universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Before the Engine: Combining Writing for Machine Translation and Post-editing of the Source-Text to Improve Neural Machine Translation Output Quality

verfasst von | submitted by

Matilde Maria Boldrin

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 070 348 331
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Translation Italienisch Deutsch
Betreut von | Supervisor:	Univ.-Prof. Dragos Ioan Ciobanu PhD

Summary

1. Introduction **4**
2. A Brief History of Machine Translation: Technology and Humans **9**
2.1 Early Developments in Machine Translation: From Origins to the 1970s 9
2.2 From Transfer Systems to Corpus-Based Approaches (1980s–1990s) 11
2.2.1 The Emergence and Evolution of Computer-Assisted Translation Tools 14
2.3 The Rise of Neutral Machine Translation (2000s - today) 16
2.4.1 Corporate Optimism vs. Translator Reality (2025) 19
3. Pre- and Post-editing **20**
3.1 Pre-editing 22
3.1.1 Controlled Language 25
3.1.2 Writing for machine translation 28
3.2 Post-editing 31
3.3 "Post-editing" the source text 34
4. Research Questions and Methodology **38**
5. Experiment **43**
5.1 Original Text: Rationale for the Choice and Description 43
5.2 First editing: Pre-Editing according to Writing for Machine Translation 45
5.3 Second editing: Post-Editing of the Source Text 52
5.4 Operational use of DeepL (NMT) and Matecat 58
5.5 Participant instructions 59
5.6 Questionnaires 60
6. Results and Analysis **62**
6.1 Data preparation and scoring 62
6.2 Descriptive statistics: what the numbers show 63
6.3 Within-participant differences (V2−V1) and effect size (Cohen's dz) 66
6.4 Qualitative data from the questionnaires 67
6.5 Quantitative–qualitative triangulation for RQ1 and RQ2 74
6.6 Sintesi e takeaway 77
7. Discussion and Conclusion **80**
7.1 Future work and limitations 81
8. Bibliography **83**

1. Introduction

In today's highly globalised and multilingual landscape, translation plays a pivotal role not only as a practical tool for communication but as a structural necessity within large institutional frameworks. Nowhere is this more evident than in the European Union, where the coexistence of "28 member states, 500 million citizens, 3 alphabets and 24 official languages" (Vardaro et al., 2019:1) gives rise to 552 possible combinations of a source and a target language. Within such a setting, high-quality translation is indispensable to ensure mutual intelligibility across legal, administrative, and technical domains. The Directorate-General for Translation (DGT) of the European Commission alone received 73,000 translation requests in 2016, producing over 2.2 million pages—more than 650,000 of which were assigned to freelance translators (Vardaro et al., 2019:43). These figures illustrate not only the volume but also the operational intensity under which contemporary translation services must function.

The accelerating pace of globalisation and the proliferation of digital content have dramatically reshaped the landscape of the translation industry. As Doherty (2016) observes, the volume of material requiring translation has increased so significantly that "traditional human translation simply cannot keep up the pace with the translation needs of today (and tomorrow)." The demand is not limited to official or literary texts, but extends across a wide range of content types such as product documentation, user manuals, marketing copy, and websites. The exponential growth of global content has placed mounting pressure on the translation industry to deliver faster, more scalable solutions. This trend is clearly reflected in recent data from the Nimdzi 100 (2025), which reports that the global language services market has reached a value of over USD 70 billion, confirming a steady expansion driven by digitalisation, international trade, and multilingual communication needs. Alongside this growth, the industry has faced increasing demands for cost efficiency, pushing both agencies and professionals to rely more heavily on automation. In this evolving context, machine translation—especially when integrated with human post-editing - has emerged as a strategic response to the need for speed, volume, and budget control. The use of computer-assisted translation tools has become not only widespread but indispensable. Arenas (2019:349) notes that translation technologies—such as machine translation and computer-assisted translation tools—can speed up workflows and lighten cognitive load. Earlier, Hutchins (2003:23) observed that the use of technologies, like machine translation, might help maintain

terminological consistency and, in domains where full automation is acceptable, may even replace human input, reduce costs and increase productivity.

The field of machine translation (MT) has evolved dramatically in response to these demands. A major shift has occurred with the move from rule-based and statistical machine translation (SMT) approaches to neural machine translation (NMT), a transformation that redefined both the capabilities and expectations of MT systems. Since its deployment in 2017, the European Commission's eTranslation platform—trained on around 1.2 billion aligned segments—has enabled automated translation between all EU official languages across a variety of formats, including plain text, XML, and TMX-compatible files (Vardaro et al., 2019:3). While this development marks a significant technological achievement, it also raises new questions about the nature of quality, responsibility, and the role of human intervention. Despite NMT's enhanced fluency and semantic precision (Zhang et al. 2020), its output still frequently requires human editing—particularly in cases involving domain-specific terminology, syntactic complexity, or culture-bound elements.

This ongoing reliance on human oversight has elevated the role of post-editing as a structured and professionalised phase within the translation workflow. Rather than being a uniform or monolithic task, post-editing today is understood as a spectrum ranging from minimal intervention for basic intelligibility - commonly referred to as light post-editing - to full post-editing, which aims to match the fluency and accuracy of human translation (Arenas 2019:355). While the need for translators to engage in revisions that go beyond basic error correction has long been part of professional practice, the advent of neural machine translation (NMT) has shifted the nature and focus of these interventions. Complex revisions — including the enhancement of stylistic cohesion, the restoration of discourse-level consistency, and the adaptation of tone or register — remain essential, but are now often applied to machine-generated drafts rather than human translations. Despite the sophistication of NMT systems, many challenges persist. Most models operate on a sentence-by-sentence basis, overlooking broader textual phenomena such as anaphora resolution, thematic progression, and overall coherence (Bawden et al., 2018; Läubli et al., 2018; Voita et al., 2019; Lopes et al., 2020; Post and Junczys-Dowmunt, 2023). These limitations are particularly evident in longer or multi-paragraph texts, where inconsistencies or contextual mismatches can occur. Consequently, post-editors are often required not only to correct surface-level issues but also to address deeper structural and contextual deficiencies in the output (Garcia, 2015).

The professional profile of the post-editor has evolved over time, shaped not only by advances in machine translation but also by the long-standing presence of Computer-Assisted Translation (CAT) tools, which have been part of professional practice since the 1990s. Their widespread adoption during that decade was strongly influenced by the introduction of Trados, a pioneering CAT environment that became an industry standard (Chan, 2023:13). What has changed in recent years is the increasing integration of these tools with neural MT systems and formal quality evaluation frameworks, making technological literacy an essential component of the translator's skill set. As a result, today's professionals operate not merely as autonomous authors of the target text but as hybrid figures combining linguistic expertise, quality management, and system-aware editing. In this sense, post-editing has moved from being a purely corrective task to a central element of the translation process in the age of neural automation (Zhang, 2025).

Compared to the extensive discussion of post-editing, scholarly work on source text interventions appears to be less developed and generally receives more limited attention. The idea that adapting the source text could enhance MT output—thus acting as a form of upstream quality control—has so far been explored to a lesser extent. Guerberof Arenas (2019), for example, addresses both pre-editing and post-editing in a widely cited industry reference, yet allocates considerably more space and theoretical depth to the latter. In concrete terms, pre-editing is introduced briefly at the start of the chapter, without a dedicated abbreviation, whereas post-editing is followed by a detailed theoretical framework and is consistently referred to in abbreviated form. This difference in treatment suggests that, within the scope of the work, post-editing is viewed as more closely aligned with current professional priorities and market realities (pp. 343–348).

When discussed, interventions on the source side are often framed through the concept of controlled language, which involves adhering to predetermined lexical and grammatical conventions to achieve maximum clarity and simplicity—particularly in technical documentation (O'Brien, 2003). Arenas (2019:334) notes that pre-editing incorporates aspects of this approach but applies them in a more targeted way, focusing on preparing the original text for translation automation in order to improve the quality of the raw output. This preparation can include correcting typographical or factual inaccuracies, simplifying sentence structures, favouring straightforward grammatical patterns over more complex ones, maintaining consistent terminology, and flagging items such as product names that should remain untranslated. By doing so, pre-editing functions as a more adaptable, task-specific

adaptation of controlled language principles, tailored to the operational needs of MT workflows.

An alternative to such static pre-editing models is offered by Somers (1997), who proposed a cyclical editing model: *post-editing the source text*. This approach involves refining the source text iteratively, in response to recurring errors detected in MT output. Somers' model introduces a more dynamic and adaptive mechanism that could prove especially relevant in the neural translation era. However, despite its theoretical appeal, this method has yet to be empirically validated in recent research, leaving its potential benefits in contemporary NMT contexts largely untested.

The present study is positioned within this underdeveloped area. It investigates the impact of *post-editing the source text* - after an initial MT output - as a means of improving subsequent translations. This practice, though unconventional, reflects the increasing integration of human expertise and machine capabilities in modern workflows. In line with the framework outlined by Guerberof Arenas (2019), this workflow includes an initial pre-editing phase that precedes the actual post-editing of the source text. As described by Arenas, pre-editing often draws on principles of controlled language—applying predefined lexical and grammatical rules to ensure clarity, consistency, and simplicity—while also incorporating task-specific adjustments aimed at optimising the text for machine translation. In this study, the pre-editing stage is guided by the recommendations contained in the Translation Centre for the Bodies of the EU's Writing for Machine Translation, which will be examined later in the thesis. This compendium offers detailed guidance on drafting source texts in ways that minimise ambiguity and enhance machine translatability, thereby laying a solid foundation for the subsequent phase of source-text post-editing. Moreover, this perspective is closely aligned with the central premise of the present study, which assumes that combining an initial phase of pre-editing with subsequent source-text post-editing—as proposed in Somers' (1997) cyclical model—can lead to measurable improvements in neural machine translation output quality. The rationale is that pre-editing reduces potential sources of ambiguity and structural complexity before the text enters the MT system, while source-text post-editing, carried out after an initial machine translation, targets recurrent issues in order to refine subsequent outputs. Based on this premise, the research addresses two core questions: first, do NMT outputs from a source edited in two steps—(i) pre-editing and (ii) post-editing of the source text—outperform outputs from the unedited source? Second, is this NMT quality gain similar for IT→EN (UK) and IT→DE, or does it differ by language pair?

The working hypothesis is that targeted interventions at the source level—through this two-step editing process—will enhance lexical precision and overall clarity, resulting in better MT outputs. Furthermore, it is anticipated that the positive effects of this approach will be comparable across both language pairs, suggesting that the methodology could operate as a language-independent enhancement to translation workflows which use NMT. To evaluate this, the study adopts a dual approach: quantitative error analysis of MT outputs in DE and EN, using a customised adaptation of Matecat's error classification categories derived from the MQM framework, and qualitative user feedback gathered through structured surveys.

The thesis unfolds over seven chapters. Chapter 2 sketches a concise history of machine translation—from rule-based and statistical approaches to neural systems—and outlines the practical implications of this shift. Chapter 3 develops the conceptual framework, discussing pre-editing and controlled language, post-editing, and Somers' proposal of post-editing of the source text. Chapter 4 states the research questions and method: a within-subject design with two factors (workflow and language), the outcome measures (quality score/EPT and severity profile), the MQM-inspired error taxonomy, and the qualitative criteria adopted for the questionnaires (accuracy, fluency, fidelity). Chapter 5 describes the experiment in detail: the source text and domain, the two-step editing procedures (Writing for MT and Somers), the DeepL/Matecat configuration, participants, operational instructions (cf. Appendix B), and the five questionnaires. Chapter 6 reports the results and analysis, covering data preparation, descriptive statistics by condition, within-participant differences, triangulation with the questionnaire evidence, and direct answers to the two research questions, followed by a short synthesis. Chapter 7 closes with the discussion and conclusion, interpreting the findings, acknowledging limitations, and outlining future work.

2. A Brief History of Machine Translation: Technology and Humans

In the contemporary world, technology permeates nearly every sphere of human activity, and the translation sector is no exception. Advances in automation have transformed the speed and scale at which translation can be produced, raising new questions about the role of human translators. For instance, in 2016, Google Translate reported processing around 143 billion words per day in 100 language combinations—equivalent to roughly 20 words per person worldwide. While such figures reflect the growing reliance on automated systems, they do not necessarily indicate whether the quality is consistently sufficient; it is equally possible that users turn to machine translation (MT) even when results are suboptimal, depending on the purpose of the translation. The debate over MT quality remains highly active. Claims of so-called "human parity"—the idea that MT output matches the quality of professional human translations—have been made in high-profile cases (Hassan et al., 2018), but researchers have cautioned that such assertions require robust and transparent evidence (Läubli et al., 2018). Within the industry, satisfaction with MT quality is far from unanimous, and questions of accuracy and usability remain central. This issue will be explored in greater depth in the section dedicated to neural machine translation (NMT) in 2.3 of this work. Among the different MT paradigms, NMT has been identified as producing the highest quality output to date (Toral & Sánchez-Cartagena, 2017), though performance varies significantly across language pairs and domains. As Yonghui (2016) observes, even the best NMT systems may only "approach the accuracy achieved by average bilingual human translators" on certain test sets. The persistence of such limitations underscores the continued need for human involvement in quality-critical contexts.

At the same time, numerous studies highlight the positive impact of technology on translation, noting that integration has often fostered the development of new workflows and market opportunities (Way, 2020; Poibeau, 2017; Hassan et al., 2018). Market research further supports this view: according to the 2025 Nimdzi 100 ranking, the global language services market continues to expand, driven in part by the strategic integration of MT and post-editing into professional workflows (Nimdzi Insights, 2025).

2.1 Early Developments in Machine Translation: From Origins to the 1970s

According to the research of the last millennium, the effectiveness of machine translation was difficult to prove and it was unthinkable that translators could be replaced by machines. Translators began to consider the implications of machines as early as the unsuspected 17th century, when the use of a 'mechanical dictionary' was first hypothesised (Hutchins 2006).
The first significant contribution to research in this field came towards the end of the 1940s when the idea of using the first computers to translate natural language came to life. This gave an impetus to the research of the following years, which tried to find an answer to the first questions in the field, although it must be taken into account that it was difficult for most research groups to have access to computers, especially in the U.S.S.R. (cf. Poibeau 2017).

In June 1952, Yehoshua Bar-Hillel recognised that a "full automation of good quality translation was a virtual impossibility" and that "human intervention either before or after computer processes (known from the beginning as pre- and post-editing respectively) would be essential" (cf. Hutchins 2006). In the 1950s, research groups began to form all over the world, in Milan, Cambridge, Russia, China and Japan. During the 1950s and 1960s, efforts were concentrated on minimising the need for editing. These years saw the birth of what would later be called 'computational linguistics', using statistical methods to discover grammatical and lexical regularities that could be applied computationally.

Three main approaches to machine translation were developed at that time. Direct translation was the most popular and involved the use of programming rules specifically designed for translation into the target language, without any particular reorganisation of syntax or detailed analysis. In this approach to machine translation there was no need for intermediate representations of language, unlike in some other later approaches, as will be explained in the following paragraph.

The second approach taken into consideration in the present work is the so-called “interlingua model”, which is based on the use of a neural, abstract intermediate language (Hutchins 2006:3). In this case, translation takes place in two stages: first from the source language (SL) to the interlingua and then from the interlingua to the target language (TL). The aim of the “transfer approach” is to use syntactic analysis of the source language structures in order to avoid the word-for-word translation typical of direct systems, thus resulting in a more idiomatic translation (Poibeau 2017:27).

The momentum of these years came to a halt at the gates of the 1960s. It was in 1960 that Bar-Hillel questioned whether the aim of research should be to create a machine that

could translate at the same level as a human being. An insurmountable semantic barrier is mentioned, with the famous phrase "The box was in the pen" being cited as an example. Six years later in 1966, an ALPAC (Automatic Language Processing Advisory Committee) report stated that machine translation is slower, less accurate and twice as expensive as human translation and that "there is no immediate or predictable prospect of useful machine translation" (ALPAC 1966).

Between 1967 and 1976 there was a standstill in research and a consequent change of direction. Whereas previously the emphasis had mainly been on a direct approach, the ALPAC report favoured 'indirect' models such as interlingua and transfer-based approaches (Hutchins 1986). Research into the creation of a syntactic transfer system for English-French translation began in Montreal in 1970, with the development of a second experiment, Météo, designed to translate weather forecasts. Météo relied on a restricted vocabulary and a limited number of syntactic structures. A third experiment similar to Météo, attempted in the field of aviation, unfortunately failed and was abandoned in 1981, mainly due to the complexity of compound vocabulary and structures. During these years, the most innovative research focused on interlingua approaches. Between 1960 and 1971, a group of researchers headed by Bernard Vauquois at the University of Grenoble developed a system to translate mathematical and physical texts from Russian into French. A 'pivot language' was created (cf. Hutchins 2006). This was not a complete interlanguage approach, as it used an existing language. In the mid-1970s, interlingua approaches were also questioned because of the loss of information during translation and the weight that each individual error in the analysis could have on the final result: one error was enough to jeopardise the entire translation.

Between 1976 and 1989, a number of international institutions began to take an interest in machine translation applications, such as the Commission of the European Communities, which purchased an English-French version of Systran installation in 1976. The first version had been completed in 1970 by Peter Toma for the USAF Foreign Technology Division in the language combination Russian-English. It was an installation that originally offered a 'direct translation' design but was refined over the years to the 1976 version "with increased modularity and greater compatibility of the analysis and synthesis components of different versions, permitting cost reductions when developing new language pairs" (ibid.: 7). Systran was later installed by many international bodies such as NATO and the International Atomic Energy Authority and also by various companies such as General Motors, Dornier, and Aérospatiale. Most interestingly, it was modified for the Xerox Corporation, in this case the post-editing phase was completely eliminated thanks to a careful

analysis of the lexical and syntactical choices of technical manuals for translations from English into French, Italian, German, Spanish, Portuguese and some Scandinavian languages. Other systems, such as Systran, were developed in those years. However, they did not have a specific dictionary per se and had to be trained as needed (ibid.). The application of machine translation to the industry continued steadily.

2.2 From Transfer Systems to Corpus-Based Approaches (1980s–1990s)

In the 1980s Japan was the stage of a new frontier in machine translation. Major computer companies such as Fujitsu, Hitachi, NEC, Sharp, Toshiba developed software for computer-aided translation, especially for translation in the combinations Japanese-English and English-Japanese. These systems were based on both direct and transfer approaches and focused on syntactic and morphological information while paying less attention to lexical analysis. The quality of the final translation was closely linked to the presence of a pre- or post-editing translator.

These systems were still rudimentary but held the potential for future market developments. The 1980s gave a new impetus to research, that almost completely abandoned interlanguage systems in favour of transfer-based approaches. Amongst these approaches, the most popular are "syntax oriented and founded on the formalisation of lexical and grammatical rules influenced by linguistic theories of the time" (cf. Hutchins 2006). Linguistics-based transfer systems began to develop, characterised by a certain flexibility and modularity, and the concept of static and dynamic grammar. Few of these systems actually became operational, mainly because of problems with lexicon management. However, they have helped to revive interest in computational linguistics.

Thanks to parallel research in the field of artificial intelligence and cognitive linguistics, interlingua became an object of study again in the second half of the 1980s. Research in this area was particularly strong in Japan: it developed from transfer-based approaches and was supported by substantial funding from international collaborations (China, Indonesia, Malaysia and Thailand). Other important research projects were also carried out in Russia, North America, Europe, Korea and China. During these years, the idea that "means for improving MT quality would come from natural language processing research within the context of artificial intelligence (AI)" became particularly widespread (cf. Hutchins 2006: 9).

Since the mid-1980s, the perspective on the subject of machine translation changed. There was no longer an attempt to build complex systems that worked on several levels and

with many transformation and mapping rules. The aim was to create simpler systems in which the focus was not only on vocabulary and morphological and grammatical structures, but also on syntactic and semantic information, as well as non-linguistic and conceptual information (Hutchins 2006).

In the 1990s, research took a new direction. Starting in the 1980s with the advent of the first personal computers, the volume of available electronic texts increased. These included original texts with their translations that could be 'aligned'. Alignment is the process of saving sentence-pairs, that is ‘aligning’ the original sentence to its corresponding translation. The pair consisting of original and translated text is called a 'parallel corpus' (Poibeau 2017). Texts can be aligned with their translations at both sentence and word level. Aligning two languages with different syntactic structures and semantic connotations on a word level can be very difficult and in some cases counterproductive. In addition to its applications, the concept of sentence alignment is important because of the new meaning of the word 'sentence', i.e. 'linguistic units that is syntactically and semantically autonomous' (ibid.). This new definition of sentence as a linguistic unit shows the change of perspective from previous rule-based systems, which focused on smaller units such as single words instead. Alignment is particularly useful in the translation of technical texts where terminology, phraseology and style need to be consistent. The aligned text pairs are usually saved in a Translation Memory (TM) that can also be consulted for monolingual research to improve the fluency of the final translation (cf. Angelone et al. 2019).

During this period, two innovative approaches emerged alongside traditional transfer models: example-based translation and statistical machine translation. The former relied on leveraging past translations stored in databases to guide new translations, while the latter was trained on large bilingual corpora to identify probable correspondences between source and target segments (Huchins 2006). The statistical approach is more popular and is a translation system that is trained with as much data as possible, and does not merely use a single translation memory. Existing parallel corpora can be exploited for this purpose or some new ones can be created automatically. To create them, a system is needed that can go from one website to another via links and search for correspondences between source and target text. It can search, for example, whether a version in another language is available for the same web page. To ensure that one website is a translation of another, the system checks the length of the text, the layout and other elements that are not strictly linguistic, but allow an initial selection. This process is automatic and therefore does not guarantee the quality of the final

product. However, the risk of poor quality alignments is reduced by the large amount of data collected or by the creation of synthetic data.

However, statistical machine translation shows some limitations. The question arises whether it is really possible to translate only by putting together data collected from bilingual corpora. Secondly, one must consider the difficulty of using a statistical system to translate between two very different languages such as Chinese and Arabic. In his 2017 research, Thierry Poibeau (2017:76) states that statistical translation is the most practised but also that:

> [...] it is easier to translate between French or German and English than between Arabic and Japanese or Chinese and English. Even between these languages there are crucial differences. For example, translating from German into English works better than translating from English into German, since compound words in German (i.e., the complex combination of several simple words into a single string of characters) remain problematic for automatic processing.

This shows that the language direction can also influence the quality of the final translation, not only the language combination per se. It is easier to guarantee quality when working with more widely used languages on the web. In its early days, as it still used a statistical system, Google Translate started using English as a pivot language. However, even this strategy did not completely eliminate errors. Achieving a good level for new language combinations still remains a challenge nowadays.

Nevertheless, statistical systems were considered such a success that, in the 2000s, hybrid systems began to proliferate, which were still partly based on large lexicons and transfer rules, but which attempted to incorporate statistical information into their approach. Systran itself is an example of this.

2.2.1 The Emergence and Evolution of Computer-Assisted Translation Tools

While the late 1980s and 1990s saw the expansion of corpus-based approaches in machine translation (MT), the same period also laid the groundwork for another technological trajectory in the translation field: the development of Computer-Assisted Translation (CAT) tools. Emerging in response to the persistent limitations of fully automated MT, CAT tools shifted the focus from replacing human translators to supporting and enhancing their work. As early as 1966, the ALPAC report had expressed scepticism about the prospects of achieving high-quality, fully automatic translation in the short term, but it encouraged research into computational linguistics and forms of machine-aided human translation (Garcia, 2015). This orientation provided the conceptual basis for CAT technology: systems designed to retain the translator's central role while leveraging computational resources to deliver translations.

As Angelone et al. (2019: 388) observe, “CAT tools refer to computer programs that are specifically designed to help translators” by segmenting the source text and storing paired translations in a translation memory (TM), which can later be reused or adapted through features such as fuzzy matching and terminology look-up. By integrating dictionaries, corpora, and terminology databases, CAT tools allow translators to work fluidly, moving between segments, revising earlier decisions, and accessing contextual examples.

Bowker and Fisher (2010) situate CAT tools within a continuum between fully manual human translation and fully automated MT. In MT, the machine generates the draft translation with optional human intervention in pre- or post-editing. In contrast, CAT tools are fundamentally human-driven, with the translator actively controlling the process, making interpretive decisions, and using the system’s suggestions to guide their choices. As Craciunescu et al. (2004) describe, these tools consist of integrated utilities aimed at assisting translators by streamlining the workflow and optimising the use of linguistic resources.

The commercialisation of CAT tools began in the early 1990s, with landmark products such as Trados setting the standard for desktop-based translation environments (Garcia 2015). According to Vukalović (2021), the evolution of CAT tools can be divided into two broad phases. The “classic era” (1995–2005) was characterised by local installation, increasing computing power, and expanding storage capacity, which made it possible to handle larger bilingual corpora and more sophisticated translation memories. The “modern era” (2005–present) brought a shift to cloud-based solutions, enriched by collaborative functionalities and user-friendly interfaces. This second phase enabled real-time teamwork across geographical locations and facilitated integration with other digital resources, including MT engines.

At the heart of CAT systems are three core components. First, the Translation Memory (TM), which stores previously translated source–target sentence pairs and retrieves them when similar segments appear, thus promoting terminological consistency and reducing turnaround times (Bowker & Fisher, 2010). Second, termbases — specialised, often domain-specific glossaries — help standardise terminology and can be shared across projects and teams (Craciunescu et al., 2004). Third, the editor interface functions as the translator’s primary workspace, displaying source and target texts in parallel and enabling the integration of TM suggestions, termbase lookups, and machine translation outputs (Garcia, 2015).

In addition to these core elements, CAT environments typically incorporate a wide range of complementary features that support both linguistic and project management tasks. As Melby et al. (2023) point out, these may include “terminology management, alignment,

analysis, conversion, and knowledge management," as well as project management modules, spellchecking, word and line-counting tools, integration with MT engines, and support for multiple export formats. The convergence of these functionalities within a single interface allows translators to handle the different stages of a project — from preparation and terminology control to translation, revision, and final delivery — without leaving the CAT environment. This integration reflects a broader shift towards multifunctional, interactive workspaces where linguistic and administrative processes are increasingly intertwined.

One of the most significant strengths of CAT tools lies in their capacity to ensure terminological and stylistic coherence, especially in technical or highly specialised texts. Reusing previously translated segments not only accelerates the process but also improves accuracy in collaborative projects involving multiple translators (Craciunescu et al., 2004). Nevertheless, these systems have intrinsic limitations: they operate at the segment level and cannot independently restructure a text or adapt it to cultural and pragmatic requirements. As Craciunescu et al. (2004) emphasise, CAT tools do not replace the translator's interpretive skills or discursive awareness but function as productivity-enhancing instruments that preserve human agency in the communicative act.

The growing integration of MT into CAT environments, particularly neural MT, has further reshaped professional workflows. Many platforms now offer MT suggestions alongside TM matches, allowing translators to decide whether to accept, modify, or reject the automated proposals (Garcia, 2023). This hybrid configuration — where MT outputs are embedded within a human-controlled workspace — has become increasingly common, blurring the boundary between CAT-assisted translation and post-editing. Looking ahead, Bowker and Fisher (2010) anticipate advances in context-sensitive TM retrieval, more sophisticated collaborative platforms, and enhanced interoperability between CAT tools and emerging language technologies.

The development of CAT tools represents a parallel but interconnected strand in the history of translation technology. While MT has pursued the goal of autonomous output generation, CAT tools have consistently prioritised human oversight, positioning the translator at the centre of a technologically enriched workflow. Their evolution — from desktop-based systems in the 1990s to today's cloud-enabled collaborative platforms — reflects broader shifts in both computing capabilities and professional translation practices.

2.3 The Rise of Neutral Machine Translation (2000s - today)

Back in 1997, Somers argued that it was simply unrealistic to expect machines to deliver "fully automatic high quality translations of unrestricted input texts" (1997:194). At the time, this view reflected the prevailing scepticism towards the possibility of achieving human-like translation quality through automation. Yet less than a decade later, in 2006, two of the world's leading technology companies — Google and Microsoft — began investing heavily in consumer-oriented machine translation (MT) tools. These early systems, while innovative in their accessibility and scope, remained limited in performance. Much like their statistical and rule-based predecessors, they often produced output that was excessively literal. Idiomatic expressions — including metaphors and culturally bound references — were frequently mistranslated or omitted altogether, while overall readability remained inconsistent (Correa 2014; Niño 2009).

A decisive shift occurred between 2016 and 2017 with the introduction and rapid adoption of Neural Machine Translation (NMT). Inspired by the architecture and functioning of the human brain, NMT represented a clear departure from both phrase-based statistical models and earlier rule-based methods. Rather than breaking sentences into smaller segments and translating each in isolation, NMT processes the sentence as a whole, allowing it to capture a broader range of syntactic and semantic relationships. This shift has been described by Bentivogli et al. (2016:1) as:

> [NMTs represent] a further step in the evolution of rule-based approaches that explicitly manipulate knowledge, to the statistical/data-driven framework, still comprehensible in its inner workings, to a sub-symbolic framework in which the translation process is totally opaque to the analysis.

Early comparative studies already suggested that NMT could outperform Statistical Machine Translation (SMT) in several key areas. Bentivogli et al. (2016) found that NMT outputs required less post-editing effort, delivered more accurate renderings of texts with richer vocabulary, and contained fewer morphological, lexical, and word order errors. These findings indicated a substantial leap forward, even though — as Castilho et al. (2017) later observed — the improvements were not uniform across all text types or language pairs.

The commercial adoption of NMT quickly followed these research advances. In 2016, Google rolled out its neural translation engine, reporting reductions in translation errors of more than half, with estimates ranging from 55% to 85% (Le & Schuster 2016). Around the same time, DeepL launched its own neural MT service, which rapidly gained attention for its

fluency and stylistic accuracy. Both platforms have since expanded their coverage and refined their models, setting new benchmarks for online MT in terms of speed, coherence, and lexical choice.

