information

Article

Machine Translation for Open Scholarly Communication: Examining the Relationship between Translation Quality and Reading Effort

Lieve Macken [1,*], Vanessa De Wilde [2] and Arda Tezcan [1]

[1] LT³ Language and Translation Technology Team, Faculty of Arts and Philosophy, Ghent University, Groot-Brittanniëlaan 45, B-9000 Ghent, Belgium; arda.tezcan@ugent.be
[2] MULTIPLES —Research Centre for Multilingual Practices and Language Learning in Society, Faculty of Arts and Philosophy, Ghent University, Groot-Brittanniëlaan 45, B-9000 Ghent, Belgium; vanessa.dewilde@ugent.be
* Correspondence: lieve.macken@ugent.be

Abstract: This study assesses the usability of machine-translated texts in scholarly communication, using self-paced reading experiments with texts from three scientific disciplines, translated from French into English and vice versa. Thirty-two participants, proficient in the target language, participated. This study uses three machine translation engines (DeepL, ModernMT, OpenNMT), which vary in translation quality. The experiments aim to determine the relationship between translation quality and readers' reception effort, measured by reading times. The results show that for two disciplines, manual and automatic translation quality measures are significant predictors of reading time. For the most technical discipline, this study could not build models that outperformed the baseline models, which only included participant and text ID as random factors. This study acknowledges the need to include reader-specific features, such as prior knowledge, in future research.

Keywords: machine translation quality; open scholarly communication; self-paced reading; reading effort

Citation: Macken, L.; De Wilde, V.; Tezcan, A. Machine Translation for Open Scholarly Communication: Examining the Relationship between Translation Quality and Reading Effort. *Information* **2024**, *15*, 427. https://doi.org/10.3390/info15080427

Academic Editor: Ivan Dunđer

Received: 3 June 2024
Revised: 5 July 2024
Accepted: 21 July 2024
Published: 23 July 2024

1. Introduction

In recent years, increasing attention has been paid to the inequalities caused by the monopoly of English as the language of science. Amano et al. [1] surveyed 908 researchers in the environmental sciences and found that non-native English speakers, particularly early in their careers, face greater challenges and spend more effort in conducting scientific activities in English than native English speakers. In addition to the negative impact on non-English-speaking researchers at an individual level, systematic reviews, which are considered a reliable form of research evidence, often neglect non-English literature as a source of important evidence [2]. Bowker et al. [3] carried out a systematic review to examine the use of machine translation (MT) tools in scholarly communication and to investigate whether such tools contribute to a more linguistically diverse ecosystem. The study found that while there is interest and positive attitudes towards these tools, the quality of MT tools is highly dependent on the data used to train the MT tool, and therefore, translation quality varies across language pairs, text types, and scientific domains. In addition, when custom-built prototypes were compared with general-purpose tools such as Google Translate, the custom-built tools designed for scholarly communication showed better performance on related tasks.

Several papers explored the usability of MT in the context of scholarly communication from different perspectives. O'Brien and colleagues [4] explored the potential of using MT and self-post-editing to support the academic writing process for authors writing in English as a foreign language. An experiment was conducted in which participants wrote

part of an abstract in their native language and part in English, and the native language section was then machine translated. The results suggest that MT and self-post-editing show promise as a tool that could help support academic writing in a foreign language without compromising quality, although further research is needed. Two studies [5,6] focus on improving the quality of MT from Russian to English for academic articles by pre-editing the source text. The first study demonstrated that by minimizing abbreviations, simplifying complex phrases, and ensuring grammatical accuracy, higher quality was obtained when comparing the original versus pre-edited text translations from Russian to English using the Google, Amazon, and DeepL machine translation systems. The second study developed fifteen pre-editing rules based on common negative translatability indicators in Russian life sciences texts. When these rules were applied, over 95% of the sentences met the generally accepted threshold for publication-level quality with minimal post-editing. Roussis and colleagues [7] present the development of domain-specific parallel and monolingual corpora focused on scientific domains and their application in the fine-tuning of general-purpose neural MT systems. They built large corpora for the language pairs Spanish–English, French–English, and Portuguese–English, covering general scientific texts as well as the specific domains of cancer research, energy research, neuroscience, and transportation research. Their results suggest that domain adaptation through targeted collection of scientific data and model fine-tuning can effectively improve the performance of NMT for translating specialized scientific and academic texts. In this study, we evaluate the performance of three different MT engines, one of which was developed using a similar methodology to that described in [7]. In this study, we do not pre-edit or post-edit the texts but evaluate the usability of the machine-translated texts by means of reading experiments, which is a new contribution to the field.

In his 2021 paper, "Investigating how we read translations. A Call to Action for Experimental Studies of Translation Reception", Walker [8] stresses the need for more research on the reception of written translation. He highlights that while there is a considerable amount of research on the reception of audiovisual translation [9], less attention has been paid to the reception of written translation. Whyatt and colleagues [10] responded to this call and conducted a small exploratory study in 2023 to investigate the relationship between the translator's effort in producing a translation, the translation quality of the final translation, and the reader's reception effort. Twenty native Polish speakers were randomly assigned to read either a high-quality or a low-quality translation of a product description translated from English into Polish. The researchers fitted a linear mixed effects model to analyze the data. The results show that the translator's effort did not have a significant effect on the readers' reception effort. But there was a significant effect of translation quality on readers' reception effort, with participants showing greater effort when reading low-quality translations compared with high-quality translations. Despite the small data set (the source text consisted of 8 sentences and 162 words), the authors suggest that reading experience can be used to evaluate the effectiveness of translation decisions, particularly in terms of translation quality, but that further research is needed to explore the impact of translation errors and the severity of their effects on the reading experience of translated texts.

Today, translation is not only performed by human translators. In many scenarios, MT systems are used to generate translations, especially in scenarios where human translation (HT) is too expensive or simply not feasible. Eye-tracking has previously been used to evaluate the quality of MT output. In 2010, Doherty et al. [11] investigated whether eye movement data reflected MT quality. Ten native French speakers were instructed to read 50 machine-translated sentences of either excellent or poor quality. The study found that when participants read MT sentences that were rated poorly by human evaluators, the number of fixations and the average gaze duration increased. In 2020, Kasperavičienė et al. [12] examined the eye movements of 14 participants when reading a news article automatically translated from English into Lithuanian. According to the authors, the language of the source text was simple and intended for a general audience. The study found that there were more fixations and longer gaze duration for segments containing MT errors compared

with correct segments. In 2022, Colman et al. [13] collected eye movements of 20 Dutch-speaking participants reading an entire novel (Agatha Christie's The Mysterious Affair at Styles), comparing the published human translation with a machine translation. The participants alternated between reading the human translation and a machine translation, or vice versa. In line with earlier research, the average reading duration was slightly higher in the MT condition compared with the HT condition, as well as the number of fixations, and the average fixation duration. In 2023, Kapsere et al. [14] conducted a study using eye-tracking to investigate the reading process of professional and nonprofessional users of MT. The participants read a 371-word recipe that had been automatically translated from English into Lithuanian. The results show that professional users spent more time reading the machine-translated text and assessed the quality of the MT more critically than nonprofessional users.

Although the papers we discussed above work on different language pairs, use different MT architectures and technologies, and vary in the type and size of the texts that were read, there is a consensus that reading machine-translated output requires more processing effort compared with reading human-translated text. Most studies, however, have compared machine-translated texts with human-translated versions, compared high- and low-quality translations, or focused on specific errors in the machine-translated texts. We are not aware of any studies that have examined the reader's effort when processing the output of different MT systems of varying translation quality levels.

