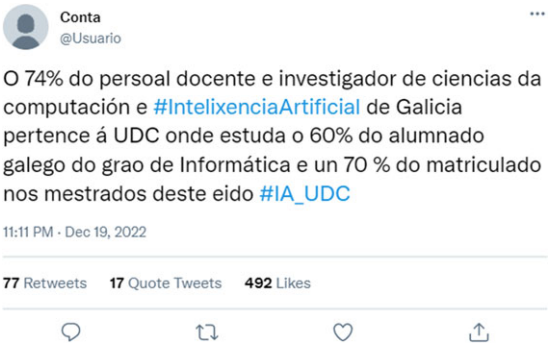

Figure A2. Machine-translated long-sentence tweet.

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

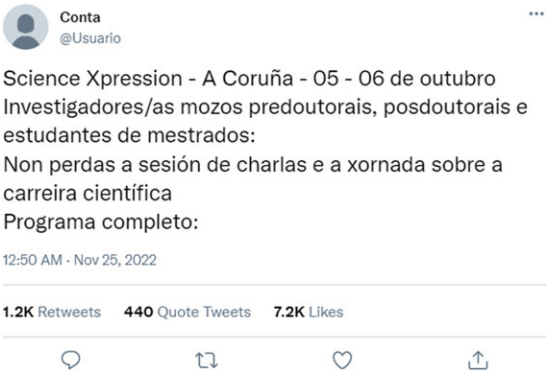

Figure A3. Machine-translated short paragraph tweet.

Figure A4. Machine-translated long paragraph tweet.

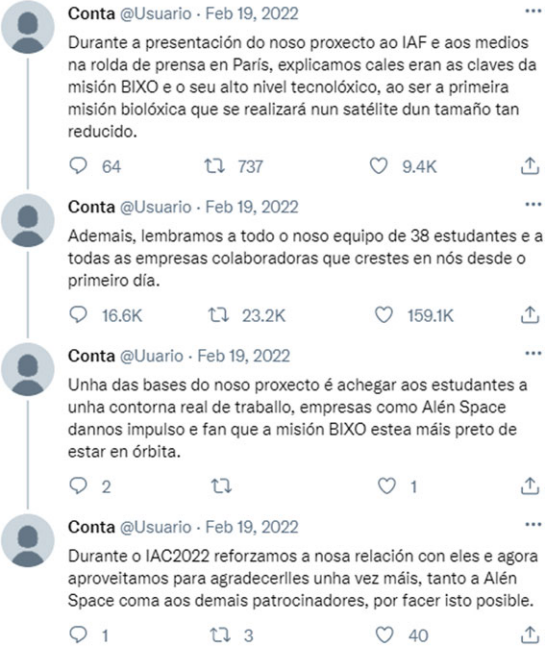

Figure A5. Machine-translated thread.

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

Figure A6. Originally written short-sentence tweet.

Figure A7. Originally written long-sentence tweet.

Figure A8. Originally written short paragraph tweet.

Figure A9. Originally written long paragraph tweet.

https://doi.org/10.1017/nlp.2024.4 Published online by Cambridge University Press

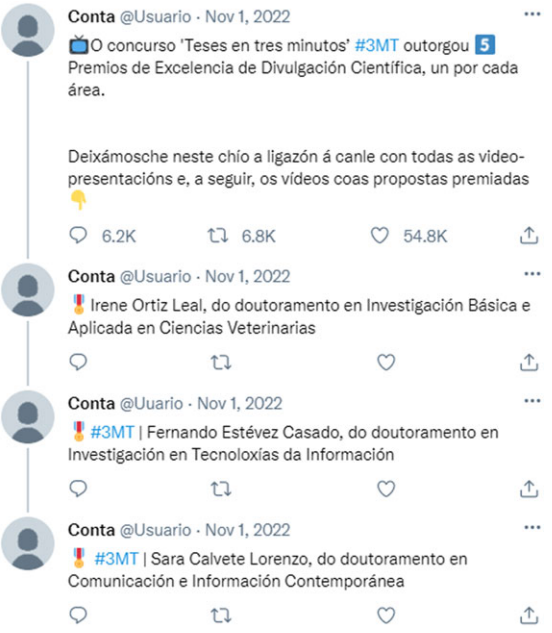

Figure A10. Originally written thread.

Cite this article: do Campo Bayón M. and Sánchez-Gijón P. Evaluating NMT using the non-inferiority principle. *Natural Language Processing* https://doi.org/10.1017/nlp.2024.4