While these developments have been widely welcomed, researchers also stress that NMT remains a relatively young technology, particularly from the perspective of post-editing (Bentivogli et al. 2016; Toral, Wieling & Way 2018). Also Arenas (2019: 349) points out that "it is still early days for NMT from a PE perspective, and even from a developer's perspective NMT is still in its infancy." Its strengths lie in improved fluency and more natural-sounding output, but persistent challenges remain, such as handling document-level context, managing domain-specific terminology, and maintaining accuracy in low-resource language pairs. As such, NMT is better understood as a significant milestone in the evolution of MT rather than a definitive endpoint. Its trajectory continues to be shaped by ongoing research into both system design and the role of human intervention, reinforcing the relevance of editing practices — pre- and post-editing alike — in maximising translation quality.

Over the past two decades, MT has undergone a deep transformation, driven by the transition from rule-based and statistical models to approaches based on deep learning. SMT, which dominated before the advent of NMT, relied on probabilistic models that calculated the likelihood of word and phrase correspondences across large bilingual corpora. These systems divided sentences into smaller units — words or short phrases — and translated them piece by piece, recombining them according to statistical rules derived from the training data. Although SMT represented a notable advance over purely rule-based systems, it struggled with phenomena such as long-distance dependencies, agreement, and idiomatic expressions, particularly because it could not model full-sentence context (Way 2020; Zhang et al. 2018).

NMT overcame many of these constraints by adopting an encoder–decoder architecture (Sutskever et al. 2014), later enhanced by Bahdanau et al. (2014) with attention mechanisms. The encoder transforms the source sentence into a dense, continuous "context vector" (Jia 2019) containing syntactic and semantic information. This process relies on word embeddings (Mikolov et al. 2013), which arrange words in a multidimensional space such that semantically or contextually similar terms are placed close to one another. As Kocmi et al. (2017: 1) note, "the vectors that embeddings assign to each word type are almost never provided manually but always discovered automatically in a neural network trained to carry out a particular task."

The decoder then generates the target sentence one token at a time, drawing on both the context vector and the sequence of previously generated words. Early encoder–decoder

models, however, experienced bottlenecks when dealing with long or complex sentences, as the fixed-length context vector could not retain all relevant information (Goldfeld 2020; Jia 2019). The introduction of attention mechanisms resolved much of this by allowing the model to focus on the most relevant portions of the source sentence for each output word (Way 2020).

The subsequent arrival of the Transformer model (Vaswani et al. 2017) replaced recurrence entirely with multi-head self-attention, enabling the model to consider dependencies between any two words regardless of their distance in the sentence. As Zhang et al. (2018) observe, Transformers not only improved translation accuracy but also enabled far greater parallelisation, significantly reducing training time. This architecture is now the foundation for commercial NMT systems like Google Translate and DeepL, whose capacity to handle complex syntactic structures and produce fluent output has raised user expectations.

Nevertheless, significant challenges remain. Most NMT systems still operate at the sentence level, which limits their ability to handle document-level phenomena such as cross-sentence coreference, lexical cohesion, and discourse structure. Zhang et al. (2018) propose augmenting Transformers with additional attention layers to incorporate context beyond the sentence boundary, but this requires large-scale document-level parallel corpora, which are scarce, and entails high computational costs.

Other persistent issues include low-resource languages, domain-specific terminology, and the treatment of out-of-vocabulary terms. Subword modelling techniques alleviate some of these problems but do not eliminate mistranslations or omissions, particularly with proper nouns or newly coined terms. Poibeau (2017) notes that deep learning-based systems are especially effective with short sentences but can experience degraded performance on longer, more complex inputs.

Despite these open questions, the advances brought about by NMT are undeniable. Its holistic processing of sentences, dynamic attention mechanisms, and capacity for stylistically natural output have redefined the landscape of automated translation. Yet, as Way (2020) cautions, NMT is not the endpoint of MT's evolution. The coming years will likely focus on expanding context awareness, improving interpretability, and enhancing robustness through techniques such as pre-trained language models, transfer learning, and hybrid architectures. The interplay between technological innovation and human linguistic expertise will remain central to achieving translations that meet the nuanced demands of cross-cultural communication.

2.4.1 Corporate Optimism vs. Translator Reality (2025)

In order to contextualise the present study within the broader translation market, it is useful to compare recent industry analyses that capture how machine translation (MT) and post-editing are shaping professional practices. Two widely cited sources — the Nimdzi 100 Report (2025) and the European Language Industry Survey (ELIS 2025) — offer sharply contrasting perspectives. While Nimdzi presents an optimistic outlook focused on growth, efficiency, and technological integration, ELIS documents the economic pressures and professional concerns that these same developments generate for smaller companies and independent translators.

From Nimdzi's standpoint, large language service providers (LSPs) view MT, combined with human post-editing, as an established productivity driver. Many of the top-ranked providers report sustained revenue growth — in some cases exceeding USD 1 billion — and attribute part of this success to hybrid workflows that enable the handling of larger volumes at reduced costs. This "human-in-the-loop" approach is framed not as a threat to human labour but as a strategic asset, enabling scalability and maintaining competitiveness in a global market.

By contrast, the ELIS 2025 Report paints a more critical picture from the perspective of small to medium-sized LSPs and freelance professionals. While MT is now used in over half of all professional translation projects (ELIS 2025:6), respondents report that this integration often serves as a cost-cutting measure rather than a quality-enhancing one. Price pressure, reduced work volumes, and the shift from human translation to discounted post-editing have led more than half of freelance respondents to question their professional future (ibid.:20), with almost a quarter considering leaving freelancing altogether.

This divergence underscores a central challenge for the industry: the benefits of MT adoption are unevenly distributed. Large LSPs can invest in proprietary engines and advanced workflows, while smaller players and individuals often adapt reactively, without the same resources or bargaining power. For many freelancers, the growing share of MTPE — now representing 27% of total revenue for language companies (ibid.:28) — translates into lower remuneration and limited control over working conditions (ibid.: 31).

In light of these findings, the significance of research into editing practices — both pre-editing and post-editing — becomes evident. Academic and industry interest in how human intervention can enhance MT output has been ongoing for decades, evolving alongside technological developments. The trends highlighted by Nimdzi and ELIS suggest that, as MT continues to occupy a central role in professional workflows, the ability to refine and adapt

texts remains a critical skill for sustaining quality and competitiveness. This broader context underlines why editing strategies, whether applied upstream to the source or downstream to the target, continue to be a relevant and active field of investigation within translation studies.

3. Pre- and Post-editing

Machine translation (MT) has long been recognised as a key research direction in the field of natural language processing, and it has evolved considerably over the decades. Early MT systems were predominantly rule-based, relying on extensive sets of linguistic rules and bilingual dictionaries manually crafted by experts. This approach, while innovative for its time, often failed to handle the full complexity and ambiguity inherent in natural language. Later developments saw the introduction of instance-based or example-based approaches, which tried to translate by analogy using large databases of translated sentence pairs. Subsequently, the advent of statistical machine translation (SMT) marked an important milestone, leveraging probabilistic models and large parallel corpora to generate translations based on statistical patterns.

Among the many innovations explored in the history of MT, the concept of a pivot language deserves mention, even though it is not the focus of the present study. As noted by Craciunescu et al. (2004), the aim of a pivot language is to represent the source text in a conceptual format that is neutral and independent of any particular language pair. This involves a two-step process: first, converting the source into a set of abstract semantic units; second, generating a coherent target-language text from these units. Unlike direct or transfer-based translation, this approach reformulates meaning at a deeper, more universal level. The now-historic DLT project, which employed Esperanto as a pivot between twelve European languages, exemplifies the theoretical ambition behind this method. Despite its elegance, practical implementation proved difficult, particularly due to the semantic disambiguation required. Nonetheless, pivot-based strategies remain a compelling research direction in multilingual MT systems (Craciunescu et al. 2004:6), especially when considered alongside transfer models.

The latest and arguably most transformative step has been the rise of neural machine translation (NMT), which utilises deep learning architectures to produce more fluent and natural output. Yet, as early as 1997, Somers already argued that editing was an unavoidable step, pointing out that “the original aim of MT – to provide fully automatic high quality

translations of unrestricted input texts – is beyond the current means of computer science and computational linguistics" (Somers 1997). At that time, the machine was clearly incapable of replicating the complexity and nuance of human language, and computers could not offer the same level of contextual understanding and pragmatic awareness as a human translator. Consequently, human intervention, in the form of revision and post-editing, remained indispensable, although less traditional than a decade earlier thanks to increasing computerisation of translation processes.

By the turn of the millennium, translators could already rely on a combination of computer-assisted tools, such as word processors, electronic dictionaries, and linguistic resources on CD-ROMs, alongside more traditional aids. Somers (1997) nevertheless acknowledged specific situations where machine translation alone could be sufficient, without any human editing. This mainly occurs when "the language and style of the input to the process (the source text) is restricted so that it is within the limitations of the MT system" (Somers, 1997:194), or where "output of mediocre quality is acceptable." The use of controlled language was, and still is, a practical solution to this challenge. Controlled language implies the deliberate simplification and standardisation of source texts to minimise ambiguity and make the text more amenable to machine translation. It is widely used in technical writing, especially in industries such as aerospace and engineering. Standards such as ASD-STE100 (Simplified Technical English) ensure that documents are clear, consistent and suitable for machine translation.

In addition to controlled inputs, there are also contexts in which a rough translation is entirely sufficient — for example, when a company only needs to grasp the general meaning of foreign-language documents to decide if further, more precise translation is necessary. Such gisting translations allow users to quickly identify relevant information without the need for fully polished output. Fast-forwarding to the present day, despite the impressive technological advances in NMT, the need for human editing remains. As Zahng (2025) argues, reaching the same level of quality as human translation — particularly in terms of accuracy and fluency of language expression — remains a challenge. One critical limitation is that most NMT models still translate sentences in isolation, without adequately capturing document-level phenomena such as textual cohesion, discourse coherence, or intersentential anaphora resolution (Bawden et al., 2018; Läubli et al., 2018; Voita et al., 2019; Lopes et al., 2020; Post and Junczys-Dowmunt, 2023). This shortcoming can lead to inconsistencies or misunderstandings, especially in longer texts where context is vital.

In response to these limitations, the practice of post-editing has become increasingly structured and professionalised. Today, translators often distinguish between light post-editing, which aims for basic comprehensibility and minimal corrections, and full post-editing, which seeks to ensure the final text reads as naturally and accurately as a human translation. This shift means that modern translators must acquire additional skills beyond language proficiency, including familiarity with MT systems, quality assessment, and even a certain level of technical expertise in working with translation memories and CAT (Computer-Assisted Translation) tools.

In conclusion, while the ultimate goal of fully automatic high-quality translation remains elusive for unrestricted input, the synergy between advanced MT systems and skilled human editors continues to define the state of the art. As MT technologies evolve, so too does the role of the translator, who increasingly acts not only as a linguistic expert but also as a quality manager, reviser, and technology-savvy language professional.

3.1 Pre-editing

Within the broader framework of Machine Translation (MT), pre-editing plays a crucial role in ensuring the overall quality and efficiency of the translation process. Unlike post-editing, which focuses on correcting the machine-generated output, pre-editing is concerned with the careful preparation and refinement of the source text before translation takes place. This strategic stage aims to minimise ambiguities, reduce linguistic complexity, and provide an input that is optimally structured for computational processing. When pre-editing is effectively implemented, it can significantly lower the time and costs associated with post-editing while also contributing to more consistent and reliable translations (Boix 2016).

The advancement of natural language processing and deep learning techniques has further reinforced the practical relevance of pre-editing by offering more robust tools for analysing and modifying source texts. According to Asma (2020), pre-editing extends beyond mere surface-level pre-processing and instead encompasses in-depth analysis, refinement, and targeted adjustments. This comprehensive intervention serves to make the source text more accessible for MT systems, thereby enhancing the accuracy and fluency of the target output.

One of the fundamental requirements for effective pre-editing is the correction of basic linguistic errors present in the source text. Editors engaged in this task must possess solid linguistic competence in order to identify and rectify issues such as grammatical mistakes, spelling errors, incorrect punctuation, and typographical inconsistencies. Although

these problems may appear minor, they often function as significant “stumbling blocks” during translation, disrupting the system’s processing and generating errors that require further correction at the output stage (Bernth & Gdaniec, 2001).

Beyond the removal of superficial errors, pre-editing involves a deeper level of structural optimisation. This includes revising syntactic constructions, simplifying unnecessarily complex sentences, clarifying potentially ambiguous expressions, and avoiding the use of idioms or slang that the MT system may struggle to interpret accurately (Zhang 2025). Such refinements are intended to preserve the original meaning while reshaping the text to align with the operational ‘preferences’ of the translation engine. Bahdanau (2014) emphasises that this step entails more than stylistic improvements; it requires targeted revisions that help the machine to process the content more efficiently without distorting its intended message.

Terminological consistency represents another core element of pre-editing, especially in technical or domain-specific contexts. Ensuring that key terms are used uniformly throughout the source text is vital to avoid misunderstandings or mistranslations that could compromise the accuracy of the final product. Previous research has highlighted the importance of this practice, demonstrating how the standardisation of terminology prior to translation reduces the risk of inconsistent outputs and enhances overall coherence (Bouillon et al. 2014). This step often involves consulting specialised glossaries or term banks to verify correct usage and maintain alignment with industry standards.

Equally important is the need to respect the stylistic features and register of the source text. Different genres and text types demand varying degrees of formality, tone, and rhetorical devices. For example, technical documentation may prioritise clarity and conciseness, whereas literary works place greater emphasis on stylistic nuance and figurative language. Pre-editing thus demands a delicate equilibrium: the text must be reformulated in a way that facilitates machine translation, while still preserving the stylistic character that ought to be conveyed in the final output (Zhang 2025).

A persistent challenge in machine translation is the system’s limited capacity to understand complex contextual and rhetorical elements. Figurative language, metaphors, and other forms of implicit meaning often pose difficulties for automated systems that lack true semantic awareness. As noted by Cho et al. (2014), addressing these challenges during pre-editing demands a thorough contextual analysis. This may involve rephrasing figurative expressions in more explicit terms or providing clarifying details to reduce the likelihood of misinterpretation in the output.

Studies such as Ive (2018) illustrate how pre-editing can be integrated systematically into professional translation workflows. A structured approach typically involves identifying segments that are likely to present difficulties, implementing targeted corrections, and incorporating the refined version into the MT process. This method enables more stable and predictable translation results by minimising the need for repetitive corrections during post-editing. In large-scale projects, particularly those involving repetitive or formulaic language, the impact of this approach becomes even more pronounced.

Moreover, pre-editing is highly context-dependent and must be adapted to the specific characteristics of the MT system in use. Different systems may vary significantly in their handling of syntactic complexity, idiomatic language, or domain-specific terminology. Editors therefore benefit from a sound understanding of the system's limitations and error patterns, allowing them to anticipate problematic constructions and adjust the source text accordingly (Ive, 2018). This adaptability contributes directly to the overall cost-effectiveness and quality of the translation process.

An additional consideration is the professional expertise required to carry out pre-editing effectively. Ive (2018) points out that linguistic proficiency alone is insufficient; editors must also possess practical experience with machine translation systems and an awareness of the target language's syntactic and semantic structures. This combination of skills ensures that revisions support the system's strengths while avoiding unnecessary distortions of meaning.

Although pre-editing entails an investment of time and effort at the initial stage of the workflow, this upfront commitment frequently results in significant downstream savings. Once recurring errors are addressed in the source text, there is no need to correct them multiple times in the output, as would be necessary in workflows relying solely on post-editing. This is particularly relevant in domains such as legal or medical translation, where even minor inconsistencies can lead to serious consequences if left uncorrected (Ive, 2018).

Furthermore, pre-editing aligns well with the growing emphasis on hybrid models that integrate human expertise and adaptive MT technologies. When combined with interactive feedback loops, pre-editing can contribute to more dynamic learning processes within the system, enabling gradual improvements in its handling of complex or domain-specific texts. Nevertheless, caution is needed to ensure that excessive human intervention does not inadvertently introduce inconsistencies or degrade system performance over time (ibid.).

Although pre-editing offers evident benefits, its application in practice presents several difficulties. Resource constraints, tight deadlines, and varying levels of editorial expertise can limit the extent to which this approach can be applied in everyday practice. As the literature suggests, the effectiveness of pre-editing depends heavily on the editor's capacity to identify high-risk segments and make appropriate adjustments. While technological advancements in quality estimation tools offer valuable support, they are unlikely to fully replace the nuanced judgment that experienced language professionals bring to the process.

In conclusion, pre-editing represents a vital element within contemporary MT workflows, serving as a proactive strategy to enhance translation quality. By addressing linguistic errors, optimising sentence structures, ensuring terminological coherence, and preserving stylistic integrity, pre-editing minimises the risk of downstream issues that would otherwise demand more extensive post-editing interventions. As research continues to refine best practices and develop supporting tools, the role of pre-editing is likely to expand further, especially in fields where precision and cost-efficiency are critical. The broader significance of pre-editing lies in its ability to bridge the inherent gap between human linguistic insight and machine processing capabilities. Rather than viewing MT as a purely automated solution, pre-editing underscores the value of strategic human involvement as a means of leveraging the full potential of advanced translation systems. By preparing source texts that align with the system's operational strengths, translators and editors ensure that machine translation can deliver outputs that meet increasingly demanding quality standards across diverse professional contexts.

3.1.1 Controlled Language

The concept of Controlled Language (CL) originally emerged as a practical solution to facilitate the reading and comprehension of technical or specialized texts, especially for non-native readers. By relying on carefully defined stylistic rules and specific terminological conventions, CL aims to reduce ambiguity and ensure that texts remain clear, consistent, and easily understandable. These key features made controlled language an attractive tool for improving the efficiency and quality of machine translation (MT) as well (cf. Elliston 1979; Pym 1990; CLAW 1996).

When examining the development and application of CL more closely, it becomes clear that its intended functionality is essentially twofold: it targets both the human reader and the machine as end-users. For human readers, controlled language seeks to enhance

readability, comprehensibility, clarity, and consistency, aspects that are well supported by the findings of Lehrndorfer (1996) and Spyridakis et al. (1997). As noted by Reuther (2003), this approach proves advantageous not only for non-native speakers, who may struggle with idiomatic language, but also for native speakers working with specialised or highly technical texts, where accuracy and clarity take precedence. In all these contexts, the design and implementation of a controlled language system are shaped by principles from text linguistics and cognitive processing, ensuring that the target audience can easily interpret the information provided.

The second key target of controlled language is the machine itself. CL could in fact purposefully be developed to effectively translate with a machine. Here, it is vital to consider the type of technological tool in use, whether it is a translation memory system or a machine translation system, since different technologies present different constraints and affordances. According to Reuther's research (2003), there exists an additional cross-dimensional aspect that intersects both readability and translatability. This dimension aims to balance the requirements of producing a text that is accessible to humans yet also sufficiently structured to be processed accurately by a machine, without appearing overly mechanical or, conversely, too idiomatic. In his study, which focuses on the German language, Reuther develops a tailored controlled language system, but the general categories he outlines can be adapted for other languages, including Italian, which is relevant to the present research.

Reuther (2003) identifies several levels of intervention when applying controlled language principles: these include lexical, formatting, phrase, and sentence levels. At the lexical level, CL aims to minimize ambiguities that may arise from different spelling variants. For example, he points out that while a human reader would have no trouble interpreting both "Lambdasonde" and "Lambda-Sonde," a machine might produce two different translations for the same concept (Reuther 2003:2). Similarly, the use of morphological variants like "Abkühlungsvorgang" versus "Abkühlvorgang" may pose no issue for human comprehension but could result in inconsistent output in MT systems. Moreover, the excessive use of synonyms is discouraged, as it may lead to ambiguity for both human readers and machine translation systems.

Formatting interventions are less significant for human translation but become essential when working with MT systems. This is particularly relevant when dealing with specific translation software inputs that require precise formatting conventions. Reuther (2003:2) provides useful guidelines in this regard, highlighting, for example, the role of punctuation marks: inconsistent punctuation may negatively affect the output quality of MT.

Similarly, incorrect spacing or inconsistent typographical elements—such as various ways of presenting lists—can complicate automatic processing.

Perhaps the most visible interventions when employing controlled language occur at the phrase and sentence levels. Reuther (2003) lists several key recommendations in this area, such as avoiding ambiguous constructions. A classic example is the phrase "the chickens are ready to eat," which in German could be rendered either as "Die Hühner sind bereit zu fressen," implying the animals are about to eat, or "Die Hühner sind fertig zum Essen," meaning the chickens are ready to be eaten. Such ambiguities, while possibly resolved by human translators through context, pose significant challenges for MT. Another rule is to avoid pronouns that can refer to multiple antecedents. The example "Die Kanne traf die Fensterscheibe, dabei zerbrach sie" illustrates this well: the pronoun "sie" could refer to either the jug or the window, leading to different translation outcomes.

Complex formulations are likewise discouraged. Both humans and machines may struggle to parse highly intricate sentences, and readability can suffer, especially in MT contexts. While human translators or translation memory systems may handle complex sentences more effectively—particularly if matching segments already exist in the memory—the likelihood of finding an exact match decreases as sentence complexity increases. For this reason, elliptical constructions should also be avoided. Reuther (2003) provides an example with the German sentence "Ist die Betriebstemperatur erreicht, erlischt die Kontrolllampe," which lacks the conjunction "wenn," leading to potential parsing issues for both human readers and MT systems.

An additional factor to take into account is the arrangement of words within a sentence. Even when grammatically correct, awkward word order can reduce naturalness and affect the quality of MT output. For example, "Today I am going to visit my grandma" is grammatically acceptable, but in English, time adverbs are typically placed at the end of the sentence, and deviation from this pattern might sound unnatural, even if the MT system produces a technically correct output.

Stylistic guidelines are also crucial. Reuther (2003:3) recommends avoiding passive constructions, future tense verbs, and negations, as these can impact readability due to the cognitive demands they place on human processing. While such stylistic choices are less consequential for human translators or translation memory systems, they can significantly influence the quality of fully automated translation output. An illustrative example is "Es ist darauf zu achten, dass alle Ventile geschlossen sind." Here, the combination of an infinitive

verb (“sein”) with a prepositional construction (“zu achten”) and a subordinate clause (“dass”) makes the sentence unnecessarily complex for MT.

In conclusion, it is evident that there is no single, universally applicable definition of controlled language. Its characteristics are entirely dependent on its intended use, the nature of the text, and the type of translation technology employed. As will be explained in detail in the chapter dedicated to the experimental framework of this thesis, the specific approach adopted for this research includes pre-editing, post-editing, and controlled language interventions. Although the categories defined by controlled language remain highly valuable, it is important to recognise that the traditional framework has certain limitations, as it tends to constrain both the range of text types and the stylistic flexibility of the final output.

An illustrative historical example of controlled language is the Caterpillar Technical English (CTE), which, although no longer widely in use, remains significant from a research perspective (Kamprath et al. 1998). Caterpillar Inc., a heavy equipment manufacturing company based in Peoria, Illinois, produces and distributes large quantities of equipment and parts worldwide. Each product involves a range of systems—such as engines, hydraulic systems, drive systems, implements, and electrical components—for which detailed technical documentation must be produced in multiple languages. To streamline this process, Caterpillar developed CTE in collaboration with Carnegie Mellon University’s Center for Machine Translation (CMT) and Carnegie Group Incorporated (CGI). This controlled English system was built around a carefully curated lexicon of approximately 70,000 unambiguous terms, each with a specific and limited semantic scope. This approach offered two key advantages: on the one hand, it provided the company’s automated MT systems with highly controlled input, thus improving translation quality and reducing the costs of human translation and post-editing; on the other hand, it ensured consistency in terminology and writing style across all documentation.

In the context of the present research, it is crucial to emphasize that CL is just one element within a broader strategy that combines pre-editing, post-editing, and other quality assurance measures. While the categories proposed by controlled language frameworks remain useful, the concept itself must be adapted to modern needs, as strict application can overly restrict the style and flexibility of the resulting texts. Therefore, CL should be understood as a dynamic tool that, when combined with editing techniques and context-sensitive interventions, contributes to the overall goal of producing texts that are both comprehensible to humans and effectively processable by machines.

3.1.2 Writing for machine translation

In the previous chapter, the concept of Controlled Language (CL) was examined. Originally developed to enhance the readability and comprehensibility of technical or specialised texts—particularly for non-native readers—CL is characterised by the application of rigorous stylistic and terminological rules across various linguistic levels, including lexis, syntax, formatting, and style. Its primary objectives are to reduce ambiguity, ensure consistency, and facilitate processing by both human readers and machine translation (MT) systems.

While writing for machine translation (WMT) shares with CL a focus on clarity and the reduction of ambiguity, its orientation is more specific: it is designed to optimise the source text for processing by MT systems, particularly those based on neural architectures. Whereas CL can represent a broad and potentially rigid normative framework, WMT adopts a more flexible and adaptive approach, tailoring its recommendations to the domain, the technology in use, and the intended workflow.

The present chapter therefore explores WMT as a set of operational practices that, while partly inspired by the principles of CL, represent a more targeted and pragmatic application. The discussion has two main objectives: first, to identify the general principles guiding the drafting of "machine-friendly" texts, and second, to compare two relevant approaches—one institutional, from the Translation Centre for the Bodies of the European Union, and one corporate, from Red Hat Localization Services—in order to understand how these guidelines are applied in different contexts. The concept of WMT arises from the need to adapt texts, particularly technical and institutional content, to the characteristics and limitations of MT systems. With the widespread adoption of neural machine translation (NMT), the quality of the source text plays a decisive role in determining the quality of the final output.

As noted by the Translation Centre for the Bodies of the European Union (2021), in a multilingual environment such as the European Union, it is increasingly likely that a text will be processed by a machine at some stage. In this context, the source text must be "machine-friendly"—that is, structured in a way that facilitates NMT processing and minimises the risk of systematic errors. The Centre emphasises that NMT processes sentences as isolated units, without drawing on the broader context of surrounding sentences; hence, clear and linear writing is essential to ensure coherence and accuracy. The Centre's view of WMT is therefore closely tied to preventing common NMT errors, whereas the approach

outlined by R. Hardin et al. focuses on the seamless integration of MT into the overall production process.

The work of R. Hardin, Ito, and Sasaki (2020), developed in the corporate setting of Red Hat Localization Services, positions WMT within a collaborative workflow between technical writers and translators, where MT is combined with translation memories and human revision. In this model, producing MT-friendly content is not simply a matter of following general rules; rather, it is an integral part of documentation design. A technical writer familiar with MT processes, and working in close cooperation with translators, can make design decisions that maximise MT efficiency and reduce the need for subsequent editing.

Both R. Hardin et al. and the Translation Centre identify key objectives for WMT. For R. Hardin et al., the priority is the quality of the translated output: the Red Hat guidelines stress that translation quality is closely linked to the quality of the source text, and that authors should make linguistic choices that enhance MT performance. The Translation Centre, by contrast, places greater emphasis on process efficiency: clear and consistent source texts speed up MT processing and reduce the need for post-editing. Both approaches agree on accessibility and clarity as fundamental goals, recognising that the text must be easy to interpret and free from ambiguity for both humans and machines.

In terms of underlying principles, the two approaches display substantial convergence, albeit with subtle differences. Both highlight the importance of clarity and ambiguity reduction, as well as terminological and stylistic consistency throughout the text. The idea of well-structured sentences containing a single main idea is also shared, though expressed differently: the Translation Centre formulates it as the explicit rule “not too long / not too short”, while R. Hardin et al. frame it in terms of “checking sentence logic” and “avoiding ambiguity”, in a less prescriptive manner. Similarly, both acknowledge the relevance of grammatical and orthographic correctness: in the Translation Centre’s compendium, it appears as part of the general rules, whereas in the Red Hat document it is presented as a best practice rather than a universal principle.

Rules from the Translation Centre for the Bodies of the European Union (2021):

1. Use consistent terminology for the same concept, avoiding synonyms.
2. Limit the use of pronouns and prefer explicit nouns.
3. Avoid unnecessary abbreviations and acronyms; when used, accompany them with the full form.
4. Keep sentences of moderate length (neither longer than 60 words nor shorter than 7).

5. Avoid line breaks in headings and tables.
6. Use punctuation correctly.
7. Standardise the use of quotation marks, bullet points, and symbols.
8. Avoid sentences written entirely in uppercase.
9. Other tips, e.g. prefer full words over signs or symbols that MT systems may fail to recognise.

Rules from Red Hat Localization Services – R. Hardin, Ito, Sasaki (2020):

1. Apply the principles of minimalism:
    - focus on reader goals,
    - use clear headings,
    - remove unnecessary content (fluff).
2. Check sentence logic to remove ambiguity.
3. Eliminate ambiguous words or expressions.
4. Avoid longer sentences (under 25 words).
5. Maintain correct grammar (agreement, structure, punctuation).
6. Follow corporate style and terminology guidelines for consistency.

When compared directly, the differences between the two approaches become clearer. The Translation Centre offers a highly normative and prescriptive model, designed to reduce MT errors through a universal set of writing rules. In this model, the author bears primary responsibility for the quality of the MT output and is expected to follow nine operational recommendations, based on common NMT error patterns. The technological focus is explicitly on the inherent limitations of NMT, particularly its sentence-by-sentence processing, and the quality strategy relies on preventing errors at the drafting stage. This model is intended to be transferable across contexts, especially for institutional and administrative texts requiring uniformity and stylistic consistency.