In recent years, the field of MT has seen rapid advancements, beginning with the introduction of encoder–decoder transformer models [15] and more recently with the development of decoder-only large language models (LLMs) such as, Mistral [16] and Llama 3 (https://ai.meta.com/blog/meta-llama-3/, accessed on 5 July 2024). While there is increasing interest in leveraging LLMs for MT due to their additional capabilities, such as instruction following [17], their use does not necessarily lead to superior performance in translation tasks, particularly in specialized domains (e.g., legal) or when translating into less-represented languages (e.g., Czech or Russian) [18]. Several toolkits are available to train MT models. Open-source NMT toolkits, such as OpenNMT [19] and Marian [20], allow users to train NMT models from scratch with existing bilingual data sets, providing flexibility while requiring substantial data and computational power. Additionally, open-source NMT models, such as Opus-MT [21] and Microsoft's NLLB (No Language Left Behind) [22], and LLMs, such as Llama 3 and Mistral, offer pretrained models that can be further fine-tuned with language- or domain-specific data sets, balancing ease of use with customization potential. Commercial off-the-shelf MT solutions, like Google Translate (https://translate.google.com/, accessed on 5 July 2024) and DeepL (https://www.deepl.com/en/translator, accessed on 5 July 2024), offer user-friendly, high-quality translation services without the need for any model training. However, they are limited in their ability to be customized for specialized domains, specific styles, or terminology. Some commercial toolkits, such as ModernMT (https://www.modernmt.com/, accessed on 5 July 2024) and Google Cloud AI (https://cloud.google.com/, accessed on 5 July 2024), bridge this gap by allowing further fine-tuning of user-specific data, combining the benefits of high baseline performance and adaptability to particular needs.

MT output can be assessed through various methods, each providing insights into different aspects of translation quality. Existing MT assessment techniques can be categorized into two types: human evaluation and automatic evaluation. Human evaluation typically involves human annotators or evaluators who assess translations based on predefined criteria. To this end, common human evaluation techniques consist of (i) direct assessment, where MT output is scored on a predefined scale (e.g., from 0 to 100); (ii) ranking, where translations produced by different systems are compared and ranked from best to worst; (iii) intrinsic evaluation of fluency (i.e., well-formedness of translation in target language) or accuracy (i.e., to what extent source text meaning is conveyed in the target text); and (iv) fine-grained error annotation, where common errors in MT output are identified and categorized on the word/phrase level. Error annotation tasks are based on existing MT

error taxonomies, such as the MQM [23] or the SCATE taxonomy [24]. Notably, both the MQM and the SCATE taxonomies categorize MT errors hierarchically within accuracy and fluency as the two main error categories. While direct assessment, ranking, and intrinsic evaluation techniques provide insights into the overall quality of MT output from different perspectives, fine-grained error annotation allows for a detailed analysis of translation quality, pinpointing error patterns and determining the relative strengths and weaknesses of different MT systems, according to various linguistic aspects.

Despite that human evaluation offers a nuanced and more detailed linguistic assessment of MT quality, this approach is also often resource-intensive and time-consuming, which necessitates the use of automatic evaluation methods to provide efficient and cost-effective assessments, particularly in large-scale evaluation tasks or real-time monitoring scenarios. Automatic evaluation of MT involves the use of computational techniques to assess the quality of translated texts without human involvement. Unlike human assessment techniques, which compare source texts with machine-translated texts, traditional automatic evaluation metrics often rely on reference translations (i.e., gold-standard translations) and assign a final quality score to a given MT output based on the degree of similarity or divergence between them.

Existing automatic evaluation metrics vary in their methods for measuring the similarity between MT output and reference translations. While some metrics, such as BLEU [25] and METEOR [26], rely on word-level n-gram precision, others, such as chrF [27], utilize character-level n-gram F-score, and some like TER [28] calculate word-level edit distance. More recently developed metrics that rely on neural-based machine learning techniques, such as BERTScore [29], BLEURT [30], and COMET [31], focus on computing the similarity of sentence-level vector representations. Moreover, certain neural-based metrics are tailored for specific evaluation tasks; for example, COMET, which additionally incorporates source-text information, was specifically trained to predict various human judgments, including post-editing effort, direct assessment, or translation error analysis [31].

In many evaluation scenarios where reference translations are not consistently available, the dependency on such reference translations restricts the usability of automatic evaluation metrics. Consequently, there has been a growing interest directed toward developing quality estimation (QE) systems, metrics that aim to assess MT quality in the absence of reference translations, relying solely on the source text and MT output. Examples of recently developed neural-based QE metrics include COMETKiwi [32] and MetricX [33]. Besides enhancing the applicability of automatic metrics, by comparing source text with MT output, QE systems can be argued to capture different characteristics of MT quality compared with reference-based metrics.

While automated evaluation is generally faster and more cost-effective than human evaluation, their performance is often measured by to what extent they emulate human assessment. Hence, a reliable metric is often characterized by its ability to strongly correlate with human judgments [34]. Viewed from this standpoint, neural-based automatic evaluation (and QE) metrics continue to stand out as the state-of-the-art approach for automatic quality assessment, having achieved notably higher correlations with human assessments compared with non-neural metrics across different domains and language pairs in recent years [35,36]. In addition to employing advanced machine learning techniques, the effectiveness of neural metrics can also be attributed to their superior ability to capture semantic similarity between texts. By using word and sentence embeddings, these metrics are not confined to surface form comparisons with the reference translation, unlike lexical-based metrics, allowing them to more effectively identify semantically related translations (e.g., paraphrases and synonyms) [37]. On the other hand, while neural-based metrics generally offer better performance than non-neural metrics, they require a pre-trained language model, human-labeled data, and additional training, which limits their use for low-resource languages [37]. In addition to enabling comparisons with past research to some degree, as Lee et al. [37] further argues, this is one reason why lexical-based metrics like BLEU continue to be commonly used for MT assessment.

Current Study

This study is part of the 'Translations and Open Science' project (https://operas-eu.org/projects/translations-and-open-science/, accessed on 5 July 2024), which explores the possibility of using MT to disseminate research results in different languages. More specifically, it is part of a larger study on the evaluation of MT in the context of scholarly communication, focusing on the language pair English–French in three scientific disciplines:

- D1: Human Mobility, Environment, and Space (Social Sciences and Humanities, translated from French into English)
- D2: Neurosciences (Life Sciences, translated from English into French)
- D3: Climatology and Climate Change (Physical Sciences, translated from English into French)

Three different engines were selected for evaluation: two commercial systems (DeepL and ModernMT) and one open-source system (OpenNMT), which was trained on various data sets [38]. ModernMT was further customized by uploading a domain-specific translation memory. The OpenNMT system was trained from scratch on publicly available data from the OPUS repository [39] and fine-tuned on collected in-domain data and the SciPar data set [40]. In the larger study, automatic evaluations on held-out test sets revealed that MT quality varied across MT systems, with DeepL achieving the best performance and ModernMT the second best. For more information, we refer to the reports that are publicly available.

The project evaluated the usability of the raw MT output in three different use cases: (i) researchers specialized in the domain in question using MT to translate publications, as a writing aid, or for gisting purposes; (ii) professional translators using MT for post-editing to speed up the translation task; and (iii) nonexpert readers using MT to get an idea of the content of a scientific publication. This study focuses on the assessment of MT quality for nonexpert readers. The other use cases are described in Fiorini et al. [38]. Detailed reports and resources are available in a Zenodo repository (https://zenodo.org/records/10972872, accessed on 5 July 2024).

To assess the usability of the MT output for nonexpert readers, reading experiments were combined with automatic evaluation and fine-grained error annotations using the MQM framework [23]. The whole project was on a very tight schedule, which did not allow for time-consuming eye-tracking studies. Instead, self-paced reading experiments were set up [41]. Participants read translations in a cumulative, self-paced reading view, where each key press revealed the next sentence, while the previous sentences of the text remained in view. Although less fine-grained than eye-tracking, this set-up allowed us to collect reading times per text.

