Data Ethics and Machine Translation – Basic Level

Table of Contents

1 Introduction 2
2 Machine Translation Training and Translation Data 2
3 Personal Data and Translation Data 4
4 Data Protection, Copyright and Translation Data 4
5 Linguistic Issues and Translation Data 6
6 Automatic Evaluation of Translation Data 8
7 Conclusion 9
References 9

1 Introduction

This paper provides a theoretical discussion of several ethical issues we need to consider when handling translation data in the context of machine translation. With neural machine translation (NMT) systems gaining ground in professional translation workflows, more and more translators have to implement NMT systems into their translation projects – alongside 'traditional' translation technologies and resources such as translation memories or terminology databases. The fact that NMT systems rely on large quantities of high-quality translation data for training raises legal and ethical questions about how to collect, use, and evaluate these data.

In the following sections, we will take a look at some ethical and legal issues related to different kinds of translation data involved in the use of NMT systems. You will learn about:

- pairs of source and target language sentences used for NMT training,
- translation productivity data, which contain information about the way translators work on a particular project, and
- NMT output.

As described above, NMT training depends on the availability of high-quality texts translated by humans (although synthetic data creation is becoming a major trend in NMT training, cf. Moslem et al. 2022). The following section covers some ethical aspects translators can think about when collecting these translation data for NMT training (or when their own translation data are collected by third parties for NMT training). Other topics to be discussed in this paper include personal data and its role in machine translation workflows, different approaches to machine translation copyright as well as ethical issues, which can arise from the linguistic quality of NMT training data. The paper ends with a summary of ethical aspects to be considered when applying automatic quality evaluation metrics to assess the quality of machine translations, followed by a conclusion.

2 Machine Translation Training and Translation Data

The driving force behind high-quality NMT is the availability of large quantities of source texts and their high-quality translations (cf. Moorkens/Lewis 2019:474). This is because NMT systems are trained on large collections of bilingual (and sometimes multilingual) text. Using such training data in the form of source text segments and their corresponding translations, NMT systems can learn to reproduce translations similar to the examples seen during training (cf. Pérez-Ortiz et al. 2022:148). Only after training has been completed, NMT systems can be expected to generate high-quality translations even for unknown source texts.

The translation data required for NMT training may be gathered from web searches (for example, the Common Crawl internet corpus) or by consulting translation databases such as the translation memory provided by the European Commission's Directorate-General for Translation. In practice, however, NMT systems are often trained with texts that are not publicly available and that have not originally been translated with subsequent NMT training purposes in mind. For example, some buyers of translation services may have an interest in using the translations they receive from their translators for NMT training.

According to Moorkens (2022), ethical questions arise when we think about how to remunerate the professional translators who have produced the training data. Translators are not always aware of clients repurposing their translations. Thus, they sometimes contribute to NMT training without being paid for providing their translations for this purpose. Freelance work arrangements are common in the translation industry and fuel these practices. Translators working in freelance setting often lack the support which employees of (large) companies can expect from their employers. Not only is freelance work characterised by uncertain working conditions due to contracts limited to single projects, but asymmetric power relations between translators and employers may give translators little space to negotiate suitable conditions for the reuse of their translation data. (cf. Moorkens 2022:132)

Another aspect to consider when collecting NMT training data arises from the fact that there are language pairs for which we do not have enough translated texts at hand. A shortage in high-quality texts available for training may result in a decline in NMT output quality. One strategy that can be applied to generate more translations in low-resource languages is called back-translation. This method involves creating new synthetic training data through machine translation. In a first step, high-resource language texts are machine translated into low-resource languages in order to generate the needed amounts of text in these low-resource languages. The second step consists of using the machine translated low-resource language texts as NMT training data. (cf. Moorkens 2022:134)

The practice where NMT system training relies on machine translated texts instead of using high-quality human translated texts may introduce ethical issues: Analyses of collections of translated texts have shown that translations usually show linguistic features that differ from those found in original texts. For instance, it was found that, on average, human translations are often longer than texts originally produced in the target language (cf. Baker 1996:180). Research also investigates typical linguistic features of machine-translated texts (*machine-translationese*). For example, Vanmassenhove et al. (2021:2211) found that machine-translated texts are less lexically diverse than human translations. With these findings in mind, we can imagine that training NMT systems on synthetic texts may result in the systems learning different linguistic patterns than would have been the case if training was based on human-produced texts. As research shows, NMT systems trained on machine-translated texts are likely to produce linguistically less diverse translations to a greater extent (cf. Vanmassenhove et al. 2021). Moorkens (2022:135) warns that "[…] poorly-resourced languages will be impoverished in the long run." (See section 5 for more details on ethical issues related to machine translationese.)