By contrast, the Red Hat Localization Services approach, as described by R. Hardin, Ito, and Sasaki, is more strategic and collaborative. Its aim is not only to improve MT output quality but also to optimise the entire workflow in which MT is integrated with translation memories and human post-editing. Here, the author is seen as part of a team engaged in continuous feedback with translators and reviewers. Instead of adhering to a fixed set of universal rules, the approach is grounded in the principles of minimalism and in-house style guidelines, tailored to specific operational needs. Quality is achieved not solely through error prevention at the source, but through a shared process that spans the full translation cycle.

This more flexible perspective allows strategies to be adapted to the requirements of each project, integrating WMT practices with targeted pre-editing and terminology optimisation.

In summary, the Translation Centre's model prioritises standardisation and prescriptive rules to ensure universally "machine-friendly" texts, whereas Red Hat emphasises collaboration among stakeholders and the adaptation of guidelines to the operational context, viewing WMT as part of a broader ecosystem of tools and human expertise. The methodology chapter will detail the specific approach adopted in the present study, explaining the rationale behind this choice and how WMT practices were applied to the source text used in the experiment.

3.2 Post-editing

Post-editing, the process of revising and improving machine translation (MT) output by human intervention, represents one of the most longstanding forms of collaboration between human cognition and computational assistance in the field of translation. This hybrid process, despite its recent surge in scholarly and industrial interest, has roots that reach back to the very inception of operational MT systems. Winther Balling (2014) asserts that post-editing is perhaps the oldest example of human–machine symbiosis in translation practices, emerging concurrently with the first practical MT systems. In the mid-twentieth century, early research aimed to minimise human involvement entirely, seeking fully automatic high-quality translations (Reifler 1952). Nevertheless, some scholars such as Bar-Hillel (1951) recognised that the human element could not be completely excluded, coining the term "human partners" to describe post-editors even when they lacked proficiency in the source language.

This process has evolved gradually over time. Initially, research and practice focused on Automatic Post-Editing (APE), a method that aimed to correct MT output through a secondary automated system. The concept was first discussed by Knight and Chander in 1994, who envisioned the use of an automatic editing tool operating alongside MT engines. In the years that followed, it became evident that certain recurrent errors could be mitigated through the use of controlled language. At the International Controlled Language Application Workshop in 2000, Allen and Hogan presented an APE module that relied precisely on this principle. Their system had been trained using three aligned components: the source text, the raw MT output, and the corresponding post-edited version. From this aligned dataset, rules for post-editing could be extracted and then applied to new texts.

According to Escribe and Mitkov (2021), this same training approach was still being used in APE systems as recently as 2021. In their work, they also note that more sophisticated methods are gradually emerging to improve the training phase. The evolution of APE runs parallel to that of MT itself, as both rely on similar underlying technologies. With the growing use of Neural Machine Translation (NMT), APE methods have undergone significant changes, and notable improvements have been recorded since the adoption of neural approaches. However, as Carmo et al. (2021) observe, APE is not always the optimal solution: in some cases, it may lead to overcorrections or may fail to address certain errors altogether. They further argue that APE is only worthwhile if it guarantees a measurable improvement over the raw MT output. In the case of NMT, Carmo et al. maintain that working with neural systems is generally advantageous, but that the margin of improvement offered by APE is not always substantial enough to justify its application.

Regardless of the potential benefits, APE is not the final stage in the production of a translation. A proofreading phase remains necessary, as Escribe and Mitkov (2021:168) explicitly point out: “an APE output necessitates more effort than post-editing alone, as extra attention should be paid to overcorrections and unspotted errors.” This underscores a key discrepancy between the theoretical promise of APE systems and the practical needs of MT users.

To address some of these limitations, interactive models for MT have been developed. These systems use an interface to collect and reuse the corrections made by post-editors during the translation process. One of the most notable features of interactive MT is a form of autocompletion, in which the system proposes suffixes based on translation prefixes already validated by the post-editor. The post-editor may choose to accept these suggestions or modify them, creating a dynamic feedback loop between human and machine (Escribe & Mitkov 2021). In 2019, Peris and Casacuberta focused their research precisely on the issue of reducing effort in interactive neural machine translation, aiming to refine these systems to better support translators.

Today, the landscape of post-editing has radically shifted. The proliferation of free and commercially accessible MT software, coupled with the rising quality of machine-generated output, has made post-editing a central aspect of modern translation workflows. In contrast to the initial MT paradigms where humans assisted machines, the current model repositions the translator at the heart of the process, supported by a suite of digital tools such as terminology management systems, translation memories (TMs), and computer-assisted translation (CAT) environments.

Significantly, post-editing is no longer conceptualised merely as a corrective phase after MT output. Instead, it is recognised as a distinct professional process with its own international standards, including ISO 18587 (2017), which delineates norms for post-editing services and outcomes. The increasing sophistication of MT systems—particularly with the advent of NMT—has transformed post-editing from a linear correction task to a multifaceted activity that interfaces dynamically with interactive translation tools.

NMT represents a major technological leap from Statistical Machine Translation (SMT), offering enhanced fluency and improved automatic evaluation metrics (Castilho et al. 2018). Nonetheless, results from human evaluators remain mixed. NMT outputs, while smoother in linguistic surface, tend to conceal semantic errors that are harder to detect (Ustaszewski 2019), especially without the original source text. This shift underscores the growing complexity of post-editing, as translators are now required to assess not just surface fluency but also deep semantic accuracy. In some cases, NMT output is so well-formed linguistically that errors pass unnoticed, which may increase the cognitive demands on post-editors (Castilho et al. 2017).

Despite clear gains in productivity, many professional translators express resistance to post-editing assignments. One key explanation lies in the heightened cognitive demands of the task (Béchara et al. 2021). Unlike human translation, where the translator works directly from the source, post-editing involves constant comparison and evaluation between the MT output and the original text. This dual-layered operation introduces additional mental load, as the translator must both identify and correct errors while maintaining awareness of the intended meaning in both source and target languages. A crucial moderator of this cognitive burden is the quality of the raw MT output. Productivity gains materialise only when the system delivers reasonably adequate drafts; if the output is weak, the time and mental energy spent hunting down and repairing errors can rival—or even exceed—the demands of translating from scratch. Choosing PE is therefore not only an economic or time-management decision, but also a judgement about the total load placed on the translator.

Following Koponen (2012), it is useful to distinguish three intertwined dimensions of effort. Cognitive effort covers the mental work of comprehension, error detection, hypothesis testing, and decision-making. Technical effort refers to the physical actions needed to implement changes—typing, cursor movement, switching panes, or operating CAT features. Temporal effort is simply the time required to finish the task. These dimensions rise and fall together but not always in lockstep; a segment may be quick to fix yet mentally taxing, or vice

versa. They also give us a practical basis for comparing post-editing to human translation and for deciding when PE is a viable choice in a given workflow (Koponen, 2012).

Task configuration matters as well. In monolingual PE, where post-editors only see the MT output, visual complexity is lower and navigation is simpler, which can reduce technical effort. The trade-off is a higher risk of missing adequacy problems that only become visible when the source is available. Bilingual PE—the default in most commercial CAT environments—supports a fuller adequacy check because the source and target are aligned, but it typically adds to cognitive effort by forcing continuous cross-lingual comparison (Mitchell et al., 2013; Nitzke, 2016). The choice between the two modes should therefore be driven by text criticality, risk tolerance, and the expected error profile of the MT system.

Moreover, post-editing tasks are often remunerated at lower rates than human translation, reinforcing perceptions of lower status and value. This has led to debates within the translation community regarding the fair recognition of post-editing as a professional activity in its own right (Bowker and Buitrago Ciro 2015). As MT output continues to improve, the boundaries between light and full post-editing become less distinct, and expectations regarding translator expertise and compensation must be re-evaluated.

Taken together, the evidence recasts post-editing as a situated professional practice. Automatic post-editing delivers value only where it reliably improves raw MT and, even then, human intervention remains essential to catch overcorrections and omissions. With neural systems, greater fluency raises a new risk: well-formed sentences can mask adequacy errors, shifting the post-editor's work toward deeper semantic verification. Whether post-editing is defensible therefore hinges on input quality and on a triad of effort—cognitive (analysis and decisions), technical (operations in tools), and temporal (time on task)—that does not always rise or fall in tandem. Workflow design further shapes outcomes: monolingual setups simplify interaction but can hide source-related problems; bilingual setups enable fuller adequacy checks at a higher mental cost; static pipelines separate generation and revision, while interactive systems can reduce friction when well configured and familiar to translators. Persistent professional resistance is thus understandable, given the heightened cognitive demands and remuneration models that often understate risk and responsibility. In practice, post-editing is advisable when machine output clears a quality threshold, the domain tolerates residual error, and the tooling demonstrably lowers effort; otherwise, full human translation is the safer and, in the end, more efficient choice.

3.3 “Post-editing” the source text

Within the domain of machine translation (MT), the role of human intervention has long been recognized as crucial for achieving acceptable quality, especially in contexts where fully automatic, high-quality translations remain beyond reach. Traditionally, this human involvement has been described as taking place in three distinct phases: pre-editing, interactive intervention during the translation process, and post-editing the system’s output (Somers 1997). However, Harold Somers’ (1997) contribution introduces an innovative perspective by proposing a hybrid approach that blurs the boundaries between these established stages: post-editing the source text. This idea invites translators to look at the source text not just before, but also after, seeing how the MT system handles it.

At first glance, this notion may seem counterintuitive. After all, post-editing has conventionally referred to the revision of the output text — the translation that the machine produces — to correct errors, improve readability, and ensure that the final version serves its communicative purpose. Yet Somers argues that sometimes it can be more efficient to adjust the input text in reaction to errors that have emerged during the process, rather than fixing each instance of the same mistake in the output. In his own words, “the activity known as ‘post-editing’ […] can equally involve adjusting the input to the system” (1997:193). The key benefit of this approach is that it allows the translator to exploit the flexibility of Human-Aided Machine Translation (HAMT) systems, which also at the time typically displayed the source and target texts side by side in an interactive environment, enabling users to make changes and instantly re-run segments through the translation engine.

A fundamental aspect of this practice is the distinction Somers draws between the source text — the original, authentic version provided by the client or author — and the input text, which may become an adapted, simplified, or otherwise manipulated version crafted specifically to generate better MT output. While pre-editing aims to prepare the source text before any translation takes place — for example, by removing ambiguous constructions or standardizing terminology — post-editing the source text is a reactive process. It emerges only after the user has seen what kinds of errors the MT system produces, so the modifications are targeted and informed by concrete evidence rather than general assumptions.

This subtle but significant difference changes how the translator relates to the text. It also frees the process from the constraints usually associated with linguistic correctness in the source language. As Somers explains, it may be perfectly acceptable — even desirable — for

the input text to become “stylistically clumsy or even ungrammatical” if doing so helps the system produce a higher-quality translation in the target language (1997:203). In other words, the goal is not to maintain the elegance or authenticity of the source text at every step but rather to maximize the efficiency and accuracy of the final output.

This approach offers a flexible, iterative workflow. The translator can repeatedly refine the input text, run new segments through the system, and evaluate whether each round of edits brings the desired improvements. This cycle acknowledges what Martin Kay (1980) pointed out decades ago: that there is often no single ‘correct’ point for human intervention in the translation process. Instead, translators should be empowered to decide whether an issue is best addressed before, during, or after the machine’s processing — or, as Somers shows, by looping back to adjust the input once the system’s output reveals consistent patterns of error.

To illustrate this idea, Somers presents a series of concrete examples drawn from practical use cases. One case focuses on the French translation of maritime weather forecasts originally written in English. In this domain, the word “fog” should often be rendered as “brume” rather than “brouillard”, because the latter may lead to incorrect gender agreement with adjectives that follow. If the system defaults to “brouillard”, the translator might have to fix multiple instances of mismatched gender throughout the output text. A more strategic solution is to replace “fog” with “mist” in the input, which naturally guides the system toward “brume” and ensures the correct grammatical concord: “A better strategy is to change the input […] and let the system do the gender agreement for us” (1997:203).

Another example involves French recipes being translated into English. Recipes in French commonly use the infinitive form (e.g., “Peler les pêches”) to indicate instructions. Unfortunately, some MT systems tend to translate these literally into present participles in English (“Peeling the peaches”) rather than the imperative form that native English speakers expect (“Peel the peaches”). Correcting each instance in the output can become tedious and error-prone, especially when multiple verbs follow the same pattern. Instead, by modifying the source text from the infinitive “Peler” to the imperative “Pelez”, the translator prompts the system to generate the proper English imperative automatically. This simple input adjustment can eliminate repetitive manual fixes, saving time and cognitive effort.

Such examples reveal the broader principle: translators must weigh whether changing the input or the output is the more practical option in each context. As Somers notes, “The key point is that it is up to the user to determine which is easier, editing the input or the output” (1997:204). This decision depends on the nature and frequency of the error, the MT

system's known strengths and weaknesses, and the user's familiarity with the software's behavior.

This strategy also highlights the interesting relationship between post-editing the source text and the concept of controlled language. While both approaches aim to make machine translation more accurate, they differ in timing and intent. Controlled language is usually applied during the original drafting stage, often by technical writers following strict style guidelines. In contrast, post-editing the source text arises later, only after the machine's limitations have become visible, and the changes may depart from normal stylistic or grammatical conventions if this leads to a better result.

Somers' additional examples further illustrate the practical value of this method. In cases where MT systems mishandle comparatives and superlatives — for instance, misinterpreting une plus grande intégration as a biggest integration instead of a greater integration — it may be more efficient to adjust determiners or phrasing in the input rather than hunt for scattered mistakes throughout the target text. Similarly, systematic issues with prepositions, definite articles, or domain-specific terminology can often be corrected more effectively upstream, through targeted tweaks to the input.

This approach inevitably raises questions about the skills required for post-editing. Somers (1997:207) argues that experienced translators are generally better suited to this task than domain specialists or monolingual editors because they possess the linguistic knowledge and practical instincts needed to spot errors quickly and fix them efficiently. Crucially, post-editing MT output is not simply a matter of 'proofreading'. It is a form of intervention that demands an ability to detect patterns in system-generated errors and the practical know-how to use editing tools, search-and-replace functions, and interactive features to minimize repetitive work. As Vasconcellos (1987) and Wagner (1985) note, translators who specialize in post-editing develop a unique mindset: they focus not on rewriting the entire text for style but on identifying the minimum necessary changes to make the translation fit for its purpose.

Of course, this strategy also has technical implications. Somers points out that it is vital to keep clear records of which versions of the input text have been altered for MT purposes. Since these modified inputs may be stylistically awkward or partially ungrammatical, they are not suitable for other uses — such as republishing the source text. Good practice requires separating the original source document from any 'machine-optimized' variants to prevent accidental reuse. Somers suggests that translation tools could help by allowing users to save these variants with clear markup or annotations,

making it easier to manage translation memory assets and ensure consistency in future projects.

Ultimately, post-editing the source text stands as a pragmatic and adaptable response to the enduring limitations of machine translation, especially for systems that still produce rough or error-prone output. While advances in neural MT have undoubtedly reduced the need for heavy human intervention in many contexts, there remain countless scenarios — especially in technical or domain-specific translation — where targeted human editing can dramatically improve the final result. By giving translators the option to intervene both before and after they see how the system handles a text, this method bridges the gap between traditional pre-editing and conventional post-editing.

In summary, the practice of post-editing the source text, as proposed by Somers, offers a pragmatic extension of traditional human involvement in machine translation workflows. By allowing the translator to make targeted adjustments to the input after analysing the system's recurring errors, this approach complements pre-editing and conventional post-editing in a flexible and iterative manner. In certain contexts, direct intervention in the target text may be the most efficient solution to eliminate residual errors; in other cases, strategically modifying the input can help the MT system generate a cleaner, more accurate draft from the outset. This method demonstrates how the translator's expertise, supported by a thorough understanding of the system's capabilities and limitations, remains vital for achieving high-quality translations, especially when fully automatic solutions fall short. Ultimately, post-editing the source text reinforces the importance of adaptive strategies that balance time, cost, and linguistic quality, ensuring that human skills and machine resources are employed to their maximum potential within contemporary translation practice.

4. Research Questions and Methodology

This thesis investigates whether applying pre-editing together with post-editing of the source text increases the translatability of the input and thereby improves the quality of neural machine translation (NMT) output. In this study, pre-editing is understood as the targeted application of a selected subset of rules drawn from the Translation Centre for the Bodies of the European Union's 2021 compendium Writing for Machine Translation and from Hardin et al. (2020) on "Writing for Machine Translation"; Section 3.1.2 reviews these two frameworks, while Chapter 5 specifies the subset adopted and its operationalisation within the experiment. Post-editing of the source text is theoretically grounded in Somers (1997); the rationale and concepts are outlined in Section 3.3, and Chapter 5 presents their concrete implementation in the experimental workflow. Within this framework, translatable is construed as a gradable property of the source whereby lexical and syntactic ambiguities are minimised; segmentation and punctuation have been adjusted; domain-appropriate terminology and style are enforced; under such conditions, an NMT system is required to make fewer uncertain choices and can map source units onto target structures with more predictability. Empirically, one and the same Italian source (317 words) underwent two consecutive interventions: after pre-editing it comprised 322 words, and after source-side post-editing it comprised 333 words. Translations into English (UK) and German (DE) were then produced from the unedited source (Version 1, V1) and from the twice-edited source (Version 2, V2) using DeepL (free edition) outside Matecat and were imported segment by segment into four participant-specific projects labelled X_V1_EN, X_V1_DE, X_V2_EN, X_V2_DE (where X is a unique letter A–J). The review phase took place exclusively in Matecat. Immediately after each project, participants completed a short post-task questionnaire capturing qualitative judgements of accuracy, fluency, and fidelity, together with an open item for issues not covered by the error taxonomy and a space for comments. If the combination of pre-editing and source-side post-editing does increase translatability, V2 was expected to yield a lower quality score and fewer errors than V1, and the questionnaires were expected to reflect the same pattern through higher qualitative appraisals of V2.

The three questionnaire dimensions were chosen to mirror established quality constructs. Accuracy and fluency align with the widely used dimensions in MQM-based assessment, where accuracy targets semantic adequacy/transfer and fluency targets well-formedness and naturalness of the target text. Fidelity follows the evaluative tradition stemming from the ALPAC report, where intelligibility and fidelity captured, respectively,

ease of understanding and faithfulness to the source. In this study, fidelity is used as a concise label for perceived adherence to the source's communicative intent and stylistic stance (as distinct from strict semantic accuracy and local fluency). Using these three lenses keeps the questionnaires interpretable for raters while connecting the qualitative strand to recognised evaluation frameworks (MQM; ALPAC).

The study pursues two objectives. First, it evaluates whether the combination of pre-editing and post-editing of the source text yields higher-quality translation than the automatic translation of the unedited source, when translating with a neural machine translation system. Secondly, it assesses whether any improvement is comparable across IT-EN (UK) and IT-DE or whether the two directions diverge. Accordingly, the research questions are:

1. Do NMT outputs from a source edited in two steps—(i) pre-editing and (ii) post-editing of the source text—outperform outputs from the unedited source?
2. Is this NMT quality gain similar for IT-EN (UK) and IT-DE, or does it differ by language pair?

The working hypotheses mirror these questions:

Hypothesis 1: V2 supplies the NMT system with an input that reduces errors relative to V1;

Hypothesis 2: the quality attained on V2 is comparable (or, at minimum, acceptable in both cases) for IT-EN (UK) and IT-DE.

The original text is an Italian excerpt from the trail guide "Le scritte dei pastori, Etnoarcheologia della pastorizia in Val di Fiemme" by Bazzanella e Kezich (2013:6). While the book as a whole documents routes and cultural features on Monte Cornon (Trentino, Italy), the selected passage focuses on the pastoral inscriptions—ochre-based markings left by herders on the mountain's rock faces. The text interweaves culture-specific references tied to the Val di Fiemme with technical terminology spanning ethnography, archaeology and local geography, and is written with complex hypotaxis (long, highly embedded sentences).

The study adopts a within-subjects design in which each participant completes all four conditions (V1 and V2 in EN and DE), permitting contrasts within the same individual and limiting between-person variance. Two independent variables are fixed ex ante: the source-preparation workflow (V1 unedited versus V2 edited in two steps) and the target language (EN-UK versus DE). Outcomes are organised as primary and secondary. The primary quantitative outcomes comprise Matecat's quality score (EPT) and the raw error points by severity. To ensure like-for-like comparisons across conditions with different

lengths, the quality score serves as the primary indicator. It is a length-normalised summary of the total weighted error load: it takes the points assigned to each annotated error (Neutral = 0; Minor = 0.5; Major = 2), accumulates them, and expresses the result on a common base of one thousand words using the reviewed-word count for that translation. Because Version 1 projects contain 325 reviewed words and Version 2 projects 344, this scaling is essential; without it, raw totals would conflate how many errors there are with how long the text happens to be. Alongside the quality score, the chapter also reports raw error points by severity—the direct weighted totals recorded under Minor and Major (Neutral remains visible for completeness but contributes zero). These severity totals are not normalised per thousand words and are not used for cross-condition comparisons; rather, they serve a profiling function, showing whether changes in the quality score are driven predominantly by reductions in Major errors (substantive gains) or by reductions in Minor errors (refinements). In short: the quality score (EPT) answers "How heavy is the error burden per standard length?", while the severity-level raw points answer "What kind of errors is that burden made of?". Secondary quantitative outcomes characterise the distribution of errors by category under an MQM-inspired labelling scheme (Style; Translation Errors; Terminology Consistency; Language Quality).

Materials and procedures are held constant to privilege interpretability. All translations were generated with DeepL (free edition) under default settings—no custom glossaries, no explicit register control, no advanced options—and no post-correction was applied prior to review, so that the texts examined by participants reflect the engine's output under the stated conditions. Using a single engine across V1/V2 and EN/DE ensures that any differences can be attributed to the treatment on the source rather than to system variation; the constraints of the free edition apply uniformly and therefore do not compromise the comparison. All reviewing took place in Matecat, a browser-based CAT environment that provides a uniform interface and centralised data capture. Built-in suggestions from machine translation and translation memories were disabled in every project to prevent pre-filling or cross-session interference. The standard quality-review profile was employed with MQM-inspired labels adjusted for clarity—Style, Translation Errors, Terminology Consistency, Language Quality—and the "tag" category was removed as not pertinent to the text type and task. Severities and weights were fixed as Neutral = 0, Minor = 0.5, Major = 2. Segmentation and one-to-one alignment between source and target were preserved on import to support reliable accounting. Because target texts do not have identical length, analyses use the actual reviewed-word counts recorded by Matecat (325 for V1; 344 for V2).

Interactive MT and interactive post-editing systems enable the translation engine to integrate user revisions in real time, thereby establishing a feedback loop between editor and system (e.g., prefix-based autocompletion and adaptive suggestions), creating a live feedback loop between human and machine (see Escribe & Mitkov, 2021; cf. Langlais et al., 2002). In this thesis the focus is deliberately on post-editing of the source text rather than interactive post-editing, because the former intervenes directly on the input—stabilising segmentation, clarifying syntax and logic—whereas interactive post-editing typically modifies the output. This design choice isolates the contribution of source-side preparation to the downstream quality observed in a static NMT pipeline; potential future integrations with interactive post-editing are discussed in the concluding chapter.

The quality score is accompanied by a clearly stated operational threshold of 8 per 1,000 words, adopted as a practical acceptability cut-off. Given the reviewed-word counts, this corresponds to a maximum of 2.6 error points for Version 1 (8 × 325 / 1000) and 2.752 error points for Version 2 (8 × 344 / 1000). The threshold is provided as an intelligible reference for readers; it is not treated as a hard pass/fail boundary in the inferential work. In addressing RQ1 and RQ2, the quality score and the severity-level error points are analysed as continuous outcomes, which better capture the magnitude and direction of effects than a dichotomous classification would. Normalising to a one-thousand-word base is a methodological necessity whenever units of analysis differ in length across conditions. Reporting both the normalised quality score and the unnormalised severity-level points preserves comparability while retaining insight into the profile of errors that drive the aggregate.

Quantitative exports from Matecat (per project and participant) include severity-level error points (Neutral, Minor, Major), the quality score, reviewed-word counts and segment identifiers. These files were organised into four analytical tables to provide complementary cross-views of the results. Two tables align both languages within the same source version—one table for Version 1 and one table for Version 2—with rows listing participants (A–J) and columns comprising the quality score and three severity columns: Neutral (0-weight), Minor (0.5 points per instance), Major (2 points per instance). Two further tables fix the language and contrast the two versions: an EN table displaying EN-V1 and EN-V2, and a DE table displaying DE-V1 and DE-V2, with the same row/column structure. Importantly, the severity columns contain raw weighted points, not per-thousand values; the only length-normalised indicator is the quality score (EPT). Read together, the four tables allow the workflow effect (V2 versus V1 at fixed language) and any language-specific modulation (EN-UK versus DE

at fixed workflow) to be inspected at a glance. In parallel, the questionnaire materials are collated qualitatively by condition and language, with representative excerpts and concise summaries of comparative judgements (V1 vs V2; EN vs DE); no numerical inference is performed on these materials. These tables are to be found in the Appendix A of this work.

The procedure is straightforward and replicable. Participants first completed the background survey documenting self-reported proficiency, training and exposure to CAT/NMT tools, received a concise briefing with operational instructions and the error-annotation scheme (cf. Appendix B), then reviewed their four projects in Matecat (X_V1_EN, X_V1_DE, X_V2_EN, X_V2_DE) and, immediately after each project, completed the corresponding post-task questionnaire. Participants saw only the project names (letter, version, language), so the V1/V2 distinction was visible but uniform; the order of tasks followed written instructions for comparability; sessions were conducted remotely without strict time limits; technical assistance was available on request. Participants were informed in advance about the study's aims, tasks and intended use of data; participation was voluntary; data were anonymised for analysis and reporting; no formal information sheet was issued.

In summary, the independent variables are the two design factors set in advance—workflow (V1 versus V2) and language (EN-UK versus DE)—while the dependent variables comprise the quality score (EPT), severity-level raw error points, category profiles, and the qualitative judgements collected through the five questionnaires. The expectation that V2 would outperform V1—namely, a lower quality score together with fewer error points—provides the benchmark for RQ1, with RQ2 asking whether this expected advantage holds to a similar extent in English and in German. The quantitative strand addresses these questions through changes in the quality score and in the Minor/Major distribution; the questionnaires, treated expressly as qualitative evidence, are intended to corroborate (or problematise) the same pattern in participants' own terms.

5. Experiment

This chapter documents, in the order in which the work was carried out, the experimental pipeline used to test the effect of pre-editing and post-editing of the source text on the quality of neural machine translation (NMT) output. It operationalises the methodological framework outlined earlier by (i) specifying the stimulus text and its domain profile; (ii) detailing the two-step source preparation (pre-editing, then source-side post-editing); (iii) describing the generation of the four NMT outputs with a single engine under constant conditions; (iv) reporting the set-up of the Matecat review environment and the standardised instructions given to participants (cf. Appendix B); and (v) explaining how qualitative perception data were collected through post-task questionnaires. At each stage, the artefacts produced (edited source versions, NMT outputs, Matecat projects, exports and qualitative responses) are made explicit so that the path from intervention to evidence is fully traceable.

The chapter is anchored to the two research questions. Accordingly, the narrative closes each phase with a brief statement of its analytical relevance (what the phase contributes to answering RQ1 and RQ2) and with a pointer to the metrics that will be examined in the Results and Analysis chapter—primarily the quality score (EPT) and the severity-level raw error points, complemented by qualitative judgements of accuracy, fluency and fidelity.

5.1 Original Text: Rationale for the Choice and Description

The original text is an Italian excerpt drawn from Marta Bazzanella and Giovanni Kezich, “Le scritte dei pastori. Etnoarcheologia della pastorizia in Val di Fiemme” (2013:6-7). The volume is organised broadly in two parts: the first half surveys customs and pastoral life in the Val di Fiemme (Trentino, Italy) across centuries; the second half offers route descriptions for visitors to Monte Cornon, whose rock faces bear ochre inscriptions and figurative markings produced by local herders over time. The selected passage belongs to the introductory chapter authored by Giovanni Kezich. It explains what these inscriptions are and why they were made, situating them within the wider tradition of rock painting from prehistoric times to the present. Because the excerpt is taken in *medias res* and carries no stand-alone title, this contextualisation clarifies its position within the book’s architecture and its connection to the subsequent, route-oriented chapters.

With respect to domain and genre, the passage is best described as descriptive–interpretive ethnography for a general readership. It weaves culture-specific

references anchored in the Val di Fiemme with a technical–conceptual vocabulary drawn from ethnography, archaeology, and cultural history, while mentioning places and contexts relevant to the local setting. Two axes of interest structure the content and terminology. First, a content-driven axis: the writings function as diary-like notations on rock, recording names, episodes, and motifs that mirror the economic, cultural, religious, and social evolution of the community, with remarks on how subjects and modes of depiction changed over time. Second, a historical–cultural axis: the author argues that the inscriptions display iconographic features and stand in direct continuity with prehistoric rock art. These claims mobilise a register that is at once accessible and conceptually dense, increasing the proportion of domain-specific terms and culture-bound references.

Linguistically, the excerpt is characterised by complex hypotaxis: long, highly embedded sentences with multiple subordinate clauses and clause-initial or clause-medial insertions that modulate information structure. Even in Italian, the prose is syntactically heavy, with coordination and subordination patterns that blur segment boundaries and create non-local dependencies. This profile challenges NMT in both IT→EN (UK) and IT→DE.

These properties motivate the two-step source editing. When writing for machine translation (see § 3.1.2), operations target segmentation and punctuation stability, harmonisation of terminological choices, expansion or standardisation of abbreviations, and explicit marking of logical and spatial relations that would otherwise remain implicit—thereby reducing the number of uncertain decisions the engine must make. By post-editing the source text (Somers, 1997; see § 3.3), the same passage receives a second, source-side pass to resolve residual lexical or syntactic ambiguities, make referential links explicit across sentences, and fine-tune sentence shaping so as to encourage stable, parallel segmentation for the MT engine. Crucially, this second pass implements a managed compromise: the edited Italian remains faithful to the cultural content while being adequately serviceable for both target languages (EN-UK and DE). In practice, this entails preserving culturally marked items (and occasionally contextualising them), moderating clause embedding, and making head–modifier relations overt.