In this study, we were not only interested in how readers rated the quality of MT but also in the relationship between translation quality (either measured by fine-grained error annotations or by automatic metrics) and the readers' reception effort, measured by reading times. More specifically, we were interested in whether different measures of translation quality could predict reading times.

2. Materials and Methods

2.1. Data Collection

For each discipline, short text excerpts of 120–200 words were selected from the evaluation sets created in the project. The evaluation sets contain held-out texts that were not included in the training materials for the OpenNMT engine or in the translation memory used to adapt the ModernMT system. For discipline 1, the texts were selected from three different text types (thesis abstracts, journal articles and journal abstracts). For both disciplines 2 and 3, it was rather difficult to select suitable texts for nonexpert readers as both disciplines contained highly technical texts. No abstracts were available for discipline 2. We selected text fragments from Cochrane reviews (https://www.cochranelibrary.com/, accessed on 5 July 2024), and excerpts from journal articles. As the texts had to be similar in length, coherent text fragments were occasionally manipulated by deleting intermediate

sentences. There were no full texts available for discipline 3. The text excerpts were selected from different sources of abstracts (project, journal, and thesis abstracts). In addition, the human reference translations for discipline 3 were often too divergent from the source or occasionally incorrect. Therefore, we were forced to manually correct 3 sentences of the reference translation in one text of discipline 3. A summary of the source text characteristics for the French source texts of discipline 1 is presented in Table 1 and for the English source texts of disciplines 2 and 3 in Table 2.

Table 1. Source text characteristics of the selected text excerpts for discipline 1 (French originals).

Discipline	Text Id	No. Words	No. Sent.	Avg. Sent. Length	Text Type
D1	T1	137	5	27.4	Thesis abstract
D1	T2	145	8	18.1	Thesis abstract
D1	T3	146	8	18.3	Thesis abstract
D1	T4	154	8	19.3	Thesis abstract
D1	T5	222	9	24.7	Journal article
D1	T6	152	8	19.0	Journal article
D1	T7	171	7	24.4	Journal article
D1	T8	198	7	28.3	Journal article
D1	T9	164	6	27.3	Journal abstract
D1	T10	186	6	31.0	Journal abstract
D1	T11	179	6	29.8	Journal abstract
D1	T12	201	5	40.2	Journal abstract
D1	**Total**	**2055**	**83**	**24.8**	

Table 2. Source text characteristics of the selected text excerpts for disciplines 2 and 3 (English originals).

Discipline	Text Id	No. Words	No. Sent.	Avg. Sent. Length	Text Type
D2	T1	136	7	19.4	Cochrane review
D2	T2	134	6	22.3	Cochrane review
D2	T3	125	5	25.0	Cochrane review
D2	T4	140	5	28.0	Cochrane review
D2	T5	204	8	25.5	Journal article
D2	T6	200	8	25.0	Journal article
D2	T7	201	8	25.1	Journal article
D2	T8	191	9	21.2	Journal article
D2	T9	132	6	22.0	Journal article
D2	T10	128	7	18.3	Journal article
D2	T11	123	6	20.5	Journal article
D2	T12	139	6	23.2	Journal article
D2	**Total**	**1853**	**81**	**22.9**	
D3	T13	143	5	28.6	Project abstract
D3	T14	142	5	28.4	Project abstract
D3	T15	112	5	22.4	Project abstract
D3	T16	131	6	21.8	Project abstract
D3	T17	108	4	27.0	Journal abstract
D3	T18	104	5	20.8	Journal abstract
D3	T19	123	5	24.6	Journal abstract
D3	T20	106	4	26.5	Journal abstract
D3	T21	119	6	19.8	Thesis abstract
D3	T22	138	6	23.0	Thesis abstract
D3	T23	137	5	27.4	Thesis abstract
D3	T24	127	6	21.2	Thesis abstract
D3	**Total**	**1490**	**62**	**24.0**	

Per discipline, the selected texts were divided into four different sets. For each set, we made sure that all conditions (Human Reference, DeepL, ModernMT, and OpenNMT) were evenly distributed (see the balanced design provided in Table 3). For each discipline, sixteen participants took part in the self-paced reading experiments, which means that four participants read all the texts of one set. All participants, aged between 24 and 58 years, were staff members of the Faculty of Arts and Philosophy at Ghent University and were highly proficient in the target language (English for discipline 1 and French for disciplines 2 and 3). All participants can be considered nonexpert readers, in the sense that they are used to reading academic articles, but in other scientific fields. For practical reasons, the experiments for disciplines 2 and 3 were combined in one session. Participants were allowed to take breaks during the experiment. All participants signed an informed consent form and received a financial reward of EUR 10 for discipline 1 and EUR 20 for the combined discipline 2 and 3 experiments.

Table 3. Balanced design used for the self-paced reading experiments. All participants read the texts of one specific set.

	Set 1	**Set 2**	**Set 3**	**Set 4**
T1	ModernMT	OpenNMT	Reference	DeepL
T2	OpenNMT	Reference	DeepL	ModernMT
T3	Reference	DeepL	ModernMT	OpenNMT
T4	DeepL	ModernMT	OpenNMT	Reference
T5	ModernMT	OpenNMT	Reference	DeepL
T6	OpenNMT	Reference	DeepL	ModernMT
T7	Reference	DeepL	ModernMT	OpenNMT
T8	DeepL	ModernMT	OpenNMT	Reference
T9	ModernMT	OpenNMT	Reference	DeepL
T10	OpenNMT	Reference	DeepL	ModernMT
T11	Reference	DeepL	ModernMT	OpenNMT
T12	DeepL	ModernMT	OpenNMT	Reference

The self-paced reading experiments were conducted using the Zep Experiment Control Application (version 2) (https://www.beexy.nl/zep/wiki/doku.php?id=home, accessed on 5 July 2024), a toolkit used for experimental psycholinguistics. Reading times were measured per sentence and aggregated at text level. During the experiments, participants were alone in a room to minimize distractions. The experiment started with a practice text to familiarize the participants with the task. Participants read the texts in a cumulative, self-paced reading view in which each key press revealed the next sentence while the previous sentences of the text remained in view. After each text, participants were asked to answer a multiple-choice comprehension question as an incentive to read the text attentively and answer a yes/no question to assess the perceived usability of the translation for gisting purposes.

To assess the translation quality of the machine translations, we used both manual and automatic metrics. A linguist used the MQM framework to annotate all errors in the machine-translated texts. The MQM core error taxonomy (https://themqm.org/the-mqm-typology/, accessed on 5 July 2024) and the MQM decision tree (https://themqm.org/error-types-2/decisiontree/, accessed on 5 July 2024) were used as the basis for the annotation guidelines. A fine-grained error annotation was carried out using all the main and subcategories of the MQM core taxonomy, with the exception of the two main error categories, 'Audience appropriateness' and 'Design and markup', which were not considered relevant for this study. An overview of the error categories of the MQM core taxonomy is shown in Figure 1.

As is common practice in the MQM framework, all errors have additionally been given a severity label, namely 'neutral', 'minor', 'major', and 'critical', to reflect the effect of a particular error on the usability of the text. The severity labels of the annotated

errors were used as penalties when calculating the MQM quality score for each given text (https://themqm.org/error-types-2/the-mqm-scoring-models/, accessed on 5 July 2024).