Collecting translation data may present further ethical pitfalls when we think about the current technological advancements in the development of large language models (LLMs, for an overview, cf. Zhao et al. 2023). A prominent example of a powerful LLM is ChatGPT, which was developed by the US-based company OpenAI. ChatGPT uses artificial intelligence to answer a wide range of questions. Users can interact with the model in a chat conversation.

LLMs and state-of-the-art NMT systems share basic working principles as both technologies are based on the same system architecture, the *transformer* (cf. Vaswani et al. 2017). Just like NMT systems, LLMs' performance improves through training and the adjustment of billions of parameters during the training process. Also, the technology requires large amounts of training data in order to produce high-quality content. Bender et al. (2021:611) write that increasingly larger models "[…] have pushed the boundaries of the possible both through architectural innovations and through sheer size." The authors provide an insightful discussion of the potential societal implications of LLMs that go beyond the ethical issues discussed in this paper.

Reading on, you will learn about some ethical issues concerning the large quantities of data which need to be processed for LLM training. For instance, social biases in training data are an issue for both NMT systems and LLMs, whereby the latter pose new challenges due to a different composition of training data required (see section 5 on linguistic issues and translation data).

3 Personal Data and Translation Data

Translation memories used for NMT training usually do not only contain pairs of source texts and translated texts, but also additional information about the translators who have worked produced these translations. The aligned source and target language sentences may be associated with so-called "metadata", such as creation and modification dates and even the name or a coded identifier of the translator (cf. Moorkens 2022:125). From a legal perspective, data that allow identifying an individual are called "personal data" (Moorkens 2022:127).

These data help project managers and translators working on a translation project to answer questions like: Is this already completed translation helpful for a new translation project? Was the previous translation done by a translator we know and trust? What quality can we expect from this translation?

Moorkens (2022) points out a downside of this approach: Metadata enable employers to draw conclusions which may impact their relationship with the translators working for them. For instance, with metadata revealing a decrease in productivity, an employer may want to exclude the respective translator from future projects. Of course, such a decrease in productivity can be caused by multiple factors and an ethically responsible decision in this case should not solely be based on metadata, but always include a discussion of the situation with all relevant stakeholders. (cf. Moorkens 2022:128)

For NMT training, metadata is typically deleted from the aligned sentences stored in translation memories. Consequently, the data used for training is not considered personal data. This practice allows sharing and reusing the translation data. Yet, deleting all metadata makes it impossible to identify the translators who have contributed to the translation at hand. Without identification however, no remuneration is possible. (cf. Moorkens 2022:126)

Conversely, Moorkens et al. (cf. 2016:52) have suggested to maintain metadata – and hence personal information about the creators of translations – when training NMT systems to help to retrace the translators who have worked on creating the training data. This could be one way to ensure that translators can be rewarded for the crucial role they play in providing the necessary high-quality translation data for NMT training.

4 Data Protection, Copyright and Translation Data

Despite the practices of reusing translations in the translation industry as described above, translations are subject to legal regulations on copyright in various countries. The underlying interpretation of copyright, however, may differ between the individual countries.

According to article 2 of the Berne Convention, translations are protected by copyright if they are a derivative work of a source text whose translation requires some degree of creativity. Of course, this definition does not include all kinds of translations. Some fields of specialised translation such as technical translation, where texts are not notably creative, may be excluded from copyright.

Also, personal data containing sensitive information about individuals are subject to legal regulations. In the European Union, the General Data Protection Regulation (GDPR) restricts the re-usage of personal data to its original purpose. Under this law, a translation memory that contains translation data labelled with metadata can only be used for NMT training if the data included have been translated precisely for NMT training purposes. Regarding the suggestion of keeping metadata in translation memories in order to be able to identify contributions of individual translators to texts used as training data (see section 3), the law makes this alternative approach more difficult. As Moorkens (cf. 2022:473) writes, using anonymised translation data for NMT system training makes it easier to avoid the restrictions otherwise imposed upon members of the European Union by the GDPR.