The segmentation history reflects these aims. The original excerpt comprises 7 sentences; after pre-editing it was restructured into 14 sentences, and after post-editing of the source text into 15 sentences. In Matecat, this yields 7 segments for Version 1 (V1)—the unedited source—and 15 segments for Version 2 (V2)—the twice-edited source. The increase in sentence and segment count is intentional: segmentation is treated as a design lever to stabilise punctuation, align segment boundaries with plausible MT splitting points, and reduce

attachment ambiguities, thereby promoting more predictable mapping of source units into target structures.

Finally, the length of the excerpt was a pragmatic choice. A longer passage could have broadened domain coverage but would have placed an undue burden on participants, each of whom reviewed four translations (EN-V1, DE-V1, EN-V2, DE-V2). The selected size balances operational feasibility with linguistic complexity: it is short enough for a standardised review session, yet sufficiently dense in domain-specific lexicon and intricate syntax to reveal editing-driven gains both in the quality score (EPT) and in the composition of error points.

Section 5.2 documents the pre-editing interventions with reference to Writing for Machine Translation (Translation Centre, 2021; Hardin et al., 2020), using a first table that juxtaposes the original text, its IT→DE and IT→EN (UK) NMT outputs, the pre-edited version, and the corresponding outputs on that version. Section 5.3 then documents the post-editing of the source text (Somers, 1997), using a second table that juxtaposes the pre-edited and post-edited versions with their IT→DE and IT→EN (UK) outputs.

5.2 First editing: Pre-Editing according to Writing for Machine Translation

This section documents the first editing pass performed on the Italian source in accordance with Writing for Machine Translation (Translation Centre for the Bodies of the European Union, 2021) and the recommendations in Hardin et al. (2020). The shared objective is to increase the translatability of the source before machine translation, primarily by stabilising structure and information flow while postponing language-specific lexical decisions to the subsequent source-side post-editing stage.

A small set of rules was applied consistently. From the EU Translation Centre compendium, three principles were central: (1) consistent use of terminology—avoiding needless synonyms for domain terms; (2) minimise pronouns where they may generate ambiguous reference; and (4) avoid excessively long sentences by reducing clause embedding and favouring clear, linear structures. From Hardin et al. (2020), the first guideline—minimalism—was adopted explicitly, removing units that add rhetorical weight but no informational value. In addition, Hardin's guidance on simplifying complex phrasing and disambiguating structures closely aligns with the compendium's emphasis on shortening and clarifying long sentences; and Hardin's focus on terminological coherence is directly comparable to the compendium's rule on terminological consistency. In short, the chosen

rules are mutually reinforcing: reduce length and complexity, make reference explicit without pronouns where feasible, and keep specialised terms uniform.

Operationally, the pre-editing focused on syntax and segmentation. Terminology was researched in advance to identify likely pressure points, but lexical substitutions were largely deferred to the second editing phase (post-editing of the source text). This deliberate choice avoids baking in language-pair bias at this stage: inserting target-oriented corrections in the source would have required creating two different edited sources (one tailored to EN and one to DE), which is outside the scope of this study. The only lexical adjustment undertaken here concerned terminological uniformity in Italian where the original used competing labels for the same concept.

Original		Neural Machine Translation		First level of editing	Neural Machine Translation	
		German DE	English UK		German DE	English UK
1	Almeno due, sono infatti i motivi di interesse e le ragioni dell'importanza di questi graffiti.	Mindestens zwei sind in der Tat die Gründe für das Interesse und die Gründe für die Bedeutung dieser Graffiti.	In fact, at least two are the reasons of interest and the reasons for the importance of these graffiti.	Almeno due, sono i motivi di interesse e le ragioni dell'importanza di queste scritte.	Mindestens zwei Gründe sind es, die das Interesse und die Bedeutung dieser Schriften begründen.	At least two, are the reasons for the interest and importance of these writings.
2	Il primo è la testimonianza che essi rendono di un mondo pastorale appena trapassato, della sua economia, della sua relativa precoce alfabetizzazione, della sua sfera devozionale,	Das erste ist das Zeugnis, das sie von einer gerade durchbrochenen pastoralen Welt, ihrer Wirtschaft, ihrer relativen frühen Alphabetisierung, ihrer hingebungsvollen, ästhetischen	The first is the testimony they bear of a pastoral world that had just passed away, of its economy, its relative early literacy, its devotional, aesthetic and moral sphere.		Der erste ist das Zeugnis, das sie von einer gerade untergegangenen pastoralen Welt, ihrer Wirtschaft, ihrer relativ frühen Alphabetisierung, ihrer religiösen, ästhetischen und moralischen	The first is the testimony they bear to a pastoral world that had just passed away, its economy, its relative early literacy, its devotional, aesthetic and moral sphere.

	estetica e morale.	und moralischen Sphäre geben.			Sphäre geben.	
3	Le scritte infatti, in un primo tempo racchiuse all'interno di ordinate edicolette ornate e iscritte con i caratteri propri dei cippi di confine in montagna, e in un secondo tempo debordanti a tutto campo con una certa dovizia di annotazioni diaristiche, contengono firme, date, conteggi del bestiame, «segni di casa» mentre il loro stesso posizionarsi sulle falesie testimonia degli incessanti andirivieni estivi dei singoli pastori, all'interno di un sistema ancora in parte vigente di diritti territoriali di	In der Tat enthalten die Inschriften, die zunächst in hübsche kleine verzierte Edicolettes eingefasst und mit den für Grenzsteine in den Bergen typischen Schriftzeichen beschriftet sind und später mit einer gewissen Fülle von Tagebucheinträgen überladen sind, Unterschriften, Daten, Sie enthalten Unterschriften, Daten, Viehzählungen, „Hauszeichen", und ihre Positionierung auf den Felsen zeugt von den unablässigen sommerlichen Wanderungen der einzelnen	In fact, the inscriptions, at first enclosed within neat little ornate edicolettes and inscribed with the characters typical of boundary stones in the mountains, and later overflowing with a certain abundance of diary entries, contain signatures, dates, livestock counts, "house signs", while their very positioning on the cliffs testifies to the incessant summer comings and goings of individual shepherds, within a still partly existing system of disputed and always rather complex territorial	Le scritte infatti contengono firme date, conteggi del bestiame, «segni di casa». In un primo tempo le scritte erano racchiuse all'interno di ordinate edicolette ornate e iscritte con i caratteri propri dei cippi di confine in montagna. In un secondo tempo erano debordanti a tutto campo con una certa dovizia di annotazioni diaristiche. La loro posizione sulle rocce testimonia gli incessanti andirivieni estivi dei singoli pastori. I pastori erano parte di un sistema di diritti territoriali di	Die Inschriften enthalten in der Tat datierte Unterschriften, Viehzählungen, „Hauszeichen". Zunächst waren die Inschriften in saubere, verzierte Edicolettes mit den für Grenzsteine in den Bergen typischen Schriftzeichen eingeschlossen. Später waren sie mit einer gewissen Fülle von tagebuchartigen Vermerken übersät. Ihre Lage auf den Felsen zeugt von den unablässigen sommerlichen Wanderungen der einzelnen Hirten. Die	The inscriptions in fact contain dated signatures, livestock counts, "house signs". At first, the inscriptions were enclosed within neat, ornate edicolettes and inscribed with the characters typical of boundary stones in the mountains. At a later stage, they were overflowing with a certain abundance of diary entries. Their position on the rocks testifies to the incessant summer comings and goings of individual shepherds. The shepherds were part of a system of

	pascolo contesi e sempre piuttosto complessi.	Hirten im Rahmen eines teilweise noch bestehenden Systems umstrittener und stets recht komplexer territorialer Weiderechte.	grazing rights.	pascolo contesi e sempre piuttosto complessi. Questi diritti sono ancora parzialmente in vigore.	Hirten waren Teil eines Systems von umstrittenen und stets recht komplexen territorialen Weiderechten. Diese Rechte sind teilweise noch in Kraft.	disputed and always rather complex territorial grazing rights. These rights are still partially in force.
4	Così, a partire dalla sequenza cronologica espressa dalle date, dalla georeferenziazione delle scritte stesse e dal riconoscersi dei singoli pastori dietro alle sigle e ai «segni di casa», sarà certamente possibile utilizzare il complesso di queste scritte pastorali come uno straordinario archivio notarile per ricostruire una pagina importante della storia della pastorizia nella valle di Fiemme, e	Ausgehend von der chronologischen Abfolge der Daten, der Georeferenzierung der Schriften selbst und der Erkennung einzelner Hirten hinter den Akronymen und „Hauszeichen“ wird es sicherlich möglich sein, den Komplex dieser pastoralen Schriften als außergewöhnliches notarielles Archiv zu nutzen, um eine wichtige Seite der Geschichte des Hirtenwesen	Thus, starting from the chronological sequence expressed by the dates, from the georeferencing of the inscriptions themselves and from the recognition of the individual shepherds behind the acronyms and “house signs”, it will certainly be possible to use the complex of these pastoral inscriptions as an extraordinary notarial archive to reconstruct an important page in the	Il complesso di queste scritte pastorali rappresenta uno straordinario archivio notarile per ricostruire una pagina importante della storia della pastorizia nella valle di Fiemme, e dunque della storia economica, sociale e ambientale della valle. L’archivio può essere ricostruito a partire dalla sequenza cronologica espressa dalle date, dalla georeferenziazione delle scritte stesse	Der Komplex dieser Hirteninschriften stellt ein außergewöhnliches notarielles Archiv dar, mit dem sich eine wichtige Seite der Geschichte des Hirtenwesens im Fleimstal und damit der Wirtschafts-, Sozial- und Umweltgeschichte des Tals rekonstruieren lässt. Das Archiv lässt sich anhand der chronologischen Abfolge der Daten, der Georeferenzierung der	The complex of these pastoral inscriptions represents an extraordinary notarial archive for reconstructing an important page in the history of pastoralism in the Fiemme Valley, and thus of the economic, social and environmental history of the valley. The archive can be reconstructed on the basis of the chronological sequence expressed by the dates, the georeferencing of the inscriptions

	dunque della storia economica, sociale e ambientale della valle.	s im Fleimstal und damit der Wirtschafts-, Sozial- und Umweltgesc hichte des Tals zu rekonstruiere n.	history of pastoralism in the Fiemme valley, and thus of the economic, social and environment al history of the valley.	e dal riconoscersi dei singoli pastori dietro alle sigle e ai «segni di casa».	Inschriften selbst und der Identifizieru ng der einzelnen Hirten hinter den Initialen und „Hauszeiche n“ rekonstruiere n.	themselves and the identification of the individual shepherds behind the initials and “house signs”.
5	Un lavoro che, previo il completamen to del database delle scritte, può dirsi ora appena iniziato.	Eine Arbeit, die nach der Fertigstellun g der Datenbank mit den Schriften als gerade erst begonnen bezeichnet werden kann.	A work that, after the completion of the database of inscriptions, can now be said to have just begun.		Eine Arbeit, von der man nach der Fertigstellun g der Inschriften-D atenbank sagen kann, dass sie gerade erst begonnen hat.	Work that, after the completion of the inscription database, can now be said to have only just begun.
6	Secondo e non minore motivo di interesse, tuttavia, è il modo in cui questi graffiti pastorali, pur inequivocabi lmente d’età moderna e quasi precontempo ranea, si ricollegano direttamente, quanto a stile, ad alcuni aspetti dell’iconogra fia, e alla qualità specifica	Der zweite und nicht minder interessante Grund ist jedoch die Art und Weise, wie diese Hirtengraffiti , obwohl sie eindeutig modernen und fast vorzeitlichen Alters sind, in Bezug auf den Stil, bestimmte Aspekte der Ikonographie und die spezifische	The second and no less interesting reason, however, is the way in which these pastoral graffiti, although unequivocall y of modern and almost pre-contemp orary age, are directly connected, in terms of style, certain aspects of iconography and the specific	Secondo e non minore motivo di interesse, tuttavia, è il modo in cui questi graffiti pastorali si ricollegano direttamente alla lunga tradizione europea dell’arte rupestre preistorica. L’arte rupestre che ha inizio nelle grotte del paleolitico dordonese e	Zweitens und nicht weniger interessant ist jedoch die Art und Weise, wie diese Hirtengraffiti direkt mit der langen europäischen Tradition der prähistorisch en Felskunst zusammenhä ngen. Diese Felskunst begann in den Höhlen des paläolithisch en	Second and no less interesting, however, is the way in which these pastoral graffiti relate directly to the long European tradition of prehistoric rock art. This rock art began in the caves of the Palaeolithic Dordonian and Cantabrian periods and continued in

	dell'ispirazione scrittoria, alla lunga tradizione europea dell'arte rupestre preistorica, quella che ha inizio nelle grotte del paleolitico dordonese e cantabrico e prosegue poi in età protostorica con i grandi siti graffitati del Monte Bego e della val Camonica.	Qualität der Schreibinspiration direkt mit der langen europäischen Tradition der prähistorischen Felskunst verbunden sind, die in den Höhlen der paläolithischen Dordogne und Kantabrien begann und dann in der protohistorischen Zeit mit den großen Graffiti-Stätten von Monte Bego und Val Camonica fortgesetzt wurde.	quality of the writing inspiration, to the long European tradition of prehistoric rock art, which began in the caves of the Palaeolithic Dordogne and Cantabrian and then continued in the protohistoric age with the great graffiti sites of Monte Bego and Val Camonica.	cantabrico e prosegue poi in età protostorica con i grandi siti graffitati del Monte Bego e della val Camonica. Le scritte si ricollegano per stile e per alcuni aspetti dell'iconografia e sono d'età moderna e quasi precontemporanea.	Dordoniums und Kantabriens und setzte sich in der protohistorischen Periode mit den großen Graffiti-Fund stellen des Monte Bego und des Val Camonica fort. Die Inschriften sind im Stil und in bestimmten Aspekten der Ikonographie verwandt und stammen aus der Neuzeit und fast aus der Vorzeit.	the Protohistoric period with the large graffiti sites of Monte Bego and Val Camonica. The inscriptions are related in style and certain aspects of iconography and are of modern and almost pre-contemporary age.
7	Infatti, al di là dell'evidente incongruità cronologica, vi è un certo numero di fattori che consentono di collocare le scritte pastorali della valle di Fiemme sul medesimo tracciato ideale della grande arte rupestre europea.	Abgesehen von der offensichtlichen chronologischen Inkongruenz gibt es eine Reihe von Faktoren, die es ermöglichen, die pastoralen Schriften des Fleimstals auf die gleiche ideale Spur der großen	In fact, beyond the obvious chronological incongruity, there are a number of factors that allow the pastoral inscriptions of the Fiemme valley to be placed on the same ideal track as the great European rock art.		Abgesehen von der offensichtlichen chronologischen Inkongruenz gibt es eine Reihe von Faktoren, die es erlauben, die Hirteninschriften des Fleimstals auf die gleiche ideale Schiene zu setzen wie	In fact, beyond the obvious chronological incongruity, there are a number of factors that allow the pastoral inscriptions of the Fiemme valley to be placed on the same ideal track as the great European rock art.

		europäischen Felskunst zu bringen.			die große europäische Felskunst.	

Table 1: Comparison between the original text and its two translations into English (UK) and German (DE) and the text after the first edit with its translations into German (DE) and English (UK)

For orientation, Table 1 juxtaposes the original text, the corresponding IT→DE and IT→EN (UK) outputs, the pre-edited version, and the outputs generated from that pre-edited source. Rather than describing the table row by row, the following remarks highlight typical interventions. In the first sentence (see Table 1, row 1), the coordinating discourse marker “infatti” (“in fact” in English; “In der Tat” in German) was removed as redundant in context; it inflated sentence weight without adding propositional content. In the same row, “graffiti” was replaced by “scritte” to enforce terminological consistency within the chosen domain frame. The second sentence (row 2) was left unchanged despite its length: many original clauses are coordinated rather than subordinated, and keeping one exemplar intact allows observation of the engine’s spontaneous segmentation. By contrast, the third sentence (row 3) underwent substantial disembedding: it was split into five short sentences, each carrying a single main idea. Given the original’s deep clause nesting (up to a third level of subordination), this restructuring was necessary for readability even in Italian. In the closing part of that edited block, the noun “pastori” (“shepherds” in English; “Hirten” in German) was intentionally repeated (rather than pronominalised) to avoid a pronoun and keep coreference overt. The fourth sentence (row 4) was split in two; subordinate clauses were flattened where feasible, with a shift to coordination at the end of the first edited sentence. In the second sentence, the noun “archivio” (“archive” in English; “Archiv” in German) was repeated to avoid pronominal reference. The fifth and seventh original sentences were left intact as additional controls, while the sixth (row 6) was split into two to prioritise structural clarity; again, lexical items were left untouched so that the second pass could target them in a principled, source-side manner. Summarising the pattern: in this first pass the interventions primarily lighten syntactic structure, stabilise segmentation, and clarify logical links, while deferring lexical changes to the subsequent stage.

The immediate NMT behaviour on the pre-edited source confirms that this phase alone does not guarantee acceptability. As visible in Table 1, both English and German outputs remain cumbersome in places, with repetitions, unclear referents, and awkward

lexicalisations. Examples include “segni di casa” rendered as “house signs” (EN) and “Hauszeichen” (DE), which are obsolete or misleading in context; “ordinate edicolette” emerging as “neat edicolettes” (EN) and “saubere Edicolettes” (DE), where the base lemma is a non-word and the adjective is ambiguous; or contextually odd solutions for “archivio notarile” and “georeferenziazione”. Culture-specific references and toponyms—e.g., “grotte del Paleolitico dordonese e cantabrico”, “Monte Bego”, “Val Camonica”—remain fragile without additional guidance in the source. These are precisely the targets scheduled for the second pass (Somers, 1997), where source-side post-editing can insert minimal but effective cues to steer the engine away from calques and out-of-register choices in both EN and DE.

In sum, the pre-editing stage applies a minimal, rule-based reshaping of the source that: (i) shortens and disentangles long sentences; (ii) favours explicit reference over pronouns when this reduces ambiguity; and (iii) enforces internal terminological consistency in Italian. This yields a more stable segmentation (from 7 to 14 sentences at this stage) and a cleaner information flow, without committing to language-specific lexical solutions. The pre-edited-only outputs were not submitted to participant review; quantitative metrics in this thesis—quality score (EPT) and severity-level raw error points—are computed exclusively for the original text V1 and for the twice-edited V2. Accordingly, what is tested quantitatively is the combined effect of pre-editing plus post-editing of the source text (V2) against the unedited source (V1). IT- EN/DE neural machine outputs of the first editing are included for process transparency and diagnosis rather than for formal assessment. It shows which structural problems remain after the first phase (e.g., residual ambiguity, unstable reference, culture-bound items) and thereby motivates the specific operations applied in the second phase (post-editing of the source text). Because participants only reviewed V1 and V2 in Matecat, no quality score is computed for the intermediate, pre-edited-only outputs, and no causal share is attributed to the first phase in isolation.

5.3 Second editing: Post-Editing of the Source Text

The second phase implements post-editing of the source text in the sense of Somers (1997)—a reactive intervention on the input performed after observing how the MT system handles the text, with the aim of preventing recurrent errors at input rather than correcting them repeatedly in output. In the present design, this principle is applied within a two-phase preparation of the Italian source: pre-editing first delivers lighter sentence structure, stable segmentation, and internal terminological consistency; post-editing of the source text then addresses what pre-editing deliberately leaves open—residual ambiguity, lexical

normalisation of culture-bound items, referential explicitness across sentences, and discourse cueing—while keeping the prose natural.

Operationally, the only iterative component of the workflow is confined to this second phase. A limited internal loop is used: generate diagnostic MT outputs with DeepL (free edition, no glossaries/registry controls) on the pre-edited draft, apply minimal source-side adjustments informed by the observed error patterns, and regenerate the outputs; once the source is finalised as Version 2 (V2), the loop is closed. Participant review occurs once and only on the four final translations (V1_EN, V1_DE, V2_EN, V2_DE) in Matecat; intermediate outputs produced during the internal loop are not part of the evaluation. As throughout the study, interventions in V2 are language-neutral—they clarify the Italian source without embedding English- or German-specific solutions—and remain within the bounds of well-formed, natural Italian (no reliance on deliberately ungrammatical input).

	First level of editing	Second level of editing	Neural Machine Translation	
			German (DE)	English (UK)
1	Almeno due, sono i motivi di interesse e le ragioni dell'importanza di queste scritte.	Questi graffiti sono importanti per almeno due motivi.	Diese Graffitis sind aus mindestens zwei Gründen wichtig.	These graffiti are important for at least two reasons.
2	Il primo è la testimonianza che essi rendono di un mondo pastorale appena trapassato, della sua economia, della sua relativa precoce alfabetizzazione, della sua sfera devozionale, estetica e morale.	Il primo è la testimonianza che danno di un mondo pastorale che non esiste più: ne descrivono l'economia, la dimensione devozionale, estetica e morale e mostrano come già all'epoca ci fosse una relativa alfabetizzazione.	Der erste Grund ist das Zeugnis, das sie von einer pastoralen Welt geben, die heute nicht mehr existiert: Sie beschreiben ihre wirtschaftlichen, religiösen, ästhetischen und moralischen Dimensionen und zeigen, dass es damals schon eine relative Alphabetisierung gab.	The first is the testimony they give of a pastoral world that no longer exists: they describe its economy, devotional, aesthetic and moral dimensions and show how there was already relative literacy at the time.
3	Le scritte infatti contengono firme	Le scritte infatti, contengono firme,	Die Schriften enthalten in der Tat	The inscriptions in fact contain

	date, conteggi del bestiame, «segni di casa».	date, conteggi del bestiame e stemmi di famiglia.	Unterschriften, Daten, Viehzahlen und Familienwappen.	signatures, dates, livestock counts and family crests.
4	In un primo tempo le scritte erano racchiuse all'interno di ordinate edicolette ornate e iscritte con i caratteri propri dei cippi di confine in montagna.	Si trovano in punti in cui i pastori passavano spesso e sono testimoni di un sistema di complessi diritti di pascolo, che sono ancora oggi in parte in vigore e sono stati spesso motivo di litigio in passato.	Sie befinden sich an Orten, an denen die Hirten häufig vorbeikamen, und zeugen von einem komplexen System von Weiderechten, von denen einige noch heute in Kraft sind und in der Vergangenheit häufig Anlass zu Streitigkeiten waren.	They are found in places where shepherds often passed and bear witness to a system of complex grazing rights, some of which are still in force today and were often the cause of disputes in the past.
5	In un secondo tempo erano debordanti a tutto campo con una certa dovizia di annotazioni diaristiche.	È interessante notare come le scritte siano cambiate nel tempo.	Es ist interessant zu beobachten, wie sich die Inschriften im Laufe der Zeit verändert haben.	It is interesting to note how the inscriptions have changed over time.
6	La loro posizione sulle rocce testimonia gli incessanti andirivieni estivi dei singoli pastori. I pastori erano parte di un sistema di diritti territoriali di pascolo contesi e sempre piuttosto complessi.	Molte delle scritte più antiche sono circondate da ordinate cornici e riportano gli stessi nomi presenti sui cippi di confine, ovvero rocce poste a delimitare i confini fra le proprietà in montagna.	Viele der älteren Inschriften sind von einem hübschen Rahmen umgeben und tragen dieselben Namen wie die Grenzsteine, die zur Abgrenzung von Berggrundstücken aufgestellt wurden.	Many of the older inscriptions are surrounded by neat frames and bear the same names found on boundary stones, i.e. rocks placed to mark the boundaries between mountain properties.
7	Questi diritti sono ancora parzialmente in vigore.	Le scritte più recenti non hanno la cornice ma sono	Die jüngeren Inschriften haben keinen Rahmen,	The more recent inscriptions do not have frames but are

		ricche di annotazioni dei pastori, come le pagine di un diario.	sind aber voll mit Anmerkungen der Hirten, wie die Seiten eines Tagebuchs.	full of shepherds' annotations, like the pages of a diary.
8	Il complesso di queste scritte pastorali rappresenta uno straordinario archivio notarile per ricostruire una pagina importante della storia della pastorizia nella valle di Fiemme, e dunque della storia economica, sociale e ambientale della valle.	L'insieme di queste scritte rappresenta un'importante testimonianza, un archivio per ricostruire un pezzo della storia della pastorizia della Val di Fiemme, dal punto di vista economico, sociale e ambientale.	Alle diese Inschriften stellen ein wichtiges Zeugnis dar, ein Archiv zur Rekonstruktion eines Teils der Geschichte des Hirtenwesens im Fleimstal unter wirtschaftlichen, sozialen und ökologischen Gesichtspunkten.	All these inscriptions represent an important testimony, an archive to reconstruct a piece of the history of shepherding in Val di Fiemme, from an economic, social and environmental point of view.
9	L'archivio può essere ricostruito a partire dalla sequenza cronologica espressa dalle date, dalla georeferenziazione delle scritte stesse e dal riconoscersi dei singoli pastori dietro alle sigle e ai «segni di casa».	Le scritte e gli stemmi di famiglia sono infatti opera di diversi pastori che nel corso dei secoli hanno affrescato la montagna in punti diversi.	Die Inschriften und Familienwappen sind in der Tat das Werk verschiedener Hirten, die den Berg im Laufe der Jahrhunderte an verschiedenen Stellen mit Fresken versehen haben.	The inscriptions and family crests are in fact the work of various shepherds who have frescoed the mountain in different places over the centuries.
10	Un lavoro che, previo il completamento del database delle scritte, può dirsi ora appena iniziato.	Non tutte le scritte sono presenti nei database, il lavoro è di fatto appena cominciato.	Nicht alle Inschriften sind in den Datenbanken vorhanden, die Arbeit hat gerade erst begonnen.	Not all inscriptions are present in the databases, the work has in fact only just begun.
11	Secondo e non minore motivo di	Il secondo motivo di interesse sono le	Der zweite interessante Grund	The second reason for interest is the

	interesse, tuttavia, è il modo in cui questi graffiti pastorali si ricollegano direttamente alla lunga tradizione europea dell'arte rupestre preistorica.	similitudini fra queste scritte e l'arte rupestre paleolitica.	sind die Ähnlichkeiten zwischen diesen Schriften und der paläolithischen Felskunst.	similarities between these writings and Palaeolithic cave art.
12	L'arte rupestre che ha inizio nelle grotte del paleolitico dordonese e cantabrico e prosegue poi in età protostorica con i grandi siti graffitati del Monte Bego e della val Camonica.	Sorge naturale il paragone ai dipinti delle grotte della Dordogna o della Cantabria, o anche a quelli dei siti archeologici del Monte Bego e della val Camonica in Italia.	Vergleiche mit den Malereien in den Höhlen der Dordogne oder Kantabriens oder auch mit denen in den archäologischen Stätten von Monte Bego und Val Camonica in Italien drängen sich auf.	It is natural to compare them to the paintings in the caves of the Dordogne or Cantabria, or even to those in the archaeological sites of Monte Bego and Val Camonica in Italy.
13	Le scritte si ricollegano per stile e per alcuni aspetti dell'iconografia e sono d'età moderna e quasi precontemporanea.	Le scritte del monte Cornon sono inequivocabilmente d'età moderna e quasi precontemporanea eppure si collocano comunque in parte all'interno della lunga tradizione europea dell'arte rupestre preistorica.	Die Inschriften auf dem Monte Cornon sind unverkennbar modernen und fast schon vorzeitlichen Alters, und doch stehen sie teilweise in der langen europäischen Tradition der prähistorischen Felskunst.	The inscriptions on Monte Cornon are unmistakably of modern and almost pre-contemporary age, and yet they are still partly within the long European tradition of prehistoric rock art.
14	Infatti, al di là dell'evidente incongruità cronologica, vi è un certo numero di fattori che consentono di collocare le scritte pastorali della valle di Fiemme sul	Condividono con le sorelle più antiche ispirazione e stile, le scritte infatti presentano alcuni aspetti dell'iconografia.	Sie teilen mit ihren älteren Schwestern Inspiration und Stil, wobei die Inschriften sogar bestimmte Aspekte der Ikonographie aufweisen.	Sharing inspiration and style with their older sisters, the inscriptions in fact feature certain aspects of iconography.

	medesimo tracciato ideale della grande arte rupestre europea.	L'unica incongruenza è quella cronologica, eppure le scritte della Val di Fiemme possono comunque considerarsi un'evoluzione naturale della grande arte rupestre europea.	Die einzige Ungereimtheit ist chronologischer Natur, dennoch können die Inschriften des Fleimstals als eine natürliche Entwicklung der großen europäischen Felskunst betrachtet werden.	The only inconsistency is chronological, yet the Val di Fiemme inscriptions can still be considered a natural evolution of the great European rock art.

Table 2: Comparison between the text after the first edit and the text after the second edit, automatic neural translation into German (DE) and English (UK) of the second edit.

The table first makes visible a structural consequence of the two-phase preparation: segmentation is stabilised and aligned to the translation grid. The sequence grows from 7 sentences in the original to 14 after pre-editing and to 15 in the final version. This increase is not merely cosmetic; it distributes information into units that minimise long-distance dependencies and scope ambiguities, which are typical sources of instability for MT.