- **Terminology**
  - Inconsistent with terminology source
  - Inconsistent use of terminology
  - Wrong term
- **Accuracy**
  - Mistranslation
  - Overtranslation
  - Undertranslation
  - Addition
  - Ommision
  - Do not translate
  - Untranslated
- **Linguistic conventions**
  - Grammar
  - Punctuation
  - Spelling
  - Unintelligible
  - Character encoding
- **Style**
  - Organization style
  - Third-party style
  - Language register
  - Awkward style
  - Unidiomatic style
  - Inconsistent style
- **Locale conventions**
  - Number format
  - Currency format
  - Measurement format
  - Time format
  - Date format
  - Address format
  - Telephone format
  - Shortcut key
- **Audience appropriateness**
- **Design and Markup**

Figure 1. Overview of the MQM taxonomy.

To assist the annotation of terminology-related errors, domain-specific terms were annotated in the source sentences of the evaluation data prior to the annotation task. This annotation was carried out, firstly, in an automated manner using provided term lists and, secondly, manually by the same annotator responsible for the error annotation.

A set of annotation guidelines was developed, drawing upon MQM error definitions and examples, severity levels of errors, and the decision tree utilized for identifying and categorizing errors. The annotations were carried out using the LabelStudio toolkit (https://labelstud.io/, accessed on 5 July 2024), an open-source platform for data labeling. The annotation guidelines incorporated sample annotations and screenshots from LabelStudio for clarity and reference.

While annotating errors, each source sentence was presented to the annotator alongside three MT outputs (referred to as SRC, MT1, MT2, and MT3), and the errors in each MT output were annotated concurrently. To prevent any bias towards specific MT engines, the order of the MT outputs was randomized for each source sentence and recorded in a separate log file. This log file was utilized to automatically calculate the MQM quality scores after the annotation process. Figure 2 provides a screenshot from the annotation platform, which includes an example annotation.

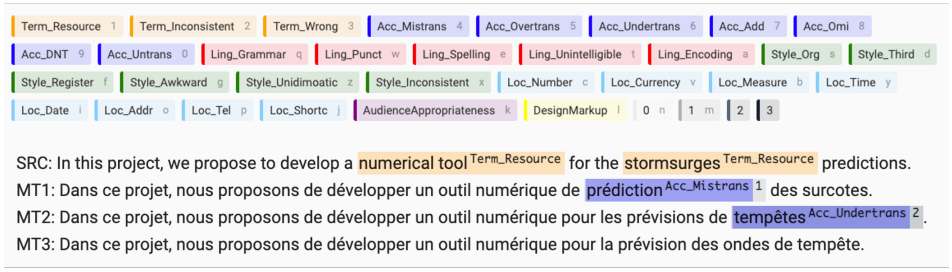

Figure 2. Example annotation on LabelStudio displaying error taxonomy and severity labels (**upper section**), alongside a source sentence paired with three translations from various MT systems, complemented by error category and severity annotations (**lower section**).

In the annotation example provided in Figure 2, two terms were identified in the source text ('numerical tool' and 'stormsurges'). 'Numerical tool' was correctly translated by all MT systems, but MT2 made an error translating the term 'stormsurges' (which refers to the rise in seawater level caused solely by a storm). As MT2 only provided a translation for the word 'storm' (Fr: 'tempête'), 'tempêtes' was labeled as a major 'accuracy-undertranslation'

error. The second error identified by the linguist was the French translation 'prédiction' instead of 'prévision'. Both translations are possible, but 'prévision' is the more specific word when referring to predictions made based on scientific or statistical analysis. The linguist labeled 'prédiction' as a minor 'accuracy-mistranslation' error.

For the automatic translation quality assessment, we opted for three reference-based neural metrics, namely BERTScore, BLEURT, and COMET. We used the MATEO web interface [42] (https://mateo.ivdnt.org/, accessed on 5 July 2024) to calculate all the automatic metric scores. Aside from the reference-based metrics, we computed QE scores using COMETKiwi (https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xl, accessed on 5 July 2024), which derives its scores from the information extracted from the source sentences and the MT output, without relying on reference translations.

As reading times are also likely to be influenced by general text characteristics, we additionally calculated the following features of the machine-translated texts:

- Average sentence length.
- Automated Readability Index (ARI). We chose ARI because it is the most comparable formula across both languages, as it is based only on the number of characters, words, and sentences in a text. We used the implementation that is available in the Textstat Python package version 0.7.3 (https://pypi.org/project/textstat, accessed on 22 January 2024)
- Moving Average Type–Token Ratio (MATTR) as a general measure of lexical diversity [43]. We used the implementation that is available in the Lexical-diversity Python package version 0.1.1 (https://pypi.org/project/lexical-diversity/, accessed on 22 January 2024) with the default window size of 50 words.
- The percentage of 3000 most frequent words (Perc-3KFreq) as a measure of lexical complexity. We made use of a custom lexical profiler [44], which relies on the subtitle-based frequency list SubtLex-UK [45] and its French equivalent Lexique [46].

2.2. Data Analysis

The data sets for each discipline were analyzed separately. In the first step, we analyzed the reading times, the answers to the comprehension questions, and the perceived usability of the whole data set (the machine translations and the reference translations). In the second step, we focused only on the machine-translated texts. We calculated descriptive statistics for all our measures of interest and fitted a series of linear mixed effects (LME) models in R with average reading time as the dependent (or response) variable.

Mixed-effects models simultaneously model random effects in addition to fixed effects. In our study, we included participant and text ID as random effects, as we expect both individual differences between participants and text-specific aspects to influence reading times. We built different models including three sets of variables as fixed effects (target text characteristics, manual translation quality scores, and automatic translation quality scores), as we expect that there is a relationship between target text characteristics (including translation quality) and reading times. In order to examine to what extent different measures of translation quality could predict reading times, different linear mixed-effect models were built and compared. As a baseline, we built a null model, which only contains the response variable and the random factors. This model is compared with more complex models in which predictor variables of three different categories are added:

- General text characteristics: Average sentence length, automated readability index, mean average type-token ratio, and the percentage of 3000 most frequent words.
- Manual translation quality metrics: MQM quality score and the number of critical errors.
- Automatic translation quality metrics: BERTScore, BLEURT, COMET, and COMETKiwi.

We visually inspected the reading times of all participants for outliers and decided to retain the outliers to ensure that the data remained as close to the original as possible and to maintain the integrity of the data set. Models were built using individual predictors, as well as combinations of predictors from the automatic or manual translation quality metrics set and the general text characteristics set. Since we built models with multiple

predictors whose scores are measured on different scales, we standardized all predictors before building the models. In addition, the average reading times per word were log-transformed in order to obtain a more normal distribution. For each discipline, we had 144 data points (16 participants, each reading 9 machine-translated texts of one specific set). Models with interaction between predictors were tested, but due to the limited size of the data set, these models did not converge.

The statistical software package R (version 4.4.0) [47] was used to analyze the data. We used the lme4 package [48] and the lmerTest package [49] to perform linear mixed effects analyses. Model comparison was carried out with the ANOVA function from the base stats package and the performance package [50]. The functions `check_singularity()`, `check_heteroscedasticity()`, `check_normality()`, and `check_collinearity()` of the performance package were used to carry out diagnostics tests. For all models, we report the coefficient estimates, standard errors, degrees of freedom, and t- and *p*-values of the fixed effects. For model comparison, we also report the marginal and conditional R^2 and the corrected Akaike Information Criterion (AICc) [51].