Despite these legal regulations concerning translation copyright, Moorkens and Lewis (2020:478) conclude that "[a]t present, the attribution of translation copyright (and the reuse of translation as data) is subject to a number of conflicting and inconsistently interpreted laws and conventions and thus remains somewhat unclear." The legal situation of translation copyright gets even more confusing when we think about the role NMT systems and other translation technology tools based on artificial intelligence play in the translation process. Until now, only humans and companies are able to be granted copyright (cf. Moorkens/Lewis 2020:475). Accordingly, NMT systems and related computer programs cannot claim ownership of a machine translation. In the case of machine translations, copyright is granted to the company or the programmers who have developed the system in use.

With neural text generative technologies such as large language models, however, we face a new form of automatic text generation and machine translation. When prompted by natural language inputs, LLMs like ChatGPT and its successor GPT-4 are able to perform various tasks including text generation and translation. In order to ask ChatGPT for a machine translation, for instance, research suggests to include information on the task to be carried out (e.g. Please translate the following English sentence into German.) and to specify the domain of the text to be translated (e.g. technical, IT, financial, or cultural domain) (cf. Peng et al. 2023:1). Also, it was found that the phrase "You are a machine translation system" is an effective way of prompting ChatGPT to generate a machine translation (cf. Peng et al.2023:4). In other words, the linguistic form of a prompt has a direct effect on an LLM's performance.

Yet, LLMs' capacities cover a wide range of fields, which distinguishes them from state-of-the-art NMT systems that are specialised on machine translation tasks. In addition to classic machine translations, LLMs are able to generate texts according to users' requirements. This text generative capacity may result in the production of texts that can be attributed a certain degree of creativity. For instance, there are poems written by ChatGPT. (See, e.g., this blog entry on the possibility to ask for ChatGPT's opinion through letting it write poems.) LLMs' writing abilities may also be applied in translation processes. Apart from many other use cases, the model's (creative) writing capabilities can be used for post-editing machine translation output (cf. Raunak et al. 2023).

Considering such writing skills, there may be a need to adapt the current copyright regulations to the latest technological advancements. Against this backdrop, the practice of excluding machines from copyright may will have to be debated.

Discussing future directions for copyright and data ownership in the translation industry and the role of computers in this context, Moorkens (2022:137) writes: "The future may see computers act as explicit ethical agents with the ability to process information about each situation and to autonomously determine the best or most ethical course of action." However, the author states that today's technology has not yet reached such a level of autonomy and clarifies that even if this happens in the future, users should not rely on technology to always act with good intentions (cf. Moorkens 2020:137). With new advancements in the field of artificial intelligence and the development of powerful tools like LLMs, we may have come a step closer to the idea of autonomous computer programs. In a recent study, Bubeck et al. (2023) test the LLM GPT-4 on tasks from various fields and come to the conclusion that the technology shows "*sparks of artificial general intelligence*" (Bubeck et al. 2023:92; original emphasis).

Yet, today's LLMs still need human intervention in order to perform at high levels of quality (this is also known as the *human/expert-in-the-loop* approach). As discussed above in this section, one way how humans can exercise control over the technology's actions is writing instructions (prompting). In the face of rapidly developing technologies it is in our hands to establish ethical approaches and legal regulations in order to use the technology's power on our behalf.

5 Linguistic Issues and Translation Data

An aspect to keep in mind when working with NMT output are specific linguistic patterns in the NMT output which can be traced back to a system's training data. As described above, NMT systems are trained on large amounts of human-translated texts. Thus, high-quality NMT output can only be achieved when the systems were trained on high-quality training data, ideally translated by professional human translators. This working principle of artificial intelligence technologies based on machine learning approaches is often referred to as "garbage in, garbage out" (Lommel et al. 2018:30).

However, the use of translation technology in the translation industry is increasingly perceived as a threat by professional translators. For instance, participants of the European Language Industry Survey (cf. ELIS Research 2023:21) mentioned reduced rates for post-editing jobs and a general fear of being replaced by machines, among others, as reasons for a negative attitude towards technology usage in the industry. Faced with such difficulties, more professional translators may decide to leave the sector (cf. Moorkens 2022:133). Together with the fact that the rise of NMT is accompanied by fewer texts being translated by professional human translators, this development may lead to a lack of high-quality human translated data needed for NMT training (cf. Moorkens 2022:133). A restricted access to training data of appropriate quality, in turn, may give rise to linguistic issues in the output of such poorly trained NMT systems. The result will most probably be a decline in NMT output quality.