Beyond segmentation, the post-editing pass focuses on referential explicitness, lexical normalisation of culture-bound items, and light discourse cueing, always within idiomatic Italian. In row 2, a sentence-initial pronoun was avoided because it would have triggered a predictable agreement error in German (neuter das cannot stand for masculine der with Grund). The Italian therefore repeats the antecedent noun rather than pronominalising—an upstream adjustment that prevents a major downstream error. In the same sentence, the list-like noun phrase was de-nominalised by inserting a finite verb, and "dare testimonianza" was preferred to "rendere testimonianza", the latter being less frequent and more prone to awkward calques. The result is a structure that supports idiomatic phrasing in EN and clean agreement in DE.

Row 3 exemplifies lexical normalisation for culture-bound items. The expression "segni di famiglia" was recast as "stemma di famiglia" ("family crest" in EN; "Familienwappen" in DE), which better captures the intended cultural concept and narrows the range of plausible mappings for the engine. The diminutive "edicolette" required special care: because "edicola" in Italian can denote both a small shrine and a newsstand, the suffix –etta multiplies readings. Left underspecified, outputs in both EN and DE tended to

misanalyse the item, occasionally producing non-words. The post-edited source therefore reformulates the passage to anchor the intended meaning more explicitly, while resisting the temptation to hard-code target renderings into the Italian.

Between rows 5–7, content is re-sequenced to make logical links locally recoverable. Material that was adjacent after pre-editing is redistributed so that anaphoric references (for instance, to previously mentioned "edicolette") appear where the antecedent is readily available in context. The effect is a tighter chain of reference that lowers the risk of misattachment in both languages. This is precisely Somers' logic: move what is cheaper to move in the input to avoid fixing the same problem repeatedly in the outputs.

In row 8, the noun "complesso" was replaced with "insieme" to avoid polysemy. In contemporary Italian, complesso can also denote a musical band, which invites misleading mappings or facetious readings. The neutral "insieme" preserves the intended "set/whole" reading and guides both EN and DE towards stable, context-appropriate phrasing. In the same row, "archivio notarile" was rephrased so that the legal nuance encoded by "notarile" is preserved as "testimonianza" ("testimony"). The underlying concept—attestation—remains intact, but the Italian syntagm is recast to avoid brittle calques such as notarial archive in English or Notariatsarchiv in German where the discourse clearly points to evidence/record rather than to an institution.

Row 9 illustrates controlled simplification. An initial clause that contributed little propositional content was removed. The deletion reduces processing load, clarifies scope and attachment, and improves the odds that English will deliver idiomatic clause shaping while German maintains consistent modifier chains in verb-final positions. In rows 11–13, content is repositioned and, where helpful, specific geographic references are added. This provides orientation for readers unfamiliar with the region and assists the engine in resolving culture-bound references (e.g., rock-art locales and local traditions) without baking target-language solutions into the source. Finally, the closing sentence is split into two units. The logical relation between ideas is surfaced rather than left implicit; attachment errors become less likely; and the overall segmentation remains consistent with the strategy adopted across V2.

Two clarifications delimit the scope of inference. First, the intermediate pre-edited outputs are used diagnostically—they help identify what still needs to be fixed at input—but they were not submitted to participant review. Secondly, quantitative evaluation—Matecat's quality score (EPT) and severity-weighted raw error points (Neutral = 0; Minor = 0.5; Major = 2)—concerns only the contrast between the unedited source (V1) and the twice-edited

source (V2). The purpose of Table 2 is therefore expository: it makes the source-side logic of V2 transparent and shows why the interventions documented here are expected to benefit both IT→EN (UK) and IT→DE. Chapter 6 (Results and Analysis) reports whether those expectations are borne out in the data and whether the perceived gains in the questionnaires converge with the objective metrics.

5.4 Operational use of DeepL (NMT) and Matecat

In the experimental design, DeepL had a clearly delimited, twofold function. During the second phase (post-editing of the source text), it was used to generate short test runs—translations produced solely to observe how the engine handled the pre-edited draft and to guide small, language-neutral adjustments to the Italian source. This internal loop—running DeepL, minimally revising the source in light of recurrent behaviours, and running the engine again—was confined to the second phase and stopped once the source was fixed as Version 2 (V2). The same system was then used to produce the final outputs in both directions, IT→EN (UK) and IT→DE, together with the baseline outputs for Version 1 (V1), the unedited text. None of these final translations was altered by hand: human intervention remained strictly source-side. All runs were carried out with DeepL (free edition), with no glossaries and no register/tone control, under the same account/session to minimise configuration drift. The only difference across input sets concerned the source condition, V1 (unedited) versus V2 (twice-edited).

Once the four final translations had been obtained for each participant (V1_EN, V1_DE, V2_EN, V2_DE), the texts were transferred manually into Matecat, segment by segment, preserving the original segmentation of the source (V1 with 7 segments, V2 with 15). Manual insertion was preferred to prevent any pretranslation or auto-completion that the tool might otherwise trigger. Each participant received a private quartet of projects labelled X_V1_EN, X_V1_DE, X_V2_EN, X_V2_DE (with X = A–J), and could view only their own four projects, ensuring anonymity and traceability.

The review environment in Matecat was configured to isolate annotation from automated aids: MT and TM suggestions were disabled, and no translation memories were linked, so that confirmed segments could not propagate elsewhere. Quality assessment used Matecat's standard Quality Review profile with labels adjusted in an MQM-inspired fashion—Style, Translation Errors, Terminology Consistency, Language Quality—and with the Tag category removed as irrelevant to the task. Severity weights were those provided by Matecat: Neutral = 0, Minor = 0.5, Major = 2; Neutral entries were recorded but carried no

points. The quality score (Matecat's term for EPT, errors per thousand words) was calculated from the sum of raw error points (Minor/Major), scaled to one thousand and divided by the project's reviewed-word count. Matecat recorded 325 words for V1 and 344 for V2 (for both language pairs), and these denominators were used consistently. Projects were marked Complete once all segments had been reviewed; per-segment logs and timestamps ensure auditability, although time-on-task was not analysed.

After completion, CSV exports were retrieved for each project, containing the quantitative measures and the distribution of errors by severity and category. The data were then organised into four analysis tables consistent with the Methodology chapter: two by version (V1 and V2, reporting EN and DE together) and two by language (English and German, each contrasting V1 vs V2). Rows list participants A–J; columns report the quality score and raw points for Neutral/Minor/Major, further broken down by error category. This arrangement enables inspection of overall performance (quality score) and of the error profile, while keeping a clear link to the operational choices outlined above.

In sum, DeepL and Matecat were used in a complementary, tightly controlled manner: the former to produce the translations (including the brief test runs that informed the second source-editing pass), the latter to collect uniform error annotations and compute the quality score with fixed weights and denominators. This configuration minimises extraneous variability and supports attributing any differences between V1 and V2 to the preparation of the source text (pre-editing plus post-editing of the source text). Chapter 6 (Results and Analysis) examines these differences in detail, bringing together the quantitative outcomes and the qualitative evidence from the questionnaires on accuracy, fluency and fidelity.

5.5 Participant instructions

Participants received a concise instruction sheet by email together with four private Matecat links and the five Google Forms. The sheet fixed a single, mandatory order for the four reviews—DE-V1 → DE-V2 → EN-V1 → EN-V2—and directed that the corresponding questionnaire be completed immediately after each review. This sequencing was selected to stabilise the evaluation context (one language at a time), reduce cognitive switching, and anchor perception judgements to the just-reviewed output. Access in Matecat was restricted to a quartet of projects per participant, labelled X_V1_DE, X_V2_DE, X_V1_EN, X_V2_EN, where X = A–J served as an anonymised identifier; only the participant's own quartet was visible.

The instruction sheet (cf. Appendix B) specified the review mode and the annotation protocol. Reviews were carried out in Matecat's Revise mode at segment level. For each segment, raters could: (i) open the error panel and select a category; (ii) add a comment where clarification was helpful; and (iii) assign a severity level. Minor layout artefacts, such as trailing spaces at the end of a segment, were to be ignored as out of scope. To avoid contamination from tool-side aids, MT and TM suggestions were disabled and no translation memories were linked; confirmed content in one project could not propagate into others. Outputs had been generated with DeepL (free edition) and were not to be altered outside the Matecat revision interface; the task was to annotate errors (and, if needed, leave some comments), not to rewrite the targets.

Error annotation followed the standard Quality Review profile with labels aligned to the MQM-inspired taxonomy used throughout the study: Style (register and textual cohesion), Translation Errors (adequacy and transfer of meaning), Terminology Consistency (use and coherence of terms), and Language Quality (spelling, grammar, syntax, punctuation). The Tag category was removed as not pertinent to this material. Severity levels and weights were fixed and communicated explicitly, as they determine the scoring downstream: Neutral = 0 points, Minor = 0.5 points, Major = 2 points. Neutral flags could be recorded (e.g., to mark a very slight infelicity or a preference) but do not contribute points; Minor and Major add, respectively, half a point and two points to the raw error total. The instruction sheet (cf. Appendix B) highlighted that Matecat's quality score (its normalised EPT) is computed from the sum of severity-weighted error points, scaled to 1,000 words and divided by the project's reviewed-word count. In this setup, V1 projects contained 7 segments and 325 reviewed words, whereas V2 projects contained 15 segments and 344 reviewed words; these denominators were used consistently when quality scores were later analysed. Making the weights explicit in the instructions was intended to align rater choices with the study's primary outcomes and to keep annotation behaviour consistent across tasks. Technical assistance was available on request throughout; no time limits were imposed on the review tasks or on the forms.

Overall, the instruction package translates the experimental design into an executable procedure: a controlled sequence of reviews, a uniform annotation protocol with category and severity explicitly defined (Neutral = 0; Minor = 0.5; Major = 2), and a survey schedule that mirrors the review timeline. By construction, this configuration aims to attribute any observed V1–V2 differences to the two-step preparation of the source rather than to tool aids,

uncontrolled order effects, or memory bias, while keeping the link clear between on-screen decisions in Matecat and the downstream quality score used in analysis.

5.6 Questionnaires

Five anonymous Google Forms were used to complement the Matecat review data. Forms were configured so that no identifying information was collected, responses were submitted once and could not be edited, and only the open comment prompts were optional. This configuration serves two purposes: it encourages candid judgements (no link back to individuals) and it preserves a clean timeline (answers captured immediately after each task, with no retroactive changes). Because Google Forms stores submissions automatically, there was no file return step; results could be monitored as they arrived.

The timeline of surveys mirrored the task order to keep recall as close as possible to the intervention just completed. Before any revision activity, a background form captured self-reported proficiency in Italian, German and English, prior exposure to CAT tools and NMT, and years of translation experience. After the first review (German, V1), a post-task form asked three five-point judgements—accuracy, fluency, fidelity—about that specific output; each scale option was accompanied by a short descriptor to standardise interpretation (for example, a “very accurate” choice indicated that the target conveyed the source meaning without salient errors). The form also included a non-mandatory slot for issues not covered by the Matecat error categories and an open comment field. After the second review (German, V2), the same three items were repeated for V2, followed by three explicit comparisons asking whether V1 or V2 performed better on accuracy, on fluency, and on fidelity; an open comment field closed the form. The pattern then repeated for English: a post-task form after EN-V1, and, after EN-V2, the same V1–V2 comparison items for English. The final form added a brief cross-linguistic section asking, again for accuracy, for fluency, and for fidelity, which language (EN or DE) was perceived to have performed better overall. This staging keeps the respondent within one language at a time (German first, then English), aligning the perceptual task with the Matecat sequence and reducing context switching.

All closed questions were mandatory; open comment prompts were optional by design, so that free-text remarks would flag genuinely salient issues rather than forced content. There were no time limits, so each form could be completed immediately after the relevant review without pressure. Each participant could access only their own submissions; results were inspected in aggregate. In line with the study’s plan, the survey corpus is treated as qualitative evidence: it characterises perceptions for each condition, articulates perceived

V1 vs V2 differences within a language, and provides a broad indication of EN vs DE preferences at the end. The full instruments are reproduced in the Appendix for transparency and replicability. A consequence of the anonymity setting is that survey responses are not linkable to individual participants (A–J); this was an intentional trade-off in favour of uninhibited judgements and a simple, low-friction workflow.

6. Results and Analysis

6.1 Data preparation and scoring

This chapter analyses four sets of observations produced by the same ten raters—V1_EN, V1_DE, V2_EN and V2_DE. Because every participant reviewed all conditions, the study adopts a within-subjects design: differences between V1 and V2 reflect the effect of editing the source rather than the personal strictness or indulgence of different evaluators. All reviews were carried out in Matecat under the same configuration, with MT/TM suggestions disabled and an identical error taxonomy and severity scheme for everyone.

The outcome measure is Matecat's quality score (its EPT), interpreted as the number of severity-weighted errors per thousand words. Each annotation carries a fixed weight by severity: Neutral is recorded but worth zero points; Minor adds half a point; Major adds two points. Matecat sums the points for a project, adjusts the total to a notional length of one thousand words, and then scales it by the actual number of words reviewed in that project. This normalisation lets scores be compared even if target texts are not exactly the same length. In our data, Version 1 projects contain 325 reviewed words and seven segments, whereas Version 2 projects contain 344 reviewed words and fifteen segments. Because Neutral issues have weight zero, they are useful as qualitative flags but do not change the metric.

In reporting results, simple totals across participants are avoided because they conceal how scores are distributed and can be dominated by a few extreme cases. Instead, each condition is summarised with standard descriptive statistics that convey both level and spread. The mean gives the plain arithmetic average but is sensitive to very high or very low values. The median identifies the middle observation once scores are ordered and is therefore robust to outliers, often offering a better picture of a "typical" outcome. Dispersion is captured in two complementary ways. The standard deviation indicates how tightly scores cluster around the mean: larger values signal more scatter, smaller values indicate compact data. The interquartile range focuses on the central half of the observations and is obtained by subtracting the first quartile from the third. Quartiles split ordered data into four equal parts; following the classical approach associated with John W. Tukey, with ten observations the median is taken between the fifth and sixth values, the first quartile is the median of the lower half, and the third quartile is the median of the upper half. Because it ignores both tails, the interquartile range is a stable summary of spread when a few observations are extreme.

Improvement from the unedited to the edited condition is quantified within each person. For every participant and for each language a difference is computed by subtracting the V1 score from the V2 score; negative values therefore indicate that the twice-edited source produced a better outcome. These ten differences are then described with a mean and a median to indicate the typical size of the change, and with a measure of spread (either the standard deviation or the interquartile range) to indicate how consistent the change is across people. To express the magnitude of the change in unit-free terms, a within-subjects effect size (Cohen's dz) is also reported; this rescale of the average difference by the variability of those differences helps judge whether the improvement is small, moderate or large relative to the noise in the data. A negative value for this effect size simply mirrors the sign convention that lower scores are better.

Alongside these continuous summaries, the analysis references Matecat's default acceptability threshold of eight points per thousand words as a practical benchmark. Given the reviewed-word counts above, this threshold corresponds to roughly 2.60 error points in V1 and 2.752 in V2. The threshold is used to communicate results in operational terms—how many outputs would be classed as acceptable—while the core analysis relies on the continuous quality score rather than a pass/fail label.

Finally, two additional sources of information are retained to contextualise the main metric. First, the distribution of error points by category (Style, Translation Errors, Terminology Consistency, Language Quality) and by severity clarifies where improvements occur—whether gains come mainly from fewer Major errors in adequacy, from lighter stylistic issues, or from other profiles. Second, the post-task questionnaires on accuracy, fluency and fidelity provide a qualitative check on the numerical patterns. All ten raters completed all four projects under identical conditions, outputs were imported consistently, and no datasets required exclusion. This preparation yields a coherent basis for the analyses that follow: four comparable conditions per rater, one normalised outcome, clear severity weights, and descriptive summaries designed to answer the two research questions without being distorted by totals or outliers.

6.2 Descriptive statistics: what the numbers show

This section presents a detailed picture of the quality scores (lower is better) across the four conditions—V1_DE, V2_DE, V1_EN and V2_EN—using descriptive statistics that convey both the typical level and the degree of dispersion of the ratings. For clarity, the mean is reported as the simple arithmetic average, while the median identifies the middle value once

scores are ordered and is therefore less affected by extreme cases. Variability is summarised with the standard deviation, which increases when observations are widely scattered around the mean, and with the interquartile range (IQR), defined as the distance between the third and first quartile in the classical Tukey sense, thus focusing on the central half of the data. These complementary summaries make it possible to avoid uninformative totals and to distinguish central tendencies from outliers. The numerical values are collected in Table 3, which can be read alongside the narrative below.

Version	**Language**	**Mean**	**Median**	**SD**	**Q1**	**Q3**	**IQR**
V1	DE	37.51	28.46	28.3	20	53.85	33.85
V1	EN	35.38	26.15	31.67	13.85	41.54	27.69
V2	DE	11.63	2.91	23.79	1.45	10.17	8.72
V2	EN	9.74	0.73	18.9	0	13.08	13.08

Table 3: Descriptive statistics by condition

For German under Version 1 (V1_DE), the mean quality score stands at 37.51 and the median at 28.46, with a standard deviation of 28.50 and an interquartile range of 27.55 (see Table 3). The mean being higher than the median indicates a right-tailed distribution, where a few particularly high scores pull the average upward. In practical terms, the profile is broadly unsatisfactory: most raters assign medium-to-high scores, and a subset assigns very high scores that emphasise the lack of readability and idiomaticity observed in the raw machine output. When the same language is assessed in Version 2 (V2_DE), both level and spread change in a way that is consistent with a substantive improvement. The mean falls to 11.63 and the median to 2.91; while the standard deviation remains relatively large at 23.79—reflecting one or two high profiles that still inflate the dispersion—the IQR contracts sharply to 8.72. Read together, these indicators show that the bulk of evaluations has shifted decisively into a low-score region, and that the German output under V2 is perceived as considerably more acceptable than under V1, even if a few difficult cases remain.

The English results display a closely parallel pattern. For Version 1 (V1_EN), the mean is 35.38 and the median 26.15, with a standard deviation of 31.67 and an interquartile range of 27.69, again suggesting a distribution with medium-to-high scores and a fair amount of scatter, driven in part by long and syntactically dense sentences in the raw output. In Version 2 (V2_EN) the shift is even more pronounced: the mean drops to 9.74 and the median

is very close to zero at 0.73. This near-zero median implies that at least half of the ratings lie at very low values, i.e. the English output is judged markedly better after the two editing phases. The standard deviation remains non-negligible at 18.90 and the IQR is 13.08—wider than in V2_DE—which suggests a slightly broader central spread for English than for German in the edited condition; nevertheless, the overall level is clearly well below the V1 range, signalling a substantial gain.

Considering medians and IQRs together provides the most robust view of change because these statistics summarise the "core" of the distributions and are not driven by a handful of extreme observations. From V1 to V2 the median collapses in both languages (DE: 28.46→2.91; EN: 26.15→0.73), which indicates that improvement is not limited to isolated cases but concerns the typical evaluation. In German, the interquartile range shrinks dramatically (33.85→8.72), so judgments become more compact; in English it also falls (27.69→13.08), though less dramatically, and remains well below the V1 spread. The persistent size of the standard deviations in the V2 conditions is compatible with the presence of one or two high profiles that keep the mean's dispersion elevated; this is precisely the context in which the median and Tukey's interquartile range yield a more faithful picture of the outcome.

A brief methodological note helps interpret these differences. The quality score is Matecat's EPT—an error-severity metric normalised to a 1,000-word basis—which assigns zero points to Neutral flags, half a point to Minor issues and two points to Major issues. The total error points in a project are rescaled by the project's reviewed-word count so that scores remain comparable even if target lengths differ. In our data, Version 1 projects comprise 325 reviewed words distributed over seven segments, whereas Version 2 projects comprise 344 reviewed words over fifteen segments. The reductions observed from V1 to V2 therefore do not simply reflect increased segmentation; the scale is comparable, and the change points to a genuine effect of source-side editing on the behaviour of the neural machine translation system.

To complement the continuous summaries with an operational benchmark that is familiar to practitioners, Table 4 reports the proportion of outputs falling below Matecat's default acceptability threshold of eight points per 1,000 words.

Version	**Language**	**Number of participants**	**Pass**	**Pass %**

V1	DE	10	1	10.0
V2	EN	10	1	10.0
V1	DE	10	7	70.0
V2	DE	10	7	70.0

Table 4: Acceptability under threshold (8 per 1,000 words)

The pass rate improves markedly in both languages: whereas in Version 1 only one out of ten translations per language meets the threshold, in Version 2 seven out of ten do so. This practical indicator aligns with the median/IQR evidence and shows that, for the majority of cases, the twice-edited source moves the output into an acceptable quality zone. Taken together, the descriptive statistics and the threshold analysis support the same conclusion: Version 2 lowers the quality score substantially and consistently in both German and English, and the improvement concerns the central mass of the distributions rather than a few favourable outliers.

6.3 Within-participant differences (V2−V1) and effect size (Cohen's dz)

This subsection quantifies how much the quality score changes for the same reviewer when moving from the unedited source (V1) to the twice-edited source (V2). For each participant and for each language direction, the paired contrast was computed as V2 quality score minus V1 quality score. Because the Matecat score increases with the number and severity of errors, lower scores indicate better quality; accordingly, negative V2−V1 values denote improvement. All ten participants completed both versions in both languages, so the comparison is genuinely within-participant and not affected by between-person differences.

To make the distribution of these paired differences easy to read, the results are summarised separately for IT→DE and IT→EN with the mean (average shift), the standard deviation (spread across raters), the median (central tendency robust to outliers), and the inter-quartile range, IQR (the width of the middle 50% of values). The share of participants who improved (percentage with V2−V1 < 0) is also reported. Finally, a compact standardised index for paired data, Cohen's dz, is included; here dz is the mean of the V2−V1 differences divided by their standard deviation. Its sign preserves direction (negative values favour V2), while the absolute magnitude describes how large the within-participant shift is. The summary by language is presented in Table 5

Language	Mean	Standard Deviation	Median	IQR	Cohen's dz	% with V2 < V1
IT - DE	-25.6	16.79	-19.99	15.53	-1.53	100.0
IT - EN (UK)	-25.65	21.81	-20.77	24.14	-1.18	100.0

Table 5: Within-participant differences (V2−V1) on the quality score by language. Negative values indicate improvement (V2 < V1). N = 10 per language; reviewed words: 325 (V1) and 344 (V2); severity weights: Neutral = 0, Minor = 0.5, Major = 2.

In IT→DE, the mean V2−V1 difference is −25.60 (SD 16.79), with a median of −19.99 and an IQR of 15.53; all participants improved (100% with V2−V1 < 0). In IT→EN (UK), the mean is −25.65 (SD 21.81), with a median of −20.77 and an IQR of 24.14; again, all participants improved. The standardised effects are dz = −1.53 for German and dz = −1.18 for English, indicating a large and a medium-to-large within-participant improvement, respectively. In practical terms, not only do average scores move in the expected direction (lower in V2), but the movement is consistent across raters and sizable in both language pairs.

These findings connect directly to the research aims. With uniformly negative within-participant differences in both directions, RQ1 is supported: the combination of pre-editing and post-editing of the source text produced better NMT output than the unedited baseline. Comparing the magnitude across languages, the means are of similar order (around −25 in both pairs), with somewhat tighter dispersion in German and a more negative dz, so RQ2 is supported in the sense that the improvement is comparable across IT→DE and IT→EN, even if the German gains appear slightly more concentrated. Any residual asymmetries can be interpreted against the linguistic profile of the source passage (long hypotactic structures, culture-bound terminology), which may have posed different challenges to the two target languages. The next subsection will triangulate these quantitative shifts with the questionnaire evidence on accuracy, fluency and fidelity.

6.4 Qualitative data from the questionnaires

Four of the five questionnaires (Questionnaire 2 to 5) were used to collect qualitative judgements on the four translations (V1/V2 × EN/DE) along three recurring dimensions—accuracy, fluency, and fidelity—together with optional free-text comments and a prompt for issues not covered by the Matecat error categories. Forms were anonymous, closed-ended items were mandatory in principle, and the administration order mirrored the

Matecat workflow so that impressions were captured immediately after each revision task. In practice, a small number of items show missing responses; below, valid sample sizes are reported per question to contextualise percentages.

For German V1 (Questionnaire 2), administered right after revising the output from the unedited source, responses were clearly unfavourable on fluency and mixed on accuracy, with fidelity difficult to appraise for several respondents. On accuracy (n = 10), only 30% clustered on the positive side (10% "very accurate", 20% "somewhat accurate"), whereas 40% expressed a negative judgement (30% "somewhat inaccurate", 10% "not accurate at all") and 30% chose "neutral". Fluency was judged particularly weak (n = 7): 42.9% selected "not fluent at all", with a further 28.6% "neutral"; only 28.6% gave positive scores ("very" or "somewhat" fluent, 14.3% each). Fidelity had fewer valid responses (n = 5) and split evenly: 60% placed the translation on the faithful side (20% "very", 40% "somewhat"), while 40% were negative ("somewhat" or "not" faithful at all), suggesting uncertainty driven by the target text's limited idiomatic profile. Comments again highlighted culture-bound references (realia)—for example the use of "Graffiti" in German—and the burden of long, heavily embedded sentences, which complicate flow and make style hard to evaluate independently of structure. (See Fig. 1–3)

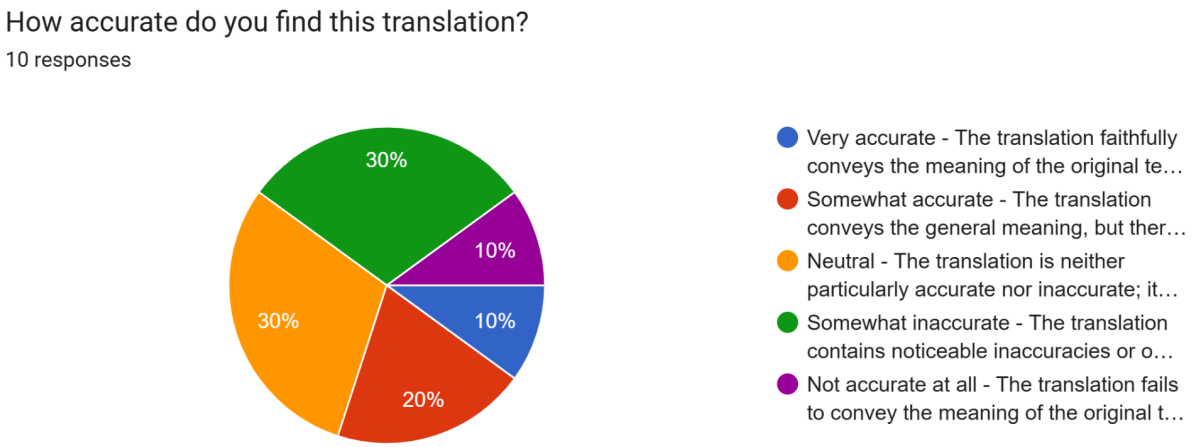

Figure 1: Distribution of accuracy judgements for the German translation from the unedited source (V1_DE)

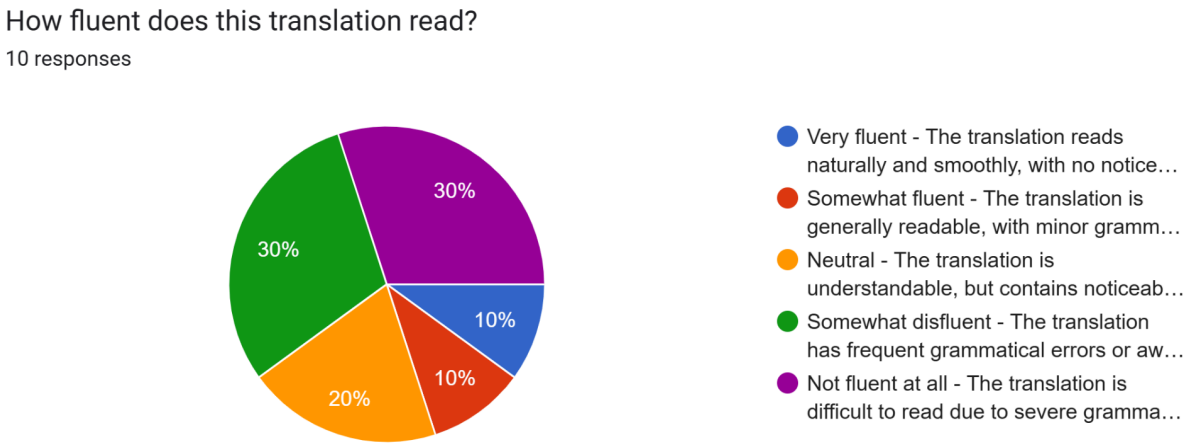

Figure 2: Distribution of fluency judgements for V1_DE

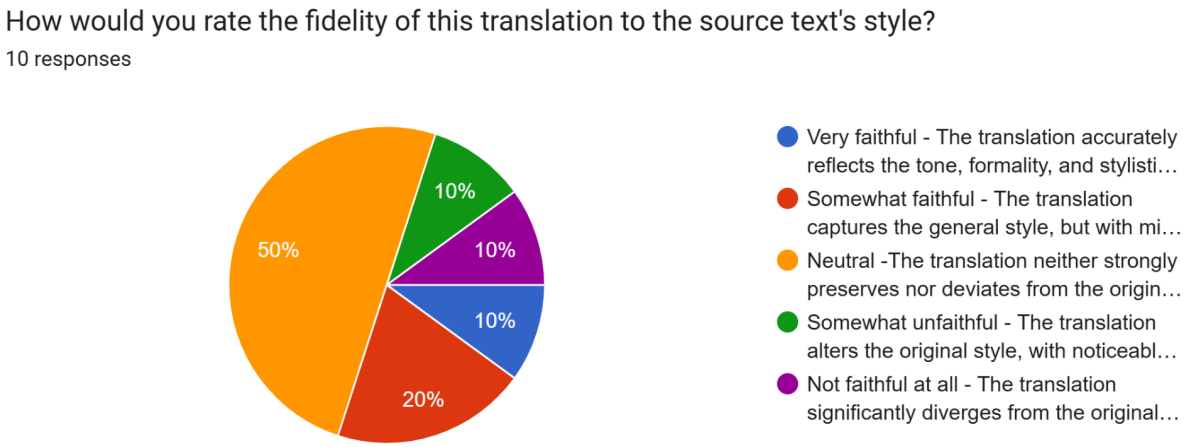

Figure 3: Distribution of fidelity-to-style judgements for V1_DE

For German V2 (Questionnaire 3), i.e. the output produced from the twice-edited source, judgements shifted markedly in a positive direction. Accuracy (n = 10) and fidelity (n = 10) concentrated on the two best categories (accuracy: 60% “very”, 30% “somewhat”; fidelity: 60% “very”, 30% “somewhat”), and fluency (n = 9) was rated “very fluent” by 55.6% and “somewhat fluent” by 44.4%, with no residual negative selections. Crucially, the comparative questions show a clear preference for Version 2 over Version 1: 90% selected V2 as more accurate (10% “about the same”, 0% V1), 100% selected V2 as more fluent, and 80% selected V2 as more faithful (10% “about the same”, 10% V1). Free-text notes still mention

terminology and cultural anchoring as areas requiring attention, but the overall readability and sentence shaping are consistently perceived as improved. (See Fig. 4–6)