3. Results

3.1. General Analysis

Figure 3 presents average reading times per word, expressed in milliseconds per word, per discipline, and per condition. In general, overall average reading times were much higher than the reading times reported in the literature. In 2019, Brysbaert [52] conducted a meta-analysis of 190 studies based on 18,573 participants and estimated that the average silent reading rate for adults in English is 238 words per minute for nonfiction, which corresponds to 252 milliseconds per word. Across all conditions, the texts of discipline 3 (Climatology and Climate Change—Physical Sciences) required considerably more processing effort than the texts of the other disciplines, with an average reading time of 472 milliseconds per word. The average reading times of the other disciplines were more on par with averages of 416 and 398 milliseconds per word for resp. discipline 1 (Human Mobility, Environment, and Space—Social Sciences and Humanities) and discipline 2 (Neurosciences–Life Sciences). Only for discipline 1, the human reference translations were read faster than the machine translations. In addition to the higher processing time, participants also gave 52% incorrect answers to the multiple-choice comprehension questions in discipline 3, whereas the results for disciplines 1 (11% incorrect answers) and 2 (25% incorrect answers) were much better. The perceived usability scores of the translations for gisting purposes were good for disciplines 1 and 2 (81%) and poorer for discipline 3 (67%). For discipline 1, DeepL (89%) and the human reference translations (83%) scored the best; for discipline 2, OpenNMT scored remarkably worse (65%), and for discipline 3, there was not much difference across translation conditions.

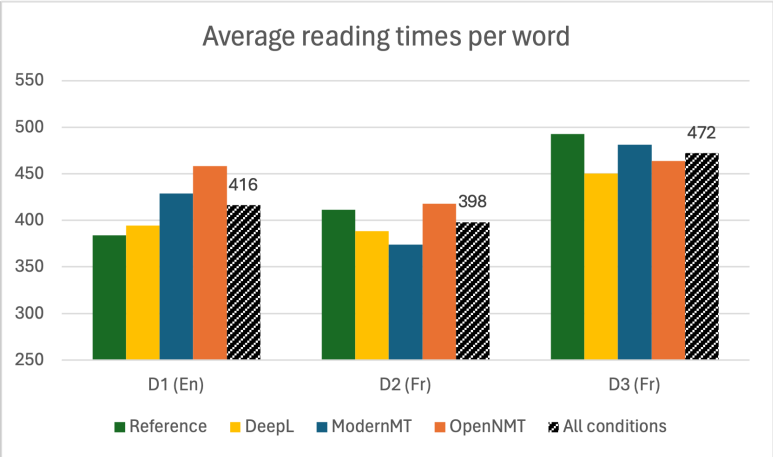

Figure 3. Average reading times per word, expressed in ms per word, per discipline and condition.

3.2. Statistical Analyses

Tables 4–6 present descriptive statistics for all the predictors included in the linear mixed effects models. As mentioned above, these data refer only to the machine-translated texts. When comparing the general target text characteristics across the three disciplines, we find that there is little difference in the mean average sentence length values. As discipline 1 contains French texts that were translated into English and disciplines 2 and 3 contain English texts translated into French, we cannot simply compare the other general target text characteristics, as these metrics are (to some extent) language-dependent. Nevertheless, if we examine the values of the Automatic Readability Index (ARI) scores and the percentage of the 3000 most frequent words in discipline 2 and discipline 3, it becomes evident that discipline 3 contains more challenging texts.

Table 4. Descriptive statistics of all predictors included in the LME models for discipline 1.

	Mean	SD	Min	Max
Average Sentence Length	24.05	6.00	15.88	38.40
ARI	16.77	3.55	11.50	25.00
MATTR	0.79	0.04	0.70	0.85
3K_subtlex	85.75	3.43	80.42	95.24
MQM_Score	77.28	23.12	0.00	99.47
Number of critical errors	1.14	1.52	0.00	5.00
BERTScore	91.71	2.43	86.64	96.40
BLEURT	78.63	3.20	72.69	87.01
COMET	88.66	1.78	84.55	92.98
COMETKiwi	0.76	0.03	0.69	0.81

Table 5. Descriptive statistics of all predictors included in the LME models for discipline 2.

	Mean	SD	Min	Max
Average Sentence Length	28.24	3.33	22.29	34.60
ARI	18.26	2.04	14.60	22.50
MATTR	0.79	0.03	0.74	0.86
3K_lexiq	84.10	4.83	76.43	93.52
MQM_Score	88.65	13.62	37.11	100.00
Number of critical errors	0.39	0.89	0.00	4.00
BERTScore	90.42	3.37	83.60	95.31
BLEURT	78.42	5.47	65.99	86.68
COMET	88.43	2.40	80.69	91.83
COMETKiwi	0.80	0.03	0.75	0.86

Table 6. Descriptive statistics of all predictors included in the LME models for discipline 3.

	Mean	SD	Min	Max
Average Sentence Length	28.81	4.11	22.60	36.25
ARI	19.70	2.01	16.60	24.30
MATTR	0.79	0.036	0.68	0.85
3K_lexiq	75.94	4.70	65.33	83.73
MQM_Score	81.56	15.55	44.08	100.00
Number of critical errors	0.58	0.80	0.00	2.00
BERTScore	90.86	2.81	85.17	95.79
BLEURT	75.17	6.24	60.40	89.02
COMET	87.33	2.62	79.91	91.93
COMETKiwi	0.79	0.05	0.61	0.86

In terms of manual translation quality assessment, discipline 1 received the lowest MQM quality scores and had on average the highest number of critical errors, while discipline 2 received the best manual translation quality scores. In terms of automatic translation quality metrics, there is not much difference between the disciplines. It should be noted that the final MQM quality scores are heavily influenced by the penalties assigned to the severity labels. A few critical errors can result in a score of zero for the whole text, as was the case in discipline 1.

To assess whether different measures of translation quality could predict reading times, we fitted a series of linear mixed-effects models in R with average reading time as the response variable. Tables 7–9 present for each discipline the results of the baseline model, including only the random factors, and all models that include significant predictors.

Table 7. Final LME models for discipline 1 (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$).

		Fixed Effects					Variance Explained		Goodness of Fit
Model	**Predictor**	**Estimate**	**SE**	**df**	**t**	***p***	**mar. R^2**	**cond. R^2**	**AICc**
M_0	(Intercept)	6.00	0.06	21.20	94.33	***	0.000	0.458	59.2
M_1	(Intercept)	6.00	0.06	20.08	101.11	***			
	ARI_z	0.08	0.04	10.44	2.29	*	0.062	0.464	56.8
M_2	(Intercept)	6.00	0.06	20.03	100.0	***			
	3K_subtlex_z	−0.10	0.04	14.82	−2.7	*	0.092	0.487	54.2
M_3	(Intercept)	6.00	0.067	21.02	90.11	***			
	MQM_Score_z	−0.07	0.03	126.68	−2.63	**	0.040	0.517	54.9
M_{3a}	(Intercept)	6.00	0.06	20.27	97.06	***			
	MQM_Score_z	−0.06	0.03	111.13	−2.27	*			
	3K_subtlex_z	−0.09	0.04	14.99	−2.37	*	0.103	0.516	51.6
M_4	(Intercept)	5.95	0.07	24.10	84.53	***			
	CriticalErrors_z	0.04	0.02	100.51	2.23	*	0.037	0.513	56.8
M_5	(Intercept)	6.00	0.06	20.76	98.71	***			
	BERTScore_z	−0.09	0.037	18.06	−2.53	*	0.073	0.485	55.1
M_{5a}	(Intercept)	6.00	0.05	18.89	109.20	***			
	BERTScore_z	−0.09	0.031	14.04	−3.08	**			
	3K_subtlex_z	−0.10	0.031	12.54	−3.26	**	0.153	0.494	47.5
M_6	(Intercept)	6.00	0.06	20.60	100.09	***			
	BLEURT_z	−0.11	0.03	43.87	−3.53	***	0.102	0.505	49.4
M_{6a}	(Intercept)	5.99	0.05	15.94	118.74	***			
	BLEURT_z	−0.12	0.024	22.86	−4.99	***			
	3K_subtlex_z	−0.11	0.024	12.95	−4.43	***	0.191	0.493	37.6
M_7	(Intercept)	6.00	0.05	15.07	117.75	***			
	COMET_z	−0.14	0.02	19.87	−6.38	***	0.171	0.482	38.5
M_8	(Intercept)	6.04	0.06	20.56	100.39	***			
	COMETKiwi_z	−0.14	0.04	17.83	−3.50	**	0.122	0.510	49.8
M_{8a}	(Intercept)	6.03	0.05	19.64	105.57	***			
	COMETKiwi_z	−0.12	0.04	13.76	−3.25	**			
	BERTScore_z	−0.07	0.03	13.09	−2.16	*	0.146	0.505	47.4
M_{8a}	(Intercept)	6.03	0.06	19.69	104.93	***			
	COMETKiwi_z	−0.10	0.04	13.02	−2.50	*			
	BLEURT_z	−0.08	0.03	24.10	−2.52	*	0.146	0.510	45.7