From an ethical perspective, poor quality NMT output may pose problems for readers. NMT occasionally struggles to translate content adequately into other languages. For instance, research has found a tendency for LLMs to "favor greater fluency at the cost of adequacy." (Hendy et al. 2023:16) This shortcoming gets even more pressing as, generally, texts translated by NMT systems are prone to be "deceptively fluent" (Way 2020:318). That is, readers may be tricked into trusting the translation due to its appealing wording. This way, quality issues regarding content are easily overlooked by layperson readers, who are not familiar with the topic of the translated text or the working principles of NMT systems.

However, ethical issues are not restricted to poor quality NMT output. Machine translated texts in general are prone to a linguistic phenomenon called "machine translationese", which comprises multiple linguistic patterns distinguishing machine translations from texts translated by humans and from original texts (cf. Vanmassenhove et al. 2021:2203). Studies on the topic of machine translationese have found machine translations to exhibit instances of "artificially impoverished language" (Vanmassenhove et al. 2021:2203). More information on machine translationese and its various forms can be found in our DataLit^{MT} learning resource on Data Evaluation: Machine Translationese & Post-Editese – How MT integration affects the target texts we produce.

One way such an artificially impoverished language manifests itself in machine translations is in the form of specific biases. From an ethical perspective, social biases in machine translation are of particular interest. We talk about a biased NMT output when stereotypes present in society are adopted in translated texts (cf. Savoldi et al. 2021:845). From an ethical point of view, biases in machine translations have to be taken seriously, as they usually come with a discrimination against a certain group of people. For instance, numerous studies have investigated gender bias in machine translations, where machine translation tends to stick to stereotypical gender roles (e.g. when assigning a gender to an occupation) (cf. Prates et al. 2019). For more information on social biases in the context of MT, see our DataLit^{MT} learning resource on Social Bias in Machine Translation – with a focus on Gender Bias.

Effects of machine translation on linguistic features of translated texts should also be considered when post-editing NMT output. In practice, the degree to which modification of raw NMT output is needed varies according to readers' expectations: Sometimes it is sufficient to only make small changes, which is referred to as "light post-editing" (Massardo et al. 2016:16). The light post-editing mode is usually applied when working on texts for which we do not expect a publishable high-quality version as an outcome. This form of post-editing typically comprises modifications linked to rather superficial changes in terms of adequacy and the correct use of terminology. (cf. Massardo et al. 2016:16)

"Full post-editing", on the other hand, is concerned with transforming raw NMT output into a high-quality translation that meets the quality standards of texts written by professional human translators. To this end, full post-editing requires steps like adapting measurements and date formats to target culture norms as well as adapting the layout of a text. (cf. Massardo et al. 2016:16)

However, the actual amount of effort needed to post-edit a machine translated text does not always match with the above mentioned categories of light and full post-editing. Post-editing effort measures how much time (temporal effort), how many deletions and insertions via the computer key board (technical effort), and how much thought (cognitive effort) it takes to post-edit raw machine translation output (cf. Krings 2001:178). However, from an ethical perspective, it should be considered that social biases in NMT output can augment the required post-editing effort for the group of people who are disadvantaged by the biased output. For instance, a woman post-editing her machine translated biography may have to correct more mistakes than a man due to a gender bias (cf. Savoldi et al. 2021:847).

In addition, biases may pose problems when collecting training data for translation technologies based on artificial intelligence. Not only state-of-the-art NMT systems rely on large amounts of training data. LLMs, too, are dependent on the quantity and quality of their training data. Yet, in order to train these models, another kind of textual data is needed. While state-of-the-art NMT systems are trained on bilingual translation data (original texts and their translations), LLMs' training data is mostly comprised of monolingual texts. For instance, only about 8% of GPT-3's training data is made up of non-English texts. It is somewhat surprising that LLMs are able to perform translation tasks despite the mostly monolingual composition of their training data (however, see Briakou et al.'s 2023 study on "incidental bilingualism" in Google's language model PaLM).