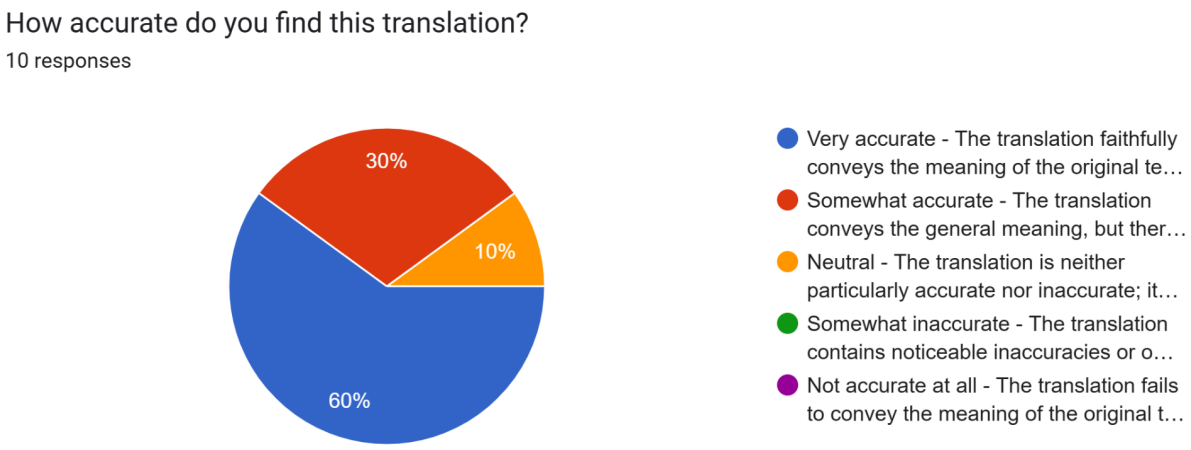

Figure 4: Distribution of accuracy judgements for the German translation from the twice-edited source (V2_DE)

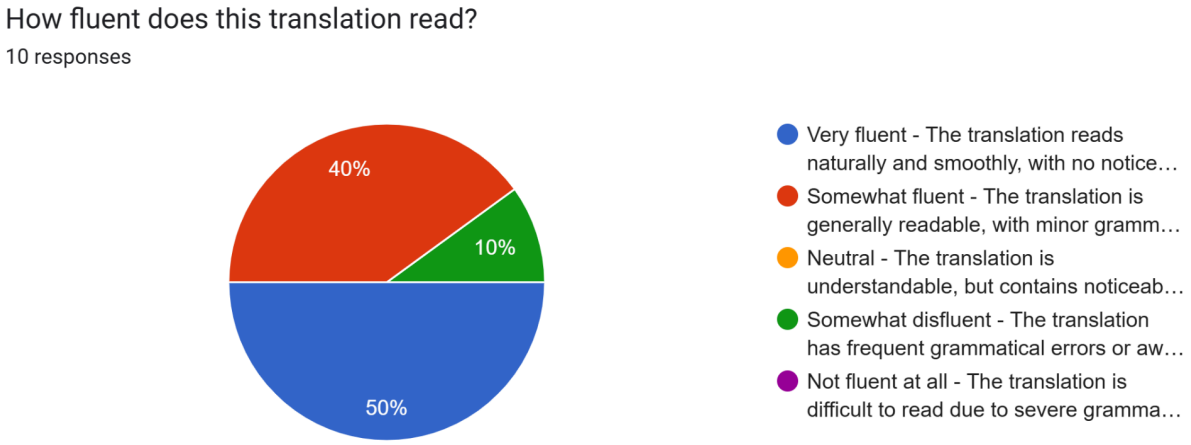

Figure 5: Distribution of fluency judgements for V2_DE

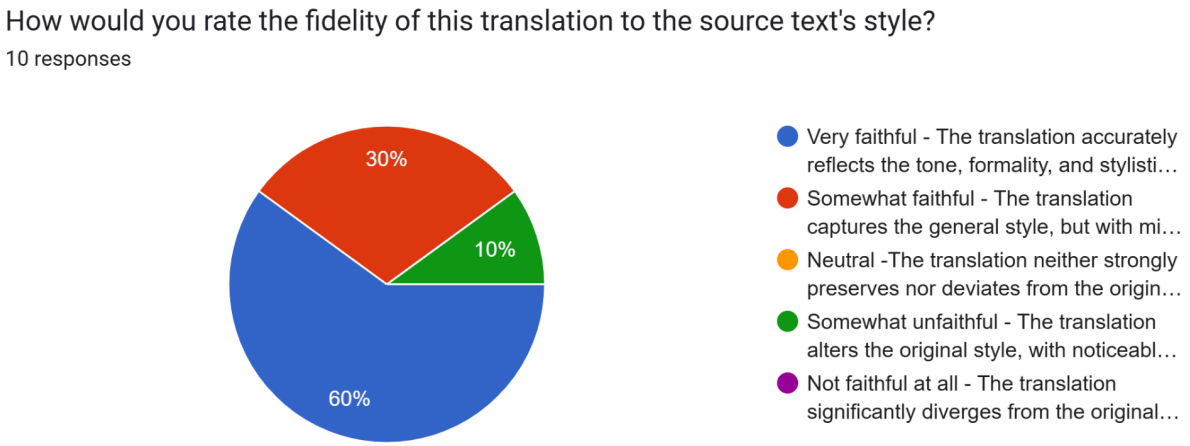

Figure 6: Distribution of fidelity-to-style judgements for V2_DE

For English V1 (Questionnaire 4), the pattern broadly mirrors German V1: the translation from the unedited source was often described as structurally heavy and insufficiently idiomatic. On accuracy (n = 10), 30% gave positive marks (20% “very”, 10% “somewhat”), 40% were “neutral”, and 30% were negative (“somewhat inaccurate”); on fidelity (n ≈ 6), half the responses were “somewhat faithful” (50%), with the remainder split between “very faithful” (16.7%) and “somewhat unfaithful” (33.3%). The fluency item shows fewer valid answers (n = 4) and is therefore interpreted cautiously, but the available data lean towards low naturalness (25% “not fluent at all”, 75% “somewhat fluent”, no “very fluent”), in line with the qualitative remarks about sentence length and calqued structure from Italian. (See Fig. 7–9)

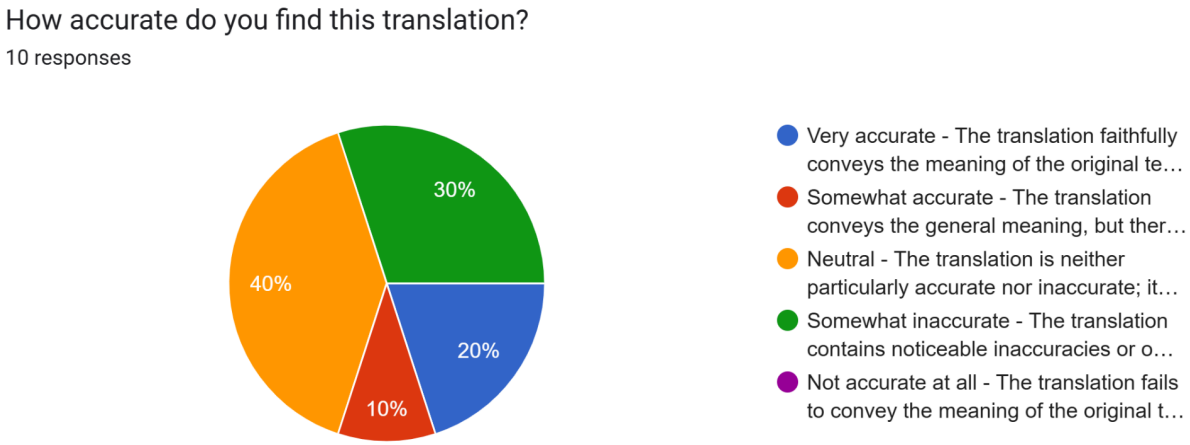

Figure 7: Distribution of accuracy judgements for the English translation from the unedited source (V1_EN)

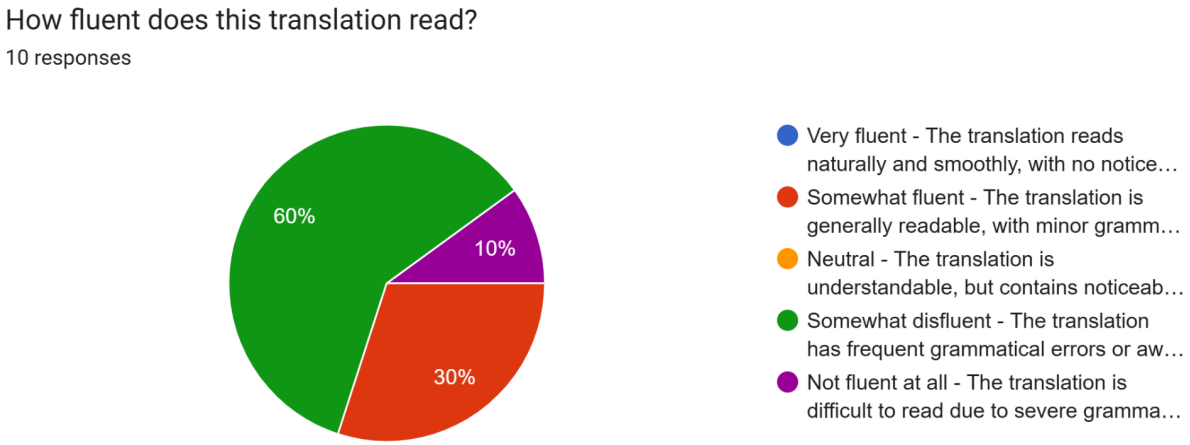

Figure 8: Distribution of fluency judgements for V1_EN

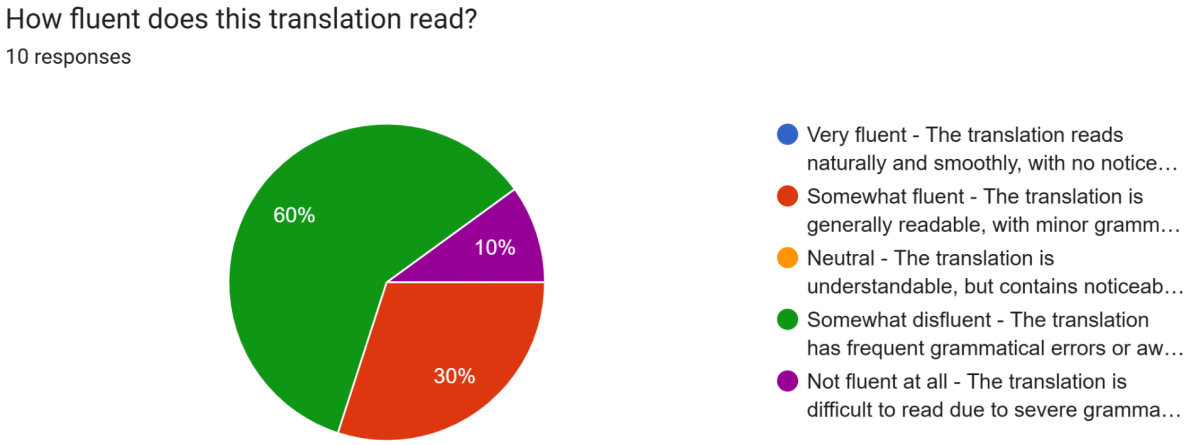

Figure 9: Distribution of fidelity-to-style judgements for V1_EN

For English V2 (Questionnaire 5), the stand-alone evaluations improve across all three dimensions and the direct comparison with V1 is unequivocal. On the V2 stand-alone items (all n = 10), accuracy is rated “very” or “somewhat” accurate by 90% (50% + 40%), fluency by 90% (70% “very”, 20% “somewhat”), and fidelity by 90% (50% “very”, 40% “somewhat”), with the remaining selections neutral and no negative tails. When explicitly asked to compare versions, 100% of respondents preferred V2 on accuracy and fluency, and

90% preferred V2 on fidelity (10% “about the same”). The final cross-language questions indicate a broad preference for English over German within this experimental setup: 70% judged English more accurate overall (30% “equal”, 0% German better), 80% judged English more fluent (20% “equal”, 0% German better), and 60% judged English more faithful (40% “equal”, 0% German better). Open comments add nuance: some residual source-language influence is still noticed in English V2, and the handling of culturally marked terms remains a point of attention, although without serious semantic failures. (See Fig.10–12)

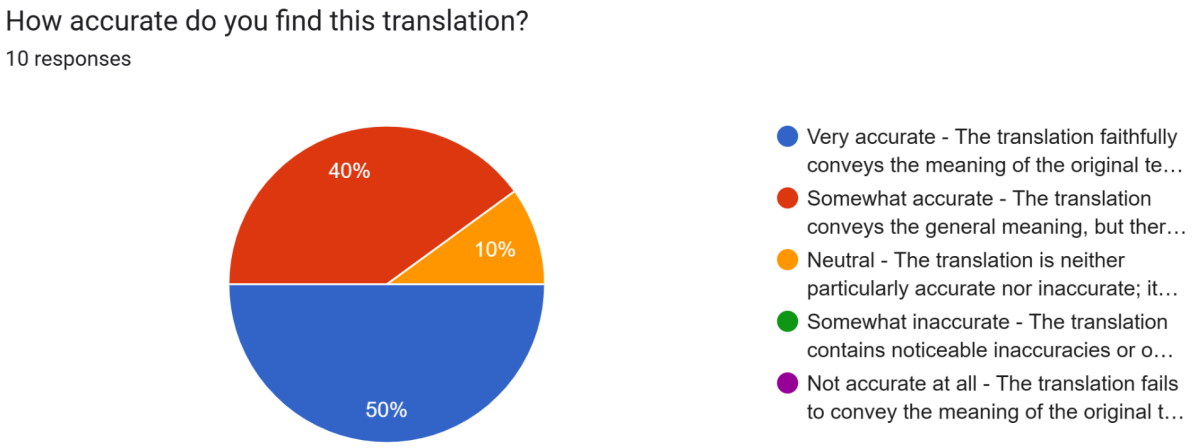

Figure 10: Distribution of accuracy judgements for the English translation from the twice-edited source (V2_EN)

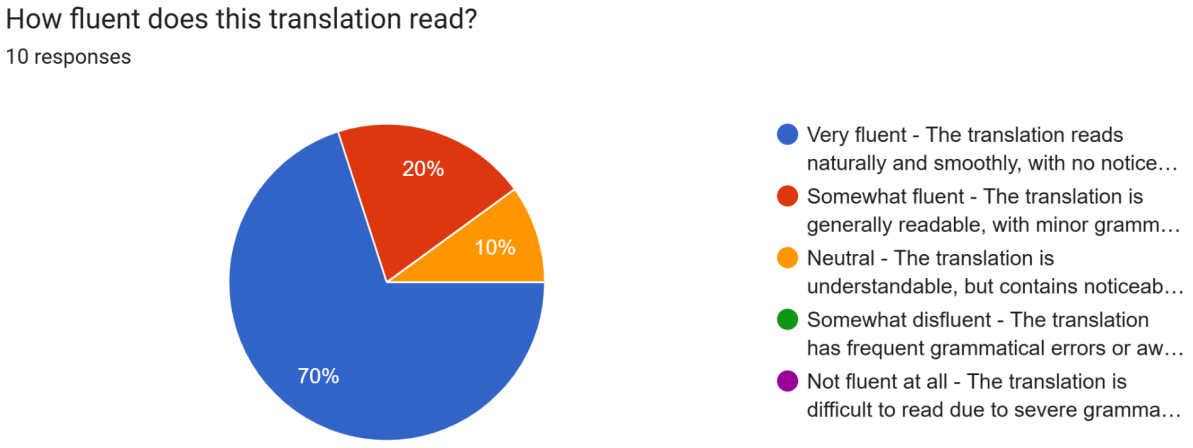

Figure 11: Distribution of fluency judgements for V2_EN

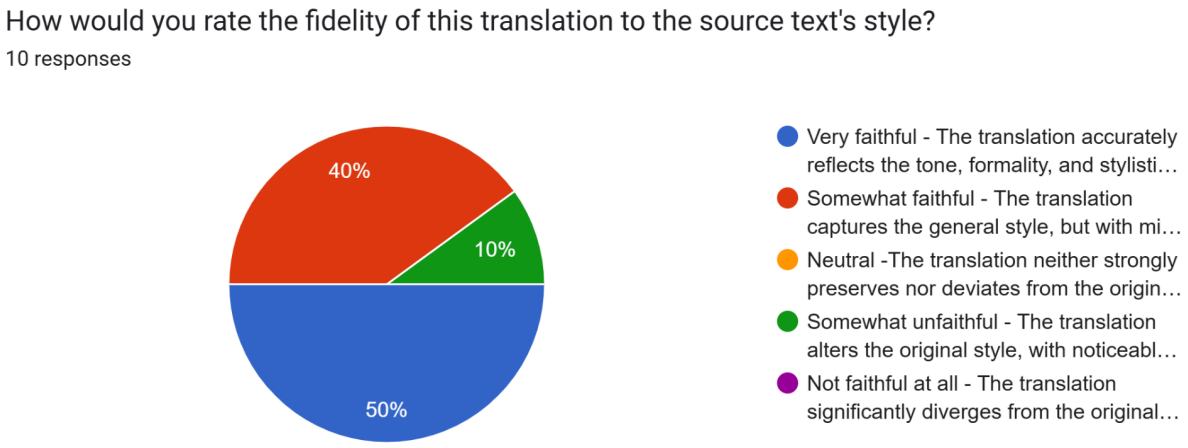

Figure 12: Distribution of fidelity-to-style judgements for V2_EN

Taken together, these qualitative findings align with the quantitative pattern from Matecat: V1 is consistently penalised for sentence-level complexity and unstable phrasing, while V2 benefits from the two-step editing of the source, which reduces structural ambiguity and stabilises segmentation. The questionnaires add interpretive depth in two ways. First, they independently corroborate the within-subject improvement from V1 to V2 in both languages (addressing RQ1) and suggest that the perceived gain is stronger in English than in German (a cue for RQ2). Second, they help explain residual error points in V2 despite lower quality scores: culture-bound items and domain-specific terminology do not vanish with structural simplification and may still depress perceived fidelity even when fluency and accuracy are rated highly. In the remainder of this chapter, these qualitative trends are woven into the descriptive statistics of the quality score and error distributions to provide a convergent answer to the two research questions.

6.5 Quantitative–qualitative triangulation for RQ1 and RQ2

This section integrates the objective indicators obtained from Matecat with the subjective judgements collected through the questionnaires, in order to address the two research questions in a convergent way. The quantitative side relies on Matecat's quality score—i.e., the Errors per Thousand words (EPT) computed from annotated error points—together with distributional summaries and within-participant contrasts from V1 to V2. The qualitative side draws on the post-task surveys that asked raters to evaluate accuracy, fluency, and fidelity and to leave open comments. The expectation is that if the two-step source-side editing truly

increases translatability, we should observe lower quality scores and fewer severe errors in V2 than in V1 and a clear shift in perceived accuracy/fluency/fidelity in favour of V2.

On the qualitative side, the comparative blocks in the post-revision surveys reinforce this picture. In German (Questionnaire 3), V2 is preferred over V1 by 90% of raters for accuracy, 100% for fluency, and 80% for fidelity (with 10% "Equal" on accuracy/fidelity). In English (Questionnaire 5), preferences are even more decisive: V2 is preferred by 100% of raters for accuracy and fluency, and by 90% for fidelity (10% "Equal"). Comments give useful texture: "Insgesamt weisen beide Übersetzungen gravierende Qualitätsmängel auf, jedoch ist die zweite Version deutlich besser lesbar." (Overall, both translations show serious quality flaws, yet the second version is clearly more readable.) Another participant notes: "Nella versione due le frasi sono più brevi e comprensibili, la traduzione ha comunque bisogno di essere controllata e corretta." (In the second version the sentences are shorter and more understandable, yet the translation still requires checking and correction.) In contrast, perceptions of the unedited outputs stress structural heaviness and poor idiomaticity: "La traduzione risulta poco scorrevole, con frasi lunghe e complesse. Si arriva anche al terzo livello di subordinazione." (The translation reads awkwardly, with long and complex sentences; it even reaches a third level of subordination.) One more nuanced remark for English V2 reads: "The text is overall fluent, I would cut down on the number of relative sentences to make it even easier to read." These testimonies are fully consistent with the numerical reduction in the quality score and with the contraction of variability from V1 to V2.

Language	Mean Δ (V2−V1)	% Improved	V1	V2	V2 preferred over V1 — Accuracy (%)	V2 preferred over V1 — Fluency (%)	V2 preferred over V1 — Fidelity (%)
IT - DE	-25.6	100.0	10.0	70.0	90.0	100.0	80.0
IT - EN (UK)	-25.65	100.0	10.0	70.0	100.0	100.0	90.0

Table 6: RQ1 synthesis by language: within-participant change (Δ V2−V1), percentage under the acceptability threshold (quality score < 8/1000), and questionnaire preferences for V2 over V1 (accuracy/fluency/fidelity)

Do NMT outputs from a source edited in two steps—(i) pre-editing and (ii) post-editing of the source text—outperform outputs from the unedited source? On the numbers alone, the answer is yes. Every rater improves in both language pairs; mean and median Δ (V2−V1) are strongly negative; the proportion under threshold jumps from ≈ 10% to ≈ 70%; major and minor errors per 1,000 both fall. The qualitative evidence points the same way. In German (Questionnaire 3), V2 is preferred over V1 by ≈ 90% of raters for accuracy, 100% for fluency, and ≈ 80% for fidelity (with the remainder "Equal" on accuracy/fidelity). In English (Questionnaire 5), preferences are even more decisive: 100% choose V2 for accuracy and fluency, and ≈ 90% for fidelity (≈ 10% "Equal"). Short comments illuminate the mechanism. For V2: "Insgesamt weisen beide Übersetzungen gravierende Qualitätsmängel auf, jedoch ist die zweite Version deutlich besser lesbar." (Overall, both translations show serious quality flaws, yet the second version is clearly more readable.) And again: "Nella versione due le frasi sono più brevi e comprensibili, la traduzione ha comunque bisogno di essere controllata e corretta." (In the second version the sentences are shorter and more understandable, yet the translation still requires checking and correction.) By contrast, unedited outputs are described as structurally heavy and non-idiomatic: "La traduzione risulta poco scorrevole, con frasi lunghe e complesse. Si arriva anche al terzo livello di subordinazione." (The translation reads awkwardly, with long and complex sentences; it even reaches a third level of subordination.) The convergence of lower scores and more positive ratings provides a consistent answer in favour of V2.

Turning to the cross-language perspective, the question is: Is this NMT quality gain similar for IT–EN (UK) and IT–DE, or does it differ by language pair? Quantitatively, the average gain is virtually identical in the two directions (≈ −25.6 in both), and the proportion under threshold increases from ≈ 10% to ≈ 70% in both languages. The interquartile range is smaller in German, indicating slightly tighter gains there, while the effect size remains medium-to-large for both language pairs. In short, the magnitude of the improvement is comparable across IT→EN (UK) and IT→DE.

Language	**Mean Δ (V2−V1)**	**Median Δ (V2−V1)**	**IQR**	**SD**	**Cohen's dz**	**% V2 < V1**
IT - DE	-25.6	-19.99	15.53	16.79	-1.53	100.0
IT - EN (UK)	-25.65	-20.77	24.14	21.81	-1.18	100.0

Table 7: Cross-language numeric comparison: mean and median Δ (V2−V1), IQR, SD, Cohen's dz, and percentage of participants with V2 < V1

Perceptions from the cross-language block in Questionnaire 5 offer an additional angle. When raters are asked which language performs better overall, responses favour English: roughly 70% select English as more accurate (≈ 30% "Equal"), 80% choose English as more fluent (≈ 20% "Equal"), and 60% select English as more faithful in style (≈ 40% "Equal"); no respondent chooses German on these items. This asymmetry in perception co-exists with the objective parity of the EPT gains and is plausibly explained by the linguistic profile of the source excerpt: dense hypotaxis, long embedded structures, and culture-bound references (e.g., realia such as "graffiti" in a historical-ethnographic sense) can make idiomatic fluency and stylistic naturalness harder to achieve in German, even when the overall error burden has been substantially reduced.

Dimension	**English (%)**	**Equal (%)**	**German (%)**
Accuracy	70.0	30.0	0.0
Fluency	80.0	20.0	0.0
Fidelity	60.0	40.0	0.0

Table 8: Cross-language preferences from Questionnaire 5: proportions selecting English, Equal, or German for accuracy, fluency, and fidelity

The cross-language items indicate a perceived advantage for English across the three dimensions; a non-trivial share of "Equal" remains, and no respondent selects German as globally superior on these items. The distribution is reported in Table 8. This asymmetry in perception co-exists with the objective parity of the EPT gains and is plausibly explained by the source's profile (dense hypotaxis and culture-bound references), which makes idiomatic fluency and stylistic naturalness more demanding in German even after source-side editing.

6.6 Sintesi e takeaway

Answer to RQ1 — Do NMT outputs from a source edited in two steps—(i) pre-editing and (ii) post-editing of the source text—outperform outputs from the unedited source?

Overall, the evidence points to a positive effect. In both directions (IT→DE and IT→EN-UK), the within-participant comparison between V1 and V2 shows that the difference in the quality score—that is, Matecat's "quality score" expressed as errors per thousand words (EPT)—is negative for every reviewer, meaning that V2 consistently scores lower (better) than V1. On average, the reduction in the quality score is about −25.6 points, and the median difference is also negative, which suggests the result is not driven by a

handful of outliers. At the same time, the share of translations below the acceptance threshold (8 per 1,000 words) rises markedly, from roughly ~10% in V1 to ~70% in V2. The severity profile moves in the same direction: Major errors per 1,000 words decrease in both language pairs and, on average, Minor errors per 1,000 words also decline (cf. Appendix A). In practical terms, V2 crosses the threshold much more often because serious errors become less frequent and lighter errors are reduced as well. This pattern is exactly what one would expect given the two source-side interventions: stabilising segmentation, easing hypotaxis, making logical relations explicit, and harmonising terminology all lower the uncertainty the NMT system has to resolve during analysis and generation. The qualitative data are consistent with this: across the comparative questionnaires, V2 is preferred over V1 for accuracy, fluency, and fidelity in both languages (see § 6.5 and Tables 6–8).

Answer to RQ2 — Is this NMT quality gain similar for IT-EN (UK) and IT-DE, or does it differ by language pair? On the objective side, the improvement appears comparable. The mean difference V2−V1 is essentially the same in the two directions (≈ −25.6), and the share below threshold increases in parallel from ~10% to ~70% for both IT→DE and IT→EN-UK. To summarise the magnitude of the within-participant change with a single standardised number, this section reports Cohen's dz—the mean difference (V2−V1) scaled by the standard deviation of those differences. In the data, dz falls in the medium-to-large range for both language pairs. The negative sign reflects direction only: it indicates that scores in V2 are lower (and therefore better) than in V1. For brevity, the term gain denotes each participant's difference Δ = V2−V1 in the quality score, so a negative gain means an improvement. A small additional point concerns how spread out these gains are: in German, the interquartile range (IQR) of Δ is smaller, which suggests that improvements were more consistent across participants in IT→DE than in IT→EN. On the perceptual side, the cross-language block of Questionnaire 5 shows a preference for English: a majority selects EN as more accurate, more fluent, and more faithful, with a non-trivial share of "Equal" responses and no explicit preference for DE (see Table 8). This asymmetry in perceived fluency and style does not contradict the broadly similar objective gains; it is plausibly linked to the typological profile of the source (dense hypotaxis, nominal chains, culture-specific realia), which makes idiomatic rendering more demanding in German even when the overall error load goes down.

Four points summarise the main effects: (i) Δ(V2−V1) in the quality score is clearly negative for all participants in both language pairs; (ii) the percentage below threshold more than doubles (~10% → ~70%); (iii) major errors/1000 decreases and, on average, minor

errors/1000 decreases as well—precisely the components that drive the quality score; (iv) the within-subject effect (Cohen's dz) is substantial in both directions, and the gains are somewhat more compact in DE (smaller IQR). Taken together, these indicators support the view that the combination of pre-editing and post-editing of the source text yields a real and reproducible advantage in this design (see Tables 6–7).