Table 8. Final LME models for discipline 2 (. $p < 0.1$; * $p < 0.05$; *** $p < 0.001$).

		Fixed Effects					Variance Explained		Goodness of Fit
Model	**Predictor**	**Estimate**	**SE**	**df**	**t**	***p***	**mar. R^2**	**cond. R^2**	**AICc**
M_0	(Intercept)	5.92	0.07	21.80	80.77	***	0.000	0.620	31.6
M_1	(Intercept)	5.92	0.07	19.17	85.74	***			
	ARI_z	0.07	0.03	12.20	2.01	.	0.036	0.605	30.6
M_2	(Intercept)	5.92	0.07	19.61	85.30	***			
	3K_lexiq_z	−0.08	0.03	12.54	−2.51	*	0.057	0.621	28.2
M_3	(Intercept)	5.92	0.07	20.87	82.93	***			
	MQM_Score_z	−0.05	0.021	125.97	−2.45	*	0.023	0.624	27.9
M_{3a}	(Intercept)	5.92	0.07	18.82	86.78	***			
	MQM_Score_z	−0.04	0.02	111.45	−2.11	*			
	3K_lexiq_z	−0.07	0.03	11.31	−2.20	*	0.068	0.623	25.9
M_4	(Intercept)	5.89	0.07	22.22	80.50	***			
	CriticalErrors_z	0.06	0.03	114.60	2.23	*	0.022	0.631	28.8
M_8	(Intercept)	5.92	0.07	21.56	81.22	***			
	COMETKiwi_z	−0.05	0.03	121.92	−1.84	.	0.015	0.628	30.4

Table 9. Final LME model for discipline 3 (*** $p < 0.001$).

		Fixed Effects					Variance Explained		Goodness of Fit
Model	**Predictor**	**Estimate**	**SE**	**df**	**t**	***p***	**mar. R^2**	**cond. R^2**	**AICc**
M_0	(Intercept)	6.09	0.06	20.24	100.3	***	0.000	0.468	45.7

For discipline 1 (Human Mobility, Environment, and Space—Social Sciences and Humanities) and discipline 2 (Neurosciences—Life Sciences), both the automatic readability index (ARI) scores and the percentage of 3000 most frequent words are significant predictors (see models M_1 and M_2). The effect of both predictors is in line with expectations, with higher ARI scores (representing more difficult texts) leading to longer average reading times, and texts with a higher percentage of frequent words leading to shorter average reading times. Average sentence length and average type-token ratio had no statistically significant effect on average reading time in any of the three data sets.

Similarly, as can be seen in models M_3 and M_4, the two manual translation quality metrics (MQM quality score and number of critical errors) are significant predictors of average reading time for both discipline 1 and discipline 2. The effect of both predictors is as expected, with higher MQM quality scores (representing higher quality translations) leading to shorter average reading times, and texts with a higher number of critical errors leading to longer average reading times. Combining the MQM quality score and the percentage of 3000 most frequent words as predictors in models M_{3a} leads to better models for both disciplines.

Only for discipline 1, the reference-based automatic translation quality metrics (BERTScore (model M_5), BLEURT (model M_6), and COMET (model M_7) were significant predictors of average reading time, with higher scores leading to shorter average reading times. When looking at the marginal and conditional R_2 values and the AICc values, the model with COMET as a predictor is the best model. Adding the percentage of the 3000 most frequent words as predictors to the models with BERTScore (model M_{5a}) and BLEURT (Model M_{6a}) improves both models, making the model with BLEURT and the percentage of the 3000 most frequent words as predictors comparable to the model with COMET only.

COMETKiwi, the automatic translation quality metric that does not rely on reference translations, was a significant predictor for discipline 1 and was almost significant for discipline 2 (see models M_8). Combining COMETKiwi with reference-based automatic quality metrics BERTScore (model M_{8a}) or BLEURT (model M_{8b}) improves the models for discipline 1, which might be an indication that the reference-based and non-reference-based translation quality metrics indeed capture different characteristics.

For discipline 3, none of the predictors we included in the linear mixed-effect models were significant.

4. Discussion

The general analysis of the data showed that the reading times in our experiment were much higher than those reported in previous studies. Nevertheless, for two of the three disciplines, the quality of the translations was judged to be sufficient for gisting purposes by the participants.

In more detailed statistical analyses in which we fitted a series of linear mixed effect models, we examined the relationship between translation quality and readers' reception effort, measured by reading times. These results show large differences across the disciplines.

For discipline 1 (translations from French into English), both the automatic readability index (ARI) scores and the percentage of the 3000 most frequent words are significant predictors. In addition, all the manual and automatic translation quality measures were significant predictors. The best models were obtained by including COMET as a predictor or a combination of BLEURT and the percentage of the 3000 most frequent words.

Also, for discipline 2 (translations from English to French), both the automatic readability index (ARI) scores and the percentage of the 3000 most frequent words are significant predictors. The manual translation quality measures and the automatic non-reference-based automatic measure (COMETKiwi) were significant predictors, but the reference-based automatic quality measures were not. This may indicate that the reference translations were not optimal (either too divergent or of poor quality). The best model was obtained by including both the MQM quality score and the percentage of the 3000 most frequent words as predictors.

For discipline 3 (also translations from English to French), we were not able to build a model that outperformed the baseline model that included only the two random factors (participant and text). There could be several reasons for this. As there were no full texts available for discipline 3, the text excerpts were selected from various sources of abstracts, which may be denser and more difficult to read than excerpts from full texts. In addition, discipline 3 contained texts from the field of Physical Sciences related to climatology and climate change, and were the most technical texts of all disciplines, as evidenced by the lowest number of high-frequency words, and the highest average reading times regardless of translation condition. As all our participants were members of a university language department, they were less familiar with the topics of discipline 3. As the impact of prior knowledge has been recognized in previous reading research, the content of the texts in discipline 3 was probably too difficult for the participants, and differences in translation quality may not be apparent in this case.

5. Conclusions

In this study, we assessed the usability of machine-translated texts in the context of scholarly communication. We conducted self-paced reading experiments with texts from three different scientific disciplines, focusing on both the translation direction from French to English and vice versa. A total of 32 participants took part in the study. For each scientific discipline, the reading times of 16 participants (nonexpert readers, highly proficient in the target language) reading 12 texts each were collected.

The machine translations were generated by three different MT engines (DeepL, ModernMT, and OpenNMT) and varied in terms of translation quality. The aim of the experiments was to investigate whether there is a relationship between translation quality and readers' reception effort. The reading times collected in the experiment were used as a proxy for the readers' reception effort. Translation quality was assessed using manual and automatic methods.