However, there are concerns regarding possible biases in LLM translation outputs due to the lopsided training data composition described above. In an online article, Walker Rettberg investigates negative effects of using mostly monolingual data for training LLMs. The author describes the LLM ChatGPT as "multilingual but monocultural" (Walker Rettberg 2022). With this description, the author points out the possible risk of ChatGPT output being biased towards values supported by a majority of the English speaking community as main contributor to the model's training data (cf. Walker Rettberg 2022). Other groups, however, may be underrepresented in the data and cannot expect their world views to be covered in the ChatGPT output to the same extent as Anglocentric views are.

6 Automatic Evaluation of Translation Data

Evaluating translations may be necessary at multiple steps in NMT-assisted translation processes. Firstly, there is a need to evaluate the output generated by an NMT system during training. Only if we can measure any changes in quality of the produced machine translations, we can get an idea of whether or not the training was effective (cf. Rossi/Carré 2022:59). Also, the performance of an NMT system is often compared with other state-of-the-art NMT systems to assess differences in their output quality, for example when a translation company wants to choose an NMT system which is suitable for a specific translation project. (cf. Rossi/Carré 2022:59)

As research begins to examine translation abilities of LLMs, automatic evaluation is also being increasingly employed to compare the output quality of LLMs and state-of-the-art NMT systems. For instance, comparing automatic quality evaluation scores for LLM and NMT output, it was found that an LLM's performance is often comparable in terms of quality to the output of state-of-the-art NMT systems, when we ask an LLM for translations of languages that are part of the training data of the model (cf. Hendy et al. 2023:22).

For such use cases, we can also employ humans as MT quality evaluators. These human evaluators may use an error typology (e. g. the Multidimensional Quality Metrics [MQM] error typology), where they can apply different error categories for judging the quality of a given translation (cf. Burchardt 2013). The MQM error typology, for example, includes the error categories *accuracy*, *fluency*, and *verity*, among others (Burchardt 2013:4). Although human evaluation is still valued as the gold standard in translation quality evaluation, it comes with multiple drawbacks: Involving humans in translation evaluation is time-consuming and may result in higher costs. Additionally, human judgements, although often more reliable and accurate than automatic metrics, may be influenced by personal preferences of the human evaluators.

To avoid some of these problems, we can apply automatic evaluation metrics for evaluating the quality of machine translations. A widely used automatic evaluation metric is called BLEU (Bilingual Evaluation Understudy). You can find the original paper here. BLEU is a so-called string-based metric, which measures the similarity between a machine translation and a human reference translation based on the number of equal words in both strings. The developers of the BLEU score describe the principle underlying the score as follows: "*The closer a machine translation is to a professional human translation, the better it is.*" (Papineni et al. 2002:1; original emphasis)

Although BLEU is widely used in machine translation research, there are more sophisticated automatic translation quality evaluation metrics, which use the concept of word embeddings to judge the quality of a machine translation. An example is the BERTScore, which is based on the language model BERT. (You can find the original paper here). In comparison to BLEU, embedding-based evaluation metrics are able to take semantic information into account. For judging the machine translation quality, such embedding-based metrics work on the premise that two translations may use different words to convey the same meaning. (More details on BLEU, BERTScore, and other automatic evaluation metrics can be found in our Jupyter Notebook on Data Evaluation – Advanced Level.)

With the rise of LLMs, we now have new opportunities for automatic translation quality evaluation. For example, Kocmi/Federmann (2023) propose the automatic evaluation metric GEMBA (GPT Estimation Metric Based Assessment), which uses the LLM ChatGPT to automatically assess machine translation quality.

When applied in machine translation research, these automatic quality evaluation methods sometimes create the impression that machine-translated texts to have the same quality as those translated by professional translators—or even surpass the work of humans in terms of quality (cf. Hassan et al. 2018; Popel et al. 2020).

Even though, at times, machine translation yields astonishing translation results, quality judgements made by automatic quality evaluation metrics should be taken with a grain of salt. One reason for this is that automatic metrics are usually restricted to the sentence level without being able to take a wider context into account (cf. Rossi/Carré 2022:75). Also, judgements made by humans and automatic metrics may differ from each other (cf. Moorkens 2022:128). From an ethical perspective, such shortcomings of automatic quality evaluation may be an issue. As Moorkens (2022:128) explains, publications of tech companies and media reports spreading claims of perfect NMT output quality may contribute to a devaluation of professional translators' skills in the eye of the public.