Responses to the questionnaires contribute to the interpretation of the quantitative findings. For V1, participants mention long, convoluted sentences, calques, and non-idiomatic phrasing—issues that align with the Style and Translation Errors categories used in annotation. For V2, they emphasise readability, shorter sentences, and greater coherence, while reasonably noting that MT output—even improved—still requires professional checking and correction. Comments on realia and terminology clarify why Minor (and occasional Major) issues may persist even when the EPT is low.

Across metrics and perceptions, the same picture emerges: editing the original text—first through pre-editing (Writing for MT) and then through post-editing of the source text—reduces uncertainty for the NMT system, stabilises segmentation, clarifies logical relations, and limits lexical and structural ambiguity. As a result, major and minor errors decrease, the quality score drops, and reviewers experience more fluent, easier-to-read texts. The preference for English in the perception data, despite similar objective gains, likely reflects different idiomatic expectations: German tends to be less tolerant of residual hypotaxis and certain rhythm choices, whereas English—especially in informative-descriptive writing—more readily accommodates prose that has been flattened and clarified at the source.

The within-subject design is a strength because each participant serves as their own control; even so, the task and survey order was fixed, so sequence effects cannot be ruled out. Time/effort measures were not collected (by design), and the severity/categorisation scheme—while MQM-inspired—was adapted for operational clarity. Language proficiency at C1 level was self-reported. The workflow also involved a substantial operational load: 40 projects (4 per each of the 10 participants), with segment-by-segment copy-paste and strict labelling (e.g., X_V1_EN, X_V2_DE). This repetitive setup is time-consuming and potentially prone to transcription or allocation errors, and it required sustained attention to avoid mistakes. No anomalies emerged in the final dataset.

The sample is focused: one source text (ethno-archaeological content with substantial realia), one NMT engine (DeepL, free version, no glossaries), two target languages, and N = 10 reviewers. Generalisability should therefore be understood as targeted to similar informative-descriptive texts with complex syntax and technical terminology. For texts with

dense hypotaxis, technical/cultural lexis, and local references, deploying pre-editing and post-editing of the source text before NMT appears defensible: it increases robustness of the draft and it may reduce the need for heavy post-editing interventions. For IT→DE, particular attention to nominalisations, relative chains, and explicit cohesion cues is advisable; for IT→EN, monitoring cohesion and avoiding unnecessary calques is useful. In operational terms, future replications in professional or teaching contexts would benefit from automating segment import (e.g., via API/structured CSV) and counterbalancing task order; if cost–benefit is of interest, it would be sensible to plan time-on-task and, where possible, an effort measure.

Session duration was outside the scope of this analysis and was not instrumented. Informally, several participants reported ~45 minutes to complete revisions and surveys, while others mentioned over 2 hours. As these are not systematically collected data, they are not included in the statistical analysis and should be read only as practical context. Without anticipating the concluding chapter, three near-term steps suggest themselves: (i) replicate with more texts and genres; (ii) counterbalance order and record time/effort; (iii) estimate inter-annotator agreement and compare multiple NMT systems and/or controlled glossaries. A broader roadmap will be outlined in the Conclusions.

7. Discussion and Conclusion

This thesis set out to examine whether two source-side interventions—pre-editing (in the sense of Writing for Machine Translation) and post-editing of the source text—can increase the translatability of the input and thereby improve neural machine translation (NMT) quality, and whether any improvement is comparable in IT→EN (UK) and IT→DE. Using a single Italian passage of high syntactic and cultural complexity, two versions of the source were prepared (V1 unedited; V2 edited in two steps), translated with the same NMT engine, and reviewed within participants in Matecat using a standardised error scheme. The results indicate a consistent reduction in error load for V2 relative to V1, alongside more favourable perceptions of accuracy, fluency, and fidelity. What follows discusses these findings in relation to prior work, outlines practical implications, and sketches directions for future research.

The observed pattern—lower error severity and higher perceived readability after source-side editing—aligns closely with the pragmatic logic of Writing for MT and with the idea of post-editing of the source text as theorised by Somers (1997). The pre-editing step—guided here by the Translation Centre's compendium (2021) and by Hardin et al. (2020)—aims to stabilise segmentation and punctuation, reduce ambiguity in lexis and syntax, and enforce domain-appropriate terminology and style. Somers' proposal complements this by allowing an additional, reactive pass on the source after observing systematic MT weaknesses, adjusting sentence shaping and making referential and logical links explicit where needed. In combination, these steps reduce the number of "risky" decisions the engine must make, which is a plausible mechanism for the measured drops in EPT and in Major/Minor errors. The fact that improvements are of similar magnitude for IT→EN and IT→DE, while English attracts somewhat higher perceived fluency and style, is consistent with typological expectations: dense hypotaxis and culture-bound nominals tend to be less easily accommodated in German prose, even when objective error counts fall by a comparable amount.

For informative–descriptive texts characterised by long, embedded sentences, culture-specific references, and technical vocabulary, two-step source editing before NMT appears to be a defensible strategy. In IT→DE, attention to nominalisations, relative chains, and explicit cohesion cues is advisable; in IT→EN, monitoring cohesion and avoiding unnecessary calques is often sufficient to achieve idiomaticity. From an operational perspective, the workflow used here is implementable with standard tools, though scaling it

would benefit from automating segment import and counterbalancing task order in evaluation settings.

7.1 Future work and limitations

Three immediate extensions follow directly from the present design:

1. Broader coverage. Replicate with more texts and genres, include additional NMT systems and controlled glossaries, and estimate inter-annotator agreement to gauge robustness across content types and reviewer populations.
2. Design refinements. Counterbalance the order of tasks and surveys and instrument time-on-task and effort (keystroke logging, lightweight workload ratings) to complement quality gains with cost indicators.
3. Ablation tests. Isolate the contribution of each source-side step—pre-editing only, post-editing of the source text only, and both—to quantify their separate and combined effects under otherwise identical conditions.

Given the shift toward neural systems, APE has itself evolved, with notable improvements under neural approaches; nevertheless, Carmo et al. (2021) caution that APE is not universally beneficial: it may over-correct or fail to address key errors, and its use is warranted only where it delivers measurable gains over raw MT. In a similar vein, Escribe & Mitkov (2021:168) note that APE outputs can demand additional checking to catch overcorrections and residual issues, potentially increasing effort relative to post-editing alone. In the context of the present findings, APE could be explored as a downstream complement to source-side editing—e.g., a lightweight neural APE layer constrained by the same error taxonomy—to test whether residual Minor issues can be reduced without harming adequacy.

A second line concerns interactive post-editing. Interactive MT systems collect and reuse corrections in real time, often via prefix-based autocompletion where the engine predicts continuations for a user-validated prefix (Escribe & Mitkov, 2021). Early studies reported mixed productivity effects (e.g., Langlais et al., 2002), but more recent work suggests that adaptive interactive setups can reduce technical effort and support faster workflows. Building on the present design, a natural extension would be to compare static (generate-then-edit) and interactive post-editing on the same V1/V2 inputs, using the same reviewer pool and error scheme, and measuring not only EPT but also keystrokes, time, and a light cognitive-effort proxy.

A complementary extension concerns effort, understood in the sense discussed for post-editing: translators often report a non-trivial mental burden when comparing source and

MT output (Béchara et al., 2021), and the literature distinguishes cognitive, technical, and temporal components of that burden (Koponen, 2012). Task configuration also matters: bilingual post-editing supports adequacy checks but typically increases cognitive load relative to simpler setups (Mitchell et al., 2013; Nitzke, 2016). Against this backdrop, a useful next step would be to test whether the two-step editing explored in this work (pre-editing plus post-editing of the source text) shifts or reduces the overall effort also when applied to post-editing, and whether interactive post-editing changes that picture.

The present conclusions are deliberately modest. They are grounded in a single text of a specific genre, one NMT engine (no glossaries), two target languages, and a small but controlled rater group, with a fixed task order and no time/effort instrumentation. Within this scope, the evidence is consistent: two-step source editing is associated with lower error rates and better perceived readability. Generalisation beyond similar informative–descriptive texts should proceed cautiously and through replication.

Taken together, the findings suggest that targeted editing on the source side—combining Writing for MT principles with post-editing of the source text—can make complex prose more tractable for NMT and yield measurable and perceived gains. The approach merits confirmation at scale and, given current trends, invites integration with automatic post-editing, interactive post-editing, and effort-aware decision support. Pursuing these lines—while extending the design across texts, engines, and languages—would help establish when and how source-side interventions most effectively translate into downstream quality.

8. Bibliography

AlOtaibi, Asma (2020). Statistical MT Training for the translation of English-Arabic UN Resolutions. *Arab World English Journal*,(239), 1–102. DOI: 10.24093/awej/th.239

Angelone, Erik; Ehrensberger-Dow, Maureen & Massey, Gary (2019). A-Z key terms and concepts. In: Angelone, Erik; Ehrensberger-Dow, Maureen & Massey, Gary (Hg.) *The Bloomsbury Companion to Language Industry Studies*. Vereinigtes Königreich: Bloomsbury Companions, Bloomsbury Academic.

Automatic Language Processing Advisory Committee (ALPAC). (1966). *Language and machines: Computers in translation and linguistics.* Washington, D.C.: National Academy of Sciences, National Research Council.

Bahdanau, Dzmitry; Cho, Kyunghyun & Bengio, Yoshua (2014). Neural Machine Translation by Jointly Learning to Align and Translate. https://doi.org/10.48550/arxiv.1409.0473.

Bar-Hillel, Yehoshua (1951). The Present State of Research on Mechanical Translation. *American Documentation*, 2(4), 229–237. https://doi.org/10.1002/asi.5090020408

Baur, Nina & Blasius, Jörg (2014). *Handbuch Methoden der empirischen Sozialforschung*. Wiesbaden: Springer.

Bawden, Rachel; Rico Sennrich; Alexandra Birch & Barry Haddow (2018). Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, 1304–1313. DOI: 10.18653/v1/n18-1118.

Béchara, Hannah; Orăsan, Constantin; Parra Escartín, Carla; Zampieri, Marcos & William Lowe (2021). The Role of Machine Translation Quality Estimation in the Post-Editing Workflow. *Informatics (Basel)*, 8(3), 61. https://doi.org/10.3390/informatics8030061.

Bentivogli, Luca; Bisazza Arianna; Cettolo Marco & Federico Marcello (2016). Neural versus phrase-based machine translation quality: a case study. arXiv preprint arXiv:1608.04631.

Bernth Arendse & Claudia Gdaniec (2001). MTranslatability. *Machine translation,* 16, 175-218.

Ortiz Boix, Carla (2016). Machine translation and post-editing in wildlife documentaries: Challenges and possible solutions. *Hermeneus* (18), 269-313.

Bouillon, Pierrette; Gaspar, Liliana; Gerlach, Johanna; Porro, Victoria & Johann Roturier (2014). Pre-editing by forum users: a case study. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland.

Bowker, Lynne; Fisher, Desmond; van Doorslaer, Luc & Gambier Yves. Computer-Aided Translation (2010). In *Handbook of Translation Studies*, 1:60–65. The Netherlands: John Benjamins Publishing Company. https://doi.org/10.1075/hts.1.comp2.

Bowker, Lynne & Jairo Buitrago, Ciro (2015). Investigating the Usefulness of Machine Translation for Newcomers at the Public Library. *Translation and Interpreting Studies,* 10(2), 165–86. https://doi.org/10.1075/tis.10.2.01bow.

Castilho, Sheila; Moorkens, Joss; Gaspari, Federico; Calixto, Iacer; Tinsley, John & Andy Way (2017). Is Neural Machine Translation the New State of the Art? *Prague Bulletin of Mathematical Linguistics,* 108(1),109–120. https://doi.org/10.1515/pralin-2017-0013.

Chan, Sin-wai. (2023). The Development of Translation Technology: 1967-2023. In *Routledge Encyclopedia of Translation Technology,* 2nd ed., 3–41. Routledge. https://doi.org/10.4324/9781003168348-2.

Cho, Kyunghyun; Van Merriënboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger & Yoshua Bengio (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv: 1406.1078.

CLAW (1996). *Proceedings of The First International Workshop on Controlled Language Applications* (CLAW 96). Leuven: Centre for Computational Linguistics, Katholieke Universiteit.

CLAW (2000). *Proceedings of the Third International Workshop on Controlled Language Applications*.

Correa, Maite (2014). Leaving the "peer" out of peer-editing: Online translators as a pedagogical tool in the Spanish as a second language classroom. *Latin American Journal of Content & Language Integrated Learning,* 7(1), 1-20. https://doi.org/10.5294/3568

Craciunescu, Olivia; Gerding-Salas, Constanza & Susan Stringer-O'Keeffe (2004). Machine translation and computer-assisted translation.

Daems, Joke; Macken, Lieve & Vandepitte, Sonia. (2017). Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality

assessment based on the translation process. *Target. International Journal of Translation Studies*, 29(2), 241–262. https://doi.org/10.1075/target.29.2.03dae

do Carmo, Félix; Dimitar, Shterionov; Moorkens, Joss; Wagner, Joachim; Hossari, Murhaf; Paquin, Eric; Schmidtke, Dag; Groves, Declan & Andy Way. A Review of the State-of-the-Art in Automatic Post-Editing (2021). *Machine Translation,* 35(2), 101–43. https://doi.org/10.1007/s10590-020-09252-y.

Doherty, Stephen (2016). The Impact of Translation Technologies on the Process and Product of Translation. *International Journal of Communication* (Online), 947.

Elliston, John S.G. (1979). Computer Aided Translation: A Business Viewpoint. In B. M. Snell (ed) *Translating and the Computer*, Amsterdam: NorthHolland, 149-58.

Escribe, Marie, and Mitkov Ruslan (2021). Interactive models for post-editing. In *Proceedings of the Translation and Interpreting Technology Online Conference*, 167-173.

European Language Industry Survey (ELIS) (2025). *ELIS Report 2025: Trends and expectations in the language industry.* https://elis-survey.org/

Garcia, Ignacio. (2015). Cloud marketplaces: Procurement of translators in the age of social media. In Minako O'Hagan (ed.), *The Routledge Handbook of Translation and Technology*, 325–339. London: Routledge.

Garcia, Ignacio & Chan, Sin-wai (2023). Computer-Aided Translation Systems. In *Routledge Encyclopedia of Translation Technology*, 2nd ed., 76–94. Routledge. https://doi.org/10.4324/9781003168348-4.

Gaspari, Federico, Toral, Antonio, Naskar, Sudip Kumar, Groves, Declan, & Way, Andy. (2015). Perception vs. reality: Measuring machine translation post-editing productivity. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation* (EAMT 2015) , 151–158.

Goldfeld, Ziv & Polyanskiy, Yury (2020). The Information Bottleneck Problem and Its Applications in Machine Learning. *IEEE Journal on Selected Areas in Information Theory*, 1(1), 19–38. https://doi.org/10.1109/JSAIT.2020.2991561.

Guerberof Arenas, Ana. (2019). Pre-editing and post-editing. In: Angelone, Erik; Ehrensberger-Dow, Maureen & Massey, Gary (Hg.) *The Bloomsbury Companion to Language Industry Studies*, 1-42.

Hassan, Hany; Aue, Anthony; Chen, Chang; Chowdhary, Vishal; Clark, Jonathan; Federmann, Christian; Huang, Xuedong; Junczys-Dowmunt, Marcin; Lewis, Williams; 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. 1-25. DOI: 10.48550/arXiv.1803.05567.

Hutchins, John W. (1986). Machine translation: past, present, future. Chichester (UK): Ellis Horwood.

Hutchins, John W. (2003). The development and use of machine translation systems and computer-based translation tools. Bahri.

Hutchins, John W. (2006). Machine translation: a concise history.

Ive, Julia, Max, Aurélien & Yvon François (2018). Reassessing the Proper Place of Man and Machine in Translation: A Pre-Translation Scenario. *Machine Translation,* 32(4), 279–308. https://doi.org/10.1007/s10590-018-9223-9.

Jia, Yuening (2018). Attention Mechanism in Machine Translation. *Journal of Physics. Conference Series,* 1314(1)*,* 12186. https://doi.org/10.1088/1742-6596/1314/1/012186.

Kamprath, Christine; Adolphson, Eric; Mitamura, Teruko & Nyberg Eric (1998). Controlled language for multilingual document production: Experience with Caterpillar technical English. In *Proceedings of the Second International Workshop on Controlled Language Applications* (Vol. 146).

Kay, Martin (1997). The Proper Place of Men and Machines in Language Translation. *Machine Translation,* 12(1-2), 3–23. https://doi.org/10.1023/a:1007911416676.

Knight, Kevin, and Chander, Ishwar (1994). Automated Postediting of Documents. https://doi.org/10.48550/arxiv.cmp-lg/9407028.

Kocmi, Tom & Ondřej, Bojar (2017), An Exploration of Word Embedding Initialization in Deep-Learning Tasks. https://doi.org/10.48550/arxiv.1711.09160.

Koponen, Maarit (2012). Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the seventh workshop on statistical machine translation*, 181-190.

Langlais, Philippe & Lapalme, Guy (2002). Trans type: Development-evaluation cycles to boost translator's productivity. *Machine Translation* 17.2, 77-98.

Läubli, Samuel; Sennrich, Rico & Volk, Martin (2018). Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4791–4796.

Le, Quoc V., & Schuster, Mike. (2016). A neural network for machine translation, at production scale. *Google Research Blog.* https://research.googleblog.com/2016/09/a-neural-network-for-machine.html

Lehrndorfer, Anne (1996). Kontrolliertes Deutsch : linguistische und sprachpsychologische Leitlinien für eine (maschinell) kontrollierte Sprache in der technischen Dokumentation. Tübingen: Narr.

Lopes, António, Amin Farajian, M.; Bawden, Rachel; Zhang, Michael & F. T. Martins, André (2020). Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 225–234, Lisboa, Portugal, November. European Association for Machine Translation.

Melby, Alan K; Ellen Wright, Sue & Chan, Sin-wai (2023). Translation Memory. In *Routledge Encyclopedia of Translation Technology*, 2nd ed., 766–81. https://doi.org/10.4324/9781003168348-48.

Mikolov, Tomas; Chen, Kai; Corrado, Greg & Jeffrey Dean (2013). Efficient Estimation of Word Representations in Vector Space, in *Proceedings of a Workshop at the International Conference on Learning Representations (ICLR-13)*, Scottsdale.

Mitkov, Ruslan; Spiessens, Anneleen & Deane-Cox, Sharon (2022). Translation Memory Systems. *The Routledge Handbook of Translation and Memory*, 1st ed., 1:364–80. Routledge. https://doi.org/10.4324/9781003273417-27.

Mitchell, Linda; Roturier, Johann & Sharon O'Brien (2013). Community-based post-editing of machine-translated content: monolingual vs. bilingual. In *Proceedings of the 2nd Workshop on Post-editing Technology and Practice*.

Mohamed, Yasir Abdelgadir; Khanan, Akbar; Bashir, Mohamed; Hakim H. M Mohamed, Abdul; A. E Adiel, Mousab & A Elsadig, Muawia (2024). The Impact of Artificial Intelligence on Language Translation: A Review. *IEEE Access,* 12, 118923–118923. https://doi.org/10.1109/ACCESS.2024.3448788.

Nimdzi Insights. (2025). The Nimdzi 100: The ranking of top LSPs in 2025. Retrieved from https://www.nimdzi.com/nimdzi-100-2025

Niño, Ana (2009). Machine translation in foreign language learning: Language learners and tutors perceptions of its advantages and disadvantages. ReCALL, 21(2), 241-258. https://doi. org/10.1017/S0958344009000172

Nitzke, Jean, & Oster, Katharina (2016). Comparing translation and post-editing: An annotation schema for activity units. In *New directions in empirical translation process research: Exploring the CRITT TPR-DB*, 293-308. Springer International Publishing.

Peris, Álvaro, & Casacuberta, Francisco (2019). Online Learning for Effort Reduction in Interactive Neural Machine Translation. *Computer Speech & Language,* 58, 98–126. https://doi.org/10.1016/j.csl.2019.04.001.

Poibeau, Thierry (2017) *Machine translation*. Cambridge, Massachusetts: The MIT Press.

Pym, Peter (1988). Pre-editing and the Use of Simplified Writing for MT: An Engineer's Experience of Operating an MT System. In Pamela Mayorcas (ed) *Translating and the Computer 10: The Translation Environment 10 Years On*, London: Aslib, 80-95.

Post, Matt & Junczys-Dowmunt, Marcin (2023). Escaping the sentence-level paradigm in machine translation. arXiv preprint arXiv:2304.12959v1.

Reifler, Eugene (1952). The First Conference on Mechanical Translation. In *Proceedings of the First MT Conference*, Massachusetts Institute of Technology, Cambridge, Mass., 17-20.

Reuther, Ursula (2003). Two in one-can it work? Readability and translatability by means of controlled language. In *Proceedings of the 4th International Workshop on Controlled Language Applications*, Dublin, Ireland.

Hardin, Ashley R.; Ito, Junko; Sasaki, Aiko; Pigg, Stacey; Hocutt, Daniel & Josephine Walwema (2020). Writing for Human and Machine Translation: Best Practices for Technical Writers. In *Proceedings of the 38th ACM International Conference on Design of Communication*, 1–3. New York, NY, USA: ACM. https://doi.org/10.1145/3380851.3416741.

Spyridakis, Jan H.; Holmback, Heather & Serena K. Shubert (1997). Measuring the Translatability of Simplified English in Procedural Documents. In *IEEE Transactions on Professional Communication* 40(1), 4–12. https://doi.org/10.1109/47.557512.

Somers, Harold (1997). A Practical Approach to Using Machine Translation Software: "Post-editing" the Source Text', *Translator*, 3(2), 193–212. https://doi.org/10.1080/13556509.1997.10798998.

Somers, Harold (2003). Translation Memory Systems. In *Benjamins Translation Library*, 35:31–47. The Netherlands: John Benjamins Publishing Company. https://doi.org/10.1075/btl.35.06som.

Sutskever, Ilya; Vinyals, Oriol & Quoc V. Le (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems;* 27.

Toral, Antonio; Castilho, Sheila; Hu, Ke & Way,Andy (2018). Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. https://doi.org/10.48550/arxiv.1808.10432.

Toral, Antonio, & M Sánchez-Cartagena, Víctor (2017). A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. https://doi.org/10.48550/arxiv.1701.02901.

Toral, Antonio; Wieling, Martijn & Way,Andy (2018). Post-Editing Effort of a Novel With Statistical and Neural Machine Translation. In *Frontiers in Digital Humanities 5*. https://doi.org/10.3389/fdigh.2018.00009.

Tukey, John W. (1977). Exploratory Data Analysis. Addison-Wesley.

Ustaszewski, Michael (2019). Exploring Adequacy Errors in Neural Machine Translation with the Help of Cross-Language Aligned Word Embeddings. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, 122-128. 2019.

Vardaro, Jennifer; Schaeffer, Moritz & Hansen-Schirra,Silvia (2019). Translation Quality and Error Recognition in *Professional Neural Machine Translation Post-Editing*. Informatics (Basel) 6(3), 41. https://doi.org/10.3390/informatics6030041.

Vasconcellos, Muriel (1987). Post -editing On-screen: Machine Translation from Spanish into English. In Catriona Picken (ed) *Translating and the Computer 8: A Profession on the Move*, London: Aslib, 133-46.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukasz & Polosukhin, Illia (2017). Attention is all you need. In *Advances in neural information processing systems,* 5998-6008.

Vela, Mihaela, Pal; Santanu Zampieri Marcos; Kumar Naskar, Sudip & van Genabith, Josef (2019). Improving CAT Tools in the Translation Workflow: New Approaches and Evaluation. https://doi.org/10.48550/arxiv.1908.06140.

Vieira, Lucas Nunes (2014). Indices of Cognitive Effort in Machine Translation Post-Editing. *Machine Translation,* 28(3-4), 187–216. https://doi.org/10.1007/s10590-014-9156-x.

Voita, Elena; Sennrich, Rico & Titov, Ivan (2019). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1198–1212, Association for Computational Linguistics.

Vukalović, Nino (2021). An Analysis of Computer-Assisted Translation (CAT) Tools (Doctoral dissertation, University of Rijeka. Faculty of Humanities and Social Sciences. Department of English Language and Literature).

Wagner, Elizabeth (1985). Rapid Post-editing ofSystran. In Veronica Lawson (ed) *Tools for the Trade: Translating and the Computer* 5, London: Aslib, 199--213.

Way, Andy (2020). Machine translation: Where are we at today. In *The Bloomsbury companion to language industry studies*: 311-332.

Winther Balling, Laura, O'Brien, Sharon; Winther Balling, Laura; Carl, Michael; Simard, Michel & Specia, Lucia (2014). Post-Editing of Machine Translation: Processes and Applications. Newcastle upon Tyne : Cambridge Scholars Publishing.

Wu, Yonghui; Schuster, Mike; Chen, Zhifeng; Le, Quoc V.; Norouzi, Mohammad; Macherey, Wolfgang; Krikun, Maxim; Cao, Yuan; Gao, Qin; Macherey, Klaus; Klingner, Jeff; Shah, Apurva; Johnson, Melvin; Liu, Xiaobing; Kaiser, Łukasz; Gouws, Stephan; Kato, Yoshikiyo; Kudo, Taku; Kazawa, Hideto; Stevens, Keith; Kurian, George; Patil, Nishant; Wang, Wei; Young, Cliff; Smith, Jason; Riesa, Jason; Rudnick, Alex; Vinyals, Oriol; Corrado, Greg; Hughes, Macduff; Dean, Jeffrey (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. Computing Research Repository, arXiv:1609.08144.

Youdale, Roy, Rothwell, Andrew; Spiessens, Anneleen & Deane-Cox, Sharon (2022). Computer-Assisted Translation (CAT) Tools, Translation Memory, and Literary Translation. In *The Routledge Handbook of Translation and Memory*, 1:381–402. Routledge. https://doi.org/10.4324/9781003273417-28.

Zhang, Biao, Xiong, Deyi & Su, Jinsong (2020). Neural Machine Translation with Deep Attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 42(1): 154–63. https://doi.org/10.1109/TPAMI.2018.2876404.

Zhang, Jiacheng; Luan, Huanbo; Sun, Maosong; Zhai, Feifei; Xu, Jingfang; Zhang, Min & Liu, Yang (2018). Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 533–542, Brussels, Belgium, October-November. Association for Computational Linguistics.

Zhang, Xue-yao (2025). A Review of Research on Pre-editing of Machine Translation. *Journal of Literature and Art Studies*, 15(1), 36-43.

Appendix

Appendix A

Four tables with data collected from Matecat Quality Reports. Divided by participant, quality score and type of error. Table A1 looks at the data considering Version 1, A2 considers Version 2, while A3 and A4 evaluate errors from a linguistic perspective, German (DE) and English (UK) respectively.

		VERSION 1											
		Quality score	style			translation errors			terminology consistency			language quality		
			N	m	M	N	m	M	N	m	M	N	m	M
A	DE	20		3	2		2							
	EN	27.69		4	3								2	
B	DE	36.92		5	1		3	2						1
	EN	24.62		4	2			1						
C	DE	38.46		1	4	1		1						1
	EN	41.54			4		3	2						
D	DE	20			3	1	1							
	EN	13.85		3									2	1
E	DE	56.63		4	2		3	2					4	2
	EN	89.23		1	4		1	3			4	1		3
F	DE	4.62	3	1			2							
	EN	33.85	1	3	2		2	2					1	
G	DE	15.38	3	3		2	3	1	1			3		
	EN	9.23	5	1		2	1	1	1					
H	DE	64.62	1	3	6		2	2					1	1
	EN	4.62	1	3										
I	DE	98.46		2	5		2	5		1	2		3	2
	EN	93.85			7			6					1	2
J	DE	20	1	1	1		1						3	1
	EN	15.38	1	4	1		2							

A.1

		VERSION 2												
		Quality score	style			translation errors			terminology consistency			language quality		
			N	m	M	N	m	M	N	m	M	N	m	M
A	DE	1.45	3	1										
	EN	0												
B	DE	11.63					4							1
	EN	14.53		5	1		1							
C	DE	5.81	1	2										
	EN	1.45	1				1							
D	DE	0	4											
	EN	0												
E	DE	2.91					1			1				
	EN	13.08				1		2				6	1	
F	DE	4.36		1									2	
	EN	0							3					
G	DE	1.45	11				1		9			1		
	EN	0	2			2			11					
H	DE	10.17		7										
	EN	0												
I	DE	78.49		15			7	4		2			6	2
	EN	61.05		8	4		2	2		1			3	1
J	DE	2.91	3				1						1	
	EN	7.27	1										1	1

A.2

		GERMAN												
		Quality score	style			translation errors			terminology consistency			language quality		
			N	m	M	N	m	M	N	m	M	N	m	M
A	V1	20		3	2		2							
	V2	1.45	3	1										
B	V1	36.92		5	1		3	2						1
	V2	11.63					4							1
C	V1	38.46		1	4	1		1						1
	V2	5.81	1	2										
D	V1	20			3	1	1							
	V2	0	4											
E	V1	56.63		4	2		3	2					4	2
	V2	2.91					1			1				
F	V1	4.62	3	1			2							
	V2	4.36		1									2	
G	V1	15.38	3	3		2	3	1	1			3		
	V2	1.45	11				1		9			1		
H	V1	64.62	1	3	6		2	2					1	1
	V2	10.17		7										
I	V1	98.46		2	5		2	5		1	2		3	2
	V2	78.49		15			7	4		2			6	2
J	V1	20	1	1	1		1						3	1
	V2	2.91	3				1						1	

A.3

		ENGLISH												
		Quality score	style			translation errors			terminology consistency			language quality		
			N	m	M	N	m	M	N	m	M	N	m	M
A	V1	27.69		4	3								2	
	V2	0												
B	V1	24.62		4	2			1						
	V2	14.53		5	1		1							
C	V1	41.54			4		3	2						
	V2	1.45	1				1							
D	V1	13.85		3									2	1
	V2	0												
E	V1	89.23		1	4		1	3			4	1		3
	V2	13.08				1		2				6	1	
F	V1	33.85	1	3	2		2	2					1	
	V2	0							3					
G	V1	9.23	5	1		2	1	1	1					
	V2	0	2			2			11					
H	V1	4.62	1	3										
	V2	0												
I	V1	93.85			7			6					1	2
	V2	61.05		8	4		2	2		1			3	1
J	V1	15.38	1	4	1		2							
	V2	7.27	1										1	1

A.4

Appendix B: Participant Instructions

Purpose and flow

You will review **four** machine-translated texts in **Matecat** and complete **five** short Google Forms. Please follow this exact order:

1. Complete the **Background** questionnaire (link in the email).

2. Review **X_V1_DE** (German translation from Version 1).
   Then complete **"Version 1 – German Translation"** (post-task form).