Overall, the results are promising. For two of the three scientific disciplines, both the manual and automatic quality measures proved to be significant predictors of reading time. For two disciplines, the human reference translations were not read faster than the machine translations, which contradicts the consensus in the research community that reading machine-translated texts requires more processing effort than reading human-translated texts. Furthermore, in the second discipline, the non-reference-based automatic evaluation measure was a significant predictor of reading time, whereas the reference-based measures were not. These observations suggest that the reference translations of this domain may not have been optimal, possibly because they were either too divergent or of poor quality.

For the most technical discipline (Physical Sciences), we were unable to build linear mixed effects models that outperformed the baseline model. We mainly focused on different translation quality measures and general text characteristics as predictors in our models, but we did not include reader-specific features that influence reading such as reading habits and readers' prior knowledge, which has been acknowledged as an important factor for comprehending scientific texts [53].

Due to the tight timeframe of the project, we were not able to conduct eye-tracking experiments. Although the self-paced reading experiment provided us with useful data at the text level, eye-tracking would allow us to carry out even more fine-grained analyses on specific areas of interest, e.g., specific error types or problematic fragments. It would also allow us to study the temporal aspects of reading a text. The tight timeframe also meant that we were unable to have a second annotator for the MQM error annotations. As the severity levels assigned to the errors have a strong impact on the MQM quality score, working with more than one annotator is recommended for future work. With 32 participants, this is still a small-scale study. Larger and more diverse sample sizes would of course provide more robust and more generalizable results.

Author Contributions: Conceptualization, L.M. and A.T.; methodology, L.M., V.D.W. and A.T.; software, L.M. and A.T.; validation, L.M. and A.T.; formal analysis, L.M. and V.D.W.; investigation, L.M. and A.T.; resources, L.M. and A.T.; data curation, L.M. and A.T.; writing—original draft preparation, L.M. and A.T.; writing—review and editing, L.M., V.D.W. and A.T.; visualization, L.M. and A.T.; supervision, L.M. and A.T.; project administration, L.M.; funding acquisition, L.M. and A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This study was part of a larger project funded by OPERAS (https://operas-eu.org/, accessed on 5 July 2024), on behalf of the French Ministry of Higher Education and Research. The Language and Translation Technology Team acted as a subcontractor of CrossLang in this study.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The research data will be shared when the article is accepted.

Acknowledgments: This study was carried out in close collaboration with the following colleagues from CrossLang: Tom Vanallemeersch, Sara Szoc, Kristin Migdisi, and Laurens Meeus.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AICc	Corrected Akaike Information Criterion
ARI	Automatic Readability Index
HT	Human Translation
LME	Linear Mixed Effects
LMM	Large Language Model
MATTR	Moving Average Type–Token Ratio
MT	Machine Translation
MQM	Multidimensional Quality Metrics
Perc-3KFreq	Percentage of 3000 most frequent words
SRC	Source
QE	Quality Estimation

References

1. Amano, T.; Ramírez-Castañeda, V.; Berdejo-Espinola, V.; Borokini, I.; Chowdhury, S.; Golivets, M.; González-Trujillo, J.D.; Montaño-Centellas, F.; Paudel, K.; White, R.L.; et al. The manifold costs of being a non-native English speaker in science. *PLoS Biol.* **2023**, *21*, e3002184. [CrossRef]
2. Hannah, K.; Haddaway, N.R.; Fuller, R.A.; Amano, T. Language inclusion in ecological systematic reviews and maps: Barriers and perspectives. *Res. Synth. Methods* **2024**, *15*, 466–482. [CrossRef]
3. Bowker, L.; Ayeni, P.; Kulczycki, E. *Linguistic Privilege and Marginalization in Scholarly Communication: Understanding the Role of New Language Technologies for Shifting Language Dynamics*; Technical Report; Social Sciences and Humanities Research Council of Canada: Ottawa, ON, Canada, 2023.
4. O'Brien, S.; Simard, M.; Goulet, M.J. Machine translation and self-post-editing for academic writing support: Quality explorations. In *Translation Quality Assessment: From Principles to Practice*; Moorkens, J., Castilho, S., Gaspari, F., Doherty, S.; Eds.; Springer: Berlin/Heidelberg, Germany, 2018; pp. 237–262.
5. Zhivotova, A.; Berdonosov, V.; Redkolis, E. Improving the quality of scientific articles machine translation while writing original text. In Proceedings of the 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), Vladivostok, Russia, 6–7 October 2020; pp. 1–4.
6. Simonova, V.; Patiniotaki, E. Pre-editing for the Translation of Life-Science Texts from Russian into English via Google Translate. In Proceedings of the New Trends in Translation and Technology 2022, Rhodes Island, Greece, 4–6 July 2022; p. 259.
7. Roussis, D.; Sofianopoulos, S.; Piperidis, S. Enhancing Scientific Discourse: Machine Translation for the Scientific Domain. In Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), Sheffield, UK, 24–27 June 2024; pp. 275–285.
8. Walker, C. Investigating how we read translations: A call to action for experimental studies of translation reception. *Cogn. Linguist. Stud.* **2021**, *8*, 482–512. [CrossRef]
9. Orero, P.; Doherty, S.; Kruger, J.L.; Matamala, A.; Pedersen, J.; Perego, E.; Romero-Fresco, P.; Rovira-Esteva, S.; Soler-Vilageliu, O.; Szarkowska, A. Conducting experimental research in audiovisual translation (AVT): A position paper. *JosTrans J. Spec. Transl.* **2018**, *30*, 105–126.
10. Whyatt, B.; Witczak, O.; Tomczak-Łukaszewska, E.; Lehka-Paul, O. The proof of the translation process is in the reading of the target text: An eyetracking reception study. *Ampersand* **2023**, *11*, 100149. [CrossRef]
11. Doherty, S.; O'Brien, S.; Carl, M. Eye tracking as an MT evaluation technique. *Mach. Transl.* **2010**, *24*, 1–13. [CrossRef]
12. Kasperavičienė, R.; Motiejūnienė, J.; Patašienė, I. Quality assessment of machine translation output: Cognitive evaluation approach in an eye tracking experiment. *Texto Livre Ling. Tecnol.* **2020**, *13*, 271–285. [CrossRef]
13. Colman, T.; Fonteyne, M.; Daems, J.; Dirix, N.; Macken, L. GECO-MT: The Ghent Eye-tracking Corpus of Machine Translation. In Proceedings of the Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 29–38.
14. Kasperė, R.; Motiejūnienė, J.; Patasienė, I.; Patašius, M.; Horbačauskienė, J. Is machine translation a dim technology for its users? An eye tracking study. *Front. Psychol.* **2023**, *14*, 1076379. [CrossRef] [PubMed]
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
16. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825.
17. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. In *Proceedings of the Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 27730–27744.