7 Conclusion

In this paper, you have been introduced to a range of ethical questions to think about when handling translation data in different forms. It should have become clear that present practices of NMT system training are not overly concerned with an ethically responsible behaviour towards the translators who provide the training data for these systems. Because legal regulations restrict the reuse of human translated texts when personal data are involved, these data are usually removed for training purposes. Also, with more texts being translated automatically, the share of human-translated resources for NMT training decreases. This means that output quality may suffer.

Especially, since state-of-the-art NMT systems are no longer the only major AI-based tools that can be used for machine translation, new ethical challenges emerge: For instance, is has been discussed above that LLM's demands for mostly monolingual training data may lead to biased output.

In times of increasing automation of professional translation workflows, we should be aware of the significant role professional human translators play in the translation process. Establishing practices that are ethically responsible can help the translation industry to create a rewarding working environment where humans can benefit the most from modern technologies.

References

Baker, Mona (1996): Corpus-based translation studies. In: Somers, Harold (Ed.): *Terminology, LSP, and translation. Studies in language engineering in honour of Juan C. Sager*. Amsterdam: Benjamin's, 175–186.

Bender, Emily M./Gebru, Timnit/McMillan-Major, Angelina/Shmitchell, Shmargaret (2021): On the dangers of stochastic parrots: Can language models be too big? In: *FAccT '21: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. Association for Computing Machinery, 610–623. https://doi.org/10.1145/3442188.3445922.

Briakou, Eleftheria/Cherry, Colin/Foster, George (2023): *Searching for needles in a haystack*: On the role of incidental bilingualism in PaLM's translation capability. *arXiv*. https://doi.org/10.48550/arXiv.2305.10266.

Brown, Tom/Mann, Benjamin/Ryder, Nick/Subbiah, Melanie/Kaplan, Jared D./Dhariwal, Prafulla/Neelakantan, Arvind/Shyam, Pranav/Sastry, Girish/Askell, Amanda/Agarwal, Sandhini/Herbert-Voss, Ariel/Krueger, Gretchen/Henighan, Tom/Child, Rewon/Ramesh, Aditya/Ziegler, Daniel/Wu, Jeffrey/Winter, Clemens/Hesse, Chris/Chen, Mark/Sigler, Eric/Litwin, Mateusz/Gray, Scott/Chess, Benjamin/Clark, Jack/Berner, Christopher/McCandlish, Sam/Radford, Alec/Sutskever, Ilya/Amodei, Dario (2020): Language models are few-shot learners. In: Larochelle, Hugo/Ranzato, Marc'Aurelio/Hadsell, Raja/Balcan, Maria-Florina/Lin, Hsuan-Tien (Eds.): *Advances in neural information processing systems 33 (NeurIPS 2020)*, 1–25. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (7 May 2023).

Bubeck, Sébastien/Chandrasekaran, Varun/Eldan, Ronen/Gehrke, Johannes/Horvitz, Eric/Kamar, Ece/Lee, Peter/Lee, Yin Tat/Li, Yuanzhi/Lundberg, Scott/Nori, Harsha/Palangi, Hamid/Tulio Ribeiro, Marco/Zhang, Yi (2023): Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv*. https://doi.org/10.48550/arXiv.2303.12712.

Burchardt, Aljoscha (2013): Multidimensional Quality Metrics: A flexible system for assessing translation quality. In: *Proceedings of Translating and the Computer, 35*, 1–7. https://aclanthology.org/2013.tc-1.6.pdf (04 July 2023).

ELIS Research (2023). European Language Industry Survey 2023. https://elis-survey.org/wp-content/uploads/2023/03/ELIS-2023-report.pdf (18 July 2023).

European Commission's Directorate-General for Translation (n. d.): DGT-Translation memory. *European Commission's Joint Research Centre*. https://joint-research-centre.ec.europa.eu/language-technology-resources/dgt-translation-memory_en (28 February 2023).

Hassan, Hany/Aue, Anthony/Chen, Chang/Chowdhary, Vishal/Clark, Jonathan/Federmann, Christian/Huang, Xuedong/Junczys-Dowmunt, Marcin/Lewis, William/Li, Mu/Liu, Shujie/Liu, Tie-Yan/Luo, Renqian/ Menezes,Arul/ Qin, Tao/Seide, Frank/Tan, Xu/Tian, Fei/Wu, Lijun/Wu, Shuangzhi/Xia, Yingce/Zhang, Dongdong/Zhang, Zhirui/Zhou, Ming (2018): Achieving human parity on automatic Chinese to English news translation. *arXiv*. https://doi.org/10.48550/arXiv.1803.05567.