3. Review **X_V2_DE** (German translation from Version 2).
   Then complete **"Version 2 – German Translation – Translations compared"**.

4. Review **X_V1_EN** (English translation from Version 1).
   Then complete **"Version 1 – English Translation"**.

5. Review **X_V2_EN** (English translation from Version 2).
   Then complete **"Version 2 – English Translation – translation and languages compared"**.

Access and project names

You will receive **four private links** to Matecat projects labelled **X_V1_DE, X_V2_DE, X_V1_EN, X_V2_EN** (your letter **X = A–J** identifies your set). You will only see your own four projects. Please keep the order above.

How to review in Matecat

- Open the project from the link and switch to **Revise** mode (menu next to the settings ⚙ icon → **Revise**).

- Work **segment by segment**. To flag an error, click the **red exclamation-mark** icon, choose an **error category**, and add a **comment** if clarification helps (click the speech-bubble icon, type, and press Enter to save).

- Assign a **severity** level to each error (see the severity scale below).

- When all segments have been reviewed, the project will show as **Complete**.

Error categories (choose one per error)

- **Style** — register, textual cohesion, readability.

- **Translation Errors** — adequacy, mistranslation, additions/omissions, idioms.

- **Terminology Consistency** — correct and consistent use of terms across segments.

- **Language Quality** — spelling, grammar, syntax, punctuation.

Severity scale and weighting (this affects the score)

- **Neutral = 0 points** — note a very minor issue or preference; does **not** affect the quality score.

- **Minor = 0.5 points** — issue that does not significantly affect meaning but harms clarity, style, or local correctness.

- **Major = 2 points** — error that materially distorts meaning, causes wrong reference/attachment, or breaks grammatical well-formedness at clause level.

Matecat's **quality score** is computed from the **sum of error points** (Minor and Major; Neutral adds 0), scaled to **1,000 words** and divided by the number of **reviewed words** in the project. Please keep these weights in mind when selecting the severity for each error.

What *not* to flag
Ignore trivial formatting artefacts such as **trailing spaces** at the end of a segment; they are **out of scope** for this experiment. Keep your focus on the categories and severities above.

Questionnaires
After each review, complete the Google Form for that task. Each form contains **three five-point** judgements—**accuracy**, **fluency**, **fidelity**—with a brief descriptor next to each option to standardise interpretation (e.g., "Very accurate — the translation conveys the source meaning without noticeable errors"). In the German and English **V2** forms you will also find **comparative** items (V1 vs V2 within the same language), and in the final form a short **EN vs DE** section. Closed questions are **mandatory**;

open comment prompts are **optional**. There is **no time limit**; answers are recorded automatically on submission.

Privacy and support
Forms are **anonymous** and cannot be tied back to your Matecat identity; only aggregated results are analysed. If you encounter technical issues (login, saving, connectivity), use the contact details provided in your invitation email.

Appendix C: link to folder with the 40 Matecat Quality Reports (CVS files)

https://drive.google.com/drive/folders/1xahSVQeih4fSpWx9UVE4t12SBpYT4gAT?usp=sharing

Appendix D: link to the folder with the 5 questionnaires (CVS files)

https://drive.google.com/drive/folders/1IxcEck_zlKkYI4noKfZYlZgnNLRW6MGD?usp=sharing

Appendix E: original text, first editing, second editing, NMT German (DE) and English (UK) translation

E.1: Original Text

Almeno due, sono infatti i motivi di interesse e le ragioni dell'importanza di questi graffiti.
Il primo è la testimonianza che essi rendono di un mondo pastorale appena trapassato, della sua economia, della sua relativa precoce alfabetizzazione, della sua sfera devozionale, estetica e morale. Le scritte infatti, in un primo tempo racchiuse all'interno di ordinate edicolette ornate e iscritte con i caratteri propri dei cippi di confine in montagna, e in un secondo tempo debordanti a tutto campo con una certa dovizia di annotazioni diaristiche, contengono firme, date, conteggi del bestiame, «segni di casa» mentre il loro stesso posizionarsi sulle falesie testimonia degli incessanti andirivieni estivi dei singoli pastori, all'interno di un sistema ancora in parte vigente di diritti territoriali di pascolo contesi e sempre piuttosto complessi.
Così, a partire dalla sequenza cronologica espressa dalle date, dalla georeferenziazione delle scritte stesse e dal riconoscersi dei singoli pastori dietro alle sigle e ai «segni di casa», sarà certamente possibile utilizzare il complesso di queste scritte pastorali come uno straordinario archivio notarile per ricostruire una pagina importante della storia della pastorizia nella valle di Fiemme, e dunque della storia economica, sociale e ambientale della valle. Un lavoro che,

previo il completamento del database delle scritte, può dirsi ora appena iniziato. Secondo e non minore motivo di interesse, tuttavia, è il modo in cui questi graffiti pastorali, pur inequivocabilmente d'età moderna e quasi precontemporanea, si ricollegano direttamente, quanto a stile, ad alcuni aspetti dell'iconografia, e alla qualità specifica dell'ispirazione scrittoria, alla lunga tradizione europea dell'arte rupestre preistorica, quella che ha inizio nelle grotte del paleolitico dordonese e cantabrico e prosegue poi in età protostorica con i grandi siti graffitati del Monte Bego e della val Camonica. Infatti, al di là dell'evidente incongruità cronologica, vi è un certo numero di fattori che consentono di collocare le scritte pastorali della valle di Fiemme sul medesimo tracciato ideale della grande arte rupestre europea.

E.2: Neural machine translation of the original text

E.2.1 German (DE)

Mindestens zwei sind in der Tat die Gründe für das Interesse und die Gründe für die Bedeutung dieser Graffiti. Das erste ist das Zeugnis, das sie von einer gerade durchbrochenen pastoralen Welt, ihrer Wirtschaft, ihrer relativen frühen Alphabetisierung, ihrer hingebungsvollen, ästhetischen und moralischen Sphäre geben. In der Tat enthalten die Inschriften, die zunächst in hübsche kleine verzierte Edicolettes eingefasst und mit den für Grenzsteine in den Bergen typischen Schriftzeichen beschriftet sind und später mit einer gewissen Fülle von Tagebucheinträgen überladen sind, Unterschriften, Daten, Sie enthalten Unterschriften, Daten, Viehzählungen, „Hauszeichen", und ihre Positionierung auf den Felsen zeugt von den unablässigen sommerlichen Wanderungen der einzelnen Hirten im Rahmen eines teilweise noch bestehenden Systems umstrittener und stets recht komplexer territorialer Weiderechte. Ausgehend von der chronologischen Abfolge der Daten, der Georeferenzierung der Schriften selbst und der Erkennung einzelner Hirten hinter den Akronymen und „Hauszeichen" wird es sicherlich möglich sein, den Komplex dieser pastoralen Schriften als außergewöhnliches notarielles Archiv zu nutzen, um eine wichtige Seite der Geschichte des Hirtenwesens im Fleimstal und damit der Wirtschafts-, Sozial- und Umweltgeschichte des Tals zu rekonstruieren. Eine Arbeit, die nach der Fertigstellung der Datenbank mit den Schriften als gerade erst begonnen bezeichnet werden kann. Der zweite und nicht minder interessante Grund ist jedoch die Art und Weise, wie diese Hirtengraffiti, obwohl sie eindeutig modernen und fast vorzeitlichen Alters sind, in Bezug auf den Stil, bestimmte Aspekte der Ikonographie und die spezifische Qualität der Schreibinspiration direkt mit der langen europäischen Tradition der prähistorischen Felskunst verbunden sind, die in den Höhlen der

paläolithischen Dordogne und Kantabrien begann und dann in der protohistorischen Zeit mit den großen Graffiti-Stätten von Monte Bego und Val Camonica fortgesetzt wurde. Abgesehen von der offensichtlichen chronologischen Inkongruenz gibt es eine Reihe von Faktoren, die es ermöglichen, die pastoralen Schriften des Fleimstals auf die gleiche ideale Spur der großen europäischen Felskunst zu bringen.

E.1.2: English (UK)

In fact, at least two are the reasons of interest and the reasons for the importance of these graffiti. The first is the testimony they bear of a pastoral world that had just passed away, of its economy, its relative early literacy, its devotional, aesthetic and moral sphere. In fact, the inscriptions, at first enclosed within neat little ornate edicolettes and inscribed with the characters typical of boundary stones in the mountains, and later overflowing with a certain abundance of diary entries, contain signatures, dates, livestock counts, “house signs”, while their very positioning on the cliffs testifies to the incessant summer comings and goings of individual shepherds, within a still partly existing system of disputed and always rather complex territorial grazing rights. Thus, starting from the chronological sequence expressed by the dates, from the georeferencing of the inscriptions themselves and from the recognition of the individual shepherds behind the acronyms and “house signs”, it will certainly be possible to use the complex of these pastoral inscriptions as an extraordinary notarial archive to reconstruct an important page in the history of pastoralism in the Fiemme valley, and thus of the economic, social and environmental history of the valley. A work that, after the completion of the database of inscriptions, can now be said to have just begun. The second and no less interesting reason, however, is the way in which these pastoral graffiti, although unequivocally of modern and almost pre-contemporary age, are directly connected, in terms of style, certain aspects of iconography and the specific quality of the writing inspiration, to the long European tradition of prehistoric rock art, which began in the caves of the Palaeolithic Dordogne and Cantabrian and then continued in the protohistoric age with the great graffiti sites of Monte Bego and Val Camonica.

In fact, beyond the obvious chronological incongruity, there are a number of factors that allow the pastoral inscriptions of the Fiemme valley to be placed on the same ideal track as the great European rock art.

E.3: First editing

Almeno due, sono infatti i motivi di interesse e le ragioni dell'importanza di queste scritte.
Il primo è la testimonianza che essi rendono di un mondo pastorale appena trapassato, della sua economia, della sua relativa precoce alfabetizzazione, della sua sfera devozionale, estetica e morale. Le scritte infatti contengono firme date, conteggi del bestiame, «segni di casa».
In un primo tempo le scritte erano racchiuse all'interno di ordinate edicolette ornate e iscritte con i caratteri propri dei cippi di confine in montagna. In un secondo tempo erano debordanti a tutto campo con una certa dovizia di annotazioni diaristiche. La loro posizione sulle rocce testimonia gli incessanti andirivieni estivi dei singoli pastori. I pastori erano parte di un sistema di diritti territoriali di pascolo contesi e sempre piuttosto complessi. Questi diritti sono ancora parzialmente in vigore. Il complesso di queste scritte pastorali rappresenta uno straordinario archivio notarile per ricostruire una pagina importante della storia della pastorizia nella valle di Fiemme, e dunque della storia economica, sociale e ambientale della valle. L'archivio può essere ricostruito a partire dalla sequenza cronologica espressa dalle date, dalla georeferenziazione delle scritte stesse e dal riconoscersi dei singoli pastori dietro alle sigle e ai «segni di casa». Un lavoro che, previo il completamento del database delle scritte, può dirsi ora appena iniziato. Secondo e non minore motivo di interesse, tuttavia, è il modo in cui questi graffiti pastorali si ricollegano direttamente alla lunga tradizione europea dell'arte rupestre preistorica. L'arte rupestre che ha inizio nelle grotte del paleolitico dordonese e cantabrico e prosegue poi in età protostorica con i grandi siti graffitati del Monte Bego e della val Camonica. Le scritte si ricollegano per stile e per alcuni aspetti dell'iconografia e sono d'età moderna e quasi precontemporanea. Infatti, al di là dell'evidente incongruità cronologica, vi è un certo numero di fattori che consentono di collocare le scritte pastorali della valle di Fiemme sul medesimo tracciato ideale della grande arte rupestre europea.

E.4: Neural machine translation of the first editing

E.4.1: German (DE)

Mindestens zwei Gründe sind es, die das Interesse und die Bedeutung dieser Schriften begründen. Der erste ist das Zeugnis, das sie von einer gerade untergegangenen pastoralen Welt, ihrer Wirtschaft, ihrer relativ frühen Alphabetisierung, ihrer religiösen, ästhetischen und moralischen Sphäre geben. Die Inschriften enthalten in der Tat datierte Unterschriften, Viehzählungen, „Hauszeichen". Zunächst waren die Inschriften in saubere, verzierte Edicolettes mit den für Grenzsteine in den Bergen typischen Schriftzeichen eingeschlossen.

Später waren sie mit einer gewissen Fülle von tagebuchartigen Vermerken übersät. Ihre Lage auf den Felsen zeugt von den unablässigen sommerlichen Wanderungen der einzelnen Hirten. Die Hirten waren Teil eines Systems von umstrittenen und stets recht komplexen territorialen Weiderechten. Diese Rechte sind teilweise noch in Kraft. Der Komplex dieser Hirteninschriften stellt ein außergewöhnliches notarielles Archiv dar, mit dem sich eine wichtige Seite der Geschichte des Hirtenwesens im Fleimstal und damit der Wirtschafts-, Sozial- und Umweltgeschichte des Tals rekonstruieren lässt. Das Archiv lässt sich anhand der chronologischen Abfolge der Daten, der Georeferenzierung der Inschriften selbst und der Identifizierung der einzelnen Hirten hinter den Initialen und „Hauszeichen“ rekonstruieren.
Eine Arbeit, von der man nach der Fertigstellung der Inschriften-Datenbank sagen kann, dass sie gerade erst begonnen hat. Zweitens und nicht weniger interessant ist jedoch die Art und Weise, wie diese Hirtengraffiti direkt mit der langen europäischen Tradition der prähistorischen Felskunst zusammenhängen. Diese Felskunst begann in den Höhlen des paläolithischen Dordoniums und Kantabriens und setzte sich in der protohistorischen Periode mit den großen Graffiti-Fundstellen des Monte Bego und des Val Camonica fort. Die Inschriften sind im Stil und in bestimmten Aspekten der Ikonographie verwandt und stammen aus der Neuzeit und fast aus der Vorzeit. Abgesehen von der offensichtlichen chronologischen Inkongruenz gibt es eine Reihe von Faktoren, die es erlauben, die Hirteninschriften des Fleimstals auf die gleiche ideale Schiene zu setzen wie die große europäische Felskunst.

E.4.2: English (UK)

At least two, in fact, are the reasons for the interest and importance of these writings. The first is the testimony they bear to a pastoral world that had just passed away, its economy, its relative early literacy, its devotional, aesthetic and moral sphere. The inscriptions in fact contain dated signatures, livestock counts, “house signs”. At first, the inscriptions were enclosed within neat, ornate edicolettes and inscribed with the characters typical of boundary stones in the mountains. At a later stage, they were overflowing with a certain abundance of diary entries. Their position on the rocks testifies to the incessant summer comings and goings of individual shepherds. The shepherds were part of a system of disputed and always rather complex territorial grazing rights. These rights are still partially in force. The complex of these pastoral inscriptions represents an extraordinary notarial archive for reconstructing an important page in the history of pastoralism in the Fiemme Valley, and thus of the economic, social and environmental history of the valley. The archive can be reconstructed on the basis of the chronological sequence expressed by the dates, the georeferencing of the inscriptions

themselves and the identification of the individual shepherds behind the initials and “house signs”. Work that, after the completion of the inscription database, can now be said to have only just begun. Second and no less interesting, however, is the way in which these pastoral graffiti relate directly to the long European tradition of prehistoric rock art. This rock art began in the caves of the Palaeolithic Dordonian and Cantabrian periods and continued in the Protohistoric period with the large graffiti sites of Monte Bego and Val Camonica. The inscriptions are related in style and certain aspects of iconography and are of modern and almost pre-contemporary age. In fact, beyond the obvious chronological incongruity, there are a number of factors that allow the pastoral inscriptions of the Fiemme valley to be placed on the same ideal track as the great European rock art.

E.5: Second editing

Questi graffiti sono importanti per almeno due motivi. Il primo è la testimonianza che danno di un mondo pastorale che non esiste più: ne descrivono l'economia, la dimensione devozionale, estetica e morale e mostrano come già all’epoca ci fosse una relativa alfabetizzazione. Le scritte infatti, contengono firme, date, conteggi del bestiame e stemmi di famiglia. Si trovano in punti in cui i pastori passavano spesso e sono testimoni di un sistema di complessi diritti di pascolo, che sono ancora oggi in parte in vigore e sono stati spesso motivo di litigio in passato. È interessante notare come le scritte siano cambiate nel tempo. Molte delle scritte più antiche sono circondate da ordinate cornici e riportano gli stessi nomi presenti sui cippi di confine, ovvero rocce poste a delimitare i confini fra le proprietà in montagna. Le scritte più recenti non hanno la cornice ma sono ricche di annotazioni dei pastori, come le pagine di un diario. L’insieme di queste scritte rappresenta un’importante testimonianza, un archivio per ricostruire un pezzo della storia della pastorizia della Val di Fiemme, dal punto di vista economico, sociale e ambientale. Le scritte e gli stemmi di famiglia sono infatti opera di diversi pastori che nel corso dei secoli hanno affrescato la montagna in punti diversi. Non tutte le scritte sono presenti nei database, il lavoro è di fatto appena cominciato. Il secondo motivo di interesse sono le similitudini fra queste scritte e l’arte rupestre paleolitica. Sorge naturale il paragone ai dipinti delle grotte della Dordogna o della Cantabria, o anche a quelli dei siti archeologici del Monte Bego e della val Camonica in Italia. Le scritte del monte Cornon sono inequivocabilmente d’età moderna e quasi precontemporanea eppure si collocano comunque in parte all’interno della lunga tradizione europea dell’arte rupestre preistorica. Condividono con le sorelle più antiche ispirazione e stile, le scritte infatti presentano alcuni aspetti dell’iconografia. L’unica incongruenza è quella

cronologica, eppure le scritte della Val di Fiemme possono comunque considerarsi un'evoluzione naturale della grande arte rupestre europea.

E.6: Neural machine translation of the second editing

E.6.1: German (DE)

Diese Graffitis sind aus mindestens zwei Gründen wichtig. Der erste Grund ist das Zeugnis, das sie von einer pastoralen Welt geben, die heute nicht mehr existiert: Sie beschreiben ihre wirtschaftlichen, religiösen, ästhetischen und moralischen Dimensionen und zeigen, dass es damals schon eine relative Alphabetisierung gab. Die Schriften enthalten in der Tat Unterschriften, Daten, Viehzahlen und Familienwappen. Sie befinden sich an Orten, an denen die Hirten häufig vorbeikamen, und zeugen von einem komplexen System von Weiderechten, von denen einige noch heute in Kraft sind und in der Vergangenheit häufig Anlass zu Streitigkeiten waren. Es ist interessant zu beobachten, wie sich die Inschriften im Laufe der Zeit verändert haben. Viele der älteren Inschriften sind von einem hübschen Rahmen umgeben und tragen dieselben Namen wie die Grenzsteine, die zur Abgrenzung von Berggrundstücken aufgestellt wurden. Die jüngeren Inschriften haben keinen Rahmen, sind aber voll mit Anmerkungen der Hirten, wie die Seiten eines Tagebuchs. Alle diese Inschriften stellen ein wichtiges Zeugnis dar, ein Archiv zur Rekonstruktion eines Teils der Geschichte des Hirtenwesens im Fleimstal unter wirtschaftlichen, sozialen und ökologischen Gesichtspunkten. Die Inschriften und Familienwappen sind in der Tat das Werk verschiedener Hirten, die den Berg im Laufe der Jahrhunderte an verschiedenen Stellen mit Fresken versehen haben. Nicht alle Inschriften sind in den Datenbanken vorhanden, die Arbeit hat gerade erst begonnen.

Der zweite interessante Grund sind die Ähnlichkeiten zwischen diesen Schriften und der paläolithischen Felskunst. Vergleiche mit den Malereien in den Höhlen der Dordogne oder Kantabriens oder auch mit denen in den archäologischen Stätten von Monte Bego und Val Camonica in Italien drängen sich auf. Die Inschriften auf dem Monte Cornon sind unverkennbar modernen und fast schon vorzeitlichen Alters, und doch stehen sie teilweise in der langen europäischen Tradition der prähistorischen Felskunst. Sie teilen mit ihren älteren Schwestern Inspiration und Stil, wobei die Inschriften sogar bestimmte Aspekte der Ikonographie aufweisen. Die einzige Ungereimtheit ist chronologischer Natur, dennoch können die Inschriften des Fleimstals als eine natürliche Entwicklung der großen europäischen Felskunst betrachtet werden.

E.6.2: English (UK)

These graffiti are important for at least two reasons. The first is the testimony they give of a pastoral world that no longer exists: they describe its economy, devotional, aesthetic and moral dimensions and show how there was already relative literacy at the time. The inscriptions in fact contain signatures, dates, livestock counts and family crests. They are found in places where shepherds often passed and bear witness to a system of complex grazing rights, some of which are still in force today and were often the cause of disputes in the past. It is interesting to note how the inscriptions have changed over time. Many of the older inscriptions are surrounded by neat frames and bear the same names found on boundary stones, i.e. rocks placed to mark the boundaries between mountain properties. The more recent inscriptions do not have frames but are full of shepherds' annotations, like the pages of a diary. All these inscriptions represent an important testimony, an archive to reconstruct a piece of the history of shepherding in Val di Fiemme, from an economic, social and environmental point of view. The inscriptions and family crests are in fact the work of various shepherds who have frescoed the mountain in different places over the centuries. Not all inscriptions are present in the databases, the work has in fact only just begun.

The second reason for interest is the similarities between these writings and Palaeolithic cave art. It is natural to compare them to the paintings in the caves of the Dordogne or Cantabria, or even to those in the archaeological sites of Monte Bego and Val Camonica in Italy. The inscriptions on Monte Cornon are unmistakably of modern and almost pre-contemporary age, and yet they are still partly within the long European tradition of prehistoric rock art. Sharing inspiration and style with their older sisters, the inscriptions in fact feature certain aspects of iconography. The only inconsistency is chronological, yet the Val di Fiemme inscriptions can still be considered a natural evolution of the great European rock art.

ABSTRACT (English)

This thesis examines whether a two-stage editing process on the original text — (i) pre-editing based on the rules of Writing for Machine Translation, and (ii) post-editing of the proposed source text — can increase the translatability of an Italian input and, consequently, improve the quality of neural machine translation (NMT) output into German (DE) and English (UK). Pre-editing refers to the targeted application of a set of rules drawn from the compendium of the Translation Centre for the Bodies of the European Union (2021) and Hardin et al. (2020); the second phase follows the post-editing process of the source text formulated by Somers (1997). There are two research questions: (RQ1) Do NMT outputs from a source edited in two steps—(i) pre-editing and (ii) post-editing of the source text—outperform outputs from the unedited source? And (RQ2) is this NMT quality gain similar for IT-EN (UK) and IT-DE, or does it differ by language pair?

For the purposes of the experiment, an Italian extract (ethnographic text) was chosen and subjected to two stages of editing. For each version — unedited (V1) and twice edited (V2) — translations into English (UK) and German (DE) were produced using DeepL (free version, without glossaries) outside of Matecat and then imported segment by segment. Ten participants (self-declared ≥ C1 in IT/EN/DE) reviewed four projects each (X_V1_EN, X_V1_DE, X_V2_EN, X_V2_DE) exclusively in Matecat, with MT/TM suggestions disabled. Error annotation follows a profile inspired by MQM (Style; Translation Errors; Terminology Consistency; Language Quality) with fixed severity levels (Neutral = 0; Minor = 0.5; Major = 2). The primary quantitative outcome is Matecat's quality score (Errors per Thousand words, EPT); in addition, the sums by severity are reported to outline the error profile. Immediately after each revision, a short questionnaire collects qualitative judgements on accuracy, fluency and fidelity, as well as open comments.

The analysis compares V2 with V1 for each participant and compares the two target languages. The expectation is a lower EPT and fewer serious errors in V2, with consistent findings in the questionnaires. The results indicate that intervening on the original text — combining pre-editing and post-editing of the source text before translating with NMT — systematically improves the quality of the output. In both directions (IT→EN and IT→DE), quality scores decrease and evaluations report greater accuracy, fluency and fidelity; in many cases, the output already reaches the acceptability threshold without further post-editing.

ABSTRACT (German)

Diese Arbeit untersucht, ob eine zweistufige Bearbeitung des Originaltextes – (i) Pre-Editing auf der Grundlage der Regeln von „Writing for Machine Translation“ und (ii) Post-editing of the source text (Post-Editing des Ausgangstextes)– die Übersetzbarkeit eines italienischen Inputs erhöhen und damit die Qualität der maschinellen neuronalen Übersetzung (NMT) ins Deutsche (DE) und Englische (UK) verbessern kann. Unter Pre-Editing versteht man die gezielte Anwendung einer Reihe von Regeln, die aus dem Kompendium von Translation Centre for the Bodies of the European Union (2021) und von Hardin et al. (2020) stammen; die zweite Phase folgt dem von Somers (1997) formulierten Prozess “Post-editing of the source text”. Es gibt zwei Forschungsfragen: (RQ1) Sind die NMT-Ergebnisse eines in zwei Schritten bearbeiteten Originaltextes – Vorbearbeitung plus Nachbearbeitung des Ausgangstextes – besser als die Ergebnisse, die mit dem unbearbeiteten Text erzielt werden? (RQ2) Ist der Qualitätsgewinn bei IT→EN (UK) und IT→DE ähnlich oder variiert er je nach Sprachpaar?

Für das Experiment wurde ein italienischer Auszug (ethnografischer Text) ausgewählt, der zwei Bearbeitungsphasen durchlaufen hat. Für jede Version – unbearbeitet (V1) und zweimal bearbeitet (V2) – werden die Übersetzungen ins Englische (UK) und Deutsche (DE) mit DeepL (kostenlose Version, ohne Glossare) außerhalb von Matecat erstellt und dann segmentweise importiert. Zehn Teilnehmer (selbst erklärte Sprachkenntnisse ≥ C1 in IT/EN/DE) überprüfen jeweils vier Projekte (X_V1_EN, X_V1_DE, X_V2_EN, X_V2_DE) ausschließlich in Matecat, wobei MT/TM-Vorschläge deaktiviert sind. Die Fehlerannotation folgt einem von MQM inspirierten Profil (Stil; Übersetzungsfehler; Terminologiekonsistenz; Sprachqualität) mit festen Schweregraden (Neutral = 0; Geringfügig = 0,5; Schwerwiegend = 2). Das primäre quantitative Ergebnis ist der Qualitätswert (Quality score) von Matecat (Errors per Thousand words, EPT); zusätzlich werden die Summen nach Schweregrad angegeben, um das Fehlerprofil zu skizzieren. Unmittelbar nach jeder Überarbeitung werden in einem kurzen Fragebogen qualitative Bewertungen zu Genauigkeit (accuracy), Flüssigkeit (fluency) und Treue (fidelity) sowie offene Kommentare erfasst.

Die Analyse vergleicht V2 mit V1 für jeden Teilnehmer und stellt die beiden Zielsprachen gegenüber. Die Erwartung ist ein niedrigerer EPT und weniger schwerwiegende Fehler in V2, mit übereinstimmenden Ergebnissen in den Fragebögen. Die Ergebnisse zeigen, dass die Bearbeitung des Originaltextes – durch eine Kombination aus Pre-Editing und Post-Editing des Ausgangstextes vor der Übertragung in NMT – die Qualität der Ausgabe systematisch verbessert. In beiden Richtungen (IT→EN und IT→DE) sinken die Qualitätswerte und die

Bewertungen zeigen eine höhere Genauigkeit, Flüssigkeit und Treue; in vielen Fällen erreicht die Ausgabe bereits ohne weitere Nachbearbeitung die Akzeptanzschwelle.

Ringraziamenti

In questo ultimo capitolo, scritto poco prima di consegnare la tesi, desidero ringraziare chi mi ha accompagnato in questo ultimo sforzo accademico.

Desidero ringraziare i miei genitori, per il sostegno costante in tutti questi anni di università e per avermi spronata con pazienza soprattutto nell'ultimo periodo. Un grazie sentito va anche al Professor Ciobanu, per avermi accompagnato con competenza e disponibilità nella stesura di questa tesi.

Ringrazio i miei fratelli, in particolare Maria e Davide, per l'aiuto e il sostegno nei momenti difficili. Grazie alle mie zie — Carla, Lucia e Margherita — per l'ospitalità e il supporto che non mi hanno mai fatto mancare.

Un pensiero riconoscente va ai miei amici in Italia: grazie a Margherita, Silvia, Elena, Alison, Enea e Luca per la pazienza infinita e per aver ascoltato tutte le mie lamentele. Grazie anche agli amici d'Oltralpe per l'aiuto prezioso di questi mesi e per l'amicizia costruita negli ultimi anni: Sarah, Lisa, Burcu, Miray, Martin ed Elias. Menzione d'onore alla mia coinquilina, Jessica e a Laura, capaci di risollevarmi l'umore anche nelle giornate più difficili.

Questa tesi è stata uno sforzo corale e sono profondamente grata a tutte le persone che mi sono state accanto. È un sollievo poter dire che anche questa avventura sia alle nostre spalle.