18. Kocmi, T.; Avramidis, E.; Bawden, R.; Bojar, O.; Dvorkovich, A.; Federmann, C.; Fishel, M.; Freitag, M.; Gowda, T.; Grundkiewicz, R.; et al. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In Proceedings of the Eighth Conference on Machine Translation, Singapore, 6–7 December 2023; Koehn, P., Haddow, B., Kocmi, T., Monz, C., Eds.; Association for Computational Linguistics: Singapore, 2023; pp. 1–42. [CrossRef]
19. Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 67–72.
20. Junczys-Dowmunt, M.; Grundkiewicz, R.; Dwojak, T.; Hoang, H.; Heafield, K.; Neckermann, T.; Seide, F.; Germann, U.; Fikri Aji, A.; Bogoychev, N.; et al. Marian: Fast Neural Machine Translation in C++. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), Melbourne, Australia, 15–20 July 2018; pp. 116–121.
21. Tiedemann, J.; Thottingal, S. OPUS-MT – Building open translation services for the World. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation; Martins, A., Moniz, H., Fumega, S., Martins, B., Batista, F., Coheur, L., Parra, C., Trancoso, I., Turchi, M., Bisazza, A., et al., Eds.; European Association for Machine Translation: Lisboa, Portugal, 2020; pp. 479–480.
22. Team, N.; Costa-jussà, M.R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; et al. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv* **2022**, arXiv:2207.04672.
23. Lommel, A.; Uszkoreit, H.; Burchardt, A. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica* **2014**, *12*, 455–463. [CrossRef]
24. Tezcan, A.; Hoste, V.; Macken, L. SCATE taxonomy and corpus of machine translation errors. In *Trends in E-Tools and Resources for Translators and Interpreters*; Pastor, G.C., Durán-Muñoz, I., Eds.; Approaches to Translation Studies; Brill | Rodopi: Amsterdam, The Netherlands, 2017; Volume 45, pp. 219–244.
25. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
26. Lavie, A.; Denkowski, M.J. The METEOR metric for automatic evaluation of machine translation. *Mach. Transl.* **2009**, *23*, 105–115. [CrossRef]
27. Popović, M. chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, 17–18 September 2015; Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Hokamp, C., Huck, M., Logacheva, V., Pecina, P., Eds.; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 392–395. [CrossRef]
28. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A study of translation edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, Cambridge, MA, USA, 8–12 August 2006; pp. 223–231.
29. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv* **2019**, arXiv:1904.09675.
30. Sellam, T.; Das, D.; Parikh, A.P. BLEURT: Learning robust metrics for text generation. *arXiv* **2020**, arXiv:2004.04696.
31. Rei, R.; Stewart, C.; Farinha, A.C.; Lavie, A. COMET: A Neural Framework for MT Evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 2685–2702. [CrossRef]
32. Rei, R.; Treviso, M.; Guerreiro, N.M.; Zerva, C.; Farinha, A.C.; Maroti, C.; de Souza, J.G.C.; Glushkova, T.; Alves, D.; Coheur, L.; et al. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In Proceedings of the Seventh Conference on Machine Translation (WMT), Abu Dhabi, United Arab Emirates, 7–8 December 2022; Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M.R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., et al., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 634–645.
33. Naskar, S.; Deutsch, D.; Freitag, M. Quality Estimation Using Minimum Bayes Risk. In Proceedings of the Eighth Conference on Machine Translation, Singapore, 7–8 December 2023; Koehn, P., Haddow, B., Kocmi, T., Monz, C., Eds.; ACL Anthology: Singapore, 2023; pp. 806–811. [CrossRef]
34. Kocmi, T.; Federmann, C.; Grundkiewicz, R.; Junczys-Dowmunt, M.; Matsushita, H.; Menezes, A. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In Proceedings of the Sixth Conference on Machine Translation, Online, 10–11 November 2021; Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-Jussa, M.R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., et al., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 478–494.
35. Freitag, M.; Rei, R.; Mathur, N.; Lo, C.k.; Stewart, C.; Avramidis, E.; Kocmi, T.; Foster, G.; Lavie, A.; Martins, A.F. Results of WMT22 metrics shared task: Stop using BLEU–neural metrics are better and more robust. In Proceedings of the Seventh Conference on Machine Translation (WMT), Abu Dhabi, United Arab Emirates, 7–8 December 2022; pp. 46–68.
36. Freitag, M.; Mathur, N.; Lo, C.k.; Avramidis, E.; Rei, R.; Thompson, B.; Kocmi, T.; Blain, F.; Deutsch, D.; Stewart, C.; et al. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Proceedings of the Eighth Conference on Machine Translation, Singapore, 6–7 December 2023; pp. 578–628.
37. Lee, S.; Lee, J.; Moon, H.; Park, C.; Seo, J.; Eo, S.; Koo, S.; Lim, H. A Survey on Evaluation Metrics for Machine Translation. *Mathematics* **2023**, *11*, 6. [CrossRef]

38. Fiorini, S.; Tezcan, A.; Vanallemeersch, T.; Szoc, S.; Migdisi, K.; Meeus, L.; Macken, L. Translations and Open Science: Exploring how translation technologies can support multilingualism in scholarly communication. In Proceedings of the International Conference on Human-Informed Translation and Interpreting Technology (HiT-IT 2023), Naples, Italy, 7–9 July 2023; Orasan, C., Mitkov, R., Pastor, G.C., Monti, J., Eds.; INCOMA Ltd.: Shoumen, Bulgaria, 2023; pp. 41–51.
39. Tiedemann, J. Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 21–27 May 2012; Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., Eds.; European Language Resources Association (ELRA): Reykjavik, Iceland, 2012; pp. 2214–2218.
40. Roussis, D.; Papavassiliou, V.; Prokopidis, P.; Piperidis, S.; Katsouros, V. SciPar: A Collection of Parallel Corpora from Scientific Abstracts. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., et al., Eds.; European Language Resources Association: Marseille, France, 2022; pp. 2652–2657.
41. Jegerski, J. Self-paced reading. In *Research Methods in Second Language Psycholinguistics*; Routledge: New York, NY, USA, 2013; pp. 20–49.
42. Vanroy, B.; Tezcan, A.; Macken, L. MATEO: MAchine Translation Evaluation Online. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, Tampere, Finland, 12–15 June 2023; Mary, N., Brenner, J., Koponen, M., Latomaa, S., Mikhailov, M., Schierl, F., Ranasinghe, T., Vanmassenhove, E., Alvarez Vidal, S., Aranberri, N., et al., Eds.; European Association for Machine Translation (EAMT): Tampere, Finland, 2023; pp. 499–500.
43. Zenker, F.; Kyle, K. Investigating minimum text lengths for lexical diversity indices. *Assess. Writ.* **2021**, *47*, 100505. [CrossRef]
44. Van Parys, A.; De Wilde, V.; Macken, L.; Montero Perez, M. Vocabulary of reading materials in English and French L2 textbooks: A cross-lingual corpus study. *System* **2024**, *124*, 103396. [CrossRef]
45. Van Heuven, W.J.; Mandera, P.; Keuleers, E.; Brysbaert, M. SUBTLEX-UK: A new and improved word frequency database for British English. *Q. J. Exp. Psychol.* **2014**, *67*, 1176–1190. [CrossRef] [PubMed]
46. New, B.; Pallier, C.; Brysbaert, M.; Ferrand, L. Lexique 2: A new French lexical database. *Behav. Res. Methods Instrum. Comput.* **2004**, *36*, 516–524. [CrossRef]
47. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2024.
48. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **2015**, *67*, 1–48. [CrossRef]
49. Kuznetsova, A.; Brockhoff, P.B.; Christensen, R.H.B. lmerTest Package: Tests in Linear Mixed Effects Models. *J. Stat. Softw.* **2017**, *82*, 1–26. [CrossRef]
50. Lüdecke, D.; Ben-Shachar, M.S.; Patil, I.; Waggoner, P.; Makowski, D. performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *J. Open Source Softw.* **2021**, *6*, 3139. [CrossRef]
51. Hurvich, C.M.; Tsai, C.L. A corrected Akaike information criterion for vector autoregressive model selection. *J. Time Ser. Anal.* **1993**, *14*, 271–279. [CrossRef]
52. Brysbaert, M. How many words do we read per minute? A review and meta-analysis of reading rate. *J. Mem. Lang.* **2019**, *109*, 30. [CrossRef]
53. Jian, Y.C. Using an eye tracker to examine the effect of prior knowledge on reading processes while reading a printed scientific text with multiple representations. *Int. J. Sci. Educ.* **2022**, *44*, 1209–1229. [CrossRef]