Hendy, Amr/Abdelrehim, Mohamed/Sharaf, Amr/Raunak, Vikas/Gabr, Mohamed/Matsushita, Hitokazu/Jin Kim, Young/Afify, Mohamed/Hassan Awadalla, Hany (2023): How good are GPT models at machine translation? A comprehensive evaluation. *arXiv*. https://doi.org/10.48550/arXiv.2302.09210.

Kocmi, Tom/Federmann, Christian (2023): Large language models are state-of-the-art evaluators of translation quality. In: Nurminen, Mary/Brenner, Judith/Koponen, Maarit/Latomaa, Sirkku/Mikhailov, Mikhail/Schierl, Frederike/Ranasinghe, Tharindu/Vanmassenhove, Eva/Alvarez Vidal, Sergi/Aranberri, Nora/Nunziatini, Mara/Parra Escartìn, Carla/Forcada, Mikel/Popovic, Maja/Scarton, Carolina/Moniz, Helena (Eds.): *Proceedings of the 24th annual conference of the European Association for Machine Translation (EAMT 2023)*. European Association for Machine Translation, 193–203. https://events.tuni.fi/uploads/2023/06/a52469c0-proceedings-eamt2023.pdf (19 June 2023).

Krings, Hans Peter (2001): *Repairing texts. Empirical investigations of machine translation post-editing processes*. Kent: University Press.

Lommel, Arle/Burchardt, Aljoscha/Macketanz, Vivien (2018): Will neural technology drive MT into the mainstream? *MultiLingual* 1 2018, 28–30. https://multilingual.com/pdf-issues/2018-01.pdf (17 July 2023).

Massardo, Isabella/van der Meer, Jaap/O'Brien, Sharon/Hollowood, Fred/Aranberri, Nora/Drescher, Katrin (2016): *MT post-editing guidelines*. Amsterdam: TAUS Signature Editions. https://www.taus.net/resources/reports/mt-post-editing-guidelines (19 July 2023).

Moorkens, Joss (2022): Ethics and machine translation. In: Kenny, Dorothy (Ed.): *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Berlin: Language Science Press, 121–140. https://doi.org/10.5281/zenodo.6653405.

Moorkens, Joss/Lewis, David/Reijers, Wessel/Vanmassenhove, Eva/Way, Andy (2016): Translation resources and translator disempowerment, In: Devillers, Laurence/Schuller, Björn/Mover Provost, Emily/Robinson, Peter/Mariani, Jaseph/Delaborde, Agnes (Eds.): *Proceedings of ETHI-CA2 2016: Ethics in corpus collection, annotation and application*. 49–53. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-ETHICA2_Proceedings.pdf (28 February 2023).

Moorkens, Joss/Lewis, Dave (2019): Copyright and the reuse of translation as data, In: O'Hagan, Minako (Ed.): *The Routledge handbook of translation and technology*. London: Routledge, 469–481. https://core.ac.uk/download/pdf/264422793.pdf (28 February 2023).

Moslem, Yasmin/Haque, Rejwanul/Kelleher, John/Way, Andy (2022): Domain-specific text generation for machine translation. In: Duh, Kevin/Guzmán, Francisco (Eds.): *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (volume 1: Research track)*. Association for Machine Translation in the Americas, 14–30. https://aclanthology.org/2022.amta-research.2/ (20 July 2023).

Papineni, Kishore/Roukos, Salim/Ward, Todd/Zhu, Wei-Jing (2002): BLEU: A method for automatic evaluation of machine translation. In: Isabelle, Pierre/Charniak, Eugene/Lin, Dekang (Eds.): *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 311–318. http://dx.doi.org/10.3115/1073083.1073135.

Peng, Keqin/Ding, Liang/Zhong, Qihuang/Shen, Li/Liu, Xuebo/Zhang, Min/Ouyang, Yuanxin/Tao, Dacheng (2023): Towards making the most of ChatGPT for machine translation. *arXiv*. https://doi.org/10.48550/arXiv.2303.13780.

Pérez-Ortiz, Juan Antonio/Forcada, Mikel L./Sánchez-Martinez, Felipe (2022): How neural machine translation works. In: Kenny, Dorothy (Ed.): *Machine translation for everyone. Empowering users in the age of artificial intelligence*. Berlin: Language Science Press, 141–164. https://doi.org/10.5281/zenodo.6653406.

Popel, Martin/Tomkova, Marketa/Tomek, Jakub/Kaiser, Łukasz/Uszkoreit, Jakob/Bojar, Ondřej/Žabokrtský, Zdeněk (2020): Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. *Nature Communications* 11, 1–15. https://doi.org/10.1038/s41467-020-18073-9.

Prates, Marcelo O. R./Avelar, Pedro H. C./Lamb, Luis (2019): Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications* 32, 6363–6381. https://doi.org/10.1007/s00521-019-04144-6.

Proton Technologies AG (2023): Complete guide to GDPR compliance. *GDPR.eu*. https://gdpr.eu/ (28 February 2023).

Raunak, Vikas/Sharaf, Amr/Awadallah, Hany Hassan/Menezes, Arul (2023): Leveraging GPT-4 for automatic translation post-editing. *arXiv*. https://doi.org/10.48550/arXiv.2305.14878.

Rossi, Caroline/Carré, Alice (2020): How to choose a suitable neural machine translation solution: Evaluation of MT quality. In: Dorothy Kenny (Ed.): *Machine translation for everyone. Empowering users in the age of artificial intelligence*. Berlin: Language Science Press, 51–80. https://doi.org/10.5281/zenodo.6653405.

Savoldi, Beatrice/Gaido, Marco/Bentivogli, Luisa/Negri, Matteo/Turchi, Marco (2021): Gender bias in machine translation. *Transactions of the Association for Computational Linguistics* 9, 845–874. https://doi.org/10.1162/tacl_a_00401.

Vanmassenhove, Eva/Shterionov, Dimitar/Gwilliam, Matthew (2021): Machine transla-tionese: Effects of algorithmic bias on linguistic complexity in machine translation. In: Merlo, Paola/Tiedemann, Jörg/Tsarfaty, Reut (Eds.): *Proceedings of the 16th con-ference of the European chapter of the Association for Computational Linguistics: Main volume*. Association for Computational Linguistics, 2203–2213. http://dx.doi.org/10.18653/v1/2021.eacl-main.188.

Vaswani, Ashish/Shazeer, Noam/Parmar, Niki/Uszkoreit, Jakob/Jones, Llion/Gomez, Aidan N./Kaiser, Łukasz/Polosukhin, Łukasz (2017): Attention is all you need. In: Guyon, Isabelle/Luxburg, Ulrike von/Bengio, Samy/Wallach, Hanna M./Fergus, Rob/Vishwanathan, S. V. N./Garnett, Roman (Eds.): *Advances in neural information processing systems 30 (NIPS 2017)*, 1–11. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (06 May 2023).

Walker Rettberg, Jill (2022): ChatGPT is multilingual but monocultural, and it's learning your values. *jill/txt blog*. https://jilltxt.net/right-now-chatgpt-is-multilingual-but-monocultural-but-its-learning-your-values/ (12 June 2023).

Way, Andy (2019): Machine translation: Where are we at today? In: Angelone, Erik/Ehrensberger-Dow, Maureen/Massey, Gary (Eds.): *The Bloomsbury companion to language industry studies*. London: Bloomsbury Academic, 311–332. http://dx.doi.org/10.5040/9781350024960.0018.

World Intellectual Property Organization (1979): *Berne convention for the protection of literary and artistic works (as amended on September 28, 1979)*. Geneva: World Intellectual Property Organization. http://www.wipo.int/wipolex/en/details.jsp?id=12214 (28 February 2023).

Zhao, Wayne Xin/Zhou, Kun/Li, Junyi/Tang, Tianyi/Wang, Xiaolei/Hou, Yupeng/Min, Yingqian/Zhang, Beichen/Zhang, Junjie/Dong, Zican/Du, Yifan/Yang, Chen/Chen, Yushuo/Chen, Zhipeng/Jiang, Jinhao/Ren, Ruiyang/Li, Yifan/Tang, Xinyu/Liu, Zikang/Liu, Peiyu/Nie, Jian-Yun/Wen, Ji-Rong (2023): A survey of large language models. *arXiv*. https://doi.org/10.48550/arXiv.2303.